Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación
Centro de Investigación en Sistemas de Información

Solución de Inteligencia de Negocios basada en una arquitectura de Big Data para el Archivo Web de Venezuela

Trabajo Especial de Grado presentado ante la Ilustre
Universidad Central de Venezuela por
Br. Krystle Desiree Salazar Bolívar
Br. Teresa Georgette Tavernelli Chagua
para optar al título de Licenciado en Computación

Tutor:
Profa. Mercy Ospina

Caracas, noviembre del 2018

ACTA

Quienes suscriben, miembros del Jurado designado por el Consejo de la Escuela de Computación, para examinar el Trabajo Especial de Grado titulado **" Solución de Inteligencia de Negocios basada en una arquitectura de Big Data para el Archivo Web de Venezuela"** y presentado por la Br. Krystle Desiree Salazar Bolívar C.I. 20822923 y la Br. Teresa Georgette Tavernelli Chagua C.I. 24902933, a los fines de optar al título de **Licenciado en Computación**, dejan constancia de lo siguiente:

Leído como fue dicho trabajo, por cada uno de los miembros del jurado, se fijó el día 23 de noviembre de 2018, a las 10:30 horas, para que las autoras lo defendieran en forma pública, lo que estas hicieron en la sala de conferencias del Centro de Computación, mediante una presentación oral de su contenido, luego de lo cual respondieron a las preguntas formuladas. El prof. Andrés Sanoja estuvo presente a través de una videoconferencia debido a que se encontraba fuera del país y los jurados presentes estuvieron de acuerdo. Finalizada la defensa pública del Trabajo Especial de Grado, el jurado decidió aprobarlo con la nota de 20 puntos.

En fe de lo cual se levanta la presente acta, en Caracas a los 23 días del mes de noviembre del año 2018.

Profa. Mercy Ospina
(Tutor)

p/: prof Robinson Rivas

Prof. Andrés Sanoja
(Jurado)

Prof. Franky Uzcátegui
(Jurado)

A mis padres, Carlos y Carmen, que anhelaron ver culminado este proyecto más que nadie.

Krystle Salazar

A mi primer ángel que siempre está cuidándome desde el cielo, mi mamá, Margarette. A mis segundos ángeles, mis nonnos, Adelmo y Teresa, se cuanto querían ver este día. A mi primer pilar, mi papá, Alfredo, y mi segundo pilar, mi tía Wanda quienes han dedicado su vida a hacer de mí una persona de bien. Por últimо a mi hermana Josefina, quien siempre me apoya y está conmigo. Esto es para ustedes.

Teresa Tavernelli

Agradecimientos

Sobre todo a mis padres, porque sin su amor y apoyo incondicional este trabajo no hubiera sido posible, no podría haber deseado unos mejores.

A mi hermano Christian y al resto de mi familia, que siempre están y estarán ahí para darme una mano, sé que puedo confiar en ellos.

A Teresa, mi gran compañera que me acogió durante el último período en que trabajamos en esto, y a su linda familia. Gracias por haber realizado este trabajo conmigo y no claudicar en los momentos difíciles, se que de aquí me llevo una increíble amiga.

A nuestra tutora, Mercy, que siempre nos demostró mucho cariño y nos ayudó durante todo el proceso, una excelente profesional.

A los profesores de la Universidad Central de Venezuela, tanto los que tuve el privilegio de ser su estudiante y que me instruyeron en esta fascinante carrera, como los que no pero estuvieron apoyando en distintos ámbitos para formar profesionales íntegros. Mencionando especialmente a Adelis Nieves, Concettina Di Vasta y Robinson Rivas, con su inmensa vocación día a día hacen todo lo posible para mantener alejada la sombra de nuestra querida casa de estudio, son ejemplos a seguir.

Al resto de integrantes de CISI, que nos permitieron realizar este trabajo interrumpiendo a veces la paz del laboratorio, gracias por la paciencia y comprensión.

Al equipo de Cuantix, que me permitieron compaginar la culminación de mis estudios con el trabajo, y también estuvieron dispuestos a brindarme su apoyo, son personas maravillosas.

A los amigos que conocí durante la carrera (también las viejas amistades), que serán para toda la vida, son de las ganancias más valiosas.

Krystle Salazar

A Dios, por bendecirme y permitirme cumplir esta meta. Por cada persona que puso en mi camino durante estos años, cada uno tiene un lugar especial en mi corazón.

A mi papá, por tanto apoyo siempre, por hacer de mí una profesional y una mejor persona. Por siempre inculcarme el valor del estudio desde niña y a ser excelente en todo lo que haga, te quiero todo lo que hay.

A mi tía, quien siempre nos apoyó y ayudó durante todos estos años, gracias por tanta dedicación, sin ti esto no hubiese sido posible.

A mi hermana, por llenarme de alegría, por nuestras largas conversaciones, por ser un alma llena de luz.

A mi prima María Alejandra y familia, por ser un gran apoyo, por orientarme, por hacerme reír siempre y por preguntarme: ¿Y para cuándo la Tesis?.

A la Universidad Central de Venezuela, por ser la mejor casa de estudios del país y que a pesar de las circunstancias sigue siendo "La casa que vence la sombra". Por regalarme tantas buenas y únicas experiencias.

A nuestra tutora Mercy Ospina, gracias por tanto apoyo y paciencia, por tantos consejos y por la confianza. Por convertirnos en sus hijas.

A los profesores de la Escuela de Computación, quienes siempre dan lo mejor de sí e hicieron de mi una mejor persona, en especial Profa. Adelis Nieves y Profa. Concettina Di Vasta.

A todos mis amigos, en especial Alexander, Jhonatan, Cristina, Rafael, Ybrahin, Josseline, Samuel y Víctor por apoyarnos y ayudarnos en esta gran travesía. Gracias por tantos momentos. A mis amigas de toda la vida, Alejandra, Andreina y María Fernanda, por escucharme y por siempre estar ahí a pesar de la distancia.

Por último, pero no menos importante a mi compañera Krys y su familia. Gracias por tanta paciencia y tanto apoyo, comenzamos este proyecto siendo compañeras y hoy podemos decir que somos grandes amigas y que logramos vencer esto juntas.

Teresa Tavernelli

Resumen

El Archivo Web de Venezuela es un proyecto que representa una iniciativa para preservar, de manera histórica, el contenido web de relevancia para el país, es decir, aquel que puede llegar a ser considerado patrimonio digital de la nación. Actualmente esta iniciativa se encuentra en la segunda fase de desarrollo del prototipo, en la cual se busca facilitar el control, gestión y monitoreo del Archivo. Por lo cual se destaca con atención el proceso de rastreo de las páginas web, que se lleva a cabo con la herramienta Heritrix, la cual genera una importante cantidad de metadatos a partir de los rastreos realizados a los sitios web previamente seleccionados, que se mantienen en archivos de texto plano difíciles de analizar, y que contienen información relevante para la administración y planificación del proyecto. Es por ello que el presente trabajo especial de grado tiene como objetivo el desarrollo de una solución de inteligencia de negocio que genere indicadores que apoyen la toma de decisiones sobre la administración del sistema. Para ello se utilizaron herramientas del ecosistema Hadoop como su almacén de datos Hive, y Tableau para la generación de los indicadores, bajo la metodología de Ralph Kimball.

Palabras clave: Archivo Web, preservación, rastreo, Big Data, almacén de datos, indicadores.

Índice general

Agradecimientos I

Resumen III

Índice de figuras VII

Índice de tablas X

Introducción 1

1. Problema de Investigación 3
1.1. Planteamiento del problema . . . 3
1.2. Objetivos . . . 7
1.2.1. Objetivo general . . . 7
1.2.2. Objetivos específicos . . . 7
1.3. Justificación . . . 7
1.4. Alcance . . . 8

2. Marco Conceptual 9
2.1. Patrimonio Digital . . . 9
2.2. Preservación Web . . . 10
2.3. Archivo Web . . . 12
2.3.1. Tipos MIME . . . 12
2.3.2. Códigos de estado HTTP . . . 13
2.4. Modelo de Referencia OAIS . . . 14
2.4.1. Ambiente OAIS . . . 15
2.4.2. Modelo funcional . . . 15
2.4.3. Modelo de información . . . 16
2.4.4. Metadatos para la preservación . . . 18
2.5. Adaptación del modelo OAIS a Archivos Web . . . 19
2.6. Herramienta de rastreo Heritrix . . . 22
2.6.1. Principales características . . . 22
2.6.2. Archivos de salida . . . 23

2.6.3. Formato WARC . . . 24
2.7. Inteligencia de Negocios . . . 25
2.7.1. Características . . . 25
2.7.2. Funciones de una Solución de Inteligencia de Negocio . . . 26
2.7.3. Arquitectura de una Solución de Inteligencia de Negocio . . . 26
2.7.4. Modelo Dimensional . . . 29
2.7.5. Ventajas y desventajas . . . 30
2.8. Indicadores . . . 31
2.8.1. Objetivos . . . 31
2.8.2. Partes de un Indicador . . . 31
2.8.3. Tipos de Indicador . . . 33
2.8.4. Importancia de los Indicadores . . . 33
2.8.5. Antecedentes de uso de indicadores en la preservación Web . . . 34
2.8.6. Datos Estadísticos e Indicadores de Calidad . . . 34
2.9. Ciencia de Datos . . . 46
2.9.1. Big Data . . . 46
2.9.2. Big Data Analítica . . . 47
2.9.3. Inteligencia de Negocios Versus Ciencia de Datos . . . 48
2.9.4. Ecosistema Hadoop . . . 48
2.9.5. Arquitectura para Big Data y Data Warehouse . . . 49
2.9.6. Organización y estructura de los datos . . . 51

3. Marco Metodológico **53**
3.1. Metodología de Ralph Kimball . . . 53
3.1.1. Hitos del ciclo de vida . . . 54
3.2. Proceso de diseño dimensional en cuatro pasos . . . 57

4. Marco Aplicativo **58**
4.1. Definición de requerimientos . . . 58
4.2. Diseño de la arquitectura técnica . . . 60
4.3. Selección de productos . . . 61
4.3.1. Pentaho Data Integration . . . 61
4.3.2. MySQL . . . 61
4.3.3. Sqoop . . . 61
4.3.4. Hive . . . 62
4.3.5. Tableau Desktop . . . 62
4.4. Modelo dimensional . . . 62
4.4.1. Selección del Proceso de Negocio . . . 62
4.4.2. Identificación del nivel de granularidad . . . 63
4.4.3. Dimensiones . . . 63
4.4.4. Hechos y Tablas de Hechos . . . 66
4.5. Diseño físico . . . 69
4.5.1. Diseño del Área Intermedia . . . 70

4.6. Diseño y construcción de procesos ETC 75
4.6.1. Proceso ETC del Área Intermedia 75
4.6.2. Verificación de calidad de datos del Área Intermedia 79
4.6.3. Proceso ETC del Almacén de Datos 86
4.6.4. Verificación de calidad de datos del Almacén de Datos 87
4.7. Aplicación BI 92
4.8. Implementación 94
4.8.1. Indicadores 94
4.8.2. Cuadros de mando 97

Conclusiones y Recomendaciones **102**

Bibliografía **104**

Anexos **107**

Índice de figuras

1.1. Arquitectura del primer prototipo de archivo web. 4

2.1. Modelo de Ambiente de un OAIS. 15
2.2. Modelo funcional OAIS. 16
2.3. Datos externos de un Archivo OAIS. 17
2.4. Arquitectura funcional del IIPC basada en modelo OAIS. 20
2.5. Formato de Registro WARC. 25
2.6. Arquitectura de una solución de Inteligencia de Negocios 26
2.7. Arquitectura Big Data Appliances. 50

3.1. Diagrama del Ciclo de vida de Kimball 54

4.1. Arquitectura general 60
4.2. Matriz de Bus del AWV. 63
4.3. Modelo Dimensional (Tabla de Hechos Rastreos). 68
4.4. Modelo Dimensional (Tablas de Hechos de Codigo de estado HTTP y tipo MIME). 69
4.5. Modelo del Área Intermedia 75
4.6. Job para llenar el área intermedia. 76
4.7. Ejemplo de archivo .txt - crawl-report 77
4.8. Ejemplo de archivo .txt - response-code-report 78
4.9. Ejemplo de archivo .txt - response-code-report 78
4.10. Contenido de la tabla de colecciones en la fuente (arriba) y en el área intermedia (abajo). 79
4.11. Contenido de la tabla de semillas en la fuente (arriba) y en el área intermedia (abajo). 80
4.12. Resultado de la consulta de la muestra aleatoria de códigos de estado HTTP en el Área Intermedia. 81
4.13. Resultado de la consulta de la muestra aleatoria de tipos MIME en el Área Intermedia. 82
4.14. Muestra de resultados de consultas de rastreos en la fuente y en el Área Intermedia. 83
4.15. Archivo crawl-report.txt del rastreo con id=30 en el Área Intermedia. . 84

4.16. Resultado de consultar las métricas para el rastreo con id=30 en el Área Intermedia. 84
4.17. Archivo responsecode-report.txt del rastreo con id=30 en el Área Intermedia. 85
4.18. Resultado de consultar las métricas de código de estado HTTP para el rastreo con id=30 en el Área Intermedia. 85
4.19. Archivo mimetype-report.txt del rastreo con id=50 en el Área Intermedia. 86
4.20. Resultado de consultar las métricas de tipo MIME para el rastreo con id=50 en el Área Intermedia. 86
4.21. Job para llenar el almacén 87
4.22. Ejemplo de importación de datos con Sqoop. 87
4.23. Muestra de resultados de consultas de colecciones en el Área Intermedia y en el Almacén. 88
4.24. Muestra de resultados de consultas de semillas en el Área Intermedia y en el Almacén. 89
4.25. Muestra de resultados de consultas de tipos MIME en el Área Intermedia y en el Almacén. 89
4.26. Muestra de resultados de consultas de códigos de estado en el Área Intermedia y en el Almacén. 90
4.27. Muestra de resultados de consultas de códigos de estados y métricas en el Área Intermedia y en el Almacén. 90
4.28. Muestra de resultados de consultas de rastreos en el Área Intermedia y en el Almacén. 91
4.29. Muestra de resultados de consultas de tipos MIME en el Área Intermedia y en el Almacén. 92
4.30. Importación de datos de la Tabla de Hechos de Rastreos en Tableau . . 92
4.31. Importación de datos de la Tabla de Hechos de tipo MIME en Tableau 93
4.32. Importación de datos de la Tabla de Hechos de Código de Estado HTTP 93
4.33. Indicador: Cantidad de rastreos por Fecha 94
4.34. Indicador: Cantidad de URIs por semilla 95
4.35. Indicador: Promedio de duración por semilla 95
4.36. Indicador: Distribución de espacio por tipos de formato por Colección 96
4.37. Indicador: Distribución de URLs por código de estado por Semilla . . 97
4.38. Cuadro de mando de gráficos de torta 98
4.39. Cuadro de mando de gráficos de los URIs 99
4.40. Cuadro de mando de la distribución de URLs por código de estado http 100
4.41. Cuadro de mando de la distribución de tipo MIME 101

4.42. Indicador 1 - Cantidad de semillas por colección 107
4.43. Indicador 1 - Cantidad de semillas por fecha 108
4.44. Indicador 2 - Cantidad de rastreos por colección 108
4.45. Indicador 2 - Cantidad de rastreos por semilla 109

4.46. Indicador 3 - Cantidad de URIs por colección 109
4.47. Indicador 3 - Cantidad de URIs por fecha 110
4.48. Indicador 4 - Duración promedio de rastreo por colección 110
4.49. Indicador 4 - Duración promedio de rastreo por semilla 111
4.50. Indicador 4 - Duración promedio de rastreo por fecha 111
4.51. Indicador 5 - Cobertura cronológica por semilla 112
4.52. Indicador 6 - Distribución de URLs por código de estado http por colección 112
4.53. Indicador 6 - Distribución de URLs por código de estado http por semilla 113
4.54. Indicador 6 - Distribución de URLs por código de estado http por fecha 113
4.55. Indicador 7 - Distribución de espacio por tipos de formatos (Tipos MIME) por colección. 114
4.56. Indicador 7 - Distribución de espacio por tipos de formatos (Tipos MIME) por semilla. 114
4.57. Indicador 7 - Distribución de espacio por tipos de formatos (Tipos MIME) por fecha. 115
4.58. Indicador 8 - Distribución de URLs por tipos de formatos (Tipos MIME) por colección. 115
4.59. Indicador 8 - Distribución de URLs por tipos de formatos (Tipos MIME) por semilla. 116
4.60. Indicador 8 - Distribución de URLs por tipos de formatos (Tipos MIME) por fecha. 116
4.61. Indicador 9 - Cantidad de colecciones por fecha. 117
4.62. Máximo de ancho de banda por colección y semilla. 117
4.63. Máxima duración por colección y semilla. 118
4.64. Cantidad de Hosts visitados por colección y semilla. 118

Índice de tablas

1.1. Usuarios de un Archivo Web. 6

2.1. Elementos de Información de un Archivo Web. 21
2.2. Datos estadísticos principales para el desarrollo de una colección. . . . 35
2.3. Datos estadísticos principales para la caracterización de la colección. . 36
2.4. Datos estadísticos básicos sobre el uso del Archivo Web. 37
2.5. Datos estadísticos para la caracterización avanzada del uso del Archivo Web. 37
2.6. Datos estadísticos principales para el uso de la colección. 39
2.7. Datos estadísticos para la preservación del flujo de bits. 40
2.8. Datos estadísticos relacionados a los metadatos de preservación. 41
2.9. Datos estadísticos para la preservación lógica del Archivo Web. 42
2.10. Datos estadísticos principales para la preservación de una colección. . 42
2.11. Indicadores de calidad. 44

4.1. Indicadores del proceso de rastreo. 59
4.2. Detalle de la dimensión Fecha. 64
4.3. Detalle de la dimensión Tiempo. 64
4.4. Detalle de la dimensión Semilla. 65
4.5. Detalle de la dimensión Colección. 65
4.6. Detalle de la dimensión tipo MIME. 66
4.7. Detalle de la dimensión Código de Estado HTTP. 66
4.8. Tabla de hechos de Rastreos. 67
4.9. Tabla de hechos de Código de Estado. 67
4.10. Tabla de hechos de Tipos MIME. 67
4.11. Detalle de la tabla rastreo 70
4.12. Detalle de la tabla tipo MIME 71
4.13. Detalle de la tabla código de estado HTTP 71
4.14. Detalle de la tabla colección 72
4.15. Detalle de la tabla semilla 72
4.16. Detalle de la tabla métricas tipo MIME 73
4.17. Detalle de la tabla métricas códigos de estado 73
4.18. Detalle de la tabla métricas rastreo 74

4.19. Datos de la muestra aleatoria de tipos MIME seleccionados. 81
4.20. Datos de la muestra aleatoria de tipos MIME seleccionados. 82

Introducción

Hoy en día internet ha pasado a formar parte de la vida cotidiana de las personas, tanto en el ámbito personal como el profesional, educativo, recreativo y demás, constituyendo un medio de comunicación esencial y haciendo que a cada momento se generen nuevos contenidos digitales. Parte de este contenido tiene un valor importante para la sociedad, representa parte de su herencia histórica y su cultura, esto recibe el nombre de patrimonio cultural, el cual se divide en tangible e intangible, siendo este último donde se clasifica el patrimonio digital (que puede ser material gráfico, grabaciones sonoras, bases de datos, programas informáticos, páginas Web, entre otros), y más específicamente en este caso el patrimonio web.

Debido a la naturaleza volátil de internet, los contenidos se crean y cambian o desaparecen en cualquier momento, surgiendo la necesidad de resguardar este contenido digno de protección y conservación para las generaciones actuales y futuras. Es por esto que han aparecido distintas iniciativas en varios países para preservar el contenido web, las cuáles están basadas en el concepto de archivo histórico desarrollado por las bibliotecas tradicionales, siendo muchas de estás las principales promotoras de la preservación web.

Así se crea en el 2003 el Consorcio Internacional para la Preservación del Internet (del inglés *International Internet Preservation Consortium, IIPC*) con el objetivo de desarrollar herramientas, estándares y mejores prácticas de archivado web, al tiempo que promueve la colaboración internacional y el amplio acceso y uso de los archivos web para la investigación y el patrimonio cultural. Los Archivos Web (*Web Archive, WA*) son sistemas de información usados para resguardar el contenido Web de una comunidad determinada, permitiendo su preservación y acceso por parte de los interesados.

Por ello existe actualmente un proyecto en la Universidad Central de Venezuela para desarrollar un prototipo de Archivo Web para la preservación del patrimonio web de Venezuela, basado en modelos y estándares internacionales, y tecnología de software libre. Este prototipo ya presenta una primera versión del módulo de análisis de datos (López & Sarno, 2015), sin embargo se encontraron en éste una serie de limitaciones que pretenden resolverse en una segunda versión, reforzando el uso de tecnologías para el manejo de grandes volúmenes de datos o *Big Data*.

El objetivo de esta Trabajo Especial de Grado (TEG) se centra en el desarrollo de una plataforma para el análisis de los metadatos del Archivo Web de Venezuela, basada en una arquitectura de Big Data, que permita la visualización de indicadores y métricas relacionadas con el proceso de rastreo del Archivo Web.

Este documento se encuentra estructurado en cuatro capítulos. El capítulo 1, "Problema de investigación", describe la situación actual con respecto al archivado web en Venezuela y los problemas que se presentan para el análisis de los datos. Además se exponen el objetivo general y los objetivos específicos del presente Trabajo Especial de Grado.

En el capítulo 2, "Marco Conceptual", se definen los conceptos relevantes al proceso de preservación web, los sistemas de inteligencia de negocio tradicionales, y los aspectos relativos al análisis de grandes volúmenes de datos, la nueva tendencia conocida como *Big Data Analytics*. En el capítulo 3, "Marco Metodológico", se expone la metodología de Ralph Kimball, utilizada para desarrollar el almacén de datos.

En el capítulo 4, "Marco Aplicativo", se describe como se llevó a cabo el desarrollo de la solución, siguiendo la metodología previamente mencionada. Por último se listan las conclusiones alcanzadas y propuestas para posibles trabajos futuros.

1

Problema de Investigación

1.1. Planteamiento del problema

En la Universidad Central de Venezuela, específicamente en la Escuela de Computación de la Facultad de Ciencias, se viene desarrollando desde el 2012 una incitativa de Archivo Web para preservar páginas web de interés nacional. Este desarrollo se produce de manera modular e incremental a través de distintos trabajos que han resultado en cada una de las partes que componen el prototipo de Archivo Web de Venezuela.

Actualmente este prototipo consta de los siguientes módulos que cubren las tareas básicas de un archivo web (definidas en el estándar OAIS, sección 2.4): un módulo de adquisición (productor), un módulo de indexación y almacenamiento (archivo) y un módulo de acceso al archivo web (consumidor). Además se encuentran el módulo de Predicción de cambios, el módulo de Gestión y Control de Incidencias y el módulo de Inteligencia de Negocios (*Bussiness Intelligent, BI*). Estos se muestran en la Figura 1.1.

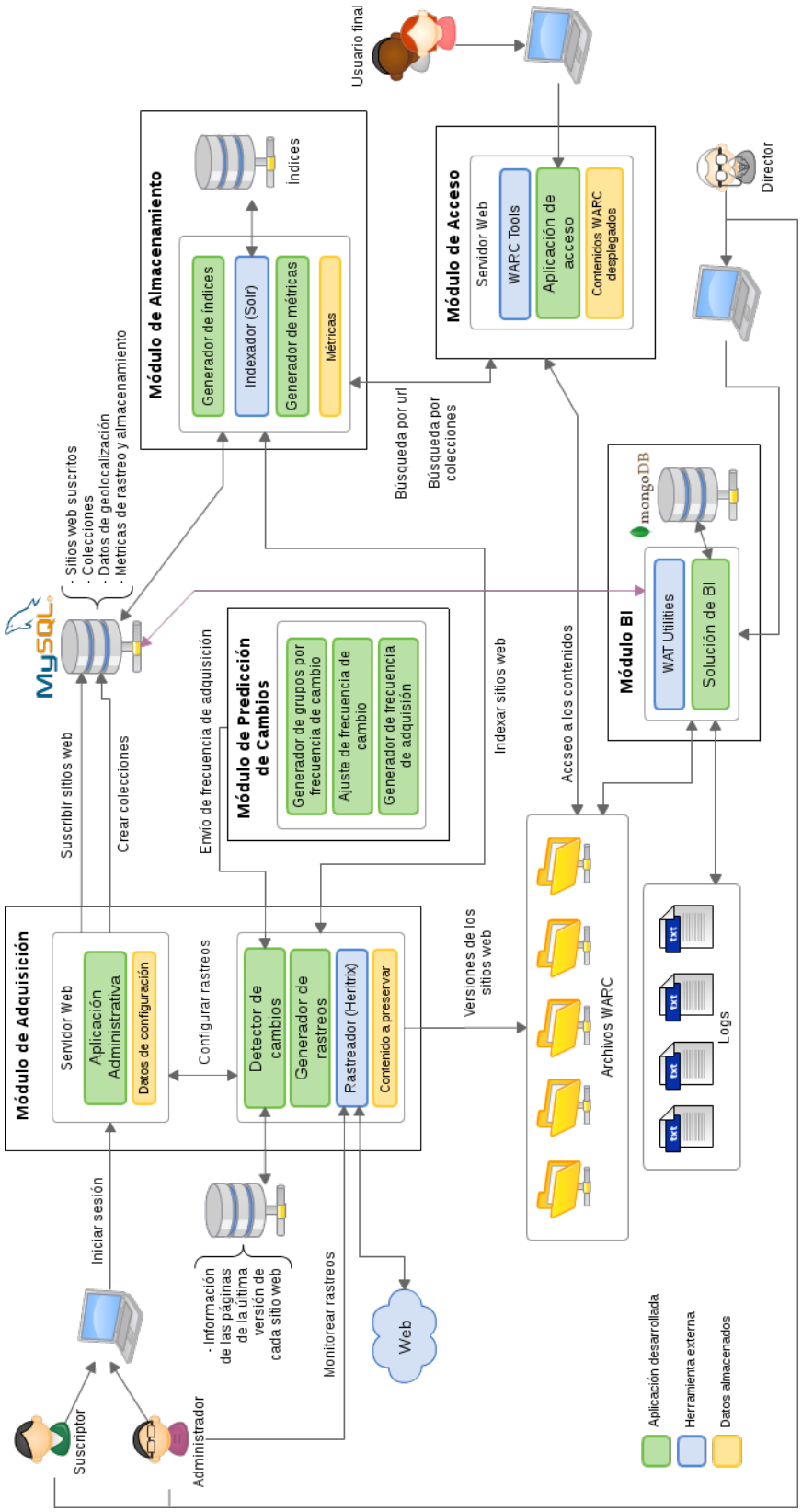

Figura 1.1: Arquitectura del primer prototipo de archivo web.
Fuente: Elaboración propia.

El módulo de adquisición engloba las herramientas de rastreo, las cuales se encargan de inspeccionar la *World Wide Web* de forma metódica y automatizada, con la finalidad de crear una copia de todas las páginas web visitadas para su preservación, a este proceso se le denomina rastreo o cosecha (Rivero & García, 2013). Posteriormente, le fue añadido un componente de predicción de cambios para mejorar la frecuencia de almacenamiento de versiones de las páginas (Casanova Diaz & Caraballi, 2015).

El módulo de indexación y almacenamiento se encarga de almacenar las páginas web cosechadas, para luego ser indexadas, y así facilitar futuros accesos a dichas páginas (Rivero & García, 2013). Este módulo fue modificado para la segunda versión del prototipo por trabajos siguientes, que trataban de manera separada y específica las tareas de almacenamiento distribuido (Estaba Fernández-Trujillo & Ciancia Biondo, 2015) y de indexación del contenido de los sitios web (Montero Hernández & Pérez Laya, 2016) con mayor enfoque en la administración de grandes volúmenes de datos.

El módulo de acceso permite al usuario ingresar y ver el contenido de los archivos almacenados, de acuerdo a la búsqueda que haya realizado (Kabchi & Martínez, 2014). El mismo fue mejorado con el desarrollo del Módulo de Gestión y de Control de Incidencias, que añade funcionalidades administrativas e implementa la división de los roles de usuarios (Sánchez & Milano, 2015). Estos usuarios que interactúan con el Archivo Web se muestran en la tabla 1.1.

El módulo de inteligencia de negocios permite al usuario observar el contenido de los metadatos almacenados en el archivo web de manera sencilla, permitiéndole así poder tomar mejores decisiones sobre los datos (Lopez & Sarno, 2015). Sin embargo, en esta primera versión del módulo se detectaron una serie de inconvenientes:

- Lentitud en la recuperación y presentación de los datos: debido a la gran cantidad de datos almacenados en el archivo web (que crece constantemente) la recuperación y acceso rápido a éstos se vuelve un reto, el cual debe mejorarse en el módulo actual ya que presenta retrasos en este proceso, y a medida que se acumulen más datos el sistema disminuirá su rendimiento.
- Poca flexibilidad: la forma como se visualizan los indicadores no puede ser modificada, limitando la forma en que los usuarios pueden obtener información para tomar decisiones. Esta situación dificulta las capacidades analíticas, las cuales requieren manipular los datos, modificar los indicadores, y la forma como estos indicadores se observan con el fin de obtener una vista de los datos que genere información útil en un momento dado.
- Limitación en la generación de nuevos informes o reportes: el actual módulo presenta un número de reportes fijo, y a medida que se usa los usuarios suelen requerir la modificación o creación de nuevas consultas o reportes, para ello se necesita de un desarrollador que realice estos cambios en el sistema, los cuales

Tabla 1.1: Usuarios de un Archivo Web.

Nombre	Descripción
Suscriptor	Rol desempeñado por las personas o los sistemas cliente, que proporcionan la información a ser conservada. Toman decisiones para incluir o excluir elementos (semillas –puntos de entrada de los sitios web a preservar–) o grupos de elementos (colecciones –conjuntos de semillas clasificados por tema–) en cada etapa del flujo, desde la adquisición hasta el almacenamiento. Tienen la responsabilidad de cumplir la política de selección.
Usuario Final	Rol desempeñado por las personas o los sistemas cliente, que interactúan con los servicios del Archivo para encontrar y adquirir información conservada de interés y estadísticas acerca de las métricas recolectadas.
Director	Rol responsable del manejo de los componentes funcionales, análisis de riesgos y costos, y definición de las políticas del Archivo a un nivel superior, así como de la coordinación entre administradores y suscriptores.
Administrador	Rol desempeñado por técnicos u operadores de rastreo, que controlan el flujo de trabajo y su operación diaria. Su tarea es desarrollar, construir, mantener y controlar el flujo de trabajo del Archivo.

Fuente: Ospina Torres (2014).

no son sencillos de realizar debido a la compleja estructura de los archivos WARC, y los procesos de extracción y transformación usados para mostrar los datos en la vista. Esto puede representar un tiempo de espera importante, lo cual implica que una necesidad analítica no puede resolverse en el momento en que esta se genera, exponiendo esto fallas en el sistema para la toma de decisiones en el momento oportuno.

- Falta de una arquitectura que soporte grandes volúmenes de datos: en vista de la ingente cantidad de datos potenciales a almacenar, se hace necesario una plataforma de fácil escalamiento que pueda manejar cada vez más datos con un rendimiento aceptable.

Tomando en cuenta sobre todo este último punto es que se da inicio a la segunda versión del prototipo de Archivo Web de Venezuela, con un enfoque de almacenamiento y procesamiento distribuido. Para esto se decide establecer el uso de la plataforma Hadoop (de la cual se habla en la sección 2.9.4) para gestionar el almacenamiento, y algunas de las herramientas de su ecosistema para apoyar el resto de los procesos.

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar una plataforma para el análisis de los metadatos del proceso de rastreo del Archivo Web de Venezuela basada en una arquitectura de Big Data.

1.2.2. Objetivos específicos

1. Comprender la información de los archivos generados por el rastreador Heritrix.
2. Elaborar indicadores en base a los datos disponibles.
3. Seleccionar las herramientas para cada componente de la arquitectura.
4. Implementar la arquitectura de Big Data para el análisis de los datos descriptivos del proceso de rastreo.
5. Desarrollar un tablero de control que permita el acceso a los indicadores y la realización de análisis definidos por el usuario.

1.3. Justificación

El prototipo de Archivo Web ya posee cierta cantidad de datos almacenados sobre páginas de sitios web venezolanos, que pudieran presentar un interés de acceso en el futuro cuando ya no estén disponibles desde su fuente original. Estos datos a pesar de que se encuentran dentro del Archivo Web, están en un formato que dificulta enormemente su manipulación.

Así, el poder obtener información descriptiva de la evolución del prototipo sería una tarea demasiado laboriosa si se realizara de forma manual, recorriendo y accediendo a cada archivo y carpeta de rastreo, y esta tarea resulta de vital importancia si se quiere saber el rendimiento del prototipo de Archivo Web, conocer de manera resumida o agregada los datos que almacena y su comportamiento en general a través del tiempo, para realizar ajustes o mejoras al prototipo. Es por esto que se decide crear una plataforma para el análisis de los metadatos, para solventar esta necesidad.

1.4. Alcance

Este trabajo se plantea establecer las bases de una arquitectura que soporte el análisis de grandes volúmenes de datos generados por el Archivo Web de Venezuela, para ello se definen un conjunto de indicadores y métricas que se toman como los requerimientos iniciales a cumplir.

Se diseñan los modelos de base de datos, tanto del área intermedia, donde se recaban inicialmente todos los datos de los archivos fuentes, como del almacén de datos. Así mismo se implementan los procesos extracción, transformación y carga inicial necesarios para adaptar los datos y alimentar cada base de datos. Por último se diseña el modelo dimensional para el área de presentación de datos y se elaboran los indicadores en un cuadro de mando con la herramienta de visualización seleccionada.

2

Marco Conceptual

En este capítulo se describirán los conceptos necesarios para comprender el contexto del trabajo de investigación. Primero se introducen los conceptos de patrimonio digital, la preservación web y los Archivos Web, incluyendo el modelo de referencia OAIS, los formatos de archivo para el almacenamiento de información histórica y sus metadatos. Luego se presenta la definición de un Sistema de Inteligencia de Negocios, sus características, funciones y la tradicional arquitectura de una solución de inteligencia de negocios. Seguidamente se define lo que es un indicador, sus partes, los tipos, y las estadísticas e indicadores especificados en el estándar ISO/DTR 14873:2012. Por último, se definen los conceptos que involucra la Ciencia de Datos, como lo son los grandes volúmenes de datos (Big Data), y la tendencia denominada Big Data Analítica.

2.1. Patrimonio Digital

El patrimonio cultural puede ser considerado como la herencia cultural propia del pasado de una comunidad, con la que ésta vive en la actualidad y que transmite a las generaciones presentes y futuras (Unesco, 2017). La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (en inglés *United Nations Educational, Scientific and Cultural Organization*), abreviado internacionalmente como Unesco, ha clasificado el patrimonio en tangible, e intangible. El patrimonio tangible está comprendido por objetos arqueológicos, etnográficos, tecnológicos, religiosos y

aquellos de origen artesanal o folclórico, y el patrimonio intangible puede ser expresiones y prácticas culturales, la medicina tradicional, la religiosidad popular y las tecnologías tradicionales de cada país.

En los documentos de la Unesco, el patrimonio se define como nuestra herencia del pasado, nuestros bienes actuales y lo que legamos a las generaciones futuras. El patrimonio es, o debería ser, algo que se transmite de generación en generación porque se valora.

Ahora bien, el patrimonio digital está formado por los materiales informáticos de valor perdurable dignos de ser conservados para las generaciones futuras, y que proceden de comunidades, industrias, sectores y regiones diferentes. No todos los materiales digitales poseen valor perdurable, pero los que lo tienen exigen metodologías de conservación activas para mantener la continuidad del patrimonio digital. Comprende recursos de carácter cultural, educativo, científico o administrativo e información técnica, jurídica, médica y de otras clases, que se generan directamente en formato digital o se convierten a éste a partir de material analógico ya existente. Los productos "de origen digital" no existen en otro formato que el electrónico.

Los objetos digitales pueden ser textos, bases de datos, imágenes fijas o en movimiento, grabaciones sonoras, material gráfico, programas informáticos o páginas Web, entre otros muchos formatos posibles dentro de un vasto repertorio de diversidad creciente. A menudo son efímeros, y su conservación requiere un trabajo específico en este sentido en los procesos de producción, mantenimiento y gestión (Unesco, 2017).

2.2. Preservación Web

La preservación web es un tipo de preservación digital y se puede definir como el proceso de recolectar información disponible de la World Wide Web, preservándola en un formato de archivo, y garantizando que el contenido pueda ser consultado posteriormente (IIPC, 2012).

Gran parte de la enorme cantidad de información que se produce en el mundo es de origen digital y existe en una gran variedad de formatos. Para las instituciones culturales que tienen a su cargo el acopio y la preservación del patrimonio cultural, se ha convertido en un problema inminente definir qué elementos deben conservarse para las generaciones futuras y cómo proceder en su selección y conservación. El enorme tesoro de información digital que se produce hoy día en prácticamente todas las áreas de las actividades humanas, y concebida para ser consultada con computadoras, podría perderse si no se elaboran técnicas y políticas específicas para su conservación.

La complejidad de los problemas que se plantean obliga a que en la tarea de preservación intervengan los productores de la información digital, comprendidos los de programas informáticos, quienes, al diseñar sus productos, deberán tener en cuenta la conservación. Ya han pasado a la historia los días en que la responsabilidad de la preservación incumbía exclusivamente a las instituciones encargadas de los archivos (Unesco, 2017).

El primer problema es el volumen de los datos, ya que se calcula que en Internet existen casi dos mil millones de sitios web (Internet Live Stats, 2018). Otro de los problemas son las tácticas tradicionales para la preservación, ya que no se pueden utilizar de la misma manera en el material digital, porque las "publicaciones" en la red habitualmente se utilizan datos almacenados en diferentes servidores ubicados en muchas partes del mundo. Es por estos problemas que se hace necesario preservar la información digital. La preservación digital, se refiere a conservar para poder asegurar la accesibilidad y la recuperación de los materiales digitales. También se puede ver como una técnica de digitalización de documentos en papel para garantizar su conservación a través del tiempo y así evitar la deterioración del documento original. De igual manera, el principal objetivo es conservar y preservar a largo plazo los objetos digitales Web, así vengan de la digitalización de un documento en papel o un objeto ya digital (REBIUN, 2009).

Los objetos digitales a ser almacenados poseen dos características principales:

1. Una fuente única (Servidor Web)
2. Un identificador único, el cual suele ser un Identificador Uniforme de Recursos (*Uniform Resource Identifier, URI*)

Así Ospina Torres (2014, pg. 14) dice que la Web desde este punto de vista, no es un contenedor de archivos fijos, sino una caja negra con recursos, de los cuales el usuario sólo recibe instancias.

Según el IIPC (2012), el internet ha dado inicio a una era sin precedente, donde se pueden compartir infinidades de cosas. Como consecuencia, las instituciones dedicadas a la preservación y documentación del conocimiento y la cultura contemporánea se enfrentan a nuevos retos. Muchas de las cosas que recolectan las instituciones dedicadas a la preservación digital son accesibles desde la web, por ejemplo las publicaciones académicas, los materiales de campañas, las obras de arte, los documentos gubernamentales, correspondencia y noticias. Las páginas son cada vez más dinámicas, su contenido cambia constantemente, por lo que es importante capturar esta información en tiempo real y asegurar su preservación para las próximas generaciones.

Según la Unesco (2017), la preservación digital debe tener las siguientes características:

1. Ser accesible para cualquier persona.

2. Garantizar la protección de información delicada o de carácter privado.
3. Disponer de un marco jurídico y técnico que proteja su autenticidad.
4. No debe estar sujeto a límites temporales, geográficos, culturales o de formato. Se debe propiciar, con el tiempo, una representación de todos los pueblos, naciones, culturas e idiomas.

2.3. Archivo Web

Los Archivos Web son Sistemas de Información que surgen para archivar de manera histórica documentos Web, considerados parte del patrimonio digital de las naciones. Entiéndase un documento web como un documento basado en el lenguaje de marcas HTML (*HiperText Markup Language*) también llamado página web, y los demás archivos asociados en su composición como imágenes, videos, hojas de estilo, scripts, entre otros, que tiene asociado un tipo MIME, que puede ser localizado a través de un URL, y que normalmente forma parte de un sitio web (Ospina Torres, 2014).

Los Archivos Web recolectan fragmentos de la World Wide Web (WWW), preservando estas colecciones en un formato de archivo específico y ofreciendo estos archivos para ser accedidos y utilizados por generaciones futuras.

El principal objetivo de los Archivos Web es preservar conjuntos seleccionados de sitios web junto a sus documentos mediante su replicación y/o migración de su formato original a otra representación. Estos sitios replicados son mantenidos completos, esto quiere decir que están acompañados de los archivos, imágenes, gráficas y aspecto visual, y son almacenados en servidores de preservación en un ambiente seguro (Masanès, 2006).

2.3.1. Tipos MIME

Los tipos MIME (*Multipurpose Internet Mail Extensions*) son una serie de convenciones o especificaciones dirigidas al intercambio a través de Internet de todo tipo de archivos (texto, audio, vídeo, etc.) de forma transparente para el usuario (Freed & Borenstein, 1996).

Estos tipos fueron definidos en el RFC 6838 por el IANA (*Internet Assigned Numbers Authority*, Autoridad de Números Asignados en Internet) el cual es el organismo oficial responsable de realizar un seguimiento de todos los tipos MIME oficiales.

Los navegadores a menudo usan el tipo MIME (y no la extensión de archivo) para determinar cómo procesar un documento; por lo tanto, es importante que los ser-

vidores estén configurados correctamente para adjuntar el tipo MIME correcto al encabezado del objeto de respuesta (Mozilla y colaboradores, 2018b).

Estructura de los Tipos MIME

La estructura de un tipo MIME es muy simple; consiste en un tipo y un subtipo, dos cadenas separadas por un '/'. No se permite espacio. El tipo representa la categoría y puede ser de tipo discreto o multiparte. El subtipo es específico para cada tipo. Un tipo MIME no distingue entre mayúsculas y minúsculas, pero tradicionalmente se escribe todo en minúsculas (Mozilla y colaboradores, 2018b).

A continuación veremos los tipos MIME discretos y multiparte:

- **Discretos:** los tipos discretos indican la categoría del documento. Por ejemplo: text, image, audio, video, application. Para documentos de texto sin subtipo específico, se debe usar text/plain. De forma similar, para los documentos binarios sin subtipo específico o conocido, se debe usar application/octet-stream.
- **Multiparte:** los tipos de partes múltiples indican una categoría de documento que está dividida en distintas partes, a menudo con diferentes tipos de MIME. Es una forma de representar un documento compuesto.

2.3.2. Códigos de estado HTTP

Los códigos de estado de respuesta HTTP indican si se ha completado satisfactoriamente una solicitud HTTP específica. Las respuestas se agrupan en cinco clases: respuestas informativas, respuestas satisfactorias, redirecciones, errores de los clientes y errores de los servidores (Mozilla y colaboradores, 2018a).

A continuación se describirá brevemente cada clase de respuesta informativa de los códigos HTTP:

- Respuestas informativas: son los que informan al navegador las acciones que se van a realizar y comienzan por el número 1.
- Respuestas satisfactorias: estos indican que la petición del navegador se ha recibido, procesado y respondido correctamente, comienzan por el número 2.
- Redirecciones: indican que el navegador debe realizar alguna acción adicional para que la petición se complete, por ejemplo redirigirse a otra página, comienzan por el número 3.
- Errores de los clientes: estos indican que se ha producido un error en el lado del cliente, comienzan por el número 4.

- Errores de los servidores: estos indican que se ha producido un error en el servidor, comienzan por el número 5.

2.4. Modelo de Referencia OAIS

El Comité Consultivo de Sistemas de Datos Espaciales (*Consultative Committee for Space Data Systems*, CCSDS) establecido en 1982, es un foro de las agencias espaciales de EEUU para el desarrollo cooperativo de estándares para el manejo de datos de las investigaciones espaciales, como parte de una iniciativa para desarrollar normas que apoyen la conservación a largo plazo de los datos obtenidos de satélites y otros tipos de misiones espaciales.

El modelo de referencia para un Sistema de Información de Archivo Abierto (*Open Archival Information System*, OAIS) es un modelo desarrollado por el Comité Consultivo de Sistemas de Datos Espaciales (*Consultative Committee for Space Data Systems*, CCSDS) en el 2002, con la aprobación de la Organización Internacional de Normalización (*International Organization for Standardization*, ISO) como un estándar internacional en el 2003, el ISO 14721. Este modelo tiene como objetivo proporcionar un marco común de alto nivel que se puede utilizar para ayudar a entender los desafíos que representan los Archivos Web.

Según Lavoie (2014) el modelo de referencia de la CCSDS definiría los componentes funcionales básicos de un sistema dedicado a la preservación a largo plazo de la información digital, detallaría las interfaces internas y externas clave del sistema y caracterizaría los objetos de información gestionados por el sistema. Estas descripciones se expresarían en términos de un conjunto bien definido de conceptos y terminología trascendente, que sin embargo, se pueden asignar a vocabularios específicos de dominio.

El término «abierto» se refiere al hecho de que el modelo fue desarrollado y publicado en foros públicos abiertos, en los cuales se alentó a cualquier parte interesada a participar. Un «sistema de información de archivo» es una organización de personas y sistemas que "asumen la responsabilidad de preservar la información y ponerla a disposición de una comunidad determinada" (CCSDS, 2012, pp. 1-1). Esta definición enfatiza las dos funciones principales para el archivo, la primera, preservar la información y la segunda, proveer acceso a la información archivada.

El modelo de referencia OAIS consiste de tres partes separadas pero relacionadas, cada parte centrada alrededor del concepto de un archivo tipo OAIS:

- El ambiente externo en el cual opera el OAIS.
- Los componentes funcionales, o mecanismos internos, que cumplen colectivamente las responsabilidades de preservación del OAIS.

- Los objetos de información que son ingeridos, gestionados y difundidos por el OAIS.

2.4.1. Ambiente OAIS

En la Figura 2.1 se muestra la interacción del Archivo con agentes externos. El modelo de referencia OAIS identifica y describe estas entidades externas, y caracteriza las interfaces entre dichas entidades y el Archivo.

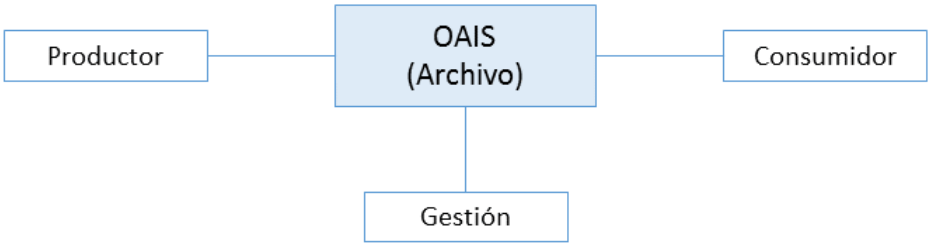

Figura 2.1: Modelo de Ambiente de un OAIS.
Fuente: CCSDS, 2012, p. 2-2.

A continuación, se describen los elementos que forman parte del ambiente OAIS:

- **Productor:** rol desempeñado por las personas o los sistemas cliente que proporcionan la información a ser conservada.
- **Consumidor:** rol desempeñado por aquellas personas o sistemas cliente que interactúan con los servicios del OAIS para encontrar y adquirir información conservada de interés.
- **Gestión:** rol responsable del manejo de los componentes funcionales y las políticas del Archivo, no así de las operaciones del día a día del archivo.

2.4.2. Modelo funcional

Describe la gama de actividades que deben llevarse a cabo por un Archivo. Define seis capas de servicio de alto nivel o componentes funcionales, como se puede ver en la Figura 2.2, que en conjunto cumplen el rol doble del OAIS: preservar y proveer acceso a la información custodiada. (Ospina Torres, 2014)

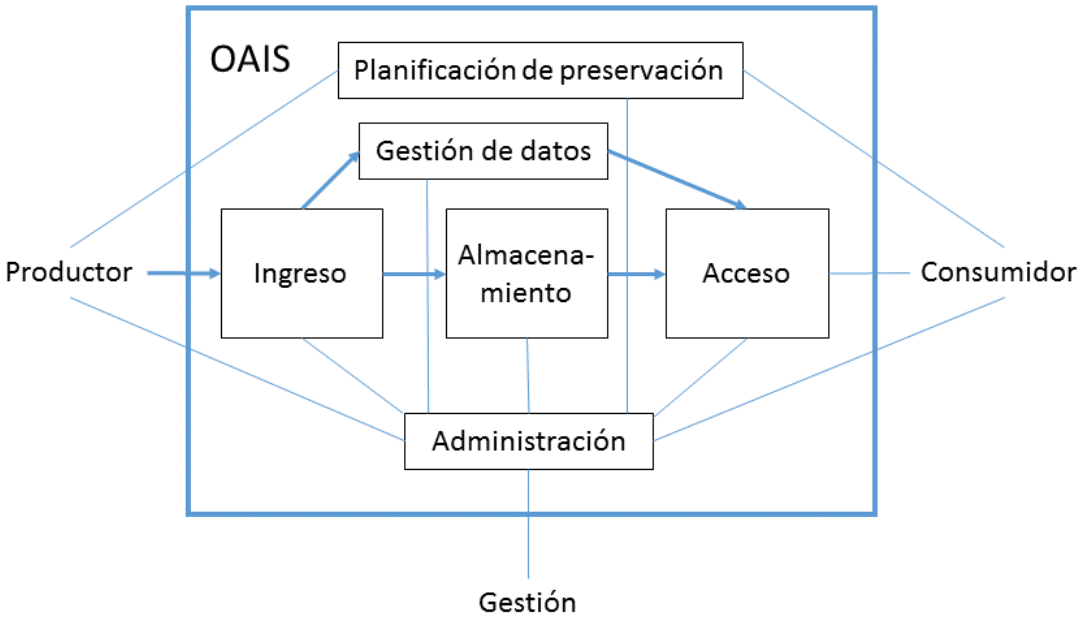

Figura 2.2: Modelo funcional OAIS.
Fuente: Traducido de Lavoie, 2014, p. 12.

- Ingreso: acepta presentaciones de los productores y los prepara para el almacenamiento y la gestión dentro del archivo.
- Almacenamiento: para el almacenamiento, mantenimiento y recuperación de contenido del Archivo.
- Gestión de datos: maneja repositorios de los metadatos que describe la información almacenada en el Archivo, realiza la gestión de información sobre el Archivo y sus propiedades.
- Administración: maneja las operaciones cotidianas del Archivo y coordina las actividades de los demás componentes funcionales del OAIS.
- Planificación de preservación: vigilancia del medio ambiente de la OAIS para garantizar la conservación a largo plazo de los contenidos del Archivo.
- Acceso: el apoyo a los consumidores (usuarios) en la búsqueda y recuperación de contenidos del Archivo.

2.4.3. Modelo de información

El modelo de referencia también provee una descripción de alto nivel de los objetos de información manejado por el archivo. El modelo de información OAIS está construido alrededor del concepto de un paquete de información (IP por sus siglas

en inglés, *Information Package*). Un paquete de información consiste del objeto que es el foco de preservación, junto con los metadatos necesarios para soportar su preservación a largo plazo, acceso y comprensibilidad, enlazados en un único paquete lógico. Existen tres variantes importantes del concepto de paquete de información. (Lavoie, 2014)

Paquetes de información

- El *Submission Information Package*, o SIP, es la versión del IP que es transferida del Productor al OAIS cuando la información es ingerida dentro del archivo. El concepto de SIP enfatiza el hecho de que la información puede no ser preservada en la forma exacta en la cual es entregada por el Productor.
- El *Archival Information Package*, o AIP, es la versión del IP que es almacenada y preservada por el OAIS. El AIP consiste de la información que es objeto de preservación, acompañada de un completo conjunto de metadatos suficientes para soportar los servicios de preservación y acceso del OAIS.
- El *Dissemination Information Package*, o DIP, es la versión del IP entregada al consumidor en respuesta a una petición de acceso. Enfatiza el hecho de que el IP diseminado por el OAIS al consumidor puede diferir en forma o en contenido al que reside en el almacenamiento en el archivo.

En la figura 2.3 se muestran como los diferentes tipos de paquetes de información interactúan con los componentes de la arquitectura.

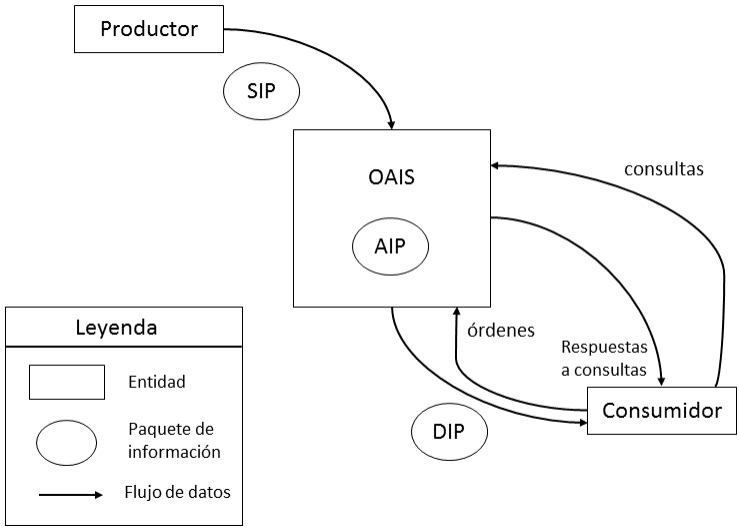

Figura 2.3: Datos externos de un Archivo OAIS.
Fuente: Traducido de CCSDS, 2012, p. 2-8.

2.4.4. Metadatos para la preservación

En primer lugar los metadatos son definidos, por la *National Information Standards Organization (NISO)* (2004), como información estructurada que describe, explica, localiza o, de cierta manera, facilita la obtención, uso o manejo de algún recurso.

Los metadatos pueden describir recursos en cualquier nivel de agregación. Pueden describir una colección, un solo recurso o un componente parte de un recurso más grande. También, los metadatos pueden ayudar a organizar recursos electrónicos, facilitar la interoperabilidad y la integración de recursos heredados (del inglés *legacy*), proveen identificación digital, y soportan el archivado y preservación de dichos recursos.

En primer lugar, la definición clásica de metadatos es literal, basada en la etimología de la palabra misma: los metadatos son "datos sobre datos". Las páginas web a menudo tienen metadatos incrustados. Los enlaces de una página Web a otras y los registros de comportamiento del usuario —selección de páginas individuales para verlas entre listas de resultados de búsqueda, por ejemplo— son también tipos de metadatos.

El modelo de información OAIS describe un amplio conjunto de requisitos de metadatos necesarios para soportar las actividades de un archivo de tipo OAIS. Los metadatos de preservación son “metadatos que soportan el proceso de preservación digital a largo plazo", incluyendo información sobre la procedencia, los derechos de propiedad intelectual y el entorno técnico e interpretativo de un objeto digital archivado (Lavoie & Gartner, 2013, p. 5).

Tipos de Metadatos

Existen diferentes tipos de metadatos, los principales según Riley (2004) son:

- **Metadatos descriptivos**: con propósitos de descubrimiento e identificación, este tipo de metadatos describe un recurso, pudiendo incluir un resumen, autor y palabras claves.
- **Metadatos estructurales**: indica como los objetos compuestos están ensamblados, como el orden de las páginas dentro de un libro.
- **Metadatos administrativos**: Provee información sobre el recurso para facilitar el manejo del mismo, tal como cuándo y cómo fue creado, tipo de archivo y otros detalles técnicos, y quién puede accederlo. Este tipo de metadato a su vez se divide en dos grandes grupos:
    - Metadatos para el manejo de derechos, el cual maneja los derechos de propiedad intelectual.

- Metadatos para la preservación, el cual contiene información requerida para archivar y preservar el recurso.

Una razón por la que los metadatos de preservación es difícil de categorizar con precisión es que no encajan perfectamente dentro de categorías bien conocidas como las anteriormente descritas, en cambio, pueden extenderse a través de las tres.

Una manera de aclarar el alcance de los metadatos de preservación es centrarse en cuál es su importancia. Los metadatos de preservación son importantes porque facilitan el proceso de lograr los objetivos generales de la mayoría de los esfuerzos de preservación digital: mantener la disponibilidad, identidad, persistencia, capacidad de representación, comprensión y autenticidad de los objetos digitales durante largos períodos de tiempo.

Así Lavoie y Gartner (2013, p. 5) explican que, fundamentalmente, los metadatos de preservación establecen un marco informativo de referencia alrededor de un objeto digital preservado que permanece conectado a ese objeto en el tiempo. La idea básica es que mantener la capacidad de explotar el valor completo de un objeto digital preservado en el futuro requiere preservar este marco de referencia en forma de metadatos de preservación bien mantenidos.

2.5. Adaptación del modelo OAIS a Archivos Web

En Masanès (2006) se hace una adaptación del modelo funcional OAIS por la IIPC, definiendo las tareas que deben ser llevadas a cabo por los Archivos Web para cumplir su objetivo de preservación, las cuales son descritas a continuación:

1. **Selección de las páginas o sitios web a resguardar**: permite limitar el ámbito del archivo, pudiendo preservar contenidos locales o de un tipo en particular, como por ejemplo, contenidos de un país o educativos solamente.
2. **Adquisición regular del contenido de dichas páginas**: logra que se puedan almacenar los cambios generados sobre los contenidos que se preservan a través del tiempo.
3. **Almacenamiento e indexación de las páginas resguardadas**: requiere estrategias que permitan preservar grandes volúmenes de información (del orden de los Terabytes), millones de archivos y diferentes formatos. Para este fin, se han desarrollado formatos de archivos contenedores específicos, cuyo objetivo principal es superar la limitación de los sistemas de archivos propios de los sistemas operativos donde se alojan los Archivos Web.
4. **Recuperación o consultas sobre la información resguardada**: el acceso o recuperación de los contenidos está estrechamente ligado a la forma en que se

encuentran almacenados, pero debido a la naturaleza hipertextual y multimedia de la web, se espera que el usuario final pueda acceder a este contenido de manera similar a cuando lo hace en los servidores originales.

La arquitectura funcional para Archivos Web propuesta por el IIPC basada en el modelo OAIS puede verse en la figura 2.4. Allí se agrupan una serie de herramientas que dan soporte a las tareas previamente definidas, de acuerdo al rol que tienen dentro del Archivo Web.

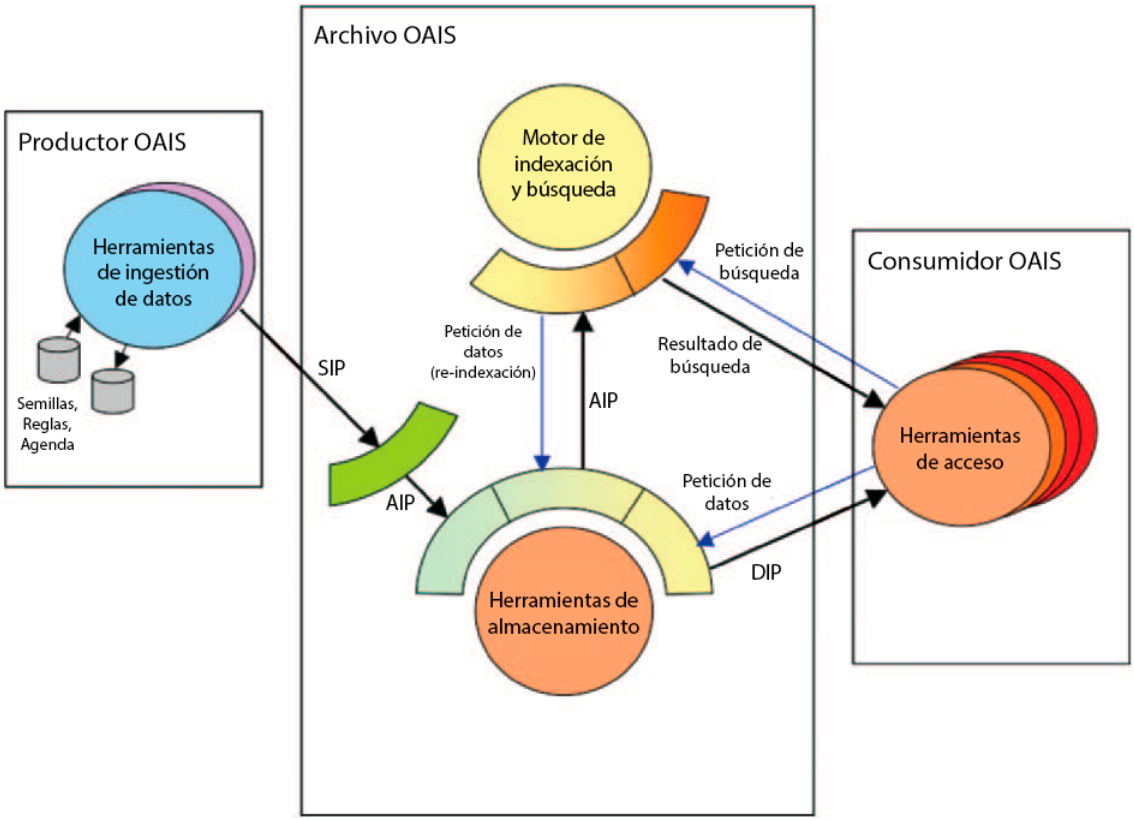

Figura 2.4: Arquitectura funcional del IIPC basada en modelo OAIS.
Fuente: Traducido de Masanès, 2006, p. 144.

- Las herramientas de ingestión de datos se encargan de la adquisición regular de los sitios web seleccionados y sus documentos web asociados.
- El proceso de selección, aunque forma parte de las políticas y lineamientos de cada Archivo, es soportado por las herramientas de ingreso a través de un repositorio de los sitios a preservar y la frecuencia de cambio para la preservación de nuevas versiones.
- Para la indexación de los documentos se hace uso de la tecnología de motores de búsqueda e indexación desarrolladas por los buscadores de internet como Google, Bing y Yahoo.

- El almacenamiento es llevado a cabo por las herramientas de almacenamiento y los manejadores de contenido.
- El acceso a los datos se hace a través de las herramientas de acceso.

Por otro lado, en Ospina Torres (2014) se definen los diferentes elementos de información (entrada, contenido almacenado y salida) en un Archivo Web y su relación con el modelo de información OAIS. Los mismos son descritos a continuación en la tabla 2.1.

Tabla 2.1: Elementos de Información de un Archivo Web.

Elementos de Información	**Descripción**	**Modelo de información Modelo OAIS**
Semillas, frecuencia de rastreo	Información que permite adquirir de manera regular el contenido a almacenar	SIP
Archivo de configuración	Archivo usado por el rastreador para realizar la adquisición	
Datos de localización geográfica del URL	Datos que permiten saber la ubicación geográfica y el ip del sitio web rastreado	
Archivos WARC	Paquete de información que contiene los contenidos preservados por sitio web	AIP
Colecciones	Agrupaciones de URL de sitios web en categorías o temas que dependen del contenido del sitio web	Grupos de AIP
Log de Heritrix	Información descriptiva de la preservación	Metadata
Registros WARC descriptivos		
Documentos web dentro de WARCS, que conforman los sitios web (HTML, CSS Imágenes, Scripts, Otros)	Información a ser desplegada al usuario final	DIP

Fuente: Ospina Torres, 2014, p. 50.

2.6. Herramienta de rastreo Heritrix

Heritrix es un rastreador web diseñado para el archivado web. Es un proyecto de código abierto escrito en Java por el Internet Archive. Se puede acceder a la interfaz principal mediante un navegador web, y existe una herramienta de línea de comandos que puede usarse opcionalmente para iniciar rastreos.

Por defecto Heritrix guarda los recursos web que rastrea en un archivo WARC, que almacena múltiples recursos archivados en un solo archivo para evitar administrar una gran cantidad de archivos pequeños. También se puede configurar para almacenar archivos en un formato de directorio similar al rastreador WGET que usa la URL para nombrar el directorio y el nombre de archivo de cada recurso.

2.6.1. Principales características

La versión 3.0 de Heritrix ofrece las siguientes características:

- Posibilidad de ejecutar varios trabajos de rastreo simultáneamente. El único límite en el número de trabajos de rastreo que se pueden ejecutar simultáneamente es la memoria asignada a Heritrix.
- Archivo de configuración XML único basado en el framework Spring.
- Posibilidad de navegar y modificar los "beans"de Spring configurados a través de un navegador.
- Extensibilidad mejorada a través del framework Spring. Por ejemplo, las anulaciones de dominios se pueden establecer en un nivel de grano muy fino.
- Consola de control de usuario más segura. Se utiliza HTTPS para acceder y manipular la consola de control de usuario.
- Mayor escalabilidad. Anteriormente, los rastreos con valores de semillas grandes (decenas o cientos de millones) podrían intentar utilizar más memoria que la asignada a Heritrix. Esto causaría la falla total del rastreador.
- Mayor flexibilidad al modificar un rastreo en ejecución. La ejecución de rastreos se puede modificar mediante el Explorador o mediante el Directorio de acciones.
- Introducción de colas paralelas. Al rastrear sitios específicos que pueden manejar grandes cantidades de tráfico, la opción de colas paralelas puede utilizarse para abrir muchas conexiones de rastreo simultáneas a un solo sitio.
- Una consola de secuencias de comandos que acepta la entrada de secuencias de comandos en varios formatos, como AppleScript y ECMAScript. Las se-

cuencias de comandos se pueden utilizar para acceder y manipular los componentes básicos de Heritrix de forma programática.

2.6.2. Archivos de salida

Los siguientes archivos son generados por Heritrix por cada rastreo ejecutado, en adición a los logs. Parte de la información en ellos también está disponible en la interfaz de usuario web.

surts.dump Este archivo contiene la forma SURT de los URI de las semillas. Donde SURT significa *Sort-friendly URI Reordering Transform*, y es una transformación aplicada a URIs que hace que su representación de izquierda a derecha coincida mejor con la jerarquía natural de los nombres de dominio.

negative-surts.dump Este archivo contiene la forma SURT de URIs que se deben excluir del rastreo.

crawl-report.txt Este archivo contiene métricas útiles sobre trabajos completados. El informe es creado por el `StatisticsTracker` bean. Este archivo se escribe al final del rastreo.

hosts-report.txt Este archivo contiene una descripción general de los hosts que se rastrearon. También muestra el número de documentos rastreados y los bytes descargados por host. Es creado por el `StatisticsTracker` bean y se escribe al final del rastreo.

mimetype-report.txt Este archivo contiene un informe que muestra el número de documentos descargados por tipo de mime. Además, se muestra la cantidad de datos descargados por tipo de mime. Es creado por el `StatisticsTracker` bean y se escribe al final del rastreo.

processors-report.txt Este archivo muestra la actividad de cada procesador Heritrix. Es creado por el `StatisticsTracker` bean y se escribe al final del rastreo.

responsecode-report.txt Este archivo contiene un informe que muestra el número de documentos descargados por código de estado. Sólo cubre códigos exitosos. Para códigos de error ver el archivo `crawl.log`. Es creado por el `StatisticsTracker` bean y se escribe al final del rastreo.

seeds-report.txt Este archivo contiene el estado de rastreo de cada semilla. Es creado por el `StatisticsTracker` bean y se escribe al final del rastreo.

frontier-summary-report.txt Este informe contiene un desglose de la actividad de la frontera en una base por hilo. Para cada subproceso en ejecución, se puede examinar el estado de la cola de frontera.

source-report.txt Este informe contiene un elemento de línea para cada host, que incluye la semilla desde la cual se llegó al host. La propiedad `sourceTagSeeds` del módulo `TextSeedModule` debe establecerse en `true` para que se genere este informe.

threads-report.txt Este informe contiene la lista de subprocesos que estaban activos al final del rastreo. También hay información detallada sobre cada hilo.

Archivos WARC Suponiendo que se esté utilizando el escritor WARC que viene con Heritrix, se generarán varios archivos WARC que contienen el contenido rastreado. Se puede especificar la ubicación de almacenamiento de los archivos WARC configurando el valor de `directory` del bean `WARCWriterProcessor`. Los archivos WARC se nombran usando como convención la siguiente plantilla con interpolación de variable:

```
${prefix}-${timestamp17}-${serialno}-${heritrix.pid}~
${heritrix.hostname}~${heritrix.port}
```

No se recomienda cambiar la plantilla a menos que el sistema de nombres alternativo también genere nombres únicos. Los archivos WARC con un sufijo `.open` están en proceso de ser escritos por Heritrix. Puede haber múltiples WARC abiertos en un momento dado. Los archivos WARC con un sufijo `.invalid` indican problemas al escribir en el archivo. Esto puede ser el resultado de un disco defectuoso o un disco completamente utilizado. En un problema de E/S, Heritrix cierra el archivo WARC problemático y le da un sufijo `.invalid`. Estos archivos deben ser revisados por coherencia.

2.6.3. Formato WARC

El formato WARC (*Web ARChive*) es un contenedor de archivos que permite concatenar múltiples registros de recursos (objetos de datos), cada uno compuesto de un conjunto de cabeceras de texto simple y un bloque de datos arbitrario en un archivo largo (ISO, 2009). En la Figura 2.5 se muestra el formato de un registro WARC perteneciente a un archivo WARC.

Según el ISO (2009), se espera que permita gestionar y almacenar billones de recursos recolectados en la Web o de cualquier otra fuente, siendo un estándar de estructura. Éste será utilizado para construir aplicaciones que obtengan, gestionen y permitan el acceso e intercambio de contenido, tales como el rastreador web, de código abierto, *Heritrix*. El modo en el que los archivos WARC van a ser creados, y los recursos almacenados y prestados, dependerá de la implementación del software y las aplicaciones.

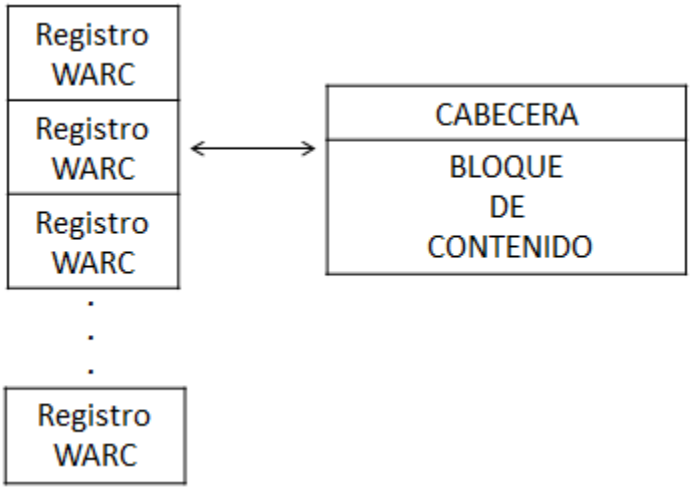

Figura 2.5: Formato de Registro WARC.
Fuente: ISO, 2009.

2.7. Inteligencia de Negocios

Según Cano (2007) la inteligencia de negocios es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un *datawarehouse*), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones. El proceso de *Business Intelligence* incluye la comunicación de los descubrimientos y efectuar los cambios.

2.7.1. Características

Según Cano (2007), toda solución de Inteligencia de Negocios debe cumplir con las siguientes características:

- Visión unificada de los datos: todos los datos deben estar localizados en un único repositorio de datos, sin importar el tipo de datos o la fuente de donde provenga, para así dar la sensación de que los datos están centralizados.
- Creación personalizada de informes y consultas: permite el desarrollo de consultas y reportes a la medida sobre información contenida en los Almacenes de Datos.
- Vistas gráficas e interactivas para la presentación de información analítica: a través de cuadros de mandos integrales y estratégicos se facilita la visualización de los indicadores de negocio.
- Capacidad de procesamiento de grandes volúmenes de datos: las soluciones

de BI permiten realizar consultas comparando los datos actuales con los históricos.

2.7.2. Funciones de una Solución de Inteligencia de Negocio

A continuación veremos las principales funciones que caracterizan a una solución de Inteligencia de Negocio. (Cano, 2007)

- Permiten reunir, estandarizar y centralizar toda la información de la empresa, mediante un almacén de datos, permitiendo así su explotación sin esfuerzo.
- Posibilita la extracción de información de los datos y el conocimiento de la información, con la utilización del software adecuado.
- Permiten el perfeccionamiento de las consultas de alto nivel, realizando las transformaciones oportunas a cada sistema (OLTP - OLAP), y liberando los servidores operacionales.

2.7.3. Arquitectura de una Solución de Inteligencia de Negocio

Una arquitectura típica de una Solución de Inteligencia de Negocio consta de cuatro elementos fundamentales, como se muestra en la figura 2.6, los cuales Cano (2007) describe como sigue a continuación.

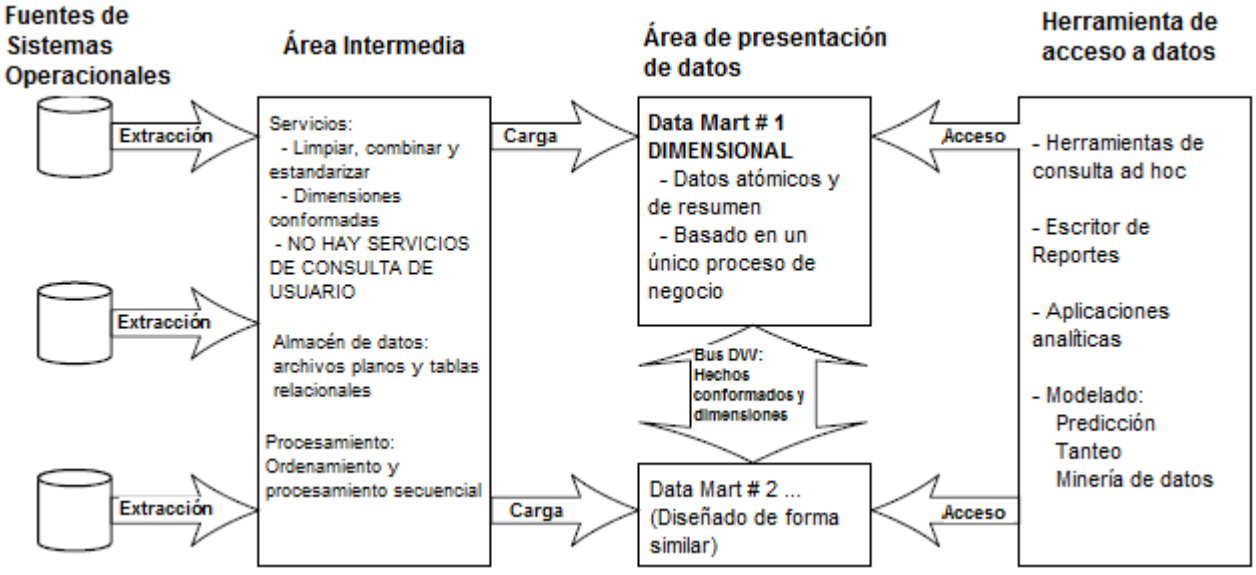

Figura 2.6: Arquitectura de una solución de Inteligencia de Negocios
Fuente: Kimball y Ross, 2002.

Fuentes de Datos

Son los datos que alimentan de información el almacén de datos, los cuales salen de los sistemas operacionales o transaccionales, incluyendo aplicaciones desarrolladas a la medida, ERP, CRM, SCM, entre otros sistemas, con información relativa a la actividad rutinaria de la empresa y de los sistemas de información departamentales, incluyendo previsiones, presupuestos y hojas de cálculo, que se requieren para el análisis del negocio.

Extracción, Transformación y Carga

Este proceso se conoce como ETL por sus siglas en inglés *Extract, Transformation, Load* y se estima que a menudo consume un 70 % del tiempo y esfuerzo para la construcción de un almacén de datos y se utiliza para migrar datos de un punto a otro.

Área Intermedia

El área intermedia es un espacio temporal y de carácter volátil, sobre el cual se van a ejecutar los procesos de ETL. Se usa para hacer una primera extracción rápida de las fuentes de datos y almacenarlos temporalmente mientras se analizan, limpian, mejoran y, posteriormente, se carga al almacén de datos, ya que generalmente, la información que se tiene en los sistemas transaccionales no está preparada para la toma de decisiones, por lo tanto se busca almacenar los datos de una forma que maximice su flexibilidad, facilidad de acceso y administración.

Almacén de Datos

Según Inmon (1996) un almacén de datos es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia. Por otro lado, para Kimball y Ross (2002) un almacén de datos son muchas copias del sistema operacional y de registros almacenados en una plataforma de hardware por separado, los cuales son utilizados para consultas y análisis.

Características de un Almacén de Datos El almacén de datos es el corazón del entorno arquitectónico y es la base de todo el proceso de DSS y según Inmon (1996) estas son las principales características:

1. **Orientado a Temas** Una de las características principales de un almacén de datos, es que es orientado a temas, esto quiere decir que un almacén de datos es clasificado por áreas temáticas basados en los elementos que son de interés para la empresa. Por esta razón el modelado es orientado a realizar consultas eficientes con respecto a la información de las actividades de la organización como por ejemplo en la corporación aseguradora pueden ser clientes, política, prima y reclamos, para un fabricante las principales áreas de interés suelen ser productos, órdenes, vendedores, facturas de los materiales y materia prima. Estos temas pueden verse como un conjunto de indicadores o medidas que son de interés para la empresa.

   Por el contrario, los sistemas operacionales almacenan los datos por aplicaciones individuales, proporcionando sólo los datos necesarios para cada requerimiento de cada sistema por separado, de forma tal que los mismos puedan ejecutar funciones específicas eficientemente. Estas aplicaciones están relacionadas con el diseño de la base de datos y del proceso. La diferencia entre la orientación de procesos o funciones de las aplicaciones y la orientación a temas, radica en el nivel de detalle e integración del contenido. Cuando se diseña y modela un almacén de datos, se debe excluir la información no relevante para el proceso de Sistemas de Soporte de Decisiones, mientras que la información que se encuentra contenida en las orientadas a las aplicaciones, posee datos para satisfacer de inmediato los requerimientos funcionales y de proceso, que no necesariamente serán usados por el analista de soporte de decisiones.

2. **Integrada** De todos las características de un almacén de datos, la integración es la más importante. Los datos se alimentan de múltiples fuentes diferentes en el almacén de datos. A medida que se alimentan los datos, se convierte, se reforma, se resume y así sucesivamente. El resultado es que los datos (una vez que ya son almacenados) tienen una única imagen corporativa física.

3. **No Volátil** La tercera característica es que el almacén de datos sea no volátil, es decir que los datos deben mantenerse invariables. Los datos en los Almacenes de Datos son cargados (Usualmente en grandes cantidades) y accesados, pero generalmente no son actualizados ni eliminados, y están cargados como foto periódica, de forma estática. Cuando se producen cambios después de la última carga, se crea un nuevo registro de foto periódica. De este modo, se guarda un registro histórico de los datos en el almacén de datos.

4. **Variable en el tiempo** La última característica de un almacén de datos es que es variable en el tiempo, esto implica que los datos que son extraídos desde los sistemas transaccionales son archivados, y por consiguiente históricos. Cada vez que se hace una nueva carga periódica los datos anteriores no son eliminados o reemplazados, estos se mantienen en el tiempo para poder hacer comparaciones y generar conocimiento.

2.7.4. Modelo Dimensional

El modelo dimensional permite tener los datos organizados entorno a hechos que tiene unos atributos o medidas, que pueden verse con mayor o menor detalle según ciertas dimensiones. (Kimball & Ross, 2002) Los conceptos mas importantes que se manejan en un modelo dimensional según Kimball y Ross (2002) y Ponniah (2001) son:

- **Hecho:** Un Hecho es una medición del negocio distinta a un atributo, tiene carácter dinámico con el fin de realizar estudios. Físicamente se representan dentro de una Tabla de Hechos y es el punto central para la toma de decisiones, es una relación multiclave es decir que cada una de las claves referenciadas está relacionada con una dimensión y la unión de las mismas compone la clave primaria de la tabla. Algunos tipos de tablas de hechos son: transaccional, de foto acumulada, de foto periódica, entre otros. (Kimball & Ross, 2002)
- **Medidas:** Son el conjunto de indicadores del hecho que se escogió para representar. Generalmente responden a la pregunta ¿Cuánto? Retomando el ejemplo anterior, las medidas para el hecho de las ventas podrían ser: ¿Cuántos productos se vendieron?, ¿Cuánto fue el total de la venta en dólares?, ¿Cuánto costaron esos productos vendidos?. (Ponniah, 2001)
- **Dimensiones:** Las dimensiones representan las diferentes formas de visualizar la información que se encuentra asociada a un hecho de acuerdo al nivel de detalle utilizado, además determina el nivel de detalle de los datos. Se representa físicamente con una Tabla Dimensión que se encuentra compuesta por una clave primaria y un conjunto de atributos descriptivos que permiten dar sentido a lo almacenado. Algunos tipos de tabla dimensión son: dimensión conformada, dimensión degenerada, dimensión role-playing, entre otros. (Kimball & Ross, 2002)
- **Granularidad:** Se refiere al nivel de detalle de los datos dentro del almacén de datos. A mayor nivel de granularidad se tiene menos detalle de los datos y a menor nivel de granularidad se tiene mayor detalle. (Kimball & Ross, 2002)
- **Jerarquía:** Es una relación en cascada de uno a muchos y está asociada a la ubicación de un atributo con respecto a otro. Una ejemplo frecuente de la jerarquía puede es la de Categoría - Subcategoría - Producto. (Kimball & Ross, 2002)
- **Esquema:** Es la representación genérica de un modelo multidimensional en una base de datos relacional, donde una tabla de hechos está unida a varias dimensiones. Existen diferentes tipos de esquemas, por ejemplo: tipo estrella en que las dimensiones están des-normalizadas, el copo de nieve en que las dimensiones están en tercera forma normal y el esquema constelación que consiste en la unión de esquemas estrella y/o copo de nieve con común. (Kimball & Ross, 2002)

2.7.5. Ventajas y desventajas

Ventajas Según Laudon y Laudon (2012), Las ventajas de una solución de Inteligencia de Negocios son las siguientes:

- Disposición de la información correcta, en el momento adecuado, para la toma de decisiones. La información está almacenada de manera sencilla y actualizada
- Provee la capacidad para evaluar distintos escenarios simultáneamente, para su análisis, y apoyo a la toma de decisiones preventivas.
- Permite definir indicadores, que miden el desempeño del negocio, a partir de los datos operacionales recolectados.
- Permite agrupar información de distintas áreas en un solo cuadro, favoreciendo el cambio de políticas o reorientación de los planes establecidos.
- Genera capacidad de reacción a situaciones imprevistas, con un nivel de riesgo menor,ya que permite analizar con anticipación el riesgo que se tendría al tomar ciertas decisiones.
- Ajustable a cambios organizacionales, es decir, permitir que el conocimiento se almacene y no sea necesario retransmitirlo directamente a las personas cuando ocurre un cambio de cargo.

Desventajas Según Laudon y Laudon (2012), Las desventajas de una solución de Inteligencia de Negocios son las siguientes:

- La creación e implantación de una solución de BI suele ser difícil para una organización, ya que se puede generar resistencia al cambio por parte de la organización.
- Falta de compromiso por parte de los sponsors, es decir, aquellos que tienen autoridad en la empresa.
- Se tiene poca disponibilidad de los representantes de negocios.
- Hay ausencia de un personal disponible y hábil.
- La base de datos de la organización se encuentra en una copia digital, vulnerable a la pérdida o al uso indebido de la información de la organización.
- No existe una apreciación del impacto que causan los datos de mala calidad en la rentabilidad del negocio.
- Es posible que la solución de BI no cumpla con las expectativas de los usuarios, por extensos tiempos de respuesta o insuficiente capacidad de cómputo.

2.8. Indicadores

Según Fuentes y Marquez Gallo (2003), que cita a J.C. Pacheco (2002) un **indicador** es una relación entre dos o más datos significativos, que tienen un nexo lógico entre ellos, y que proporcionan información sobre aspectos críticos o de importancia vital, para la conducción de la empresa

Los **indicadores de gestión**, son parámetros que sirven para medir resultados de acciones definidas, en donde las relaciones se pueden obtener de: el procesamiento de datos básicos de cada sistema organizacional y de la combinación de datos de dos sistemas. (Fuentes & Marquez Gallo, 2003)

2.8.1. Objetivos

Según la Asociación Española para la Calidad (AEC) (2013), los indicadores tienen como objetivo aportar a la empresa un camino correcto para que logre cumplir con las metas establecidas, teniendo que cumplir los siguientes objetivos al ser un sistema de medición de objetivos para la empresa:

- Comunicar la estrategia.
- Comunicar las metas.
- Identificar problemas y oportunidades.
- Diagnosticar problemas.
- Entender procesos.
- Definir responsabilidades.
- Mejorar el control de la empresa.
- Identificar iniciativas y acciones necesarias.
- Medir comportamientos.
- Facilitar la delegación en las personas.
- Integrar la compensación con la actuación.

2.8.2. Partes de un Indicador

Según Bernal (2013), a la hora de definir los indicadores, hay que fijar una serie de parámetros para cada uno de ellos. Las partes esenciales que deben definirse junto al indicador son las siguientes:

- Definición: Describe concretamente lo que se está midiendo. Ejemplo: Consumo eléctrico en una vivienda durante cada mes.
- Forma de calcularlo: Es la fórmula o ecuación que se usará para obtener el dato. Ejemplo: Si medimos un porcentaje de defectos, su fórmula será 100*(unidades defectuosas/unidades totales).
- Unidades: Junto al valor, se deben especificar las unidades en que se está midiendo. En el ejemplo del consumo eléctrico, las unidades más comunes serían kWh. En el ejemplo del porcentaje de defectos, las unidades son el tanto por ciento.
- Periodicidad: Debe fijarse cada cuánto se va a medir. Si el indicador es clave para el buen funcionamiento se deberá medir y controlar más frecuentemente que un indicador secundario menos importante.
- Proceso: La actividad o proceso que está asociado al indicador.
- Responsable: El departamento o persona que es responsable del proceso o la actividad que se está midiendo.

Sobre los resultados del indicador, debemos compararlos con un valor preestablecido: Un objetivo, una expectativa y/o un limite.

- Objetivo: Valor que queremos alcanzar. Este debe ser ambicioso, alcanzable, estar cuantificado y acotado en el tiempo.
- Expectativa: Es el valor ideal del indicador, aunque no siempre es alcanzable.
- Límites legales: Es el límite que nos impone la ley, y que no podemos propasar. Es diferente a los objetivos, porque el objetivo marca un propósito voluntario fijado por nosotros, y el límite legal es un valor que estamos obligados a cumplir.
- Límite de aceptabilidad: Aparte de lo anterior, también se puede fijar un valor límite para considerar que el proceso funciona bien. Conociendo cuál es el funcionamiento normal del proceso, fijamos un valor, por debajo del cual asumiremos el el proceso está funcionando mal y deberemos tomar acciones.

Otros conceptos que deberían definirse, pero que no siempre se hace, son estos:

- Propósito del indicador: ¿Por qué medimos este dato? ¿Para qué sirve esta medición? Todos los indicadores deben tener un propósito lo suficientemente argumentado como para que lo que ganamos obteniendo ese dato sea más valioso que el tiempo que perdemos en medirlo.
- Grupos de interés: ¿A quién beneficia que estemos controlando el aspecto medido por el indicador? Pueden ser grupos de interés los clientes, los proveedo-

res, los empleados, la dirección, los accionistas, el entorno, etc.

- Destinatarios: ¿Quién va a recibir y revisar los datos del indicador? Por lo general, los destinatarios suelen ser los responsables del proceso, los jefes de sección, y la dirección ya que muchas veces los datos se toman y luego no son observados, o no está definido quién debe tomar las decisiones ante un problema.
- Soporte: ¿En qué formato se va a almacenar? ¿Quién va a recopilar los datos? ¿Cómo se va a distribuir? Lo más común es almacenarlos en Excel o PDF y enviarlos a sus destinatarios por email, impresos o en una carpeta compartida

2.8.3. Tipos de Indicador

En teoría, se pueden establecer indicadores para cualquier aspecto medible, pero en el contexto de orientación a los procesos, podemos encontrar los siguientes tipos de indicadores (Beltrán Jaramillo, 2006):

- Indicadores de Procesos: están relacionados con el conjunto de actividades que forman parte del proceso.
- Indicadores de Resultados: se refiere al comportamiento del proceso como un todo.
- Indicadores de Eficacia: representan el cociente entre producción real y la esperada, independiente de los recursos utilizados para lograrlo.
- Indicadores de Eficiencia: el concepto de eficiencia se refiere al grado de cumplimiento de los objetivos planteados, sin garantizarlo, es decir, en qué medida la organización, está cumpliendo con sus objetivos tomando en cuenta los recursos con los que cuenta.
- Indicadores de Gestión: son medidas utilizadas para determinar el nivel de cumplimiento de los objetivos de una actividad perteneciente a un proyecto o del proyecto en sí. Los indicadores de gestión están relacionados con la administración de un proceso o actividad.

2.8.4. Importancia de los Indicadores

Mondragón (2014) señala cual es la importancia de los indicadores en los siguientes puntos:

- Elemento de planificación: durante los procesos de planificación se utilizan con frecuencia los indicadores para establecer la meta u horizonte a donde se quiere llegar.

- Estándar de seguimiento y control: el indicador ayuda a entender o muestra el estado del problema, ayuda a determinar la brecha entre lo planificado o esperado y el punto actual en el que se hace la valoración o medición. Un indicador es una señal de alerta que induce a reconocer que es necesario resolver un problema.
- Herramienta para la toma de decisiones: permite establecer métricas a través de las cuales se demuestre el cumplimiento de un objetivo o una meta en determinado proceso, proporcionando la información de apoyo para la toma de decisiones, así como también el planteamiento de políticas y estrategias para solucionar el problem.

2.8.5. Antecedentes de uso de indicadores en la preservación Web

A medida que se van involucrando más instituciones en el archivado web, una necesidad mundial surgió por directrices en la administración y evaluación de los productos y actividades del Archivo Web. Es por esto que, en 2009, el Comité Técnico 46 de la ISO (la división de información y documentación) decidió crear un grupo que trabajara en "Estadísticas e Indicadores de Calidad para Archivos Web".

En Octubre de 2012, el comité entregó un borrador del informe técnico ISO (2009) para que fuera evaluado y sometido a votación por los diferentes cuerpos pertenecientes a la ISO. Adicionalmente, hicieron público dicho borrador para obtener retroalimentación de una comunidad mayor, publicándolo en la página oficial del Consorcio Internacional de la Preservación del Internet (IIPC).

En 2013 fue publicado el documento ISO oficial, bajo el nombre ISO/TR 14873:2012, haciendo énfasis en que este reporte técnico no avala ni recomienda el uso de alguna aplicación en específico, aunque el uso de algunos puede ocasionar variaciones en los resultados. Adicionalmente, también resaltan que dentro de dicho reporte técnico se enfocan en principios y métodos para el Archivado Web, pero no abarca opciones alternativas para recolectar recursos del Internet. (Lopez & Sarno, 2015)

2.8.6. Datos Estadísticos e Indicadores de Calidad

En las tablas presentadas a continuación, se muestran todos los datos estadísticos e indicadores definidos para los Archivos Web, obtenidos del borrador del ISO/TR 14873:2013, y que son clasificados de la siguiente manera:

- Datos estadísticos para el desarrollo de una colección.
- Datos estadísticos para la caracterización de una colección.
- Datos estadísticos sobre el uso de una colección.

- Preservación del Archivo Web.
- Costo del Archivo Web.
- Indicadores de calidad.

Datos estadísticos para el desarrollo de una colección

Los datos estadísticos mostrados en la Tabla 2.8.6, son los considerados como principales y que todo Archivo Web debe poder obtener, ya que describen el estado actual de éste.

Tabla 2.2: Datos estadísticos principales para el desarrollo de una colección.

Dato Estadístico	Propósito	Ejemplo
Número de objetivos [1]	Objetivos de colección / resultados cuantitativos	8.000 objetivos
Número de capturas de objetivo [2]	Objetivos de colección / datos cuantitativos	14.000 rastreos de un objetivo
Tiempo de selección de objetivo [3]	Objetivos de colección / datos cuantitativos	2 horas
Número de URLs	Resultados cuantitativos	14 mil millones de URLs
Distribución de URLs por códigos de estatus [4]	Tipo / el número de recursos	2 millones de recursos rastreados con éxito
Número de dominios o hosts	Datos cuantitativos	3 millones de nombres de dominio
Tamaño en bytes (comprimido y descomprimido)	Datos cuantitativos	200 terabytes sin comprimir; 160 terabytes comprimidos
Número de archivos WARC	Resultados cuantitativos	18.000 archivos WARC

Fuente: ISO (2012)

[1]Un objetivo es un conjunto de recursos a ser recolectados y su alcance puede variar de recursos interrelacionados en el mismo dominio, presentado como una página web, a un único recurso. Cada rastreo es la captura de un objetivo. Para el AWV son las semillas

[2]Rastreos realizados

[3]Tiempo utilizado por el suscriptor para determinar un objetivo

[4]Respuesta del servidor al solicitar un recurso

Datos estadísticos para la caracterización de la colección

Las estadísticas propuestas en esta sección (mostrados en la Tabla 2.8.6) describen las características de los archivos web y son útiles para ayudar a comprenderlos y tomar decisiones curatoriales informadas. Si bien algunas estadísticas son específicas para la recolección selectiva o masiva, otras son genéricas y aplicables a los archivos web establecidos utilizando ambas estrategias.

Tabla 2.3: Datos estadísticos principales para la caracterización de la colección.

Dato Estadístico	Propósito	Ejemplo	Comentario
Distribución por dominios de nivel superior (TLD) o segundo nivel	Distribución geográfica	70 % del Archivo alojado utiliza el TLD .fr	Se puede considerar que un archivo web que contiene una mayor proporción de TLDs que otros dominios tiene un alcance nacional. Puede ser medido en números absolutos o en porcentaje.
Distribución por volumen de recursos por dominio	Análisis de dominio	3 % de los dominios del Archivo contiene el 30 % del total de URLs	Puede ser medido en bytes (MB, GB, TB) o contar las URLs por dominio.
Distribución por tipos de formatos [5]	Caracterización de formatos	60 % de los recursos del Archivo son html/text	
Cobertura cronológica	Análisis temporal	El Archivo contiene recursos recolectados desde 1996 hasta hoy	Se puede usar en combinación con otras estadísticas, como el tamaño del archivo o la distribución de formatos de archivo, para mostrar la tendencia o el desarrollo del archivo a lo largo del tiempo.
Número de permisos concedidos	Productividad	60 % de los permisos solicitados han sido concedidos	Único de Archivos Web selectivos.
Número de nominaciones [6]	Productividad	30 % del Archivo es seleccionado manualmente	Único de Archivos Web selectivos.

Fuente: ISO (2012)

[5]Tipos MIME

[6]Cuando las semillas tienen que ser aprobadas por algún ente, luego de ser nominadas.

Datos estadísticos sobre el uso de una colección

Los datos estadísticos, son especialmente definidos para instituciones, como Bibliotecas Nacionales, que desean conocer aproximadamente el uso del Archivo Web por parte de los usuarios.

Tabla 2.4: Datos estadísticos básicos sobre el uso del Archivo Web.

Dato estadístico	Tipo	Cálculo	Importancia
Vista de página	Cantidad	El número de veces que una página fue vista.	Alta. Indicación del uso crudo del Archivo.
Visita (Sesiones)	Cantidad	Una visita es una interacción de un individuo con un sitio web, consistiendo en una o más solicitudes a una página.	Alta. Cuenta básica de los usuarios del Archivo.
Visitantes únicos	Cantidad	El número de individuos inferidos (filtrados de arañas o robots) dentro de un plazo de reporte designado, con actividad consistiendo de una o más visitas a un sitio. Cada individuo es contado solo una vez en la medida de visitante único para el periodo de reporte.	Media
Evento	Dimensión y/o Cantidad	Cualquier acción registrada que posea una fecha específica y un tiempo asignado por el navegador o el servidor.	Baja

Fuente: ISO (2012)

Tabla 2.5: Datos estadísticos para la caracterización avanzada del uso del Archivo Web.

Dato estadístico	Tipo	Cálculo	Importancia
Página de entrada	Dimensión	La primera página de una visita.	Media
Página destino	Dimensión	Vista de una página con la intención de identificar el inicio de la experiencia del usuario, como resultado de un esfuerzo de comercialización definido.	Baja

Página de salida	Dimensión	Última página de un sitio accedida en una visita, significando el final de una visita/sesión.	Baja
Duración de visita	Cantidad	La cantidad de tiempo en una sesión. El cálculo es típicamente la marca de tiempo de la última actividad en la sesión menos la marca de tiempo de la primera actividad de la sesión.	Alta
Referente	Dimensión	Referente es un término genérico que describe la fuente del tráfico a una página o visita.	Media
Página referente	Dimensión	Página referente describe la fuente de tráfico de una página.	Media
Visitante nuevo	Cantidad	Número de visitantes únicos con actividad. Nótese que "primera visita" es con respecto a cuándo los datos comenzaron a recopilarse apropiadamente mediante el uso de la herramienta actual.	Baja
Visitante que regresa	Cantidad	Número de visitantes únicos con actividades que consisten en una visita a un sitio durante un período de tiempo y donde el visitante único también visitó el sitio antes de dicho período.	Media
Visitante repite	Cantidad	Número de visitantes únicos con actividad que consiste de una o más visitas a un sitio en un periodo de tiempo reportado.	Media
Visita de una sola página (abandono)	Dimensión o Cantidad	Una visita que consiste en la visita a una sola página	Media
Porcentaje de abandono	Proporción	Visitas a una sola página dividida por páginas de entrada	
Páginas vistas por visita	Proporción	Número de páginas vistas en un período de reporte dividido por el número de visitas en el mismo período de reporte.	Alta. Indicador de la utilidad del archivo para el usuario
Visitantes por localización geográfica	Cantidad	Reporte de Geo-IP de la dirección IP del solicitante	Baja
Términos de búsqueda para encontrar el Archivo	Cantidad	Términos utilizados en los motores de búsqueda para encontrar el sitio web de la herramienta de acceso del Archivo Web	Media

Términos de búsqueda utilizados en el Archivo	Cantidad	Términos de búsqueda utilizados en la herramienta de acceso para encontrar capturas archivadas	Alta. Indicador de *temas actuales* al buscar rastreos.

Fuente: ISO (2012)

Tabla 2.6: Datos estadísticos principales para el uso de la colección.

Dato estadístico	Propósito	Ejemplo
Número de páginas vistas	Grado de uso	48.318 páginas en el Archivo Web de UK fueron vistas entre el 1 y el 30 de Junio del 2012.
Número de visitas	Grado de uso	Hubo 11.415 visitas al Archivo Web de UK entre el 1 y el 30 de Junio del 2012.
Número de visitantes no duplicados (contados una vez)	Grado de uso	Hubo 9.434 visitantes únicos al Archivo Web de UK entre el 1 y el 30 de Junio del 2012.
Duración de visita	Utilidad / Relevancia del Archivo	En promedio, cada visita al Archivo Web de UK entre el 1 y el 30 de Junio del 2012, duró 3 minutos y 25 segundos.
Página vista por visita	Utilidad / Relevancia del Archivo	4,23 páginas fueron vistas por visita en el Archivo Web de UK fueron vistas entre el 1 y el 30 de Junio del 2012.
Términos de búsquedas utilizados dentro del Archivo	Comportamiento del usuario	La palabra clave más utilizada para buscar el archivo web de UK, en Junio de 2012, fue “goji berry”.

Fuente: ISO (2012)

Preservación del Archivo Web

La preservación digital puede hacerse en dos niveles, en el nivel básico de mantener los bits seguros, y a un nivel más sofisticado usando estrategias como migración y emulación, para preservar la apariencia, función, comportamiento e incluso la experiencia de usuario de los recursos digitales. Este primero se conoce como preservación del flujo de bits o conservación física, la segunda como conservación lógica.

Las estadísticas descritas en la tabla 2.7 están destinadas a medir la eficiencia de las actividades de preservación del flujo de bits. La tabla 2.8 propone una plantilla para

ayudar a las instituciones a informar los metadatos que se espera se conserven en un Archivo Web. Las estadísticas en la tabla 2.9 son aquellas relevantes para la preservación lógica. Finalmente, en la tabla 2.10 son resumidas las estadísticas principales para la preservación de una colección.

Tabla 2.7: Datos estadísticos para la preservación del flujo de bits.

Dato estadístico	**Propósito**	**Ejemplo**
Volumen de recursos perdidos o deteriorados	Seguridad y resistencia	Se han perdido 25 MB de datos. Se hn deteriorado 50 URLs.
Volumen de recursos replicados	Seguridad y resistencia	150 terabytes del Archivo Web están replicados.

Fuente: ISO (2012)

Tabla 2.8: Datos estadísticos relacionados a los metadatos de preservación.

Tipo de metadato	**Descripción**	**Estándar usado (si hubiere)**	**% de recursos conteniendo la metadata**	**Comentarios**
Uno de los tipos de metadatos descritos en 2.4.4, ej. Descriptiva	Descripción de los metadatos		Porcentaje de recurso conteniendo los metadatos	Cualquier comentario útil o relevante.
		Ejemplos		
Descriptiva	DCMI metadata element set, Nombre del término: Tema. Asunto del recurso.	*Dublin Core Metadata Initiative (DCMI): LCSH*	30 %	Término del tema asignado manualmente por los curadores y almacenado en la herramienta de curador web.
Proveniencia	Archivos de configuración		90 %	Los archivos de configuración de rastreos del 2004 fueron descartados.
Técnica	Formatos de archivo (Tipos MIME)	*Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*	100 %	Todos los archivos recolectados tiene infomación de tipo MIME, pero esta puede ser no fidedigna.
Derechos	Permiso de archivar y proveer acceso en línea		100 %	Requerido solo para objetivos de acceso abierto.

Fuente: ISO (2012)

Tabla 2.9: Datos estadísticos para la preservación lógica del Archivo Web.

Dato estadístico	Propósito	Ejemplo
Distribución por formatos de archivo identificados	Capacidad de preservación	60 % del archivo se encuentra en formato HTML.
Número de formatos para los cuales una estrategia de preservación ha sido definida	Capacidad de preservación y compromiso	5 formatos tienen una estrategia de preservación definida: HTML, JPEG, GIF, PNG y PDF.
Volumen de recursos por formato con una estrategia de preservación activada	Capacidad de preservación y compromiso	*

Fuente: ISO (2012)

Tabla 2.10: Datos estadísticos principales para la preservación de una colección.

Dato estadístico	Propósito	Ejemplo
Volumen de recursos relicados	Seguridad y resistencia	150 terabytes del Archivo Web están replicados.
Distribución por formatos de archivo identificados	Capacidad de preservación	60 % del archivo se encuentra en formato HTML
Número de formatos para los cuales una estrategia de preservación ha sido definida	Capacidad de preservación y compromiso	5 formatos tienen una estrategia de preservación definida: HTML, JPEG, GIF, PNG y PDF.

Fuente: ISO (2012)

Costo del Archivo Web

Al medir el costo asociado al Archivo Web, es necesario establecer seis (6) categorías por las cuales se pueden clasificar los gastos generados por las actividades de archivado web. Estas categorías son:

- **Hardware:** Incluye los gastos asociados a la adquisición y mantenimiento de la infraestructura necesaria para realizar las actividades de archivado web.
- **Procesamiento:** Los costos generados en recursos para que el Archivo Web pueda funcionar, como lo son el servicio de electricidad o el proveedor de Internet.

- **Software:** Depende de la opción de software tomada, en el caso de que no sean software libre. También se incluye el costo de desarrollo adicional u operaciones técnicas requeridas
- **Personal:** Cantidad de horas/hombre empleadas en las actividades del Archivo Web.
- **Otros:** Puede incluir:
  - **Adquisición de metadatos:** por ejemplo, la compra de listas de nombres de dominio de registradores de dominios.
  - **Legal** Tomado en consideración solo cuando se requiere asesoramiento legal.
  - **Cooperación Internacional:** Costo de membresías para organizaciones como la IIPC y/o gastos generados por viajes relacionados al Archivo Web.

Indicadores de calidad

Por último están los indicadores de calidad, los cuales están definidos en el informe técnico ISO (2012). Éstos se encuentran clasificados en: administración, calidad del proceso de recolección, accesibilidad y uso, y preservación. Estos indicadores de calidad intentan ayudar a las instituciones de recolección a contestar preguntas fundamentales como:

- ¿Sabemos qué recolectamos?
- ¿Estamos recolectando lo que queremos recolectar?
- ¿Estamos haciendo el mejor uso de nuestros recursos?
- ¿Qué tan accesible y fácil de buscar es el Archivo?
- ¿Cómo garantizamos que el Archivo Web se mantendrá accesible en el tiempo?

Tabla 2.11: Indicadores de calidad.

Categoría	Nro.	Nombre	Objetivo	Comentarios
Administración	1	Costo por URL recolectada	Evaluar la eficiencia de los procesos de archivado web	Un costo bajo por URL recolectada demuestra en general una alta eficiencia de los procesos de archivado web. Sin embargo, un costo alto también puede indicar un alto nivel de curación.
	2	Porcentaje del personal involucrado en el Archivo Web	Indicar el compromiso de la institución con el Archivo web	Especialmente para bibliotecas e instituciones que tienen personal para diversas actividades, no sólo el Archivo Web.
Calidad del proceso de recolección	3	Porcentaje de recursos desaparecidos de la Web viva durante un periodo de tiempo dado	Evaluar el valor del Archivo web	Un objetivo se considera desaparecido cuando no hay respuesta del DNS o cuando su dirección de Internet genera una respuesta 404. Si es factible, el método más confiable es la revisión manual.
	4	Porcentaje conseguido del alcance establecido	Evaluar si los resultados del Archivo Web corresponden con los establecidos.	El alcance obligatorio o la cobertura del Archivo Web puede ser establecido como los sitios web nacionales o institucionales.
	5	Porcentaje de solicitudes de acuerdos o permisos otorgados por titulares de derechos	Evaluar la efectividad de las solicitudes de permisos	Una tasa alta indica el éxito de la actividad de solicitud de permisos. También es recomendable almacenar el número de rechazos explícitos y el número de solicitudes sin respuesta.
Accesibilidad y uso	6	Porcentaje de recursos accesibles por usuarios finales	Evaluar la disponibilidad del Archivo Web	Una alta tasa indica alta visibilidad o accesibilidad del Archivo Web. La unidad en la que se calcula el indicador debería ser reportado, por ejemplo, URLs o bytes.

Fuente: ISO (2012)

Indicadores de calidad. Continuación.

Categoría	Nro.	Nombre	Objetivo	Comentarios
Accesibilidad y uso	7	Porcentaje de recursos indexados a texto completo	Evaluar la capacidad de búsqueda del Archivo Web	La búsqueda por textos completo mejorar en gran medida la accesibilidad y usabilidad del Archivo Web.
	8	Porcentaje de recursos catalogados	Evaluar la capacidad de búsqueda y el nivel de curación del Archivo Web	Catalogar recursos en Archivos Web aumenta su accesabilidad y usabilidad. Cuando se calcula este indicador es recomendable reportar también la estrategia de adquisición usada para recolectar los recursos.
	9	Porcentaje anual de recursos accedidos	Evaluar la amplitud de uso actual del Archivo Web	La razón de usar dominios en lugar de URLs es debido a la posibilidad de que solo los recursos bajo un número limitado de dominios sean activamente usados.
	10	Porcentaje de usuarios de la biblioteca utilizando el Archivo Web	Evaluar el uso del Archivo Web por parte de los usuarios de los servicios de la institución	Si es posible debería ser reportado por los usuarios que realizan consultas dentro de la institución y por los que realizan consultas desde afuera.
Preservación	11	Porcentaje de recursos con al menos una replicación	Evaluar la capacidad de preservación de la secuencia de bits	La unidad de medida usada para el cálculo debe ser reportada, por ejemplo, URLs o bytes.
	12	Porcentaje de recursos perdidos o deteriorados	Evaluar la seguridad en el almacenamiento del Archivo Web	Un bajo porcentaje indica alta seguridad. La unidad de medida usada para el cálculo debe ser reportada, por ejemplo, URLs o bytes.

Fuente: ISO (2012).

Indicadores de calidad. Continuación.

Categoría	Nro.	Nombre	Objetivo	Comentarios
Preservación	13	Porcentaje de recursos con un formato de archivo identificado	Evaluar el conocimiento institucional del Archivo Web y la capacidad de preservación	La unidad de medida usada para el cálculo debería ser reportada, por ejemplo, URLs o bytes.
	14	Porcentaje de recursos para cuyo formato ha sido definido una estrategia de preservación	Evaluar el compromiso institucional con la preservación lógica del Archivo Web	La unidad de medida usada para el cálculo debería ser reportada, por ejemplo, URLs o bytes.
	15	Porcentaje de recursos revisados por virus	Evaluar el el uso seguro del Archivo Web, por parte de otras colecciones y dispositivos de usuarios.	En la detección de virus, las instituciones pueden tener una política de no remoción de virus del Archivo Web, sino bloquear el acceso a los recursos infectados.

Fuente: ISO (2012)

2.9. Ciencia de Datos

Ciencias de datos es un campo interdisciplinario que involucra procesos y sistemas para extraer conocimiento a partir de grandes volúmenes de datos, aplicando técnicas de procesamiento paralelo y distribuido para implementar algoritmos que permitan predecir o detectar patrones sobre los datos almacenados. (Liu, 2015)

2.9.1. Big Data

Según Manyika y col. (2011), Big Data se refiere a conjuntos de datos cuyo tamaño trasciende la capacidad de las herramientas típicas de base de datos para capturarlos, administarlos y analizarlos. Permite alamcenar los datos de manera distribuida y procesarlos en forma paralela y distribuida.

Es importante destacar que Big Data no es a partir de cierto número de datos, es decir que realmente aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales, ya que tomaría demasiado tiempo procesar dichos datos o porque sería muy costoso implementar

una arquitectura que soporte esa magnitud de datos.

Para una mejor comprensión, se profundizará un poco en lo conocido como las 8 V's de Big Data, comenzando con 3 las primeras que fueron creadas. (Van Rijmenam, 2015)

- **Volumen** Big Data es capaz de gestionar un gran volumen de datos (que puede ser variante) que se generan diariamente por diferentes empresas y fuentes.
- **Velocidad** La tecnología Big Data es capaz de almacenar y trabajar en tiempo real con los millones de datos generados al segundo de diferentes fuentes, por otro lado la capacidad de análisis de dichos datos han de ser rápidos porque se reducen los largos tiempos de procesamiento que se ven en las bases de datos tradicionales.
- **Variedad** Tiene la capacidad de combinar una gran variedad de información digital en los diferentes formatos en las que se puedan presentar, por ejemplo: video, audio, texto, imagenes.

Posteriormente surgieron otro conjunto de características resaltantes, de las cuales mencionaremos otras 5 V's:

- **Veracidad** Big Data es capaz de tratar y analizar inteligentemente esta inmensa cantidad de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar nuestra toma de decisiones.
- **Valor** El valor se refiere a nuestra capacidad de convertir los datos en valor. El valor está en los análisis realizados en esos datos y cómo los datos se convierte en información productiva para la organización.
- **Variabilidad** La variabilidad se refiere a los datos cuyo significado está en constante cambio.
- **Visualización** La parte más complicada es la visualización porque esa gran cantidad de datos debe ser sencillo comprender y leer. Por otro lado, si tenemos una vista de los datos correcta se pueden utilizar para tomas de decisiones adecuadas.
- **Visión** La visión define las metas que se pretenden conseguir en el futuro. Estas metas tienen que ser realistas y alcanzables.

2.9.2. Big Data Analítica

Big Data Analytics) es donde las técnicas analíticas avanzadas operan en grandes conjuntos de datos. Por lo tanto, la gran analítica de datos se centra en dos cosas: grandes datos y análisis, y cómo ambos se han unido para crear una de las tendencias más profundas de Inteligencia de Negocios en la actualidad. (Russom, 2011)

2.9.3. Inteligencia de Negocios Versus Ciencia de Datos

Según Long y Kelly (2015) una forma de evaluar el tipo de análisis que se está realizando es examinar el horizonte temporal y el tipo de enfoques analíticos que se están utilizando. Inteligencia de Negocios tiende a proporcionar informes, cuadros de mando y consultas sobre preguntas de negocios para el período actual o en el pasado. Los sistemas de Inteligencia de Negocios facilitan la respuesta a las preguntas relacionadas con los ingresos trimestrales, el progreso hacia los objetivos trimestrales y la comprensión de la cantidad de un producto dado que se vendió en un trimestre o año anterior. Estas preguntas tienden a ser cerradas y explicar el comportamiento actual o pasado, por lo general agregando datos históricos y agruparlo de alguna manera, además proporciona una visión retrospectiva y generalmente responde a las preguntas relacionadas con "cuándo" y "dónde" ocurrieron los acontecimientos.

En comparación, Ciencia de Datos tiende a utilizar datos desagregados de una manera más prospectiva y exploratoria, centrándose en analizar el presente y permitir decisiones informadas sobre el futuro. En lugar de agregar datos históricos para ver cuántos de un producto dado se vendieron en el trimestre anterior. Además, la Ciencia de Datos tiende a ser más exploratoria por naturaleza y puede usar la optimización de escenarios para tratar preguntas más abiertas. Este enfoque proporciona una visión de la actividad actual y la previsión en los eventos futuros, mientras que generalmente se centra en las preguntas relacionadas con "cómo" y "por qué" los eventos ocurren.

Cuando los problemas de Inteligencia de Negocios tienden a requerir datos altamente estructurados organizados en filas y columnas para un informe preciso, los proyectos de Data Science tienden a utilizar muchos tipos de fuentes de datos, incluyendo conjuntos de datos grandes o no convencionales. Dependiendo de los objetivos de una organización, puede optar por embarcarse en un proyecto de BI si está haciendo informes, creando cuadros de mando o realizando visualizaciones simples, o puede elegir proyectos de *Data Science* si necesita realizar un análisis más sofisticado con conjuntos de datos desagregados o variados .

2.9.4. Ecosistema Hadoop

Apache Hadoop es un marco de trabajo de software que soporta aplicaciones distribuidas bajo una licencia libre de la comunidad Apache. Permite el procesamiento de grandes volúmenes de datos de forma distribuida a través de clusters usando modelos sencillos de programación. Está siendo construido y usado por una comunidad global de contribuyentes, mediante el lenguaje de programación Java. Está diseñado para escalar desde un servidor sencillo hasta miles de nodos, los cuales pueden ser heterogéneos. (Apache Software Foundation, 2017)

Sistema de Archivos Distribuidos de Hadoop

Más comúnmente conocido por sus siglas en inglés HDFS (*Hadoop Distributed File System*), como su nombre así lo indica es el sistema manejador de archivos sobre el cual se fundamenta Hadoop. Entre sus características podemos encontrar que es distribuido, escalable y portátil ya que el lenguaje principal que se utilizó para su desarrollo fue Java. El principal objetivo de HDFS es solucionar los problemas relacionados al almacenamiento de grandes conjuntos de datos mediante la distribución de los mismos. (Apache Software Foundation, 2018c)

Características

- Manejo de grandes archivos.
- Alta compatibilidad con hardware ya que HDFS fue diseñado con la capacidad de poder ejecutar sus procesos sin importar las características.
- Acceso a datos en flujos ya que utiliza un patrón de manejo de datos donde los mismos se escriben una única vez durante su creación pero pueden ser leídos y replicados múltiples veces.
- Tolerancia a fallos porque el software cliente almacena los datos escritos en un archivo local temporal, hasta que haya suficiente para llenar un bloque HDFS completo. Cuando se hayan guardado datos suficientes o termine la operación de escritura, los datos locales se envían a través de la red y son escritos en varios servidores del clúster, evitando así la pérdida de datos si ocurre alguna falla en el hardware. (Apache Software Foundation, 2018c)

Distribución Hortonworks

Hortonworks provee un ecosistema llamado *Hortonworks Data Plataform (HDP)* el cual se puede descargar de forma gratuita en su sitio web ya que los ingresos de esta compañía provienen del soporte para los que lo necesiten. Posee una ecosistema con una gran variedad de herramientas integradas para poder satisfacer las necesidades de una solución de Big Data. Además posee un amplio foro con expertos en el área y usuarios para hacer preguntas. (Hortonworks Inc, 2018)

2.9.5. Arquitectura para Big Data y Data Warehouse

El almacen de datos de hoy en día, se basa en los principios fundamentales de una “versión empresarial de verdad´´ y un “repositorio único de datos", pero debe

abordar las necesidades de tipos de datos nuevos, volúmenes mas grandes, nuevos niveles de calidad de datos, mayor rendimiento, nuevos metadatos y nuevos requisitos de usuario. Actualmente, hay varios problemas en los entornos actuales del almacén de datos que deben abordarse y, lo que es más importante, la infraestructura actual no puede soportar las necesidades de los nuevos datos en la misma plataforma.

La arquitectura de almacenamiento de datos de próxima generación será compleja a partir de una implementación de arquitectura física, que consta de una gran cantidad de tecnologías, extremadamente flexible y escalable desde una perspectiva de arquitectura de datos.

Aplicaciones de Big Data

En la última década los dispositivos de almacenamiento de datos, surgieron como una sólida arquitectura de caja negra para el procesamiento de cargas de trabajo para datos a gran escala. Una de las extensiones de esta arquitectura es la aparición de los dispositivos Big Data. Estos dispositivos están configurados para manejar los rigores de las cargas de trabajo, las complejidades de Big Data y la arquitectura RDBMS actual.

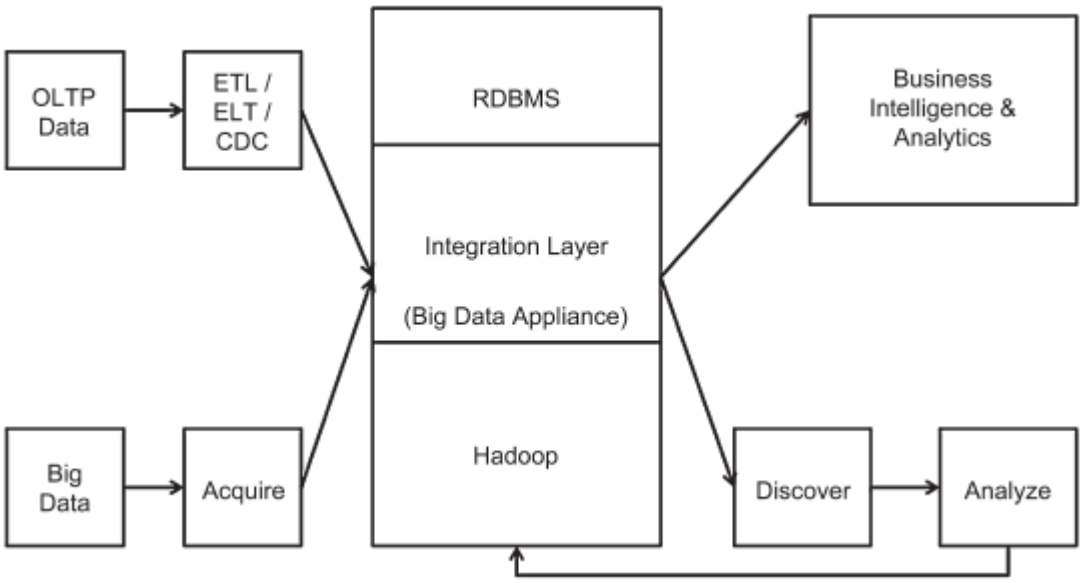

Figura 2.7: Arquitectura Big Data Appliances.
(Krishnan, 2013)

En la figura 2.7 se muestra la arquitectura de Big Data Appliances, la cual incluye una capa de Hadoop y una capa de RDBMS. Si bien la implementación de la arquitec-

tura física puede diferir entre proveedores como Teradata, Oracle, IBM y Microsoft, la arquitectura sigue siendo la misma, donde las tecnologías Hadoop y / o NoSQL se utilizarán para adquirir, preprocesar y almacenar Big Data, y las capas de RDBMS se usarán para procesar el resultado de las capas Hadoop y NoSQL. Los traductores y conectores específicos de la base de datos MapReduce y RDBMS se usarán en la arquitectura integrada para gestionar el movimiento y la transformación de datos dentro de la arquitectura.

Las áreas principales incluyen carga de datos, disponibilidad, volumen de datos, rendimiento de almacenamiento, escalabilidad, demandas de consultas cambiantes y diversas en comparación con los datos y costos operativos de la plataforma de data warehouse de próxima generación. Los riesgos se pueden aplicar tanto a los datos estructurados como a los no estructurados que estarán coexistiendo en esta plataforma.

El procesamiento de carga de trabajo en esta arquitectura se configura según los requisitos especificados por los usuarios, incluida la adquisición, el uso, la retención y el procesamiento de datos. La complejidad de esta arquitectura es la configuración y la configuración inicial, que necesitarán un gran trabajo si las especificaciones no son claras o tienden a cambiar con el tiempo, ya que la configuración inicial es personalizada.

- Ventajas
    1. Diseño escalable y arquitectura de integración de datos modular.
    2. Implementación de arquitectura física heterogénea, que proporciona la mejor integración en su clase en la capa de procesamiento de datos.
    3. Configurado a medida para adaptarse a los rigores del proceso según se requiera para cada organización.
- Desventajas
    1. La configuración personalizada es la mayor debilidad.
    2. La integración de datos y la escalabilidad de las consultas pueden volverse complejas a medida que la configuración cambia durante un período de tiempo

2.9.6. Organización y estructura de los datos

De acuerdo a la forma en que se encuentran definidos los datos podemos clasificarlos según su estructura y organización. (Sint, Schaffert, Stroka & Ferstl, s.f.)

Datos estructurados Son aquellos que hacen referencia a cualquier dato que resida en un campo dentro de un registro o archivo, por ejemplo: bases de datos

relacionales y hojas de cálculo. Esta data tiene la característica de ser de fácil registro, almacenamiento, consulta y análisis.

Datos no estructurados Se refiere a los datos que no siguen un modelo de datos predefinidos y tampoco están organizados en alguna estructura específica. La data no estructurada posee de una gran variedad de datos de diferentes tipos los cuales son difíciles de clasificar y por lo tanto difíciles de almacenar y ser recuperados con consultas. La solución para su almacenamiento son las bases de datos NoSQL debido a su estructura flexible y su rapidez en el acceso a los datos.

Datos semiestructurados Estos son datos estructurados que no corresponden a un modelo de estructura formal como las bases de datos relacionales, ya que estos datos son definidos por su propia estructura, suelen tener etiquetas o marcas para identificar y separar los elementos semánticamente y estableciendo jerarquía entre estos elementos. El concepto de esta organización de datos viene dado por lenguajes de marcado como el *Extensible Markup Language* (XML) usado en el área web y convertido en un estándar por la W3C. Otro ejemplo de datos semi-estructurados viene dado por otro formato de texto que ha sido impulsado en los últimos años debido al fácil intercambio de datos presente en las aplicaciones web llamado *Javascript Object Notation* comúnmente conocido como JSON.

3

Marco Metodológico

El desarrollo de software es un proceso complejo, por lo que para lograr su éxito se busca seguir algún método que guíe el trabajo en cada una de sus etapas, desde su concepción hasta su finalización. En este capítulo expondremos la metodología referida al Ciclo de Vida Dimensional desarrollada por Ralph Kimball para el desarrollo de un almacén de datos, la cual ya ha sido ampliamente usada y probada a lo largo de los años para la construcción rápida de una solución de inteligencia de negocio.

3.1. Metodología de Ralph Kimball

Esta metodología se basa en lo que Kimball llama el Ciclo de Vida Dimensional del Negocio (*Business Dimensional Lifecycle*). Este ciclo está basado en cuatro principios básicos:

- **Enfocarse en el proceso del negocio**: hay que centrarse en el levantamiento de requerimientos, realizando un análisis agudo de las necesidades de la organización.
- **Realizar entregas en incrementos significativos**: crear el almacén de datos de forma incremental, generalmente en plazos de seis (6) a doce (12) meses, estableciendo prioridades en los requerimientos de la organización.

- **Construir una infraestructura de información adecuada**: Diseñar un repositorio de información integrado, usable y de alto rendimiento, en el cual se aprecien los requerimientos de negocio identificados en la empresa.
- **Ofrecer la solución completa**: proporcionar todos los elementos necesarios para cumplir con las exigencias del cliente, es decir, construir un almacén de datos confiable, brindar herramientas de consultas ad-hoc, aplicaciones para informes y análisis avanzado, capacitación, soporte y documentación.

En la figura 3.1 vemos una imagen de la metodología, en la cual se puede observar que hay tres rutas, el camino superior (en color amarillo) implica las tareas relacionadas con el software, el camino del medio (color naranja) se diseña e implementa el modelo dimensional, y se desarrolla el subsistema de ETL, por último se encuentra el camino inferior (color verde) que está relacionado con tareas en las que se diseñan aplicaciones de Inteligencia de Negocios para usuarios finales.

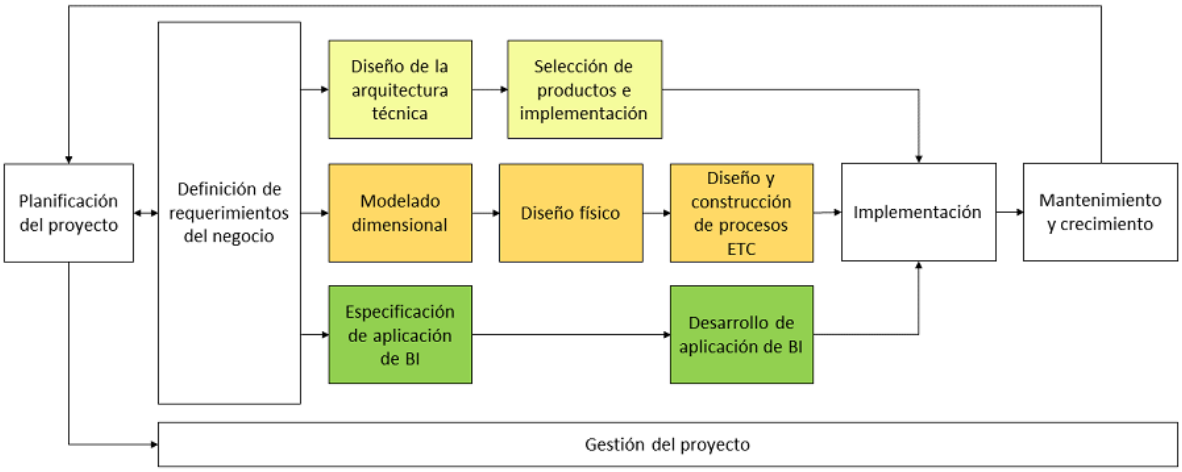

Figura 3.1: Diagrama del Ciclo de vida de Kimball
Fuente: Traducido de Kimball, 1998.

A continuación, describiremos brevemente las actividades que se realizan en el proceso de desarrollo descrito por Kimball (1998), donde nos enfocaremos mayormente en el Modelado Dimensional.

3.1.1. Hitos del ciclo de vida

- **Planificación del Proyecto:** En este primer paso se da inicio al desarrollo del proyecto. Aquí se debe definir el propósito del proyecto, elaborar un plan de trabajo y gestionar la puesta en marcha del mismo, definiendo y respetando el alcance establecido.

- **Gestión del Proyecto:** La gestión del proyecto garantiza que las actividades del ciclo de vida de Kimball permanezcan sincronizadas y por buen camino. Las actividades referentes a la gestión del proyecto se centran en el monitoreo del estado del proyecto, el seguimiento de problemas y el control de cambios para preservar los límites del alcance.
- **Definición de requerimientos del negocio:** La definición de los requerimientos es en gran medida un proceso de entrevistar al personal de negocio y técnico, pero siempre es recomendable aprender tanto como se pueda sobre el negocio, los competidores, la industria y los clientes del mismo.

Siguiendo la definición de los requisitos del negocio, hay tres caminos simultáneos que se centran en la tecnología, los datos y las aplicaciones de inteligencia negocio respectivamente. Mientras que las flechas en la figura 3.1 del diagrama del ciclo de vida designan el flujo de trabajo de las actividades a lo largo de cada una de los caminos paralelos, también hay dependencias implícitas entre las tareas, como lo ilustra la alineación vertical de los cuadros de tareas.

Camino de la tecnología

- **Diseño de la arquitectura técnica:** Los entornos DW/BI requieren la integración de numerosas tecnologías en el cual deben considerarse simultáneamente tres factores: los requisitos del negocio, el entorno técnico actual y las direcciones estratégicas planificadas, para establecer el diseño apropiado de la arquitectura técnica para un almacén de datos o un Sistema de Inteligencia de Negocios.
- **Selección de productos e instalación:** Utilizando el plan de arquitectura técnica se deben evaluar y seleccionar componentes específicos de arquitectura como la plataforma de hardware, el sistema de gestión de bases de datos, la herramienta de Extracción, Transformación y Carga (ETC) o consulta de datos e informes. Una vez que los productos han sido seleccionados, son instalados y probados para asegurar una integración apropiada dentro del entorno.

Camino de los datos

- **Modelado dimensional:** En la definición de requisitos de negocio, las necesidades de datos de la organización se determinan y documentan en una matriz de bus, la cual representa los principales procesos empresariales de la organización y su dimensionalidad asociada. Esta matriz funciona como un modelo de arquitectura de datos para asegurar que los datos puedan ser integrados y extendidos en la organización a lo largo del tiempo, y busca priorizar los requerimientos o procesos de negocio mas críticos. Luego de un análisis de datos

más detallado de una sola fila de la matriz de procesos de negocios, los modeladores identifican la granularidad de la tabla de hechos, las dimensiones, los atributos asociados y los hechos numéricos.

- **Diseño físico:** El diseño físico se centra en la definición de las estructuras físicas, incluyendo la configuración del entorno de la base de datos y la instanciación de una seguridad adecuada. Aunque el modelo de datos físicos en la base de datos relacional será prácticamente idéntico al modelo dimensional, hay problemas adicionales que abordar, como por ejemplo las estrategias de ajuste de rendimiento, desde la indexación hasta el particionamiento y las agregaciones.
- **Diseño y construcción de procesos de ETC:** El diseño y desarrollo del sistema de extracción, transformación y carga (ETC) sigue siendo uno de los desafíos enfrentados por un equipo de proyecto de Almacenes de Datos o Inteligencia de Negocios; Incluso cuando todas las otras tareas han sido bien planificadas y ejecutadas, el 70 % del riesgo y esfuerzo del proyecto proviene de este paso.

Camino de la aplicación de inteligencia de negocios

- **Especificación de Aplicaciones de Inteligencia de Negocios:** Inmediatamente después de la definición de requisitos de negocio, mientras que algunos miembros del equipo están trabajando en la arquitectura técnica y modelos dimensionales, otros deben trabajar junto con la empresa para identificar las aplicaciones candidatas de Inteligencia de Negocios, junto con interfaces de navegación apropiadas para atender las necesidades de los usuarios.
- **Desarrollo de aplicación de Inteligencia de Negocios:** Siguiendo las especificaciones de aplicaciones de Inteligencia de Negocios, las tareas de desarrollo de aplicaciones incluyen la configuración de los metadatos empresariales y la infraestructura de herramientas, y luego la construcción y validación de las aplicaciones de BI analíticas y operativas especificadas, junto con el portal de navegación.
- **Implementación:** Las tres rutas paralelas, convergen en la implementación. Se requiere una extensa planificación para asegurar que estas piezas del rompecabezas sean probadas y encajadas adecuadamente, junto con la infraestructura de apoyo apropiada. Es fundamental que el despliegue esté bien orquestado.
- **Mantenimiento y crecimiento:** Con respecto al mantenimiento, una vez que el sistema está en producción, las tareas técnicas operacionales son necesarias para mantener el sistema funcionando de manera óptima, incluida la supervisión del uso, el ajuste del rendimiento, el mantenimiento de los índices y la copia de seguridad del sistema. Referente al crecimiento, si se ha hecho un buen trabajo, el sistema puede ampliar y evolucionar. A diferencia de las iniciativas tradi-

cionales de desarrollo de sistemas, el cambio debe ser visto como una señal de éxito, no de fracaso. Primero deben establecerse procesos de priorización para hacer frente a la demanda empresarial que existe en ese momento. Luego regresamos al inicio del Ciclo de Kimball, aprovechando y construyendo sobre la base ya establecida, mientras se presta atención a los nuevos requerimientos.

3.2. Proceso de diseño dimensional en cuatro pasos

La creación de un modelo dimensional es un proceso dinámico y altamente iterativo. El proceso de diseño comienza con un modelo dimensional de alto nivel obtenido a partir de los procesos priorizados en la matriz de bus. El proceso iterativo consiste en cuatro pasos que describiremos en los siguientes puntos:

1. **Elegir proceso de negocio:** El primer paso es elegir el área a modelizar. Esta es una decisión de la dirección, y depende fundamentalmente del análisis de requerimientos y del resultado de la matriz de bus.
2. **Establecer el nivel de granularidad:** Una vez que el proceso de negocio ha sido identificado, el equipo de diseño debe declarar el grano de la tabla de hechos. La granularidad no es mas que especificar el nivel de detalle. La elección de la granularidad depende de los requerimientos del negocio y lo que es posible a partir de los datos actuales.
3. **Identificar las dimensiones:** Una vez que el grano de la tabla de hechos está establecido, la elección de las dimensiones es bastante sencilla. Es en este punto es cuando se puede empezar a pensar en las claves foráneas. El grano determinará un conjunto primario o mínimo de dimensiones. A partir de ahí, el diseño se complementa con dimensiones adicionales que toman un valor único en el grano declarado de la tabla de hechos.
4. **Identificar los hechos:** El paso final en el proceso es seleccionar cuidadosamente los hechos o métricas que son aplicables al proceso de negocio. Los hechos pueden ser capturados físicamente por el evento de medición o derivados de estas mediciones. Cada hecho debe ser fiel al grano de la tabla de hechos; No se puede mezclar hechos de otros períodos de tiempo u otros niveles de detalle que no coinciden con el grano claramente declarado.

4

Marco Aplicativo

En este capítulo se describirá todo el proceso de desarrollo utilizando la metodología descrita en el capítulo anterior. Se tomaron en cuenta principalmente los hitos relacionados a los tres caminos del Ciclo de Vida de Kimball, y se omite la gestión del proyecto y el mantenimiento y crecimiento de la aplicación, puesto que se hace énfasis en el diseño del almacén de datos, resultando en un solo proyecto manejado por un equipo reducido de integrantes.

Todos los hitos se encuentran muy relacionados, por lo que frecuentemente se hizo necesario volver a evaluar un hito anterior a medida que se avanzaba en el desarrollo. Esto provocaba que el proceso no fuera exactamente lineal sino en ciclos de constante revisión.

4.1. Definición de requerimientos

Para el desarrollo de la solución de Inteligencia de Negocios se realizó una extensa revisión de los distintos módulos del Archivo Web (listados en el capítulo 1.1) junto con el personal del negocio, con el fin de tener un panorama claro de los datos disponibles y la información requerida para analizar. De esta revisión se notó que el rastreador Heritrix producía una importante cantidad de datos estadísticos a partir de los rastreos que realiza, además de los archivos WARC. Tomando en cuenta esta información, junto con lo sugerido en el documento ISO/DTR 14873:2012 se definieron los indicadores listados en la tabla 4.1.

Tabla 4.1: Indicadores del proceso de rastreo.

Nombre del indicador	**Forma de cálculo**	**Unidad de medida**	**Frecuencia de medición**	**Criterios de clasificación**
Cantidad de semillas	Conteo de objetivos rastreados	#	Acumulado	Por Colección, Fecha y Tiempo del día
Cantidad de rastreos	Conteo de rastreos realizados a objetivos	#	Acumulado	Por Colección, Semilla, Fecha y Tiempo del día
Cantidad de recursos rastreados	Conteo de URIs procesadas	#	Acumulado	Por Colección, Semilla, Fecha y Tiempo del día
Duración promedio de rastreos	$\frac{\sum_{i=1}^{n} \text{duración de rastreo}_i}{\text{Total rastreos}}$	Minutos	Acumulado	Por Colección, Semilla, Fecha y Tiempo del día
Cobertura cronológica	Fecha de inicio de rastreo	Fecha	Acumulado	Por Colección, Semilla, Fecha y Tiempo del día
Distribución de URLs por código de estado	$\frac{\text{Cantidad de URLs}_{\text{código=A}}}{\text{Total de URLs}} \times 100$	%	Acumulado	Por Colección, Semilla, Fecha y Tiempo del día
Distribución de bytes rastreados por tipos de formatos	$\frac{\sum_{i=0}^{n} \text{bytes del formato}_i}{\text{Total bytes rastreados}} \times 100$	%	Acumulado	Top 10 por: Colección, Semilla, Fecha y Tiempo del día
Distribución de URLs por tipos de formatos	$\frac{\sum_{i=0}^{n} \text{Cantidad URLs}_{formato=A}}{\text{Total URLs}} \times 100$	%	Acumulado	Top 10 por: Colección, Semilla, Fecha y Tiempo del día

Fuente: Elaboración propia.

Cada indicador se encuentra expresado en una de las siguientes unidades de medidas:

- Cantidad (#)
- Porcentaje (%)
- Minutos
- Fecha

4.2. Diseño de la arquitectura técnica

En cuanto a la arquitectura, uno de los requerimientos principales era que la solución debía adaptarse a un entorno de grandes volúmenes de datos, por lo cuál tendría que ser desarrollada sobre el ecosistema Hadoop, al cual ya se estaban adaptando los distintos módulos del AWV.

Siguiendo dicho lineamiento y tomando como referencia el modelo básico de arquitectura de una Solución de Inteligencia de Negocio (sección 2.6) la arquitectura del sistema resultó como se observa en la figura 4.1.

Figura 4.1: Arquitectura general
Fuente: Elaboración propia.

En primer lugar se tienen las fuentes de datos, conformadas por los archivos de texto plano que genera Heritrix por cada rastreo, y la base de datos principal denominada `app` que contiene la información registrada por los usuarios para la programación de rastreos, como las colecciones, semillas y demás metadatos. Luego se tienen los sistemas creados por este proyecto, un área intermedia para integración de datos y el área de presentación, conectado finalmente a la herramienta de acceso de

datos. Cada elemento se instancia con las tecnologías mencionadas en la siguiente sección.

4.3. Selección de productos

En esta sección se describirán las herramientas que se utilizarán en la implementación de cada área del sistema.

4.3.1. Pentaho Data Integration

La herramienta utilizada para el proceso de ETC fue Spoon de Pentaho Data Integration (PDI), también llamada Kettle. Es un componente de la suite de Pentaho que se encarga de los procesos de extracción, transformación y carga (ETC).

PDI se puede utilizar como una aplicación independiente, o se puede utilizar como parte de la suite Pentaho. Es la herramienta de ETC de código abierto más popular. Admite una amplia gama de formatos de entrada y salida, incluidos archivos de texto, hojas de datos y motores de bases de datos comerciales y gratuitas, además posee integración con las distribuciones de Hadoop. (Pentaho Data Integration, 2015)

La instalación de Spoon en Windows es muy sencilla, después de descomprimir el archivo descargado, se puede iniciar Spoon navegando a la carpeta /data-integration y haciendo doble clic en Spoon.bat.

4.3.2. MySQL

Para el área intermedia se decidió utilizar la base de datos MySQL.

MySQL es el sistema de administración de bases de datos SQL de código abierto más popular. Es desarrollado, distribuido y respaldado por Oracle Corporation. Es una base de datos relacional, rápida, confiable, escalable y fácil de usar. Además que posee conectores con varios lenguajes de programación y aplicaciones (Oracle Corporation, 2018). En el proyecto se decidió utilizar MySQL gracias a que viene por defecto en la distribución de Hadoop, Hortonworks.

4.3.3. Sqoop

Sqoop es una herramienta diseñada para transferir datos entre Hadoop y bases de datos relacionales o *mainframes* (Computadores inmensos que se utilizan para procesar grandes cantidades de datos). Se puede utilizar Sqoop para importar datos de un

sistema de gestión de base de datos relacional (RDBMS) como MySQL u Oracle o un mainframe en el Sistema de Archivos Distribuidos de Hadoop (HDFS), transformar los datos en Hadoop MapReduce y luego exportar los datos de nuevo a un RDBMS. Soporta fuentes estructuradas, semi estructuradas y no estucturadas. (Apache Software Foundation, 2018b)

4.3.4. Hive

Para el almacén de datos se consideró que la mejor opción era utilizar el almacén de datos de Hadoop, Hive.

Hive facilita la lectura, escritura y administración de grandes conjuntos de datos que residen en el almacenamiento distribuido mediante SQL. Posee una herramienta de línea de comandos y un controlador JDBC para conectar a los usuarios a Hive. Está diseñado para maximizar la escalabilidad, el rendimiento, la extensibilidad, la tolerancia a fallos. (Apache Software Foundation, 2018a)

4.3.5. Tableau Desktop

Por último, como herramienta para la visualización se decidió utilizar Tableau Desktop ya que tiene conexión con Hive mediante el driver ODBC.

Tableau es una plataforma de análisis integral más eficaz, segura y flexible para sus datos. Posee varias herramientas para el análisis de datos.

Tableau Desktop posee conexiones con datos en las instalaciones físicas o en la nube. Tanto si se trata de Big Data, bases de datos SQL u hojas de cálculo. Se pueden realizar cuadros de mando con la información sumarizada y realiza cálculos eficaces rápidamente a partir de datos existentes. (Tableau Software, 2018)

4.4. Modelo dimensional

4.4.1. Selección del Proceso de Negocio

En la sección 2.5 se describieron a grandes rasgos las tareas que realizan los Archivos Web, que además se mencionan igualmente en el informe técnico ISO 14873 (ISO, 2012), éstas nos sirven para definir los procesos de negocio involucrados en una matriz de bus, la cual se presenta en la figura 4.2.

De estos cuatro procesos principales se escogió el de **Rastreo** para modelar en esta iteración de la solución BI, esto debido a que al revisar las fuentes de datos

	Dimensiones Conformadas						
Procesos de Negocio	**Fecha**	**Tiempo del día**	**Colección**	**Semilla**	**Recurso**	**Tipo MIME**	**Código Estado HTTP**
Selección							
Rastreo	X	X	X	X		X	X
Preservación							
Acceso							

Figura 4.2: Matriz de Bus del AWV.
Fuente: Elaboración propia.

disponibles se halló más información relacionada a este proceso, siendo el mejor candidato para realizar análisis.

4.4.2. Identificación del nivel de granularidad

Este paso consiste en identificar lo que se desea medir en cada indicador propuesto y definir el nivel de detalle que apoye el cumplimiento de los objetivos. Del proceso de Rastreo se derivaron tres niveles de granularidad para los datos:

- Una fila por versión de semilla rastreada en una colección en un tiempo y fecha dada.
- Una fila por tipo mime por versión de semilla rastreada en una colección en un tiempo y fecha dada.
- Una fila por código de estado http por versión de semilla rastreada en una colección en un tiempo y fecha dada.

4.4.3. Dimensiones

En este punto se procedió a estudiar cuáles podrían ser las perspectivas por las que se observarían los datos del negocio, es decir, extraer las dimensiones conformadas, que ya fueron mencionadas previamente en la matriz de bus y en el nivel de granularidad. A continuación se especifica el detalle las mismas:

Dimensión Fecha (dim_fecha) Esta dimensión contiene la información referente a fechas en general.

Tabla 4.2: Detalle de la dimensión Fecha.

Atributo	Tipo de dato	Descripción	Procedencia
fecha_id	Entero	Clave subrogada	Autoincremental
anio	Entero	Año de la fecha	Generado en ETC
mes	Entero	Mes de la fecha	Generado en ETC
mes_nombre	Cadena	Nombre del mes de la fecha	Generado en ETC
dia	Entero	Día de la fecha	Generado en ETC
dia_nombre	Cadena	Nombre del día de la fecha	Generado en ETC
dia_del_anio	Entero	Número de día en el año	Generado en ETC
fecha_completa	Fecha	Fecha completa	Generado en ETC

Fuente: Elaboración propia.

Dimensión Tiempo (dim_tiempo) Esta dimensión contiene información referente al tiempo en general.

Tabla 4.3: Detalle de la dimensión Tiempo.

Atributo	Tipo de dato	Descripción	Procedencia
tiempo_id	Entero	Clave subrogada	Autoincremental
hora	Entero	Número de la hora	Generado en ETC
minuto_de_hora	Entero	Minuto de la hora	Generado en ETC
minuto_del_dia	Entero	Minuto del día	Generado en ETC
segundo_del_minuto	Entero	Segundo del minuto	Generado en ETC
segundo_del_dia	Entero	Segundo del día	Generado en ETC
tiempo_completo	Tiempo	Hora completa (hh:mm:ss)	Generado en ETC

Fuente: Elaboración propia.

Dimensión Semilla (dim_semilla) Esta tabla contiene información sobre los puntos de entrada de los sitios web rastreados.

Tabla 4.4: Detalle de la dimensión Semilla.

Atributo	Tipo de dato	Descripción	Procedencia
semilla_id	Entero	Clave subrogada	Autoincremental
url	Cadena	URL que posee una semilla	Campo url de la tabla Semilla del Área Intermedia
nombre	Cadena	Nombre de la semilla	Campo nombre de la tabla Semilla de Área Intermedia
descripcion	Cadena	Descripción de la semilla	Campo descripción de la tabla Semilla del Área Intermedia

Fuente: Elaboración propia.

Colección (dim_coleccion) Esta tabla contiene información sobre los temas de clasificación que pueden tener las semillas.

Tabla 4.5: Detalle de la dimensión Colección.

Atributo	Tipo de dato	Descripción	Procedencia
coleccion_id	Entero	Clave subrogada	Autoincremental
nombre	Cadena	Nombre de la colección	Campo nombre de la tabla Colección del Área Intermedia
descripcion	Cadena	Descripción de la colección	Campo descripción de la tabla Colección del Área Intermedia

Fuente: Elaboración propia.

Tipo MIME (dim_tipo_mime) Esta tabla contiene información sobre los tipos MIME.

Tabla 4.6: Detalle de la dimensión tipo MIME.

Atributo	Tipo de dato	Descripción	Procedencia
tipo_mime_id	Entero	Clave subrogada	Autoincremental
tipo	Cadena	Clasificación del tipo MIME	Campo tipo de la tabla tipo_mime del Área Intermedia
subtipo	Cadena	Clasificación del tipo MIME	Campo subtipo de la tabla tipo_mime del Área Intermedia
nombre_completo	Cadena	Nombre completo del tipo MIME	Campo nombre_completo de la tabla tipo_mime del Área Intermedia
descripcion	Cadena	Descripción del tipo MIME	Campo descripción de la tabla tipo_mime del área intermedia

Fuente: Elaboración propia.

Código Estado HTTP (dim_codigo_estado_http) Esta tabla contiene información sobre los código de estado HTTP.

Tabla 4.7: Detalle de la dimensión Código de Estado HTTP.

Atributo	Tipo de dato	Descripción	Procedencia
codigo_id	Entero	Clave subrogada	Autoincremental
codigo	Entero	Código de estado HTTP	Campo codigo de la tabla codigo_estado_http del Área Intermedia
tipo	Varchar	Tipo de código de estado HTTP	Campo tipo de la tabla codigo_estado_http del Área Intermedia
descripcion	Varchar	Descripción del código estado HTTP	Campo descripcion de la tabla codigo_estado_http del Área Intermedia

Fuente: Elaboración propia.

4.4.4. Hechos y Tablas de Hechos

Ya identificada la granularidad, las dimensiones y las jerarquías, definiremos los hechos y las tablas de hechos. En las siguientes tablas (4.8, 4.9 y 4.10) podremos ver

las tablas de hechos con sus respectivas dimensiones y medidas.

Tabla 4.8: Tabla de hechos de Rastreos.

Tabla de Hechos	Descripción	Dimensiones	Medidas
Rastreos	Esta tabla se encarga de registrar los rastreos que se realizan por semilla, colección y fecha.	• Fecha • Tiempo • Semilla • Colección • Rastreo (DD)	• duración • cantidad_hosts_visitados • cantidad_uris_procesadas • cantidad_uris_fallidas • cantidsd_uris_ignoradas • total_bytes_rastreados • uris_por_segundo • ancho_de_banda

Fuente: Elaboración propia.

Tabla 4.9: Tabla de hechos de Código de Estado.

Tabla de Hechos	Descripción	Dimensiones	Medidas
Código de estado HTTP	Esta tabla de encarga de almacenar los código de estado por colección, rastreo, semilla y fecha.	• Fecha • Semilla • Colección • Rastreo (DD) • Código de estado HTTP	• cantidad_urls

Fuente: Elaboración propia.

Tabla 4.10: Tabla de hechos de Tipos MIME.

Tabla de Hechos	Descripción	Dimensiones	Medidas
Tipo MIME	Esta tabla almacena los tipos MIME por rastreo, semilla, colección y fecha.	• Fecha • Tiempo • Semilla • Colección • Rastreo (DD) • Tipo MIME	• cantidad_urls • cantidad_bytes

Fuente: Elaboración propia.

Por último se presenta el modelo dimensional resultante. Para mayor entendimiento se ilustra de manera separada por tabla de hechos, siendo el principal el que corresponde a la tabla de Rastreos (figura 4.3).

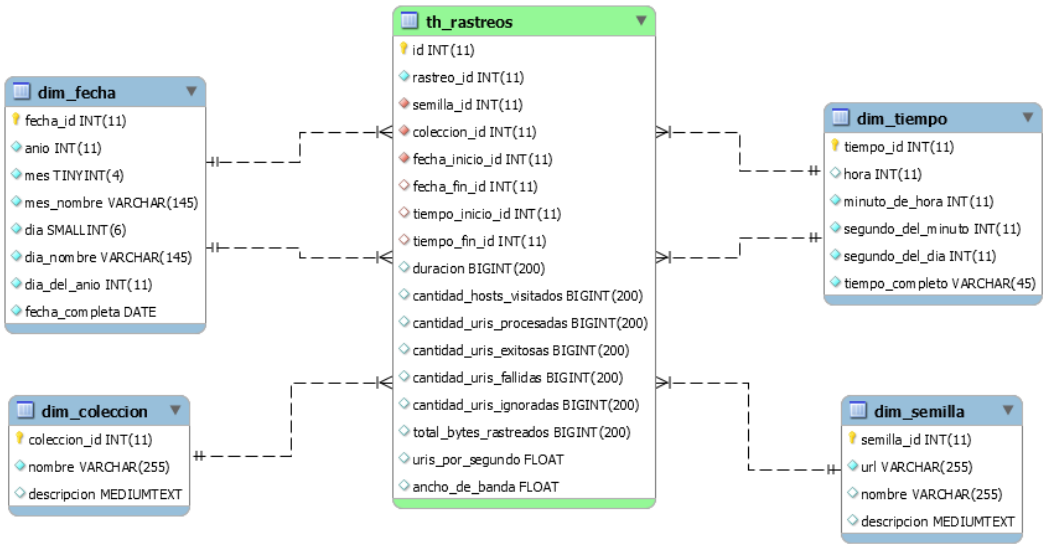

Figura 4.3: Modelo Dimensional (Tabla de Hechos Rastreos).
Fuente: Elaboración propia.

Seguidamente, se muestra en la figura 4.4 el modelo dimensional para las dos tablas de hechos restantes, correspondientes a las métricas de Tipo MIME y de Código de Estado HTTP.

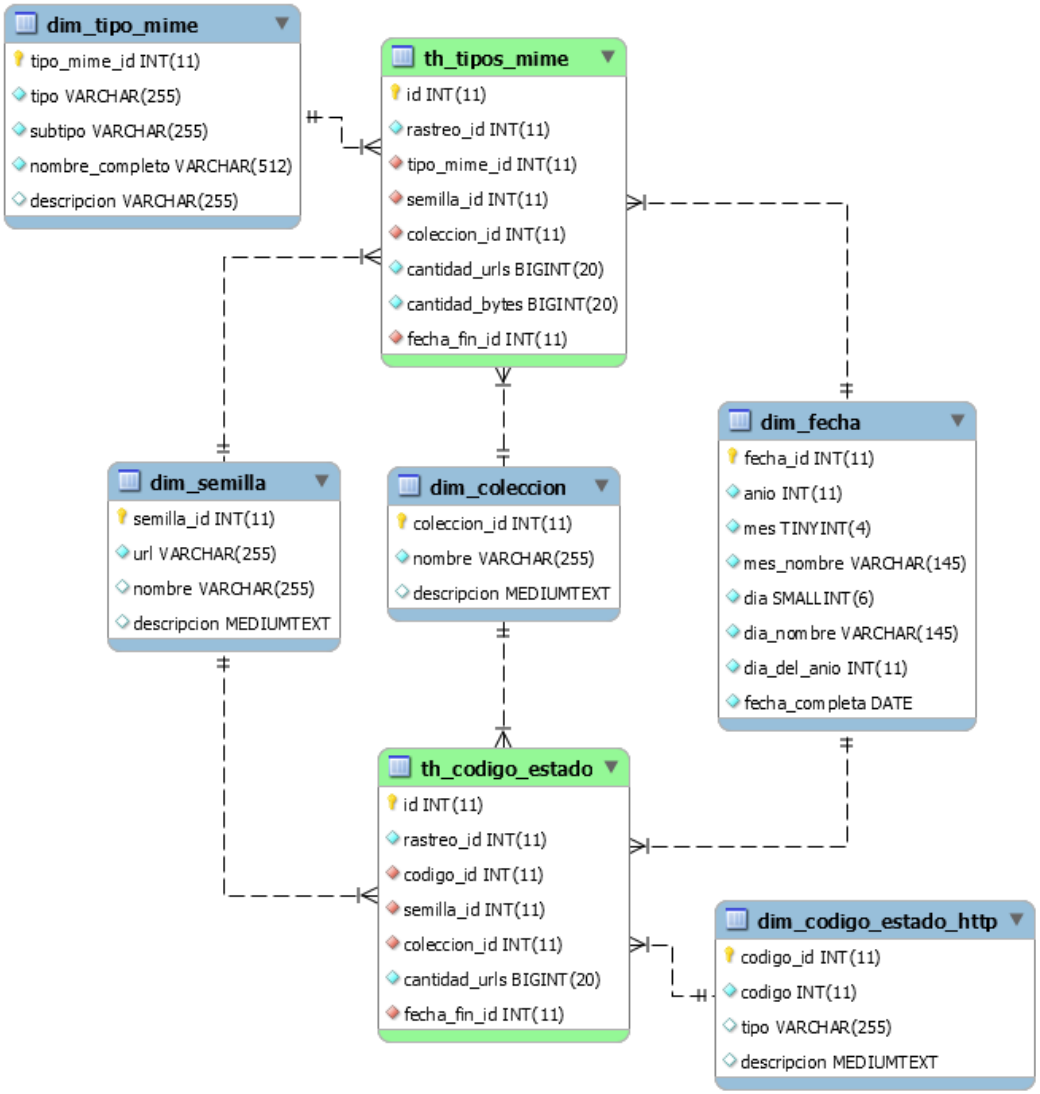

Figura 4.4: Modelo Dimensional (Tablas de Hechos de Codigo de estado HTTP y tipo MIME).

Fuente: Elaboración propia.

4.5. Diseño físico

En esta sección veremos en detalle los diseños de los almacenes.

4.5.1. Diseño del Área Intermedia

Para el diseño del Área Intermedia utilizamos dos fuentes de datos diferentes, en primer lugar tomamos la base de datos `app` que se encuentra en el servidor Solr y en segundo lugar los archivos de reporte que genera Heritrix.

A continuación veremos en detalle las tablas de la base de datos para el Área Intermedia.

Rastreo (ai_rastreo) Esta tabla contiene información de los rastreos realizados a las semillas.

Tabla 4.11: Detalle de la tabla rastreo

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada rastreo	Autoincremental
semilla_id	Varchar	Clave foránea de la tabla semilla (4.15)	-
fecha_inicio	Date	Fecha de inicio del rastreo	Tabla predictions - Base de datos App (Servidor Solr)
fecha_fin	Date	Fecha de finalización del rastreo	Tabla predictions - Base de datos App (Servidor Solr)
tiempo_inicio	Varchar	Hora de inicio del rastreo	Tabla predictions - Base de datos App (Servidor Solr)
tiempo_fin	Varchar	Hora de finalización del rastreo	Tabla predictions - Base de datos App (Servidor Solr)
ruta_carpeta	Varchar	Nombre de la carpeta en la que está almacenado el rastreo	Tabla predictions - Base de datos App (Servidor Solr)

Fuente: Elaboración propia.

Tipo MIME (ai_tipo_mime) Esta tabla contiene información sobre todos los tipos MIME.

Tabla 4.12: Detalle de la tabla tipo MIME

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada tipo MIME	Autoincremental
tipo	Varchar	Primera clasificación del tipo MIME	Lista oficial de los tipos MIME (IANA, 2018)
subtipo	Varchar	Segunda clasificación del tipo MIME	Lista oficial de los tipos MIME (IANA, 2018)
nombre_completo	Varchar	Nombre del tipo MIME	Lista oficial de los tipos MIME (IANA, 2018)
descripcion	Varchar	Descripción del tipo MIME	(IANA, 2018)

Fuente: Elaboración propia.

Código de estado HTTP (ai_codigo_estado_http) Esta tabla contiene información sobre todos los códigos de estado de HTTP.

Tabla 4.13: Detalle de la tabla código de estado HTTP

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada código de estado HTTP	Autoincremental
codigo	Entero	Código del estado HTTP	Recopilación de Mozilla (Mozilla y colaboradores, 2018a)
descripcion_corta	Varchar	Descripción del código de estado HTTP	Recopilación de Mozilla (Mozilla y colaboradores, 2018a)
tipo	Varchar	Nombre del Código de estado HTTP	Recopilación de Mozilla (Mozilla y colaboradores, 2018a)

Fuente: Elaboración propia.

Colección (ai_coleccion) Esta tabla contiene información sobre las colecciones, es decir, las posibles maneras de clasificar una Semilla.

Tabla 4.14: Detalle de la tabla colección

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada colecćon	Autoincremental
nombre	Varchar	Nombre de la colección	Tabla coleccions - Base de datos App (Servidor Solr)
descripcion	Varchar	Descripción de la coleccion	Tabla coleccions - Base de datos App (Servidor Solr)

Fuente: Elaboración propia.

Semilla (ai_semilla) Esta tabla contiene información de las semillas rastreadas.

Tabla 4.15: Detalle de la tabla semilla

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada semilla	Autoincremental
url	Varchar	URL de la semilla	Tabla traces - Base de datos App (Servidor Solr)
nombre	Varchar	Nombre de la semilla	Tabla traces - Base de datos App, Servidor Solr
descripcion	Varchar	Descripción de la semilla	Tabla traces - Base de datos App, Servidor Solr
id_coleccion	Entero	Clave foránea de la tabla colección (4.14)	-

Fuente: Elaboración propia.

Métricas tipo MIME (ai_metricas_tipo_mime) Esta tabla contiene información de las métricas de los tipos MIME.

Tabla 4.16: Detalle de la tabla métricas tipo MIME

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada métrica	Autoincremental
rastreo_id	Entero	Clave foránea de la tabla rastreo (4.11)	-
tipo_mime_id	Entero	Cláve foránea de la tabla tipo MIME (4.12)	-
cantidad_urls	Entero	Cantidad de URLs del tipo MIME por rastreo	Archivo mimetype-report.txt (Servidor Heritrix1)

Fuente: Elaboración propia.

Métricas códigos de estado HTTP (ai_metricas_codigo_estado) Esta tabla contiene información de las métricas de los códigos de estado.

Tabla 4.17: Detalle de la tabla métricas códigos de estado

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada métrica	Autoincremental
rastreo_id	Entero	Clave foránea de la tabla rastreo (4.11)	-
codigo_estado_id	Entero	Clave foránea de la tabla código HTTP (4.13)	-
cantidad_urls	Entero	Cantidad de URLs por código de estado	Archivo responsecode-report.txt (Servidor Heritrix1)

Fuente: Elaboración propia.

Métricas códigos de estado (ai_metricas_rastreo) Esta tabla contiene información de las métricas de los rastreos realizados.

Tabla 4.18: Detalle de la tabla métricas rastreo

Atributo	Tipo de dato	Descripción	Procedencia
id	Entero	Clave que identifica a cada métrica	Autoincremental
rastreo_id	Entero	Clave foránea de la tabla rastreo (4.11)	-
duracion	Tiempo	Duración de un rastreo	Archivo crawl-report.txt (Servidor Heritrix1)
hosts_visitados	Entero	Hosts que fueron visitados en el rastreo	Archivo crawl-report.txt (Servidor Heritrix1)
uris_procesadas	Entero	Total de URIs procesadas en un rastreo	Archivo crawl-report.txt (Servidor Heritrix1)
uris_exitosas	Entero	Cantidad de URIs que fueron procesadas exitosamente en un rastreo	Archivo crawl-report.txt (Servidor Heritrix1)
uris_fallidas	Entero	Cantidad de URIs que fueron abortadas en un rastreo	Archivo crawl-report.txt (Servidor Heritrix1)
uris_ignoradas	Entero	Cantidad de URIs que fueron ignoradas en un rastreo	Archivo crawl-report.txt (Servidor Heritrix1)
total_bytes_rastreados	Entero	Total de bytes que fueron rastreados	Archivo crawl-report.txt (Servidor Heritrix1)
uris_por_segundo	Real	Total de URIs rastreadas por segundo	Archivo crawl-report.txt (Servidor Heritrix1)
amcho_de_banda	Real	Ancho de banda durante el rastreo	Archivo crawl-report.txt (Servidor Heritrix1)

Fuente: Elaboración propia.

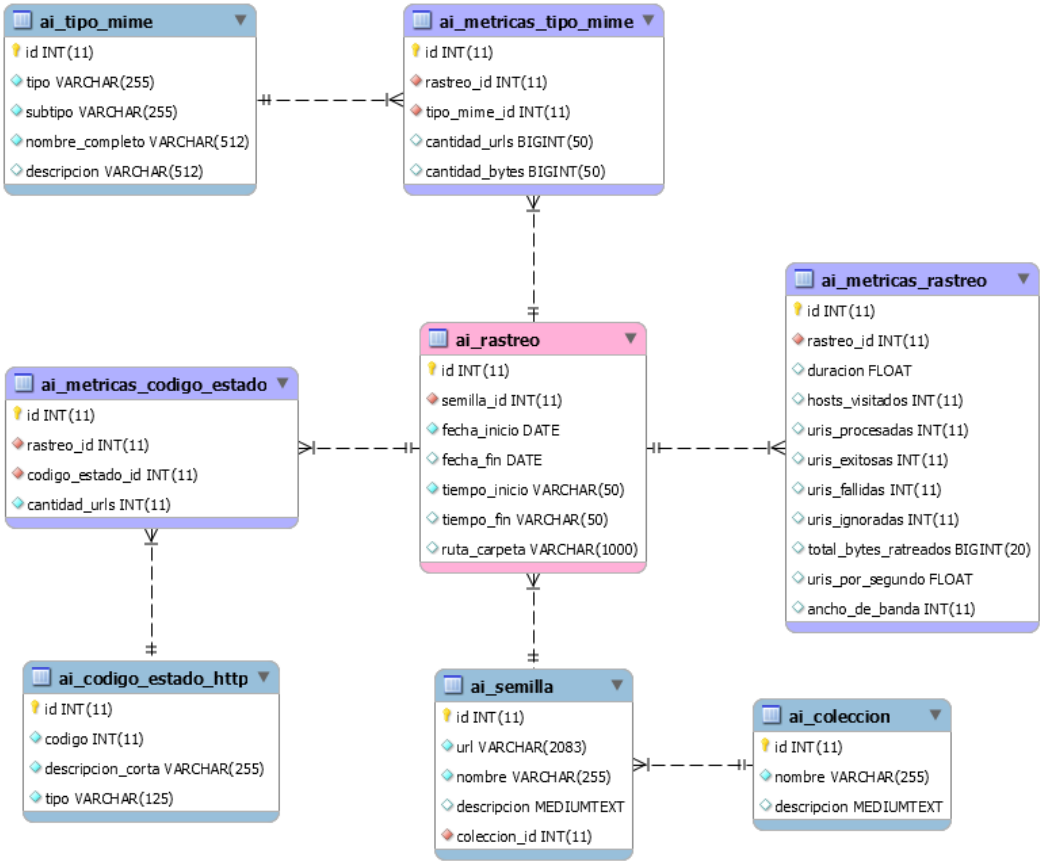

Figura 4.5: Modelo del Área Intermedia
Fuente: Elaboración propia.

4.6. Diseño y construcción de procesos ETC

En esta sección se comentará el desarrollo de los procesos de extracción, transformación y carga que se realizaron con el fin de poblar las bases de datos del área intermedia y el almacén.

4.6.1. Proceso ETC del Área Intermedia

Como fuentes de datos tenemos los archivos .txt que arroja el servidor Heritrix para cada uno de los rastreos y la base de datos `app` en el servidor de Solr. Para este proceso se utilizó la herramienta de Pentaho Data Integration, **Spoon** (anteriormente conocida como Kettle).

La figura 4.6 muestra el *job* general que se realizó para insertar los datos en la base de datos intermedia en MySQL que se encuentra en el Sandbox de Hortonworks.

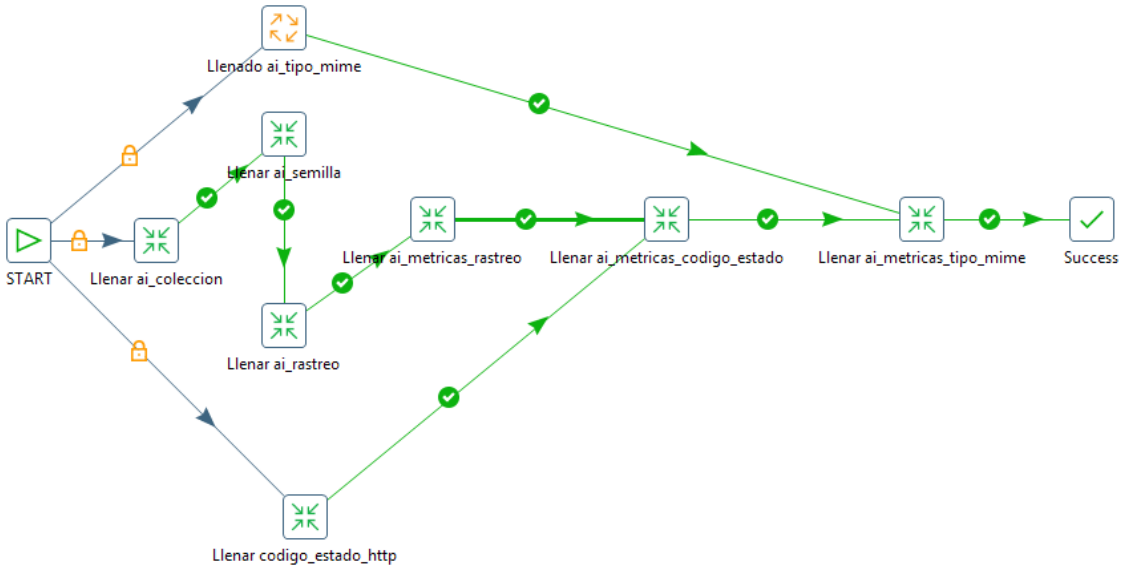

Figura 4.6: Job para llenar el área intermedia.
Fuente: Elaboración propia.

A continuación se describirán las funciones de cada elemento del *job*.

1. *Job* tipo_mime: este *job* se encarga de llenar la tabla de tipo MIME en el área intermedia. En él se carga la información por cada tipo MIME macro es decir, *application*, *text*, *image*, *video*, *multipart*, *font*, *model* y *audio* los cuales estaban en formato .csv. Como se menciona en la tabla 4.12 esta información fue extraída del sitio web de IANA, la cual es la lista oficial de todos los tipos MIME existentes.

2. Transformación ai_coleccion: esta transformación llena la tabla ai_coleccion del área intermedia y debe ir primero ya que es clave foránea en la tabla ai_semilla. Esta información fue extraida de la tabla `coleccions` de la base de datos `App` del servidor Solr, como se menciona en la tabla 4.14.

3. Transformación ai_semilla: esta transformación llena la tabla ai_semilla del área intermedia y debe ir primero ya que es clave foránea de la tabla ai_rastreo. La información de esta tabla fue extraída de la tabla `traces` de la base de datos `App` del servidor Solr, como se menciona en la tabla 4.15.

4. Transformación ai_rastreo: esta transformación llena la tabla ai_rastreo, la cual comprende todos los rastreos realizados por Heritrix. Los datos fueron extraí-

dos de la tabla `predictions` de la base de datos `App` del servidor Solr, como se menciona en la tabla 4.11.

5. Transformación ai_metricas_rastreo: esta transformación llena la tabla ai_metricas_rastreo la cual contiene información relevante de los mismos, como los URLs procesados y URLs fallidos. Una de los pasos mas importantes es la transformación de la duración de los rastreos, la cual tenía un formato '0d0h0m0s0ms' y se convirtió a un número entero estandarizandose a segundos para realizar de una manera mas sencillas los cálculos pertinentes. En los archivos crawl-report.txt no se especifica a que rastreo pertenecen dichas métricas, por lo que se se tomó la ruta del archivo, se separó el nombre de la carpeta y se hizo *join* con el campo ruta_carpeta el cual tiene el nombre de la carpeta del rastreo.

```
1  crawl name: basic
2  crawl status: Finished - Ended by operator
3  duration: 2h12m47s612ms
4
5  seeds crawled: 1
6  seeds uncrawled: 0
7
8  hosts visited: 13
9
10 URIs processed: 2229
11 URI successes: 2172
12 URI failures: 0
13 URI disregards: 57
14
15 novel URIs: 2172
16
17 total crawled bytes: 115456739 (110 MiB)
18 novel crawled bytes: 115456739 (110 MiB)
19
20 URIs/sec: 0.27
21 KB/sec: 14
22
```

Figura 4.7: Ejemplo de archivo .txt - crawl-report

6. Transformación ai_metricas_codigo_estado: se encarga de llenar la tabla ai _metricas_codigo_estado. Al igual que en la transformación anterior, el archivo response-code-report.txt no menciona a que rastreo pertenece, por lo que se tomó la ruta del archivo, se separó el nombre de la carpeta y se hizo *join* con el campo ruta_carpeta el cual tiene el nombre de la carpeta del rastreo. Además se

realizó otro *join* con la tabla ai_codigo_estado_http por los códigos de estado, el cual es el campo `rescode` del archivo .txt.

```
[#urls] [rescode]
2100 200
42 301
13 1
13 404
1 204
1 302
1 400
1 403
```

Figura 4.8: Ejemplo de archivo .txt - response-code-report

7. Transformación ai_metricas_tipo_mime: se encarga de llenar la tabla ai _metricas_tipo_mime. Al igual que en las dos transformaciones anteriores, el archivo mimetype-report.txt no menciona a que rastreo pertenece, por lo que se tomó la ruta del archivo, se separó el nombre de la carpeta y se hizo *join* con el campo ruta_carpeta el cual tiene el nombre de la carpeta del rastreo. Además se realizó otro *join* con la tabla ai_tipo_mime por los nombres de los tipos MIME, el cual es el campo `mime-type` del archivo .txt.

```
[#urls] [#bytes] [mime-types]
1659 108565110 text/html
390 424648 application/opensearchdescription+xml
27 2246503 application/atom+xml
27 947935 application/rss+xml
13 500896 application/javascript
13 2457 text/dns
13 21484 text/plain
7 559409 text/css
6 224523 font/ttf
5 1420178 text/javascript
3 455125 application/x-shockwave-flash
3 8874 image/x-icon
1 25470 application/vnd.ms-fontobject
1 25680 application/x-font-woff
1 384 image/gif
1 1397 image/jpeg
1 1405 image/vnd.microsoft.icon
1 25261 unknown

```

Figura 4.9: Ejemplo de archivo .txt - response-code-report

8. Transformación codigo_estado_http: Esta transformación llena la tabla ai_codigo_estado. Como se menciona en la tabla 4.13 esta información fue extraída del sitio web de IANA, el cual posee la lista de todos los códigos de estado HTTP.

4.6.2. Verificación de calidad de datos del Área Intermedia

Una vez realizada la inserción de datos en el área intermedia, es necesario comprobar que la carga ha sido correcta y que no fueron omitidos registros importantes en el proceso. Para esto se comparan los registros de las tablas equivalentes en los esquemas de base de datos respectivos.

Tabla de colecciones

Se encuentran tres registros en la fuente y de igual forma se encontraron tres registros en el Área Intermedia, indicando que el resultado del proceso fue correcto.

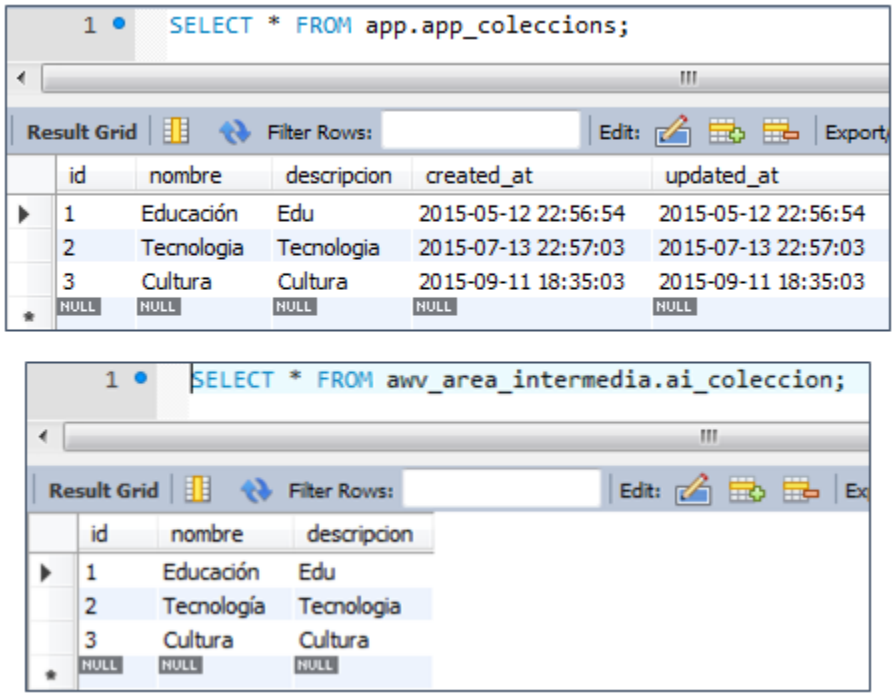

Figura 4.10: Contenido de la tabla de colecciones en la fuente (arriba) y en el área intermedia (abajo).

Tabla de semillas

Para la tabla de semillas se obtuvieron exactamente los mismos registros en igual orden, esto debido a la sencillez de su estructura.

```
1 • SELECT * FROM app.app_traces;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Ce

id	url	nombre_job	observaciones	coleccion_id
36	http://www.oncti.gob.ve	ONCTI	0	2
37	http://www.cito.com.ve	CITO	0	2
38	http://www.cida.gob.ve/cida_home/	CIDA	0	1
39	http://www.cendit.gob.ve/	CENDIT	0	2
40	http://www.fidetel.gob.ve/	FIDETEL	0	2
41	http://fundacredesa.gob.ve/	FUNDACREDESA	0	2
42	http://www.idea.gob.ve/	IDEA	0	1
43	http://www.inzit.gob.ve/	INZIT	0	1
44	http://www.ivic.gob.ve/	IVIC	0	1
45	http://www.ministeriodelacultura.gob.ve/	MINISTERIO DE CULTURA	0	3
NULL	NULL	NULL	NULL	NULL

```
1 • SELECT * FROM awv_area_intermedia.ai_semilla;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap

id	url	nombre	descripcion	coleccion_id
1	http://www.oncti.gob.ve	ONCTI	0	2
2	http://www.cito.com.ve	CITO	0	2
3	http://www.cida.gob.ve/cida_home/	CIDA	0	1
4	http://www.cendit.gob.ve/	CENDIT	0	2
5	http://www.fidetel.gob.ve/	FIDETEL	0	2
6	http://fundacredesa.gob.ve/	FUNDACREDESA	0	2
7	http://www.idea.gob.ve/	IDEA	0	1
8	http://www.inzit.gob.ve/	INZIT	0	1
9	http://www.ivic.gob.ve/	IVIC	0	1
10	http://www.ministeriodelacultura.gob.ve/	MINISTERIO DE CULTURA	0	3
NULL	NULL	NULL	NULL	NULL

Figura 4.11: Contenido de la tabla de semillas en la fuente (arriba) y en el área intermedia (abajo).

Tabla de códigos de estado HTTP

Los datos de esta tabla fueron extraídos de la fuente oficial, el IANA, por lo que al no tener acceso a los registros en un SMBD como fuente se realizó un muestreo aleatorio para su verificación con la tabla del área intermedia.

Tabla 4.19: Datos de la muestra aleatoria de tipos MIME seleccionados.

Código	Descripción corta	Tipo
100	Continue	[RFC7231, Section 6.2.1]
200	OK	[RFC7231, Section 6.3.1]
300	Multiple Choices	[RFC7231, Section 6.4.1]
400	Bad Request	[RFC7231, Section 6.5.1]
500	Internal Server Error	[RFC7231, Section 6.6.1]

Fuente: Elaboración propia.

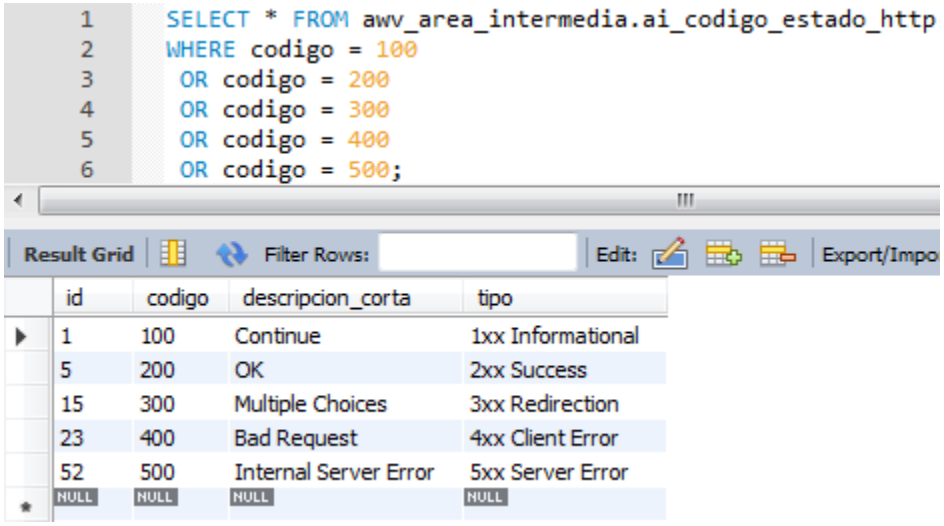

Figura 4.12: Resultado de la consulta de la muestra aleatoria de códigos de estado HTTP en el Área Intermedia.

Tabla de tipos MIME

Similar al caso anterior, los tipos MIME fueron extraídos del listado oficial mantenido por el IANA, para su comprobación de calidad se realizó un muestreo aleatorio de un registro por cada uno de los tipos (tabla 4.20), comprobando de manera manual que sus campos estuvieran correctos.

Tabla 4.20: Datos de la muestra aleatoria de tipos MIME seleccionados.

Subtipo	Nombre	Referencia
cdni	application/cdni	[RFC7736]
mp4	application/mp4	[RFC4337][RFC6381]
aac	audio/aac	[ISO-IEC_JTC1][Max_Neuendorf]
mp4	audio/mp4	[RFC4337][RFC6381]
otf	font/otf	[RFC8081]
tiff	image/tiff	[RFC3302]
3mf	model/3mf	[http://www.3mf.io/specification][_3MF][Michael_Sweet]
form-data	multipart/form-data	[RFC7578]
css	text/css	[RFC2318]
mp4	video/mp4	[RFC4337][RFC6381]

Fuente: Elaboración propia.

```
1  SELECT * FROM awv_area_intermedia.ai_tipo_mime
2  WHERE subtipo = 'cdni'   -- application
3      OR subtipo = 'aac'   -- audio
4      OR subtipo = 'otf'   -- font
5      OR subtipo = 'tiff'  -- image
6      OR subtipo = 'sip'   -- message
7      OR subtipo = '3mf'   -- model
8      OR subtipo = 'form-data'    -- multipart
9      OR subtipo = 'css'   -- text
10     OR subtipo = 'mp4'; -- video
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell C

id	tipo	subtipo	nombre_completo	descripcion
47	application	cdni	application/cdni	[RFC7736]
197	application	mp4	application/mp4	[RFC4337][RFC6381]
1319	audio	aac	audio/aac	[ISO-IEC_JTC1][Max_Neuendorf]
1389	audio	mp4	audio/mp4	[RFC4337][RFC6381]
1466	font	otf	font/otf	[RFC8081]
1495	image	tiff	image/tiff	[RFC3302]
1534	model	3mf	model/3mf	[http://www.3mf.io/specification][_3MF][Michael_Sweet]
1565	multipart	form-data	multipart/form-data	[RFC7578]
1580	text	css	text/css	[RFC2318]
1677	video	mp4	video/mp4	[RFC4337][RFC6381]
NULL	NULL	NULL	NULL	NULL

Figura 4.13: Resultado de la consulta de la muestra aleatoria de tipos MIME en el Área Intermedia.

Tabla de rastreos

Esta es una de las tablas que sufrió mayores transformaciones por lo que su estructura varió bastante en el Área Intermedia respecto a la fuente de datos, en la figura 4.14 se puede ver una muestra y el conteo realizado en ambos esquemas. Como se puede observar, se obtienen el mismo número de registros, unos cincuenta rastreos.

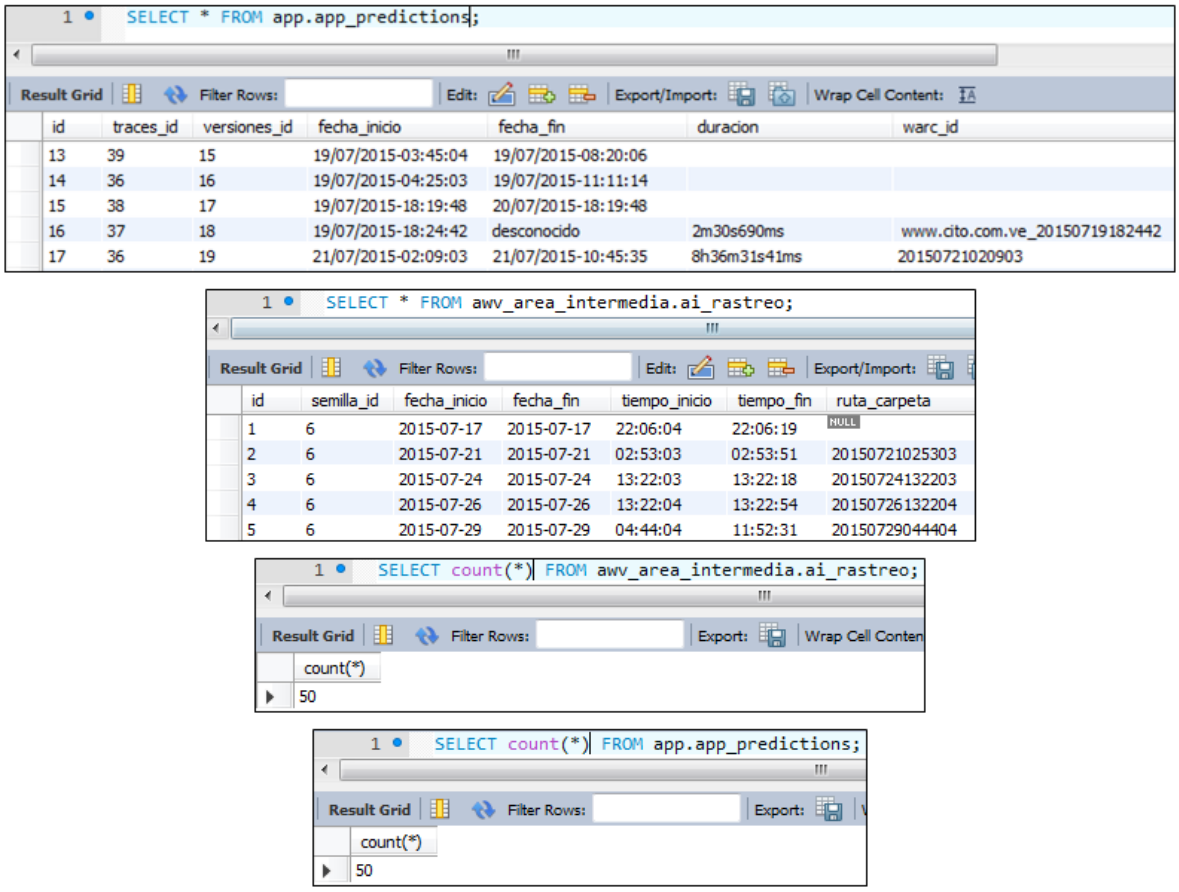

Figura 4.14: Muestra de resultados de consultas de rastreos en la fuente y en el Área Intermedia.

Tabla de métricas de rastreo

Para esta tabla que involucra información de archivos de texto se escogió un rastreo aleatorio para verificar si sus métricas fueron cargadas correctamente. Así se tomo el rastreo con **id=30**, correspondiente a la semilla **IDEA** de fecha 24-07-2015, como se puede observar en la figura 4.16, la cadena de la duración muestra el valor `7m40s773ms` que haciendo la conversión a segundos, omitiendo los milisegundos,

nos da que la conversión registrada es correcta, el resultado es `7*60+40=460`.

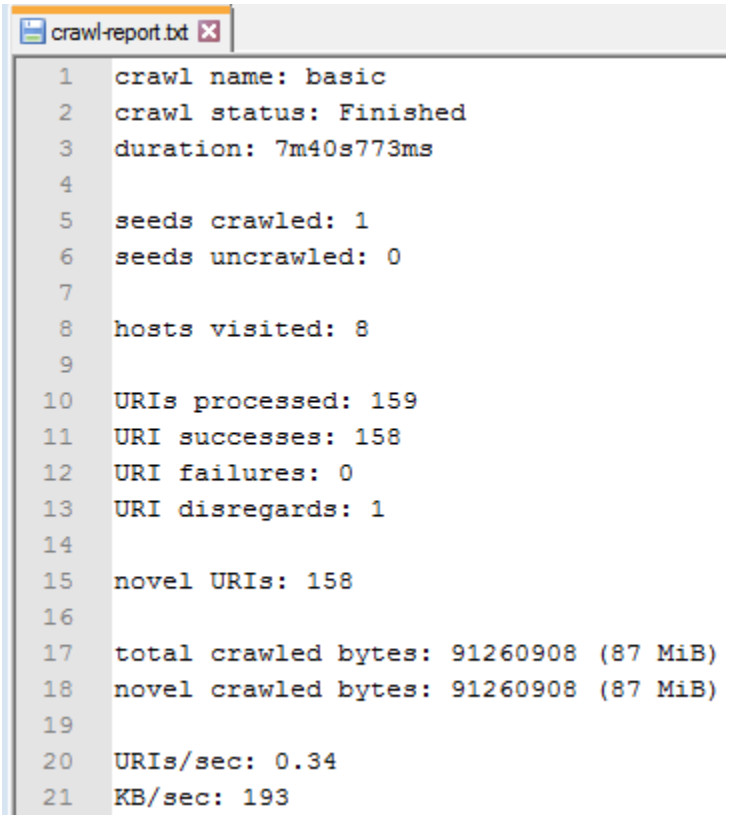

Figura 4.15: Archivo crawl-report.txt del rastreo con id=30 en el Área Intermedia.

Igualmente si comparamos el resto de los campos con el resultado de la consulta del rastreo en la figura 4.16, se comprueba que son correctos y por lo tanto los registros fueron bien insertados.

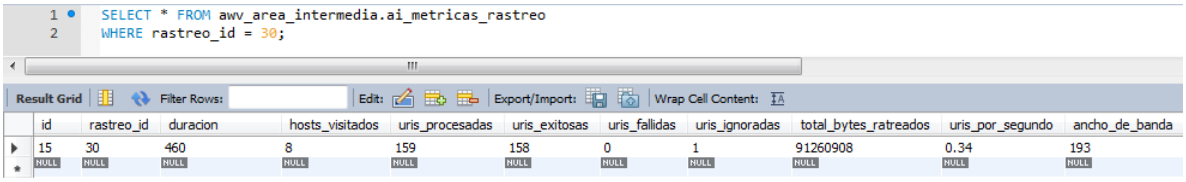

Figura 4.16: Resultado de consultar las métricas para el rastreo con id=30 en el Área Intermedia.

Tabla de métricas de código de estado

Para esta tabla se verificó igualmente el archivo correspondiente al rastreo de **id=30**, figura 4.17, con los resultados arrojados por la respectiva consulta, figura 4.18.

Omitiendo la fila con `rescode=1` debido a que no es un código de estado http oficial sino que es propio de Heritrix, se obtienen resultados correctos.

responsecode-report.txt

```
[#urls] [rescode]
125 200
18 404
8 1
7 301
```

Figura 4.17: Archivo responsecode-report.txt del rastreo con id=30 en el Área Intermedia.

```
SELECT mcod.*, cod.codigo
FROM awv_area_intermedia.ai_metricas_codigo_estado AS mcod
JOIN awv_area_intermedia.ai_codigo_estado_http AS cod
    ON mcod.codigo_estado_id = cod.id
WHERE rastreo_id = 30;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	rastreo_id	codigo_estado_id	cantidad_urls	codigo
14	30	5	125	200
38	30	16	7	301
97	30	27	18	404

Figura 4.18: Resultado de consultar las métricas de código de estado HTTP para el rastreo con id=30 en el Área Intermedia.

Tabla de métricas de tipo mime

Para esta tabla se comparó el archivo correspondiente al rastreo con **id=50**, figura 4.19, con los resultados que arrojó el mismo rastreo, figura 4.20. Como se puede observar, los resultados que se obtienen son correctos.

```
[#urls] [#bytes] [mime-types]
5921 68347569 text/html
291 531015460 application/pdf
60 67858477 image/jpeg
25 538199 application/atom+xml
25 250957 application/rss+xml
21 558644 application/javascript
8 273668 text/css
8 1231 text/dns
7 11429 text/plain
```

Figura 4.19: Archivo mimetype-report.txt del rastreo con id=50 en el Área Intermedia.

```
select m.rastreo_id, m.cantidad_bytes, m.cantidad_urls, t.nombre_completo
from ai_metricas_tipo_mime as m
join ai_tipo_mime as t on m.tipo_mime_id = t.id
join ai_rastreo as r on r.id = m.rastreo_id
where m.rastreo_id=50;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

rastreo_id	cantidad_bytes	cantidad_urls	nombre_completo
50	68347569	5921	text/html
50	531015460	291	application/pdf
50	67858477	60	image/jpeg
50	538199	25	application/atom+xml
50	558644	21	application/javascript
50	273668	8	text/css
50	1231	8	text/dns
50	11429	7	text/plain

Figura 4.20: Resultado de consultar las métricas de tipo MIME para el rastreo con id=50 en el Área Intermedia.

4.6.3. Proceso ETC del Almacén de Datos

Para este proceso se realizó un *job* de Spoon para poblar el esquema `awv_almacen` de MySQL donde, en primer lugar se llenan cada una de las dimensiones, necesarias para luego proceder con las tablas de hechos. La figura 4.21 muestra el *job* general. Este proceso fue más sencillo que el anterior ya que casi todas las transformaciones fueron realizadas en el ETC del área intermedia. A excepción de las dimensiones Fecha y Tiempo, el resto resultaron básicamente una copia del área intermedia, añadiendo una clave subrogada.

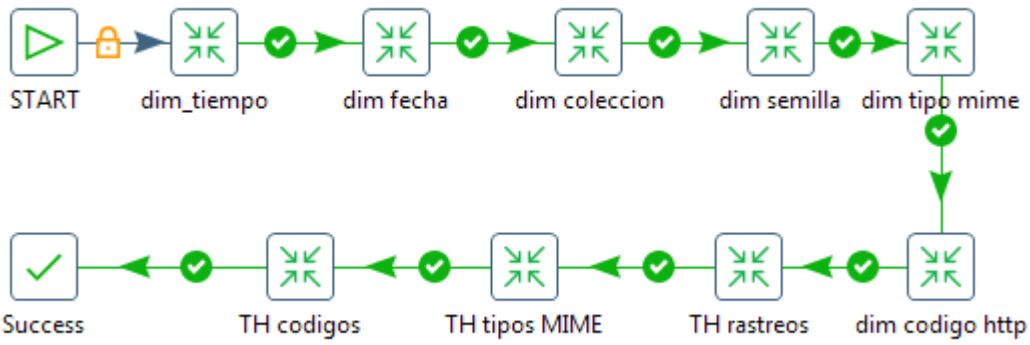

Figura 4.21: Job para llenar el almacén
Fuente: Elaboración propia.

Las dimensiones Fecha y Tiempo fueron generadas a partir de cálculos en Spoon, se ingresaron 10 mil fechas desde el 1ro de enero del 2015. Para la dimensión tiempo, además de los campos básicos de hora, minuto (de hora) y segundo (de minuto) se agregaron los de minuto y segundo del día, para ofrecer mayor variedad en las futuras consultas sobre el tiempo.

Luego de haber completado este primer proceso, se utilizó la herramienta Sqoop del ecosistema Hadoop para llevar los datos del esquema `awv_almacen` de MySQL hacia HDFS, para que pudieran leerse utilizando Hive. En la figura 4.22 podemos observar un ejemplo de comando de Sqoop para la importación de una tabla.

```
echo "===>>> importando dimensión coleccion"
sqoop-import --connect jdbc:mysql://localhost:3306/awv_almacen \
  --username root \
  --driver com.mysql.jdbc.Driver \
  --table dim_coleccion \
  --as-parquetfile \
  --password-file /user/maria_dev/passwordfile -m 1 \
  --target-dir /user/maria_dev/awv_almacen/dw/dim_coleccion
```

Figura 4.22: Ejemplo de importación de datos con Sqoop.

4.6.4. Verificación de calidad de datos del Almacén de Datos

Después de haber insertado la data en el almacén es necesario verificar que los datos insertados son correctos al igual que en el área intermedia. A continuación

verificaremos los datos del almacén con el área intermedia, la cual fue verificada anteriormente.

Dimensión colección

Fueron encontrados tres registros en el almacén al igual que en el área intermedia, por lo cual el resultado de la carga fue exitoso.

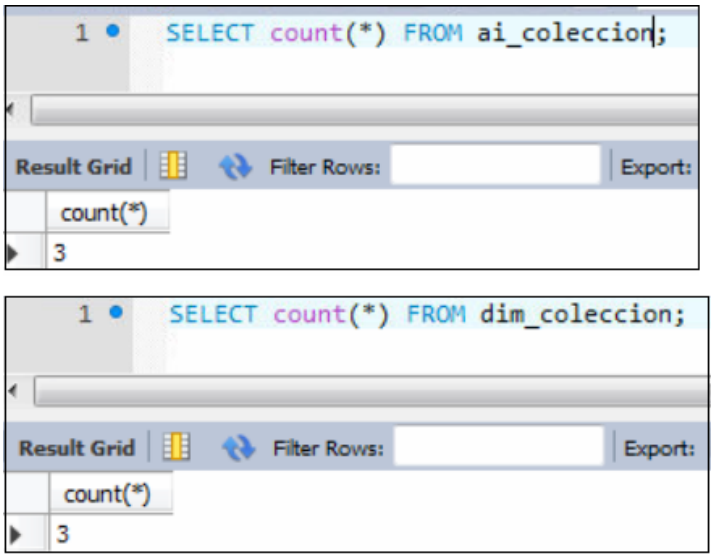

Figura 4.23: Muestra de resultados de consultas de colecciones en el Área Intermedia y en el Almacén.

Dimensión semilla

En este proceso se encontraron diez registros en el almacén y el área intermedia, por lo que el proceso de carga fue exitoso.

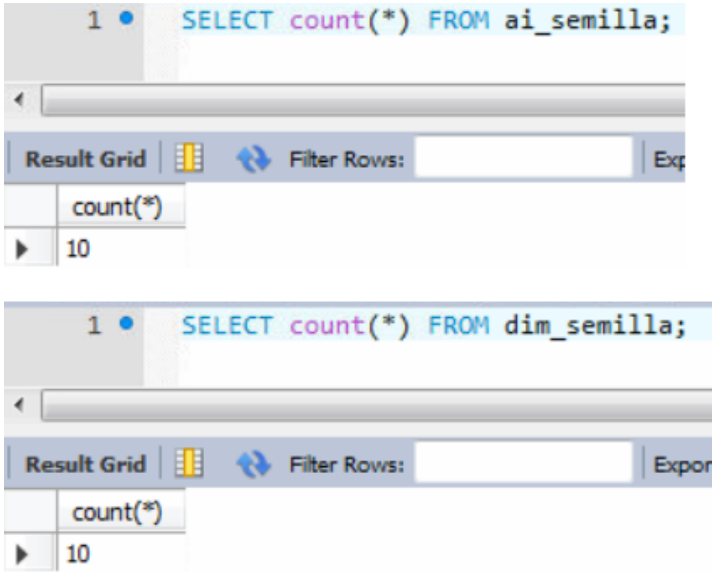

Figura 4.24: Muestra de resultados de consultas de semillas en el Área Intermedia y en el Almacén.

Dimensión tipo mime

Al igual que en los procesos anteriores, se encontraron la misma cantidad de registros en el área intermedia y en el almacén.

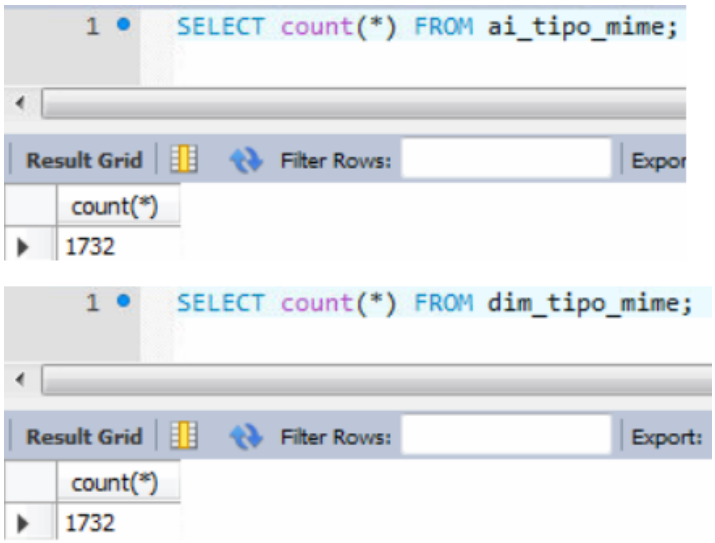

Figura 4.25: Muestra de resultados de consultas de tipos MIME en el Área Intermedia y en el Almacén.

Dimensión código de estado

En esta dimensión ocurre lo mismo que en las anteriores, se encontraron la misma cantidad de registros en ambos esquemas.

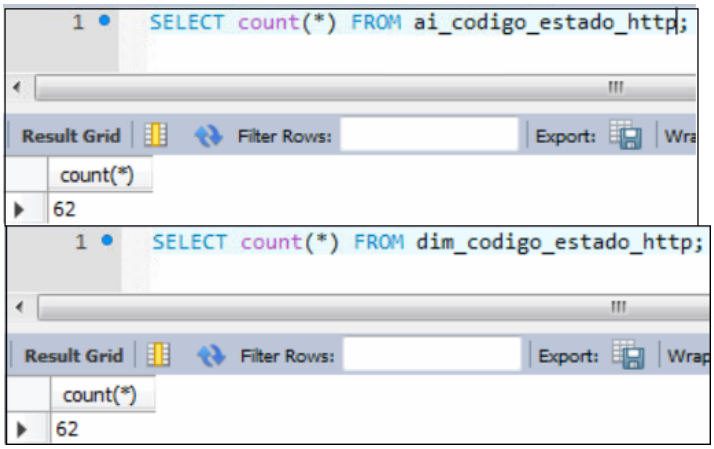

Figura 4.26: Muestra de resultados de consultas de códigos de estado en el Área Intermedia y en el Almacén.

Tabla de hechos de código de estado http

La tabla de hechos de código de estado http es producto de la intersección de varias tablas, entre ellas, la tabla del área intermedia de métricas código de estado, la dimensión fecha y la dimensión de códigos de estado, sin embargo, el número de registros en ambos esquemas da igual, por lo que el proceso fue exitoso.

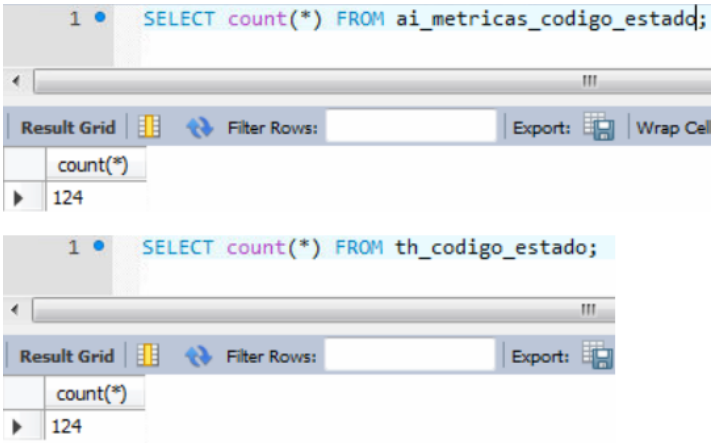

Figura 4.27: Muestra de resultados de consultas de códigos de estados y métricas en el Área Intermedia y en el Almacén.

Tabla de hechos de rastreo

Al igual que la anterior, la tabla de hechos de rastreo es producto de la intersección de varias tablas, entre ellas, las tablas rastreo y métricas de rastreo del área intermedia, y las dimensiones semilla y colección, de igual forma, el resultado de ambos esquemas es el mismo, por lo que el proceso fue exitoso.

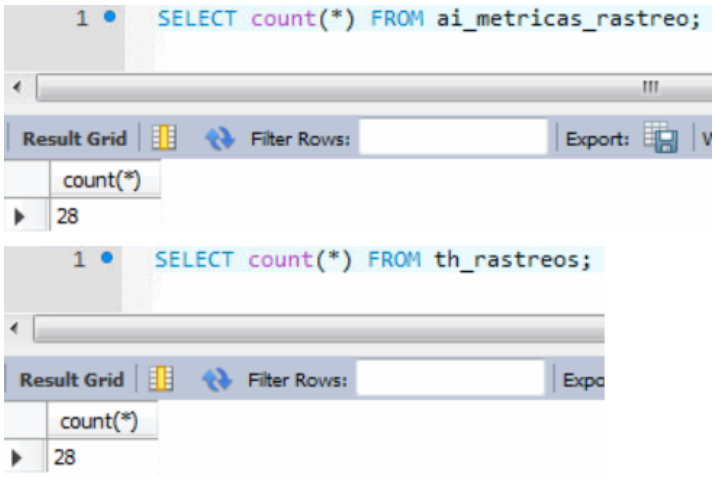

Figura 4.28: Muestra de resultados de consultas de rastreos en el Área Intermedia y en el Almacén.

Tabla de hechos de tipo mime

Por último, la tabla de hechos de tipo mime también es intersección de varias tablas. El resultado en ambos esquemas es el mismo, por lo que el proceso de carga fue exitoso.

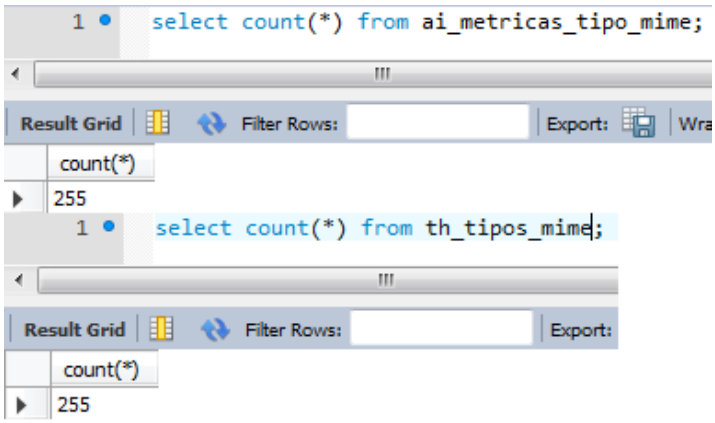

Figura 4.29: Muestra de resultados de consultas de tipos MIME en el Área Intermedia y en el Almacén.

4.7. Aplicación BI

En primer lugar se realizó una conexión entre Tableau Desktop y Hive, seguidamente se importó el almacén, creando tres libros de trabajo en (Uno por tabla de hechos) indicando las tablas de hechos y sus correspondientes dimensiones. A continuación se verá la importación de datos en Tableau de las Tablas de Hechos.

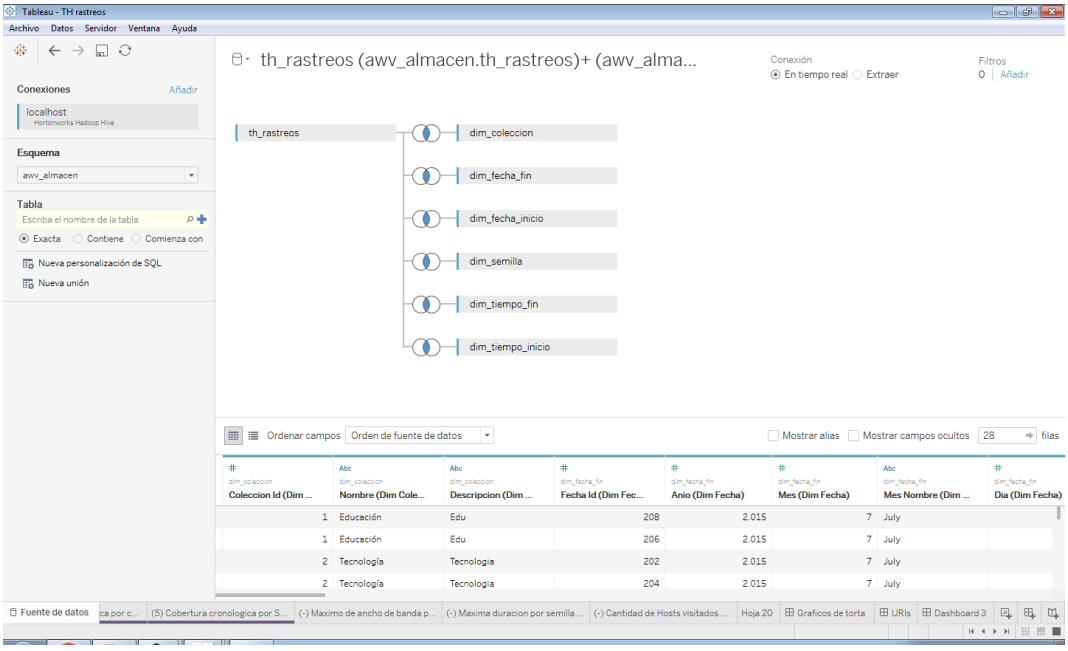

Figura 4.30: Importación de datos de la Tabla de Hechos de Rastreos en Tableau
Fuente: Elaboración propia.

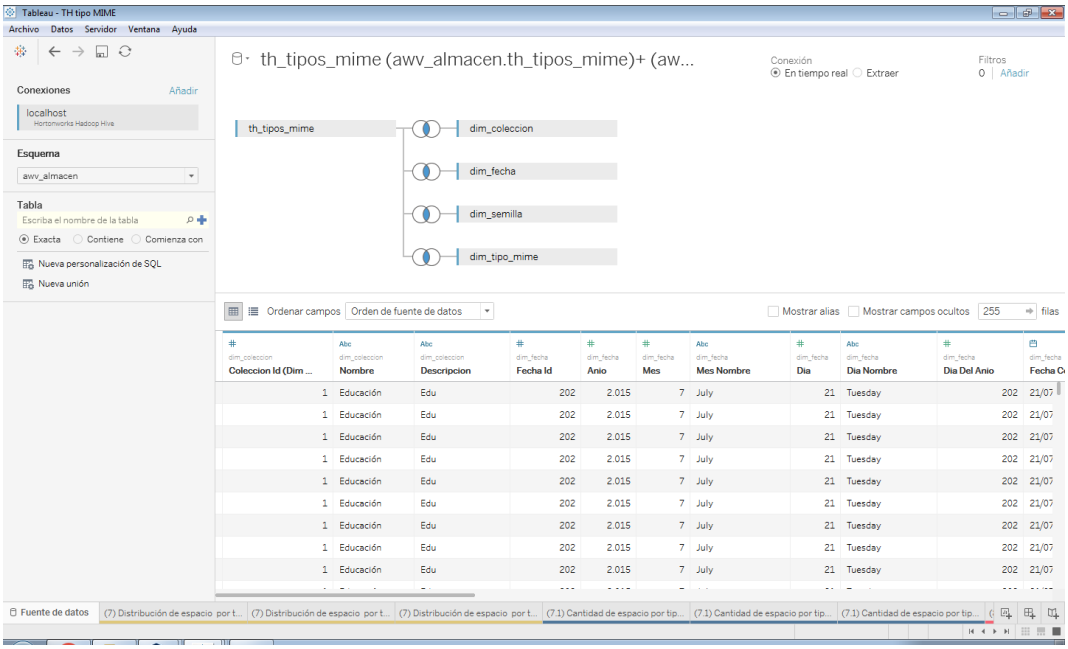

Figura 4.31: Importación de datos de la Tabla de Hechos de tipo MIME en Tableau
Fuente: Elaboración propia.

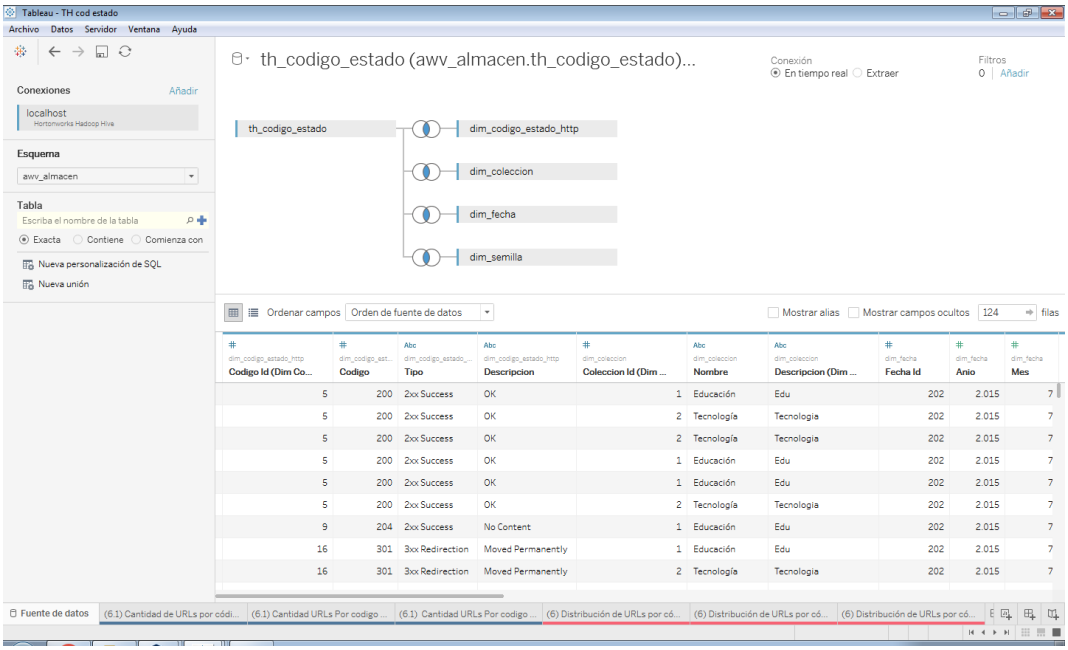

Figura 4.32: Importación de datos de la Tabla de Hechos de Código de Estado HTTP
Fuente: Elaboración propia.

4.8. Implementación

En esta sección se describirá el proceso de implementación del proyecto en la fase de acceso a los datos. Para comenzar, se utilizó un computador de 14 GB de memoria RAM y procesador Core i5. Se trabajó bajo el sistema operativo Windows 7 por ser requerido por Tableau Desktop el ambiente Windows (o Mac OS).

Para la simulación de un clúster Hadoop, se utilizó una máquina virtual, el Sandbox Hortonworks Data Platform (versión de VirtualBox) que ya incluye instaladas un conjunto de herramientas del ecosistema Hadoop. Se hizo un respaldo de la base de datos `app` hacia el MySQL de la máquina virtual, y de los demás esquemas igualmente.

4.8.1. Indicadores

A continuación se verán los resultados con las gráficas de algunos indicadores.

- Cantidad de Rastreos por fecha: En este indicador podemos ver los rastreos realizados por fecha. Se puede filtrar la búsqueda entre dos fechas, además como existe una jerarquía podemos consultar por año, trimestre y mes.

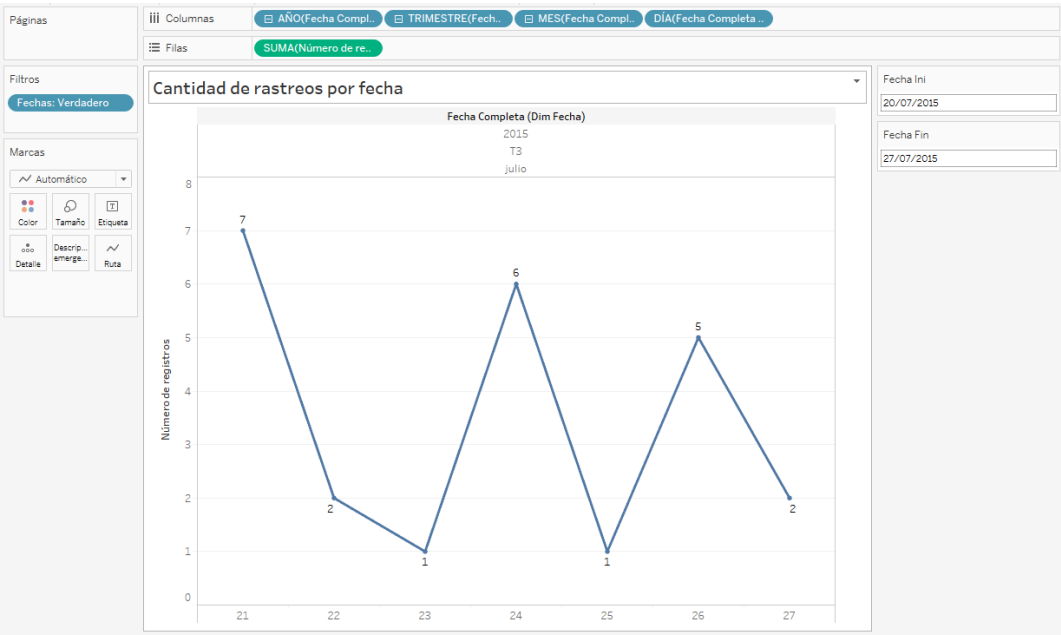

Figura 4.33: Indicador: Cantidad de rastreos por Fecha
Fuente: Elaboración propia.

- Cantidad de URIs por Semilla: En este indicador podemos ver las URIs procesadas, exitosas, ignoradas y fallidas de todos los rastreos por cada semilla.

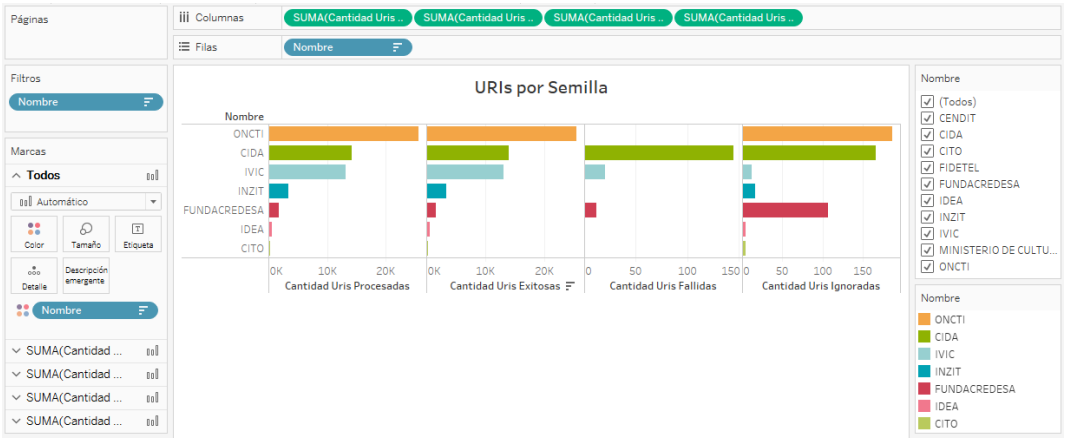

Figura 4.34: Indicador: Cantidad de URIs por semilla
Fuente: Elaboración propia.

- Promedio de Duración por Semilla: En este indicador podemos ver el promedio de duración en minutos de cada semilla. Además podemos filtrar por Colección y Semilla.

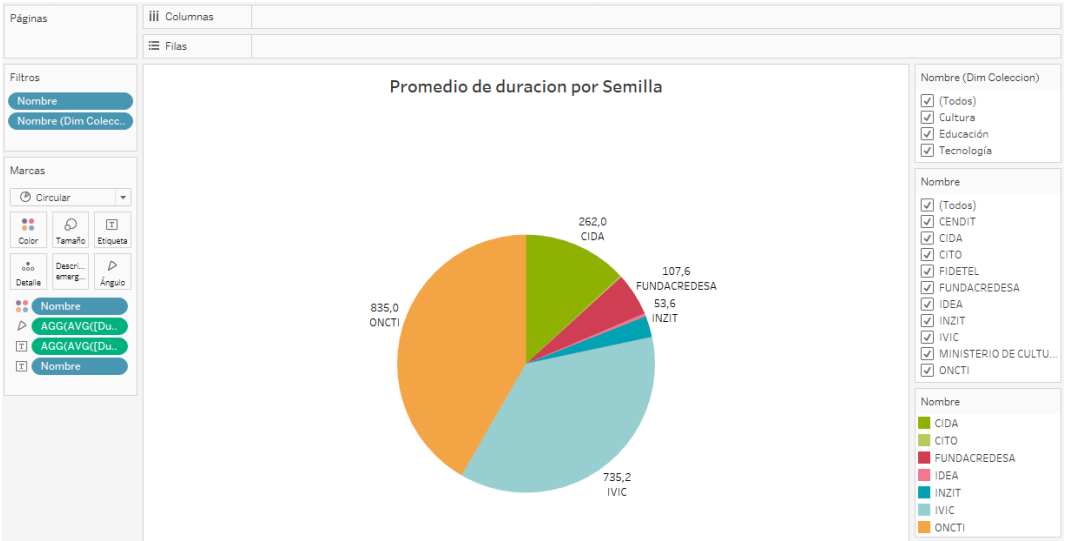

Figura 4.35: Indicador: Promedio de duración por semilla
Fuente: Elaboración propia.

- Distribución de espacio por tipos de formato por Colección: En este indicador podemos ver el porcentaje de espacio que ocupan los tipos MIME por cada colección. En este indicador podemos filtrar por tipo y subtipo MIME.

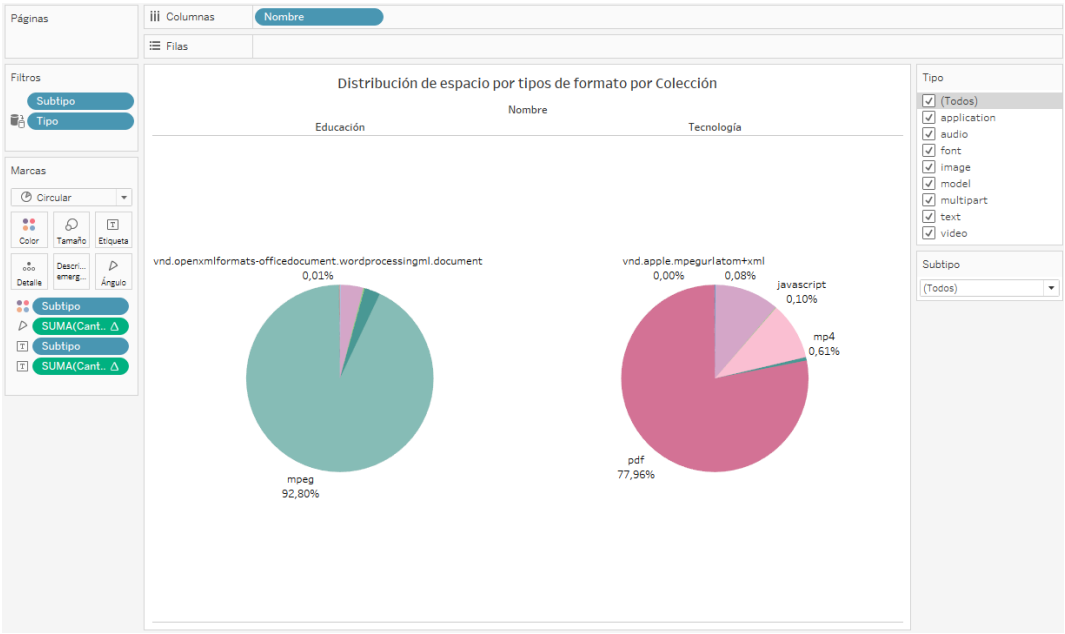

Figura 4.36: Indicador: Distribución de espacio por tipos de formato por Colección
Fuente: Elaboración propia.

- Distribución de URLs por código de estado por Semilla: En este indicador se puede ver el porcentaje de códigos de estado HTTP que arroja cada semilla. En este caso para mejor entendimiento, se filtraron las demás semillas para que se vieran las de CIDA y FUNDACREDESA. En este indicador podemos filtrar por Código de estado HTTP y por Semilla.

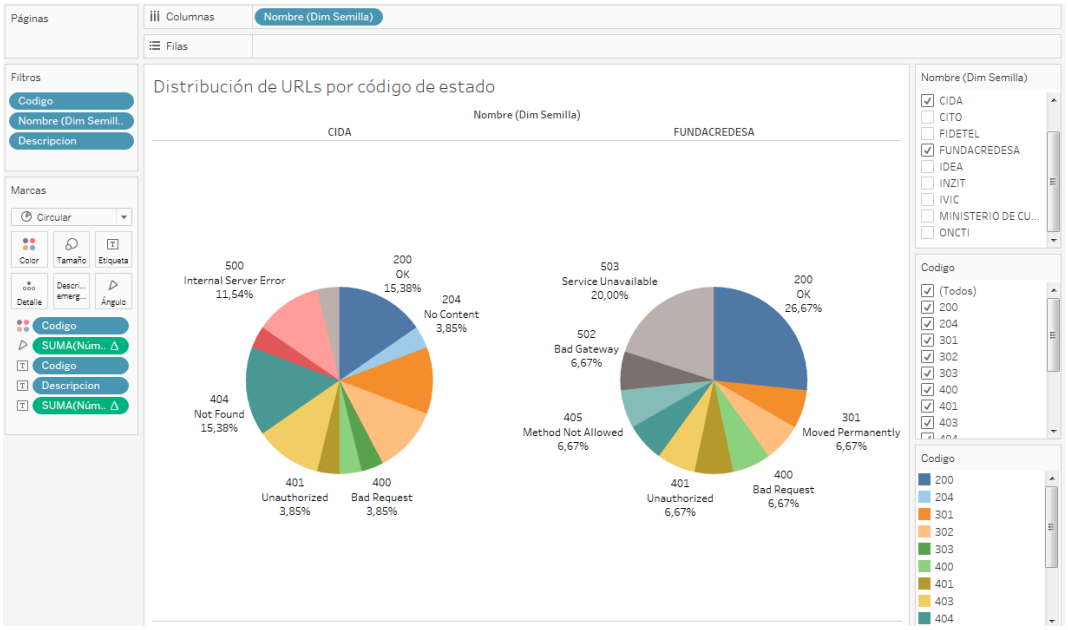

Figura 4.37: Indicador: Distribución de URLs por código de estado por Semilla
Fuente: Elaboración propia.

4.8.2. Cuadros de mando

En esta sección se verá algunos de los cuadros de mando, los cuales resumen en una sola página los indicadores que se han realizado.

En el siguiente cuadro de mando se puede observar los indicadores relacionados a la tabla de Hechos de rastreo que fueron realizados con gráficos circulares (torta). Se puede clasificar la información por semilla o colección y se pueden ver las leyendas de los indicadores del lado derecho.

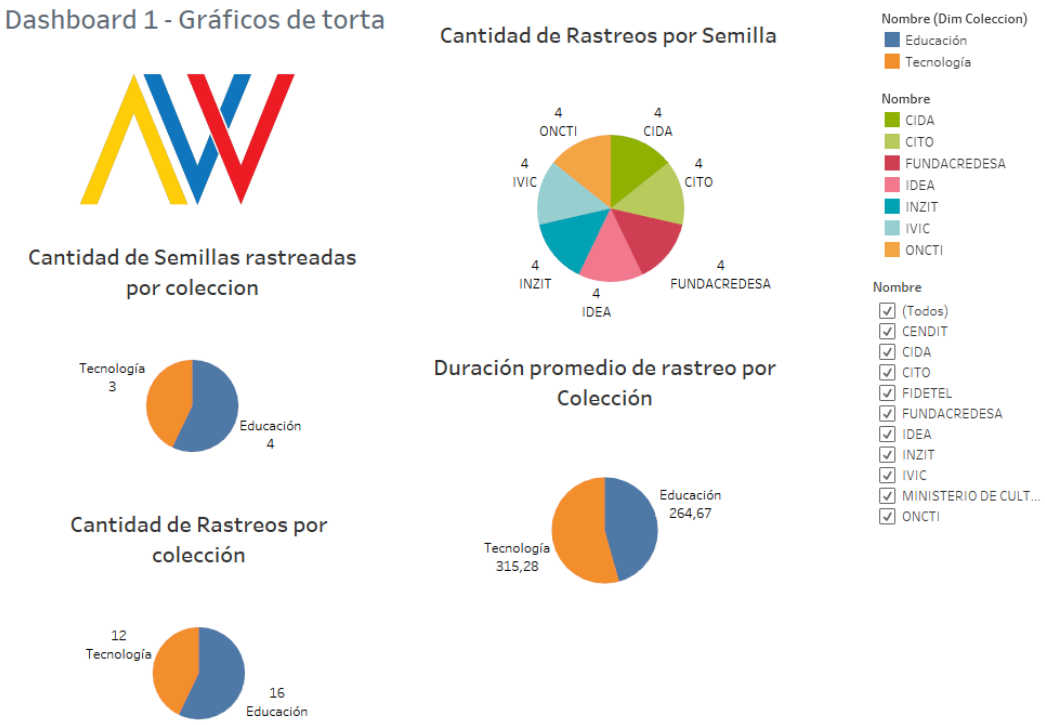

Figura 4.38: Cuadro de mando de gráficos de torta
Fuente: Elaboración propia.

En el siguiente cuadro de mando se observan indicadores relacionados a las URIs de la tabla de Hechos de Rastreo, en este caso se puede ver en la misma página las URIs procesadas, exitosas, fallidas e ignoradas por semilla y colección. Se puede clasificar la información por semilla o colección.

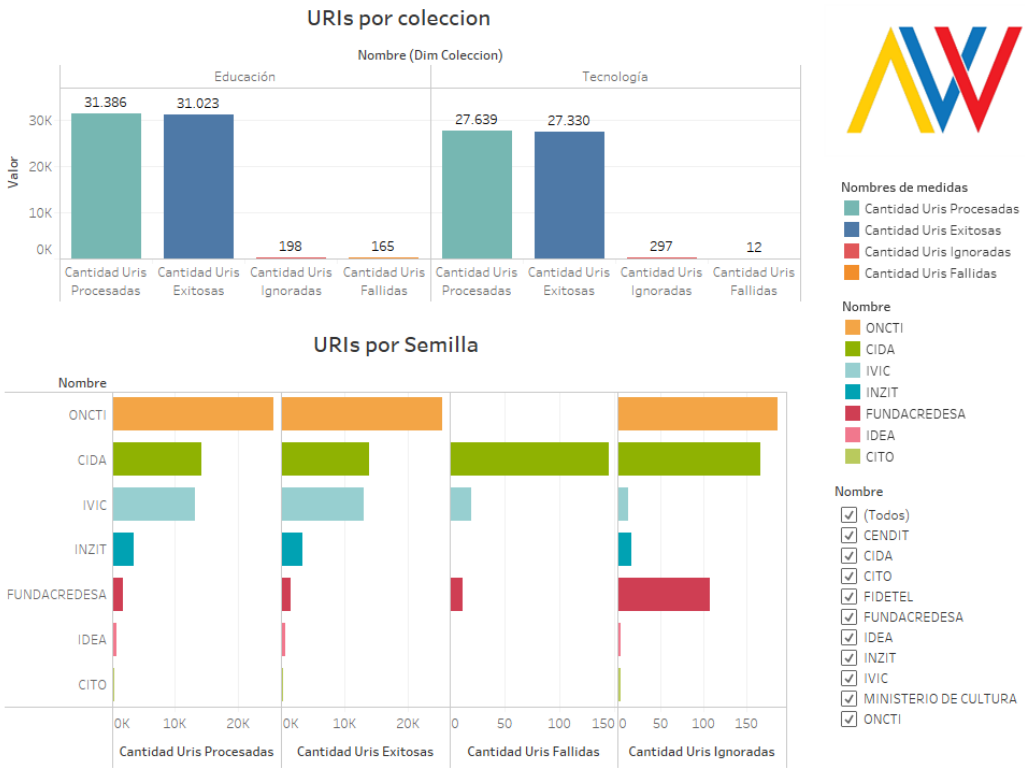

Figura 4.39: Cuadro de mando de gráficos de los URIs
Fuente: Elaboración propia.

En el siguiente cuadro de mando se pueden observar indicadores relacionados a la distribución y conteo de URLs por código de estado http, se puede clasificar por código de estado y por semilla. También se creó una jerarquía entre tipo y descripción para que se pueda ver mejor la información.

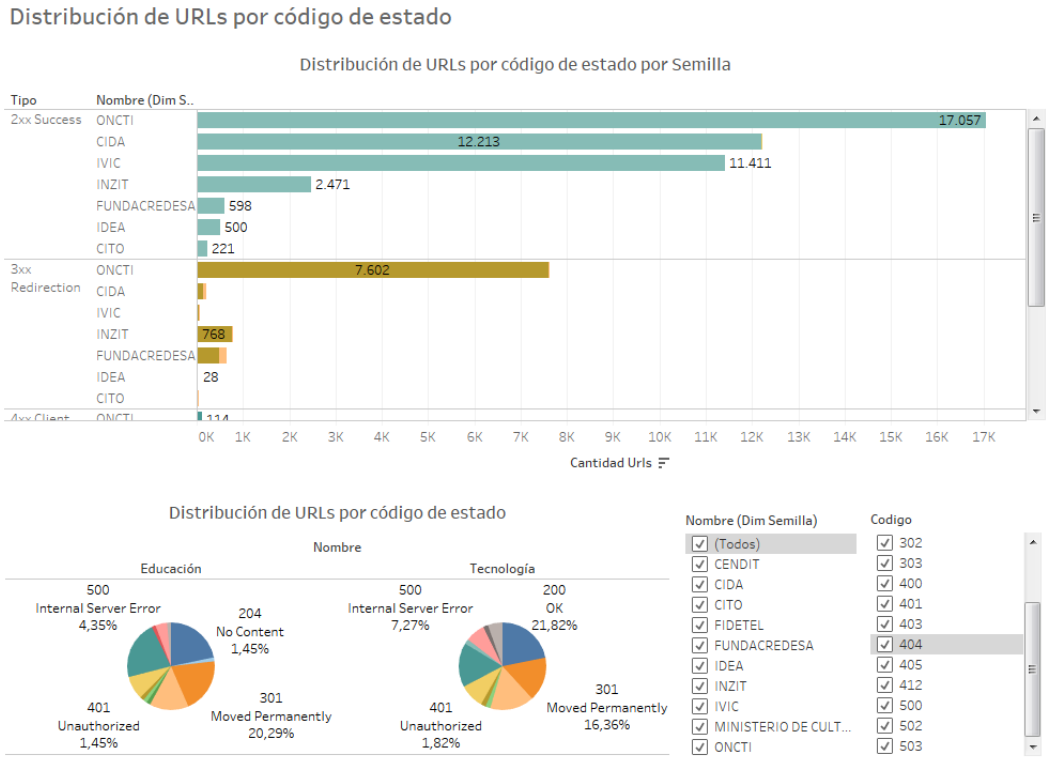

Figura 4.40: Cuadro de mando de la distribución de URLs por código de estado http
Fuente: Elaboración propia.

Por último, en el siguiente cuadro de mando se puede apreciar información acerca de la distribución de espacio y URLs por tipo MIME, y un indicador en el que se ve la cantidad de colecciones por fecha. Se puede clasificar la información por fecha, colección y tipo MIME. También se creó una jerarquía entre tipo y subtipo para que se pueda ver mejor la información.

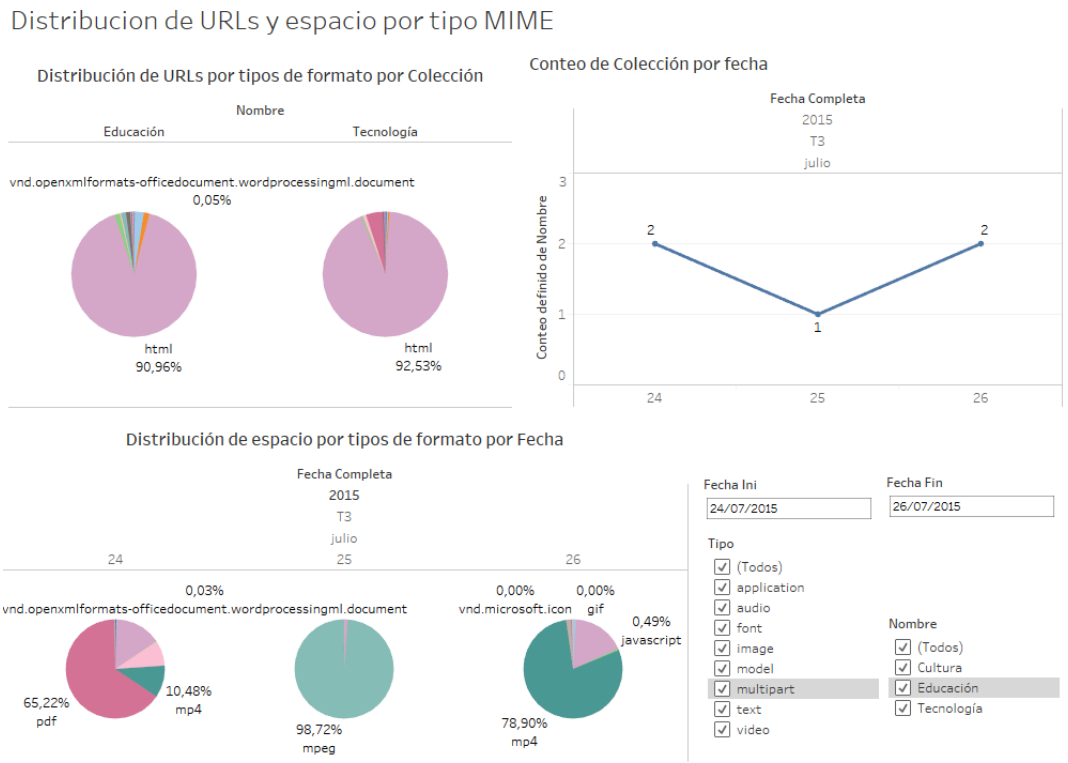

Figura 4.41: Cuadro de mando de la distribución de tipo MIME
Fuente: Elaboración propia.

Conclusiones y Recomendaciones

Al terminar el presente trabajo especial de grado, luego de un arduo periodo de investigación, se pudo cumplir exitosamente el objetivo general, que consistió en "Desarrollar una plataforma para el análisis de los metadatos del proceso de rastreo del Archivo Web de Venezuela basada en una arquitectura de Big Data", siguiendo cada uno de los objetivos específicos.

Una vez comprendida la información arrojada por el rastreador Heritrix y determinados los procesos involucrados en el Archivo Web, fue bastante clara la selección del proceso de negocio en el cual había que centrar el desarrollo de la solución. Así pues, una vez elaborados los indicadores, se continuó con el proceso natural de creación del Modelo Dimensional para un primer *data mart* o almacén de datos, dejando sentadas las bases para futuros análisis de otros procesos con la definición de las primeras dimensiones conformadas en la matriz de bus.

Se implementó una plataforma de Big Data para el análisis de los datos descriptivos del proceso de rastreo, en el cual se mejoró considerablemente el tiempo de búsqueda de las consultas. Para ello se estudiaron herramientas para cada componente de la arquitectura, seleccionando las mas aptas para el proyecto.

Así se obtuvo una plataforma que permite el acceso a indicadores y la realización de análisis definidos por el usuario, se mejoró la poca flexibilidad y la limitación en la generación de nuevos informes o reportes. Actualmente se puede generar diversas gráficas, métricas y reportes sobre el proceso de rastreo, que deriven del modelo desarrollado, además de los indicadores propuestos en principio.

Durante el desarrollo del proyecto se presentaron algunas limitantes, una de ellas fue la falta de una infraestructura distribuida real (clúster) para hacer pruebas mas certeras del tiempo de acceso a los datos para el análisis. La falta de documentación de algunas herramientas como por ejemplo Hive, la cual no se pudo conectar correctamente con la herramienta PDI, o para entender la estructura y coordinación entre los distintos módulos del AWV.

Para trabajos futuros se recomienda realizar una normalización y/o rectificación de la base de datos principal que registra lo concerniente a los trabajos de rastreo (`app`, localizada en el servidor Solr), puesto que se hallaron problemas de redundan-

cia de datos, así como de posibles pérdidas de información. También se recomienda estudiar y utilizar un motor de análisis de datos como Apache Kylin, para la agilización de las consultas, o en su defecto estudiar otras tecnologías fuera del ecosistema Hadoop, que pudieran también servir para el mismo propósito.

Como última recomendación, se sugiere ampliar el proyecto recabando datos para los demás procesos de negocio mencionados (figura 4.2, como indica Kimball, lo recomendable es centrarse en un proceso a la vez), como por ejemplo el proceso de Acceso a los datos preservados, para el estudio del comportamiento de los usuarios que hacen uso del Archivo Web de Venezuela.

Bibliografía

Apache Software Foundation. (2017). What Is Apache Hadoop? Recuperado el 12 de abril de 2017, desde http://hadoop.apache.org/

Apache Software Foundation. (2018a). Apache Hive TM. Recuperado el 18 de septiembre de 2018, desde https://hive.apache.org/index.html

Apache Software Foundation. (2018b). Apache Sqoop. Recuperado el 14 de noviembre de 2018, desde http://sqoop.apache.org/

Apache Software Foundation. (2018c). HDFS Architecture Guide. Recuperado el 14 de noviembre de 2018, desde https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Asociación Española para la Calidad (AEC). (2013). Indicadores. Recuperado el 8 de abril de 2017, desde http://www.aec.es/web/guest/centro-conocimiento/indicadores

Beltrán Jaramillo, J. M. (2006). *Indicadores de Gestión: Herramientas para lograr la competitividad*. Bogotá: 3R Editores.

Bernal, J. (2013). Gestión de procesos: Cómo definir indicadores y cuadros de mando. Recuperado el 12 de abril de 2017, desde http://www.pdcahome.com/4501/gestion-de-procesos-como-definir-indicadores-y-cuadros-de-mando/

Burner, M. & Kahle, B. (1996). Internet Archive. Recuperado el 12 de abril de 2017, desde http://archive.org/web/researcher/ArcFileFormat.php

Cano, J. L. (2007). *Business Intelligence: Competir con Información*. Madrid: Fundación Cultural Banesto.

Casanova Diaz, M. A. & Caraballi, W. (2015). *Desarrollo e implementación del módulo de predicción de cambios de sitios web para un prototipo de Archivo Web* (Tesis de licenciatura, Escuela de Computación, Universidad Central de Venezuela, Caracas, Venezuela).

CCSDS. (2012). *Reference Model for an Open Archival Information System (OAIS)*. Consultative Committee for Space Data Systems. Washington, DC, USA.

Estaba Fernández-Trujillo, J. A. & Ciancia Biondo, V. A. (2015). *Módulo de Almacenamiento y Gestión de Datos para Preservación de Archivos Web Usando Bases de Datos NoSQL* (Tesis de licenciatura, Universidad Central de Venezuela, Caracas, Venezuela).

Fuentes, E. J. & Marquez Gallo, J. L. (2003). *Sistema de Información Gerencial*.

Goel, V. (2011). Web Archive Metadata File Specification. Recuperado el 21 de mayo de 2017, desde https://webarchive.jira.com/wiki/x/R4DN

Hortonworks Inc. (2018). Hortonworks. Recuperado el 14 de noviembre de 2017, desde https://es.hortonworks.com/

IANA. (2018). Media Types. Recuperado el 7 de octubre de 2018, desde https://www.iana.org/assignments/media-types/media-types.xhtml

IIPC. (2012). Why Archive the Web? Recuperado el 9 de abril de 2017, desde http://www.netpreserve.org/web-archiving/overview

Inmon, W. H. (1996). *Building the Data Warehouse* (4ta). Indianapolis, Indiana: Wiley Publishing, Inc.

Internet Live Stats. (2018). Total number of Websites. Recuperado el 22 de mayo de 2017, desde http://www.internetlivestats.com/total-number-of-websites/

ISO. (2009). *Information and documentation – WARC file format*. International Organization for Standardization. Rep. Tec. (Borrador). Ginebra, Suiza.

ISO. (2012). *Information and documentation – Statistics and Quality Indicators for Web Archiving*. International Organization for Standardization. Rep. Tec. (Borrador). Ginebra, Suiza.

J.C. Pacheco, C. H. C., W. Castañeda. (2002). *Indicadores Integrales de Gestión*. McGraw Hill.

Kabchi, M. & Martínez, M. (2014). *Desarrollo del Módulo de Acceso a los Contenidos Preservados en formato WARC para un Prototipo de Archivo Web de Venezuela* (Tesis de licenciatura, Universidad Central de Venezuela, Caracas, Venezuela).

Kimball, R. (1998). *The Data Warehouse Lifecycle Toolkit*. Wiley Computer Publishing.

Kimball, R. & Ross, M. (2002). *The Data Warehouse*. Wiley Computer Publishing.

Krishnan, K. (2013). *Data Warehousing in the Age of Big Data*. Elsevier Inc.

Laudon, K. & Laudon, J. (2012). *Sistemas de Información Gerencial*. Pearson.

Lavoie, B. (2014). *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)*. Digital Preservation Coalition. Gran Bretaña.

Lavoie, B. & Gartner, R. (2013). *Preservation Metadata*. Digital Preservation Coalition. Gran Bretaña.

Liu, A. (2015). Data Science and Data Scientist. Recuperado el 21 de mayo de 2017, desde http://www.researchmethods.org/DataScienceDataScientists.pdf

Long, C. & Kelly, T. (2015). *Data Science & Big Data Analytics*. John Wiley & Sons, Inc.

Lopez, A. & Sarno, R. (2015). *Desarrollo de una Solución de Inteligencia de Negocio para Archivos Web* (Tesis de licenciatura, Escuela de Computación, Universidad Central de Venezuela, Caracas, Venezuela).

Manyika, J., Chui, M., Brown, B., Bughi, J., Dobbs, R., Roxburgh, C. & Hung Byers, A. (2011). *Big Data: The next frontier for innovation, competition and productivity*. Washington DC: McKinsey Global Institute.

Masanès, J. (2006). *Web Archiving*. New York: Springer.

Mondragón, A. (2014). ¿Qué son los indicadores?

Montero Hernández, E. A. & Pérez Laya, H. C. (2016). *Implementación del módulo de indexación y búsqueda para el prototipo de Archivo Web Venezuela para la búsqueda*

de los contenidos Web bajo el formato WARC (Tesis de licenciatura, Universidad Central de Venezuela, Caracas, Venezuela).

Mozilla y colaboradores. (2018a). Códigos de estado de respuesta HTTP. Recuperado el 1 de octubre de 2018, desde https://developer.mozilla.org/es/docs/Web/HTTP/Status

Mozilla y colaboradores. (2018b). Tipos MIME - HTTP. Recuperado el 9 de octubre de 2018, desde https://developer.mozilla.org/es/docs/Web/HTTP/Basics_of_HTTP/MIME_types

Oracle Corporation. (2018). What is MySQL? Recuperado el 18 de septiembre de 2018, desde https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html

Ospina Torres, M. H. (2014). *Un Marco de Referencia para la Implementación de Archivos Web* (Tesis de maestría, Universidad Central de Venezuela, Caracas, Venezuela).

Pentaho Corporation. (2017). Data Integration - Kettle. Recuperado el 2 de mayo de 2017, desde http://community.pentaho.com/projects/data-integration/

Pentaho Data Integration. (2015). Pentaho Data Integration (Kettle) Tutorial. Recuperado el 12 de septiembre de 2018, desde https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+(Kettle)+Tutorial

Ponniah, P. (2001). *DataWarehousing Fundamentals*. John Wiley & Sons, Inc.

REBIUN. (2009). Informe del objetivo operacional 1.2.1: Guía de Recursos para la Preservación Digital. Red de Bibliotecas Universitarias (REBIUN). España.

Freed, N. & Borenstein, N. (1996). *Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies*. IANA. Estados Unidos.

Riley, J. (2004). *Understanding metadata*. NISO Primer Series. Baltimore, Maryland: National Information Standards Organization (NISO).

Rivero, L. & García, J. (2013). *Implementación de los módulos de adquisición y almacenamiento de un prototipo para el archivado de sitios Web en Venezuela* (Tesis de licenciatura, Universidad Central de Venezuela, Caracas, Venezuela).

Russom, P. (2011). *Big Data Analytics*. TDWI.

Sánchez, L. & Milano, G. (2015). *Implementación del Módulo de Gestión y de Control de Incidencias del Prototipo de Archivo Web en Venezuela* (Tesis de licenciatura, Universidad Central de Venezuela, Caracas, Venezuela).

Sint, R., Schaffert, S., Stroka, S. & Ferstl, R. (s.f.). Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis.

Tableau Software. (2018). Tableau Desktop. Recuperado el 28 de septiembre de 2018, desde https://www.tableau.com/es-es/products/desktop

Unesco. (2017). Antecedentes del Patrimonio Digital. Recuperado el 9 de abril de 2017, desde http://www.unesco.org/new/es/communication-and-information/access-to-knowledge/preservation-of-documentary-heritage/digital-heritage/background/

Van Rijmenam, M. (2015). Why the 3V's are not sufficient to describe Big Data. Recuperado el 12 de abril de 2017, desde https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Anexos

A continuación, se presentarán los gráficos de los indicadores realizados. Se mostrará un gráfico por cada clasificación.

- Indicador 1: Cantidad de semillas por Colección

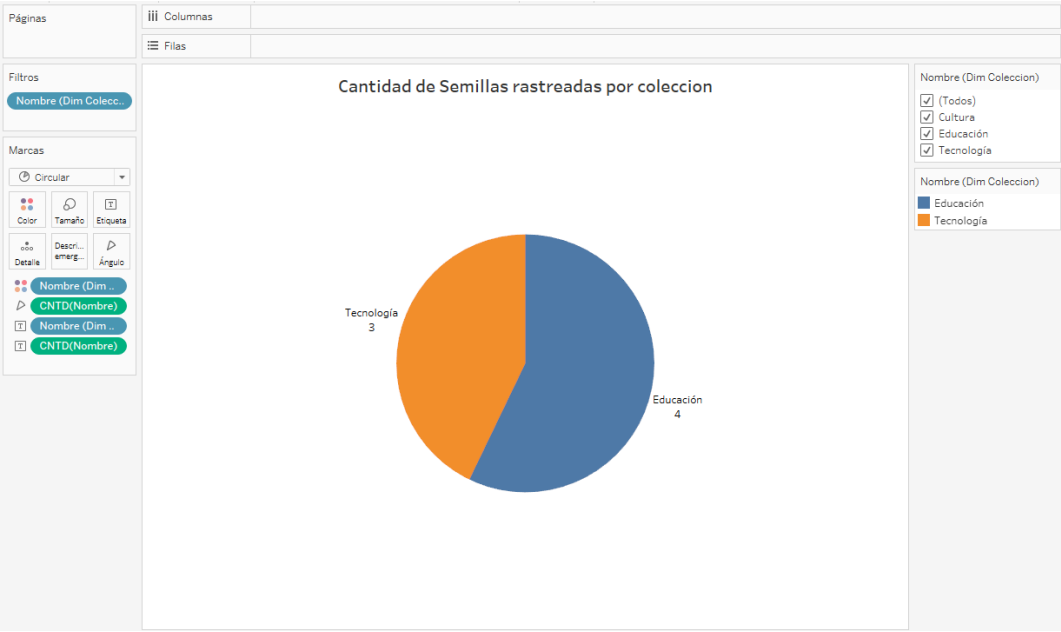

Figura 4.42: Indicador 1 - Cantidad de semillas por colección
Fuente: Elaboración propia.

- Indicador 1: Cantidad de semillas por fecha

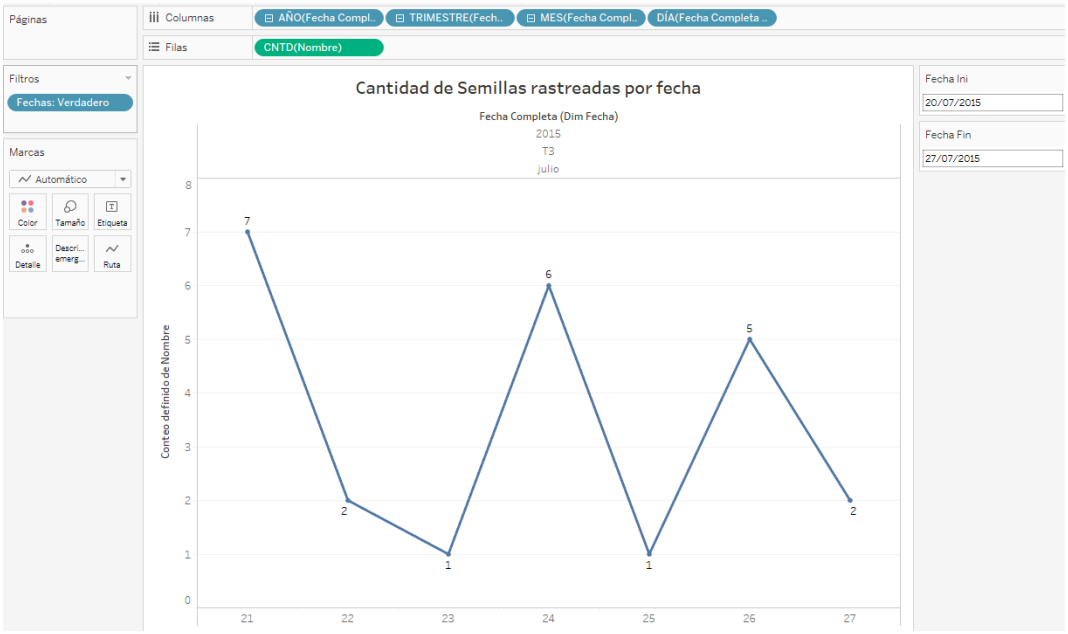

Figura 4.43: Indicador 1 - Cantidad de semillas por fecha
Fuente: Elaboración propia.

- Indicador 2: Cantidad de rastreos por colección

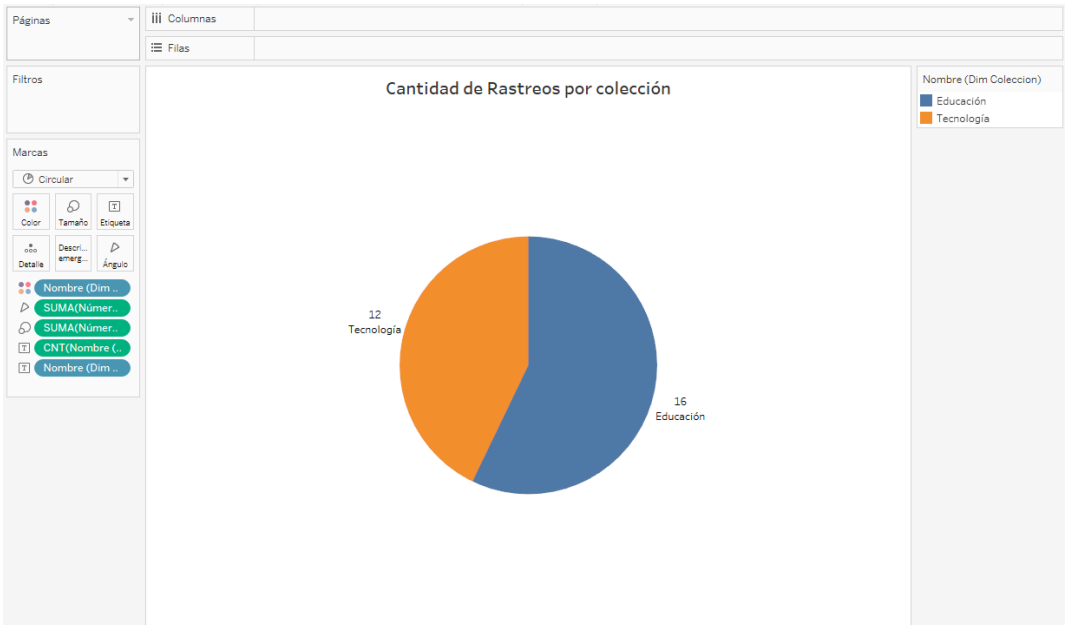

Figura 4.44: Indicador 2 - Cantidad de rastreos por colección
Fuente: Elaboración propia.

- Indicador 2: Cantidad de rastreos por semilla

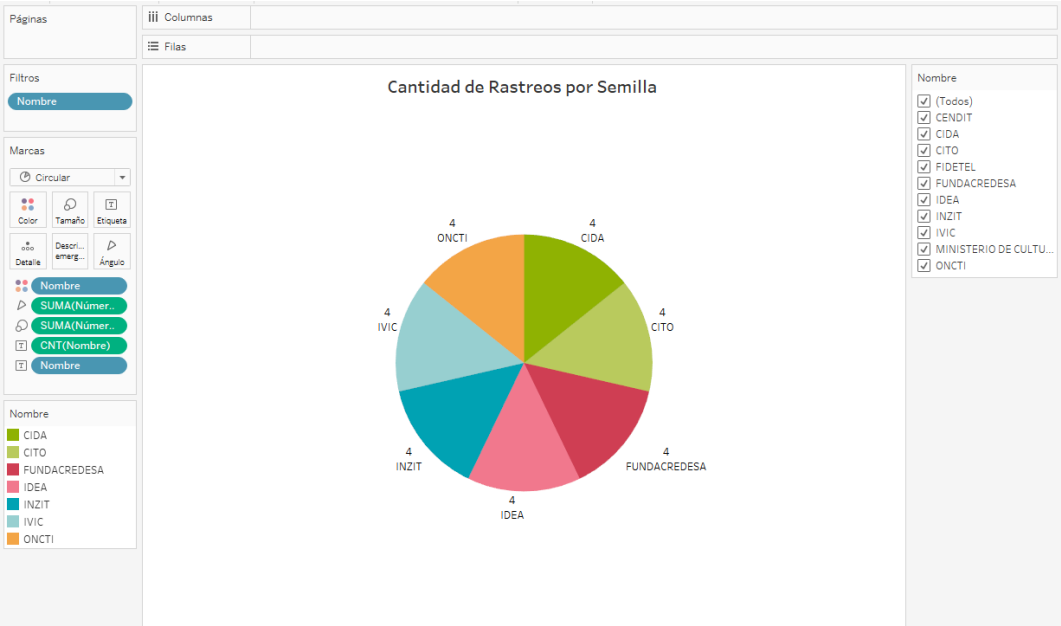

Figura 4.45: Indicador 2 - Cantidad de rastreos por semilla
Fuente: Elaboración propia.

- Indicador 3: Cantidad de recursos (URIs) por colección

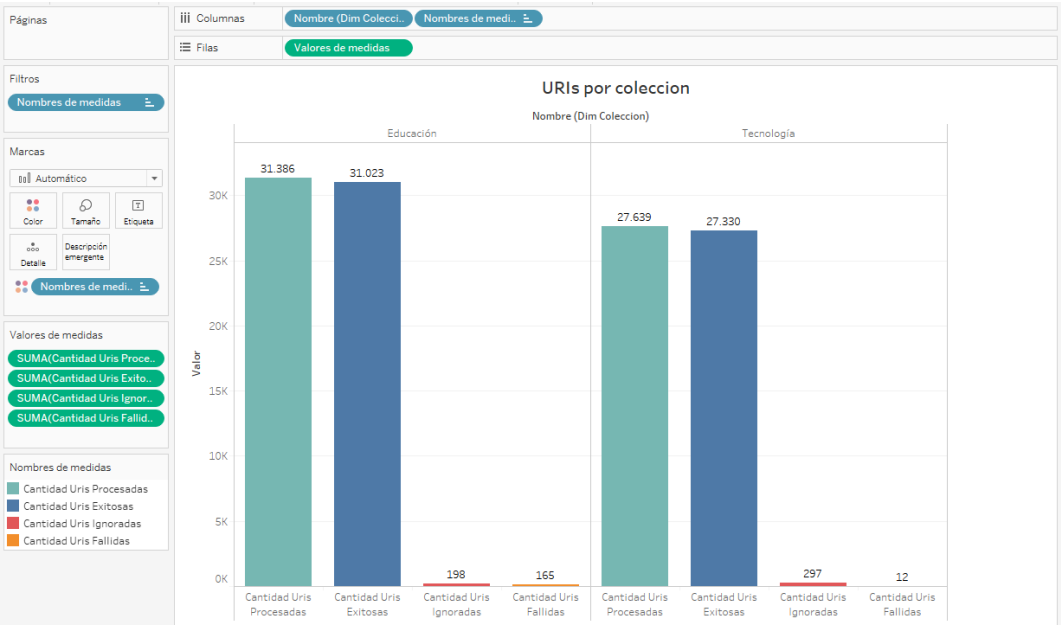

Figura 4.46: Indicador 3 - Cantidad de URIs por colección
Fuente: Elaboración propia.

- Indicador 3: Cantidad de URIs por fecha

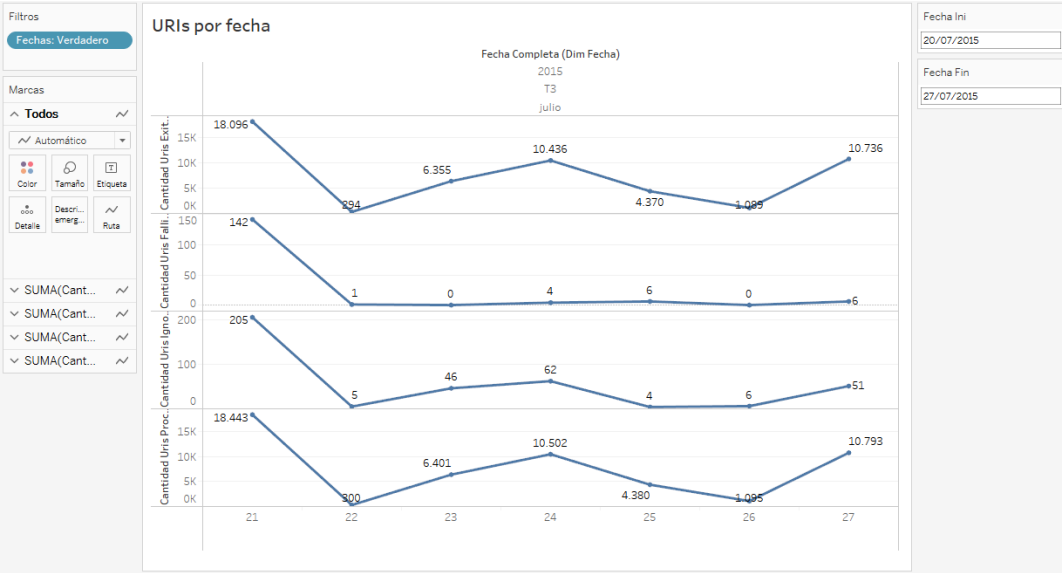

Figura 4.47: Indicador 3 - Cantidad de URIs por fecha
Fuente: Elaboración propia.

- Indicador 4: Duración promedio de rastreo por colección

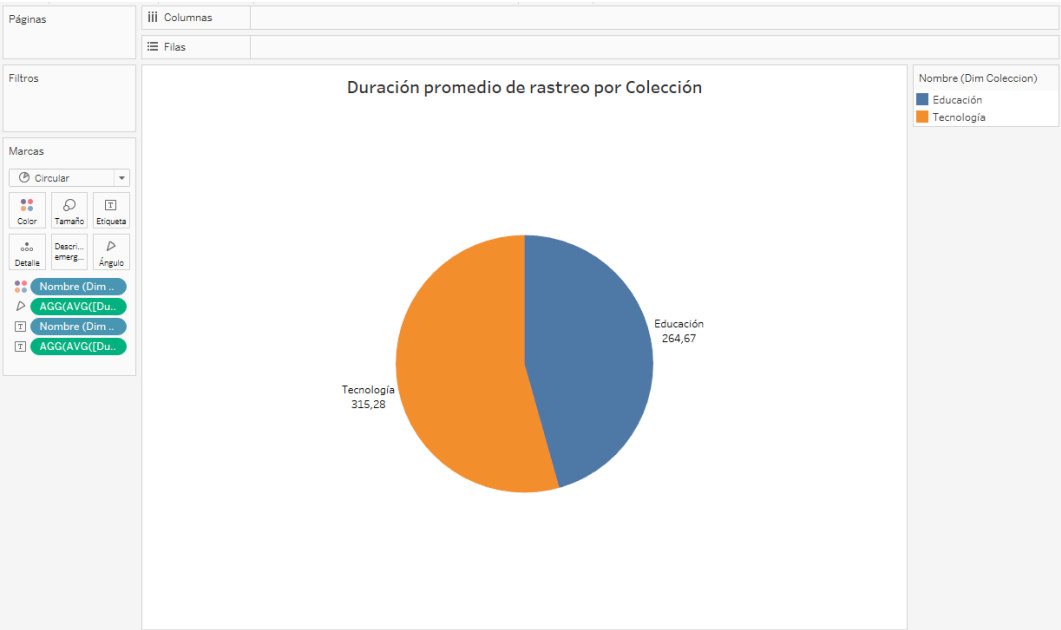

Figura 4.48: Indicador 4 - Duración promedio de rastreo por colección
Fuente: Elaboración propia.

- Indicador 4: Duración promedio de rastreo por semilla

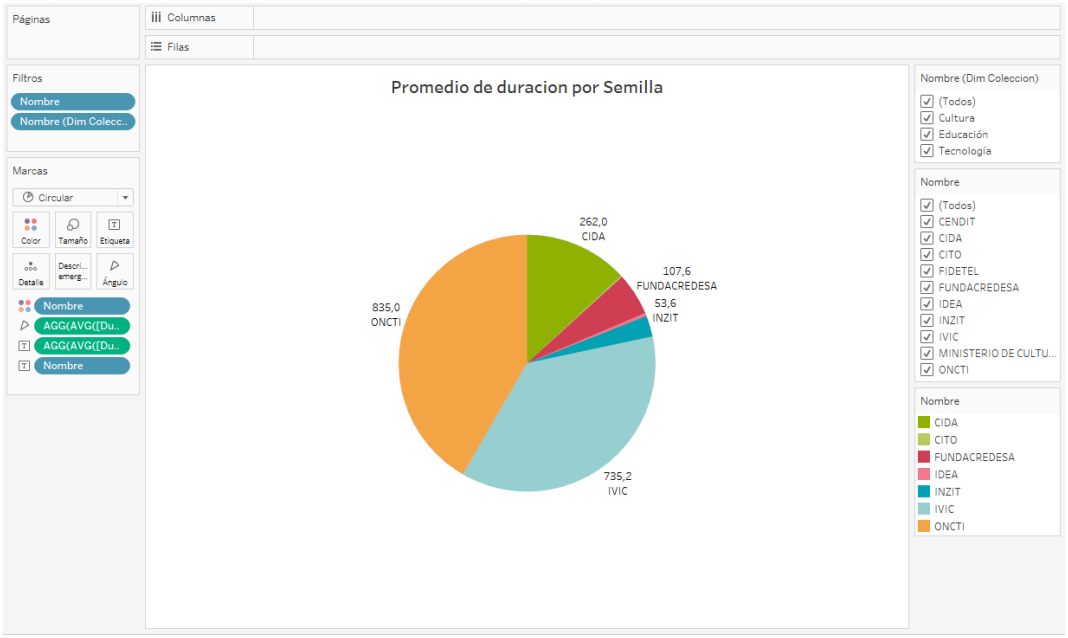

Figura 4.49: Indicador 4 - Duración promedio de rastreo por semilla
Fuente: Elaboración propia.

- Indicador 4: Duración promedio de rastreo por fecha

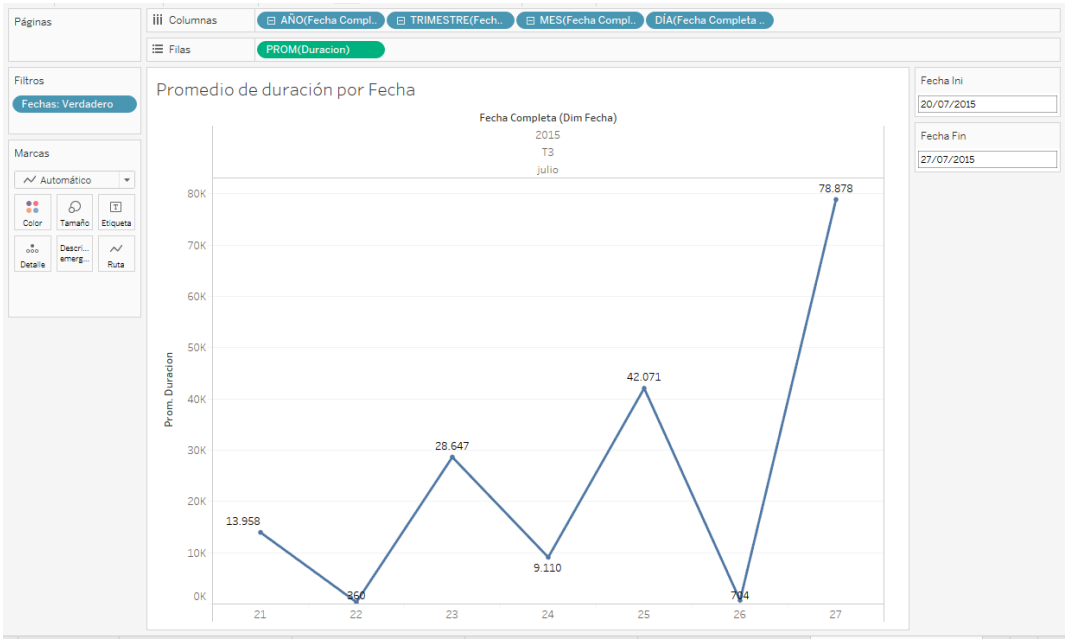

Figura 4.50: Indicador 4 - Duración promedio de rastreo por fecha
Fuente: Elaboración propia.

- Indicador 5: Cobertura cronológica por semilla

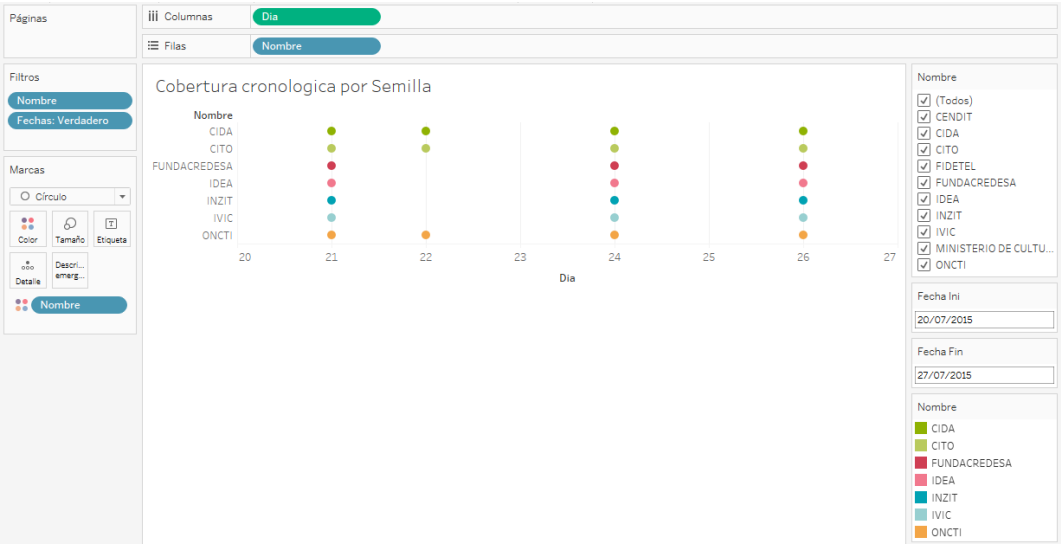

Figura 4.51: Indicador 5 - Cobertura cronológica por semilla
Fuente: Elaboración propia.

- Indicador 6: Distribución de URLs por código de estado http por colección

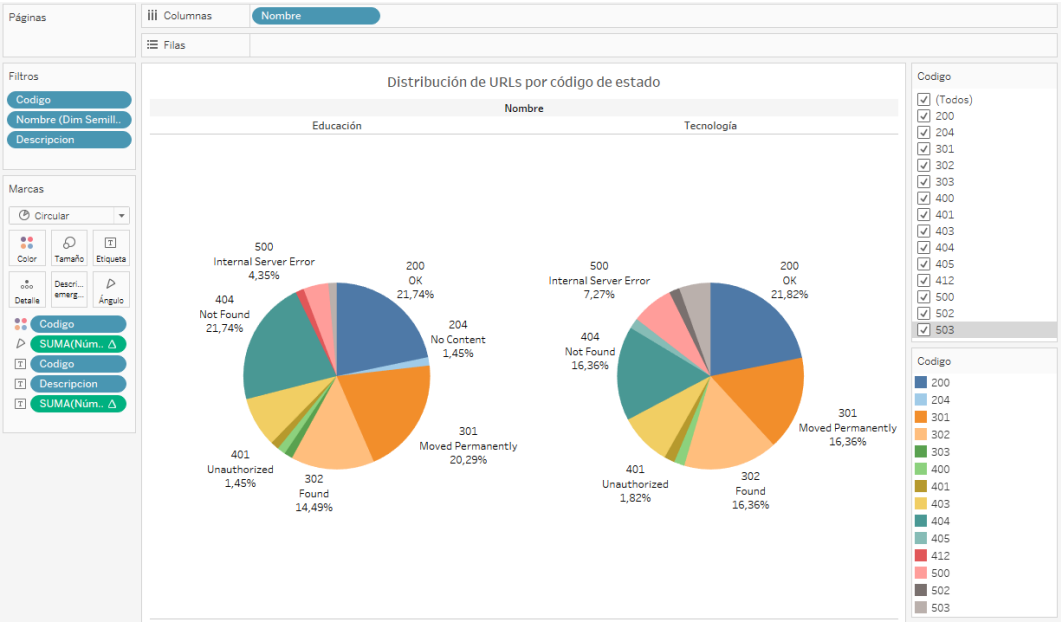

Figura 4.52: Indicador 6 - Distribución de URLs por código de estado http por colección

Fuente: Elaboración propia.

- Indicador 6: Distribución de URLs por código de estado http por semilla

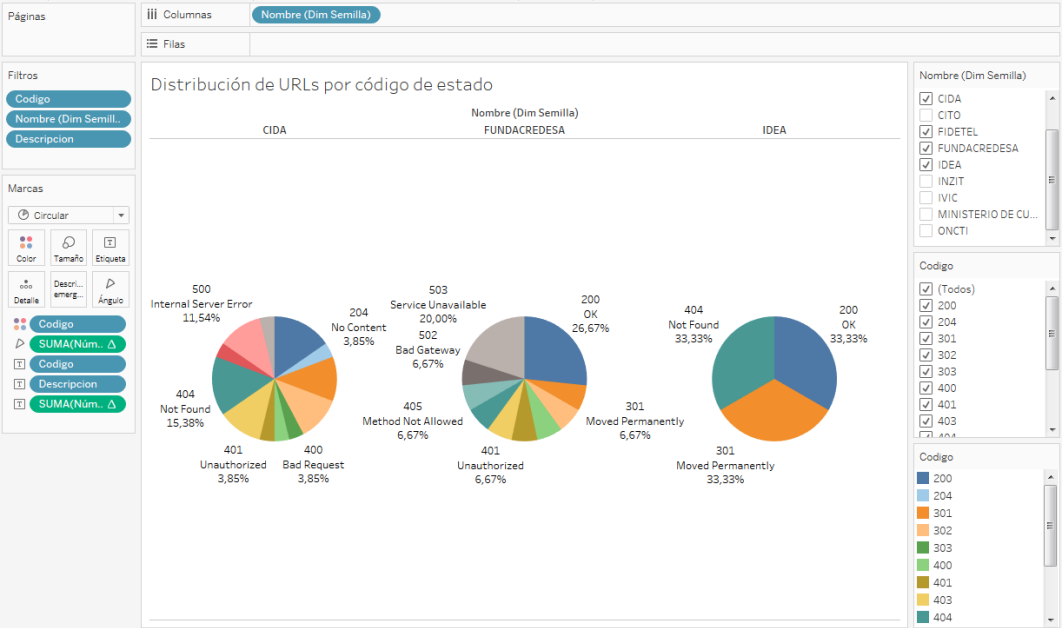

Figura 4.53: Indicador 6 - Distribución de URLs por código de estado http por semilla
Fuente: Elaboración propia.

- Indicador 6: Distribución de URLs por código de estado http por fecha

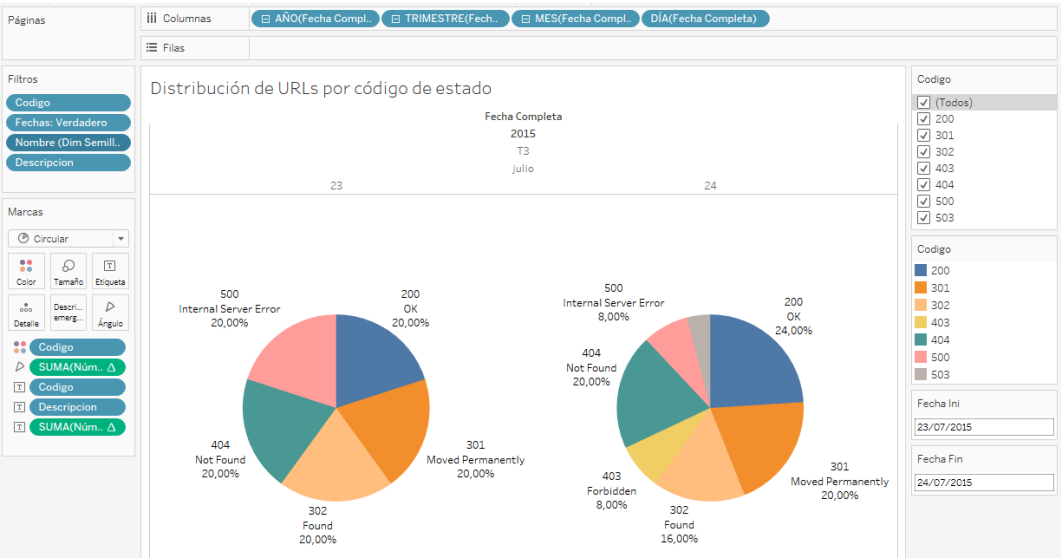

Figura 4.54: Indicador 6 - Distribución de URLs por código de estado http por fecha
Fuente: Elaboración propia.

- Indicador 7: Distribución de espacio por tipos de formatos (Tipos MIME) por colección

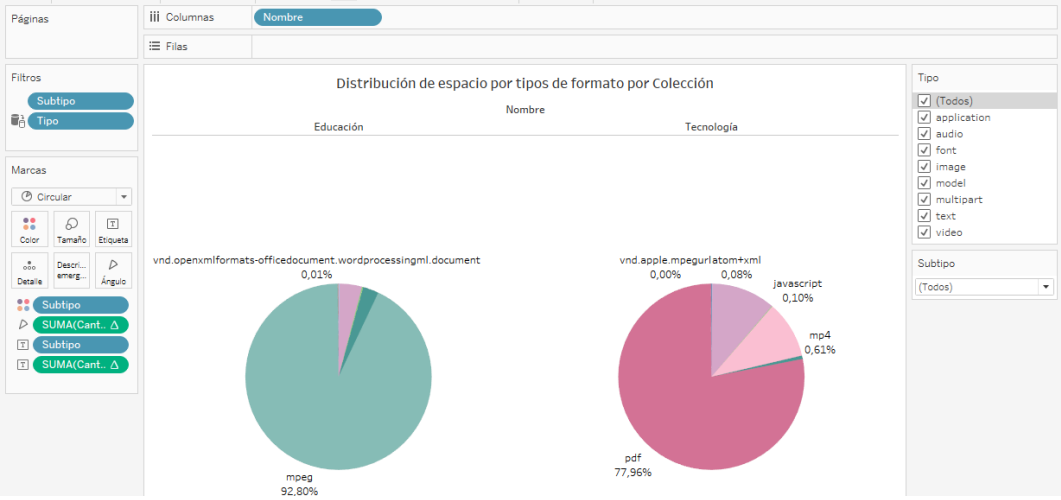

Figura 4.55: Indicador 7 - Distribución de espacio por tipos de formatos (Tipos MIME) por colección.

Fuente: Elaboración propia.

- Indicador 7: Distribución de espacio por tipos de formatos (Tipos MIME) por semilla

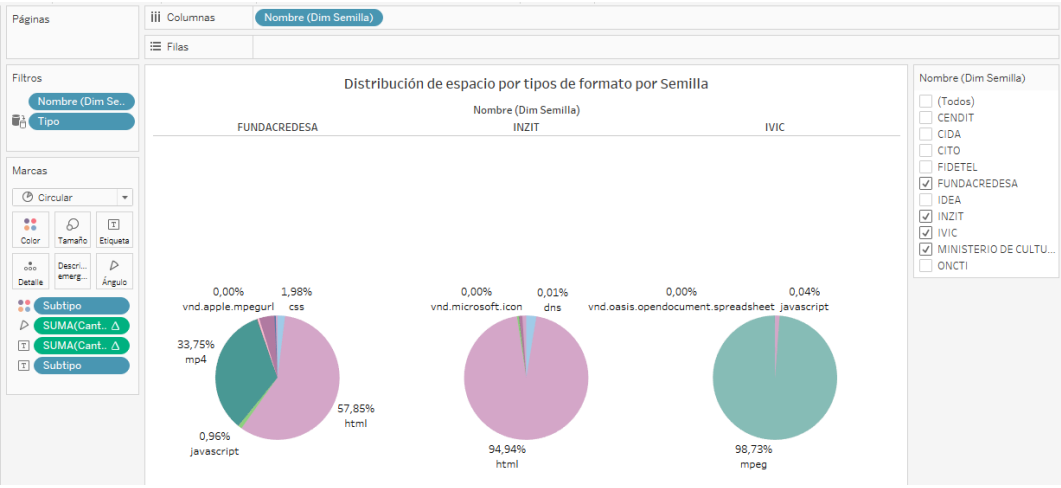

Figura 4.56: Indicador 7 - Distribución de espacio por tipos de formatos (Tipos MIME) por semilla.

Fuente: Elaboración propia.

- Indicador 7: Distribución de espacio por tipos de formatos (Tipos MIME) por fecha

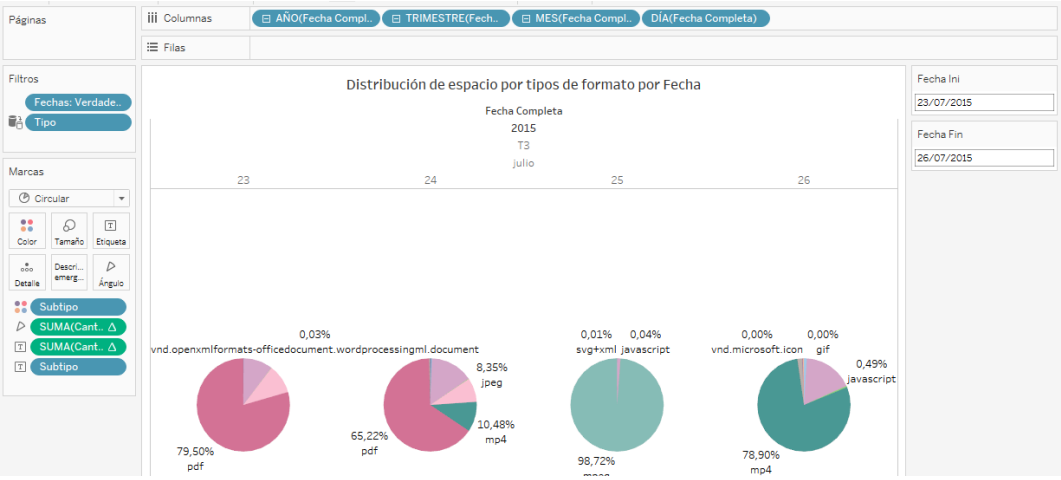

Figura 4.57: Indicador 7 - Distribución de espacio por tipos de formatos (Tipos MIME) por fecha.

Fuente: Elaboración propia.

- Indicador 8: Distribución de URLs por tipos de formatos (Tipos MIME) por colección

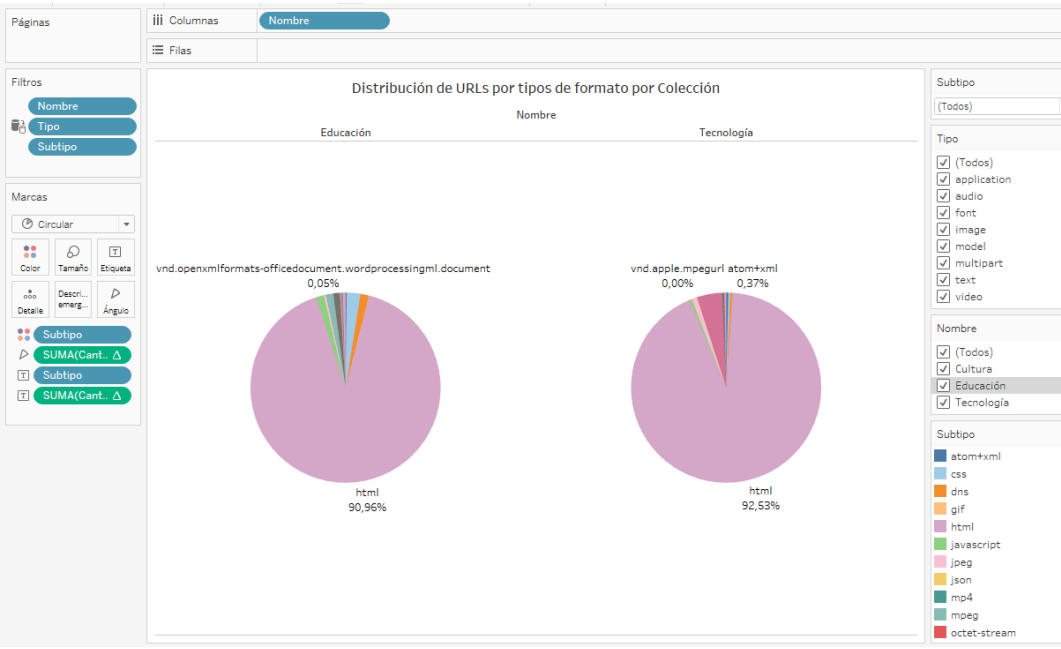

Figura 4.58: Indicador 8 - Distribución de URLs por tipos de formatos (Tipos MIME) por colección.

Fuente: Elaboración propia.

- Indicador 8: Distribución de URLs por tipos de formatos (Tipos MIME) por

semilla

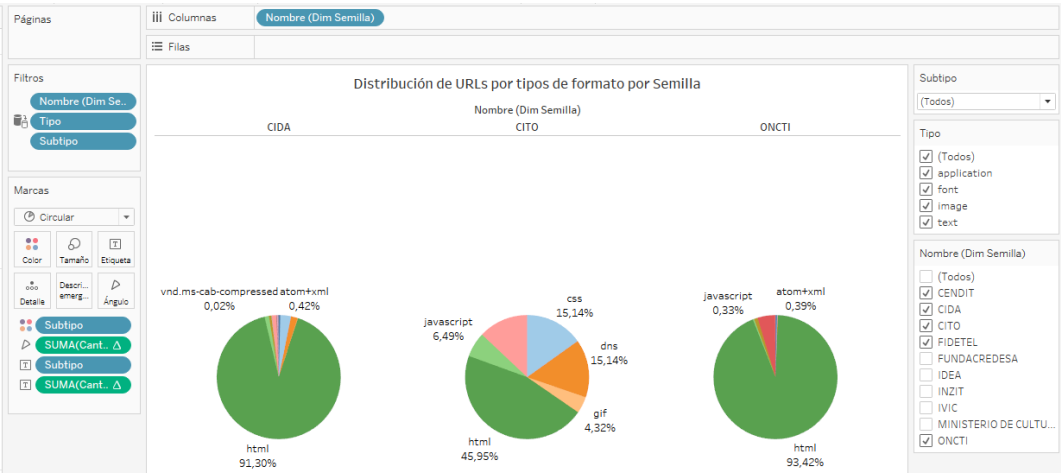

Figura 4.59: Indicador 8 - Distribución de URLs por tipos de formatos (Tipos MIME) por semilla.

Fuente: Elaboración propia.

- Indicador 8: Distribución de URLs por tipos de formatos (Tipos MIME) por fecha

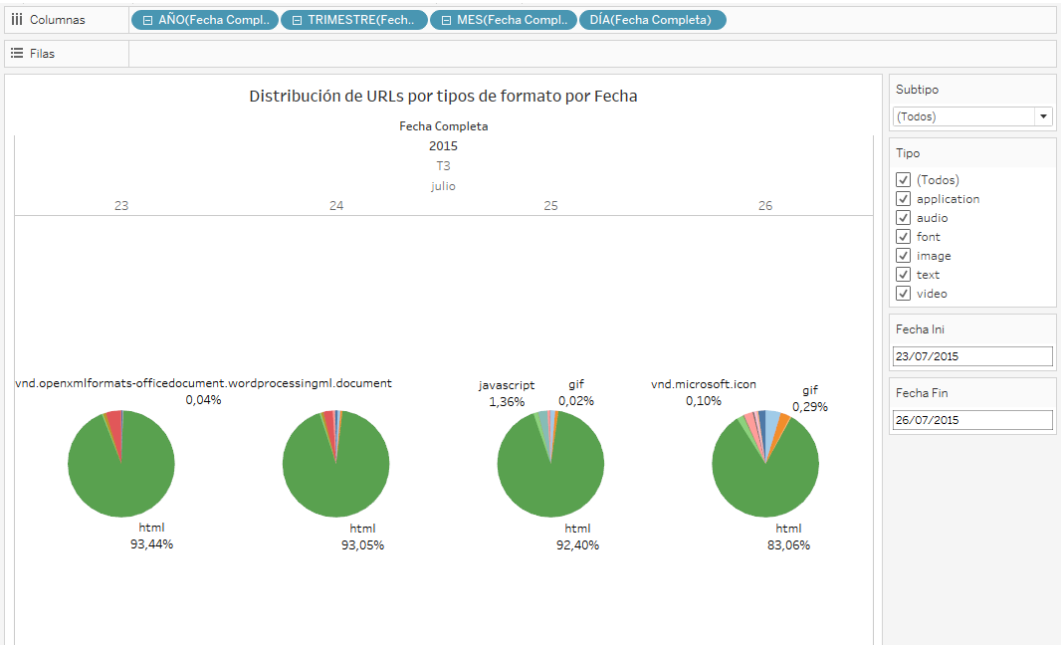

Figura 4.60: Indicador 8 - Distribución de URLs por tipos de formatos (Tipos MIME) por fecha.

Fuente: Elaboración propia.

- Indicador 9: Cantidad de colecciones por fecha

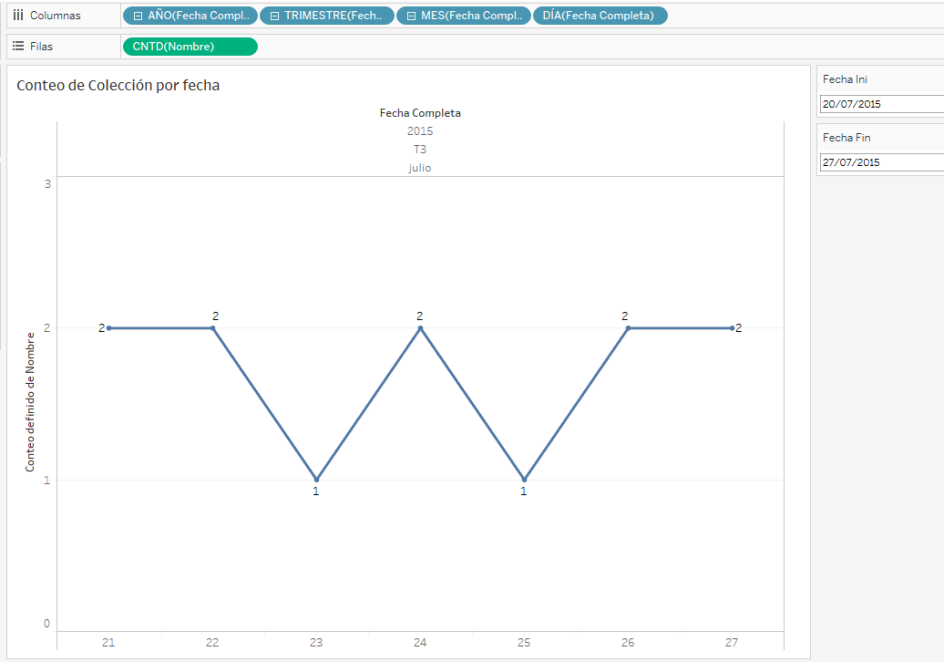

Figura 4.61: Indicador 9 - Cantidad de colecciones por fecha.
Fuente: Elaboración propia.

- Otros indicadores: A continuación se verán otros indicadores que se pudieron realizar con el modelo actual.

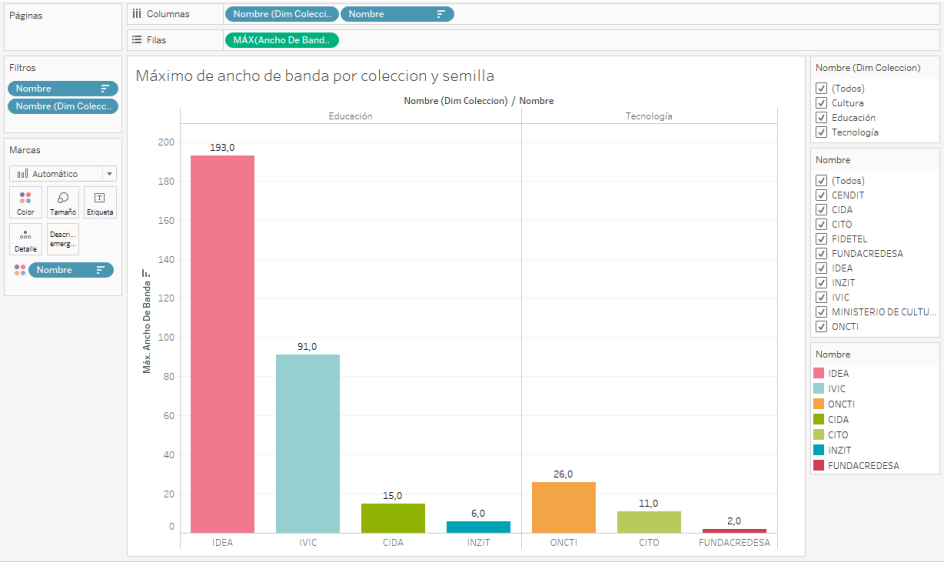

Figura 4.62: Máximo de ancho de banda por colección y semilla.
Fuente: Elaboración propia.

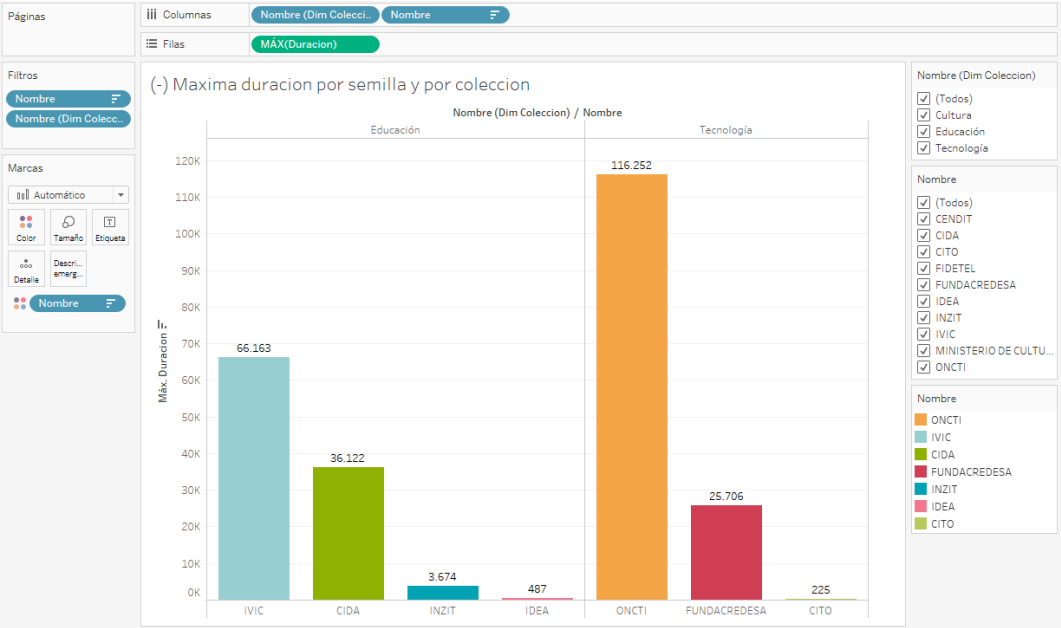

Figura 4.63: Máxima duración por colección y semilla.
Fuente: Elaboración propia.

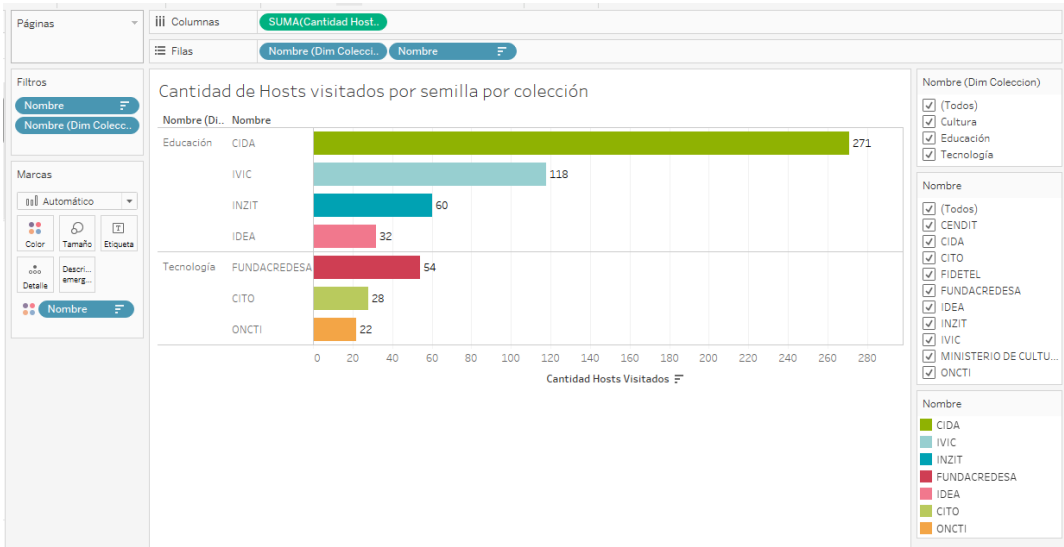

Figura 4.64: Cantidad de Hosts visitados por colección y semilla.
Fuente: Elaboración propia.