Repositorio de datos agropecuarios para el análisis del mercado de aguacate

Trabajo de grado modalidad práctica profesional

Nombre: Juan David Acosta González

Director: Dr. Ing. Juan Carlos Corrales Muñoz

Codirector: Mag. Ing. Juan David Rincón Patiño

Asesor Empresa: Mag. Ing. Juan Fernando Casanova Olaya

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones

Programa de Ingeniería Electrónica y Telecomunicaciones

Departamento de Telemática

Popayán, 2021

Dedico este triunfo a Dios y a la Virgen María,
a mis padres Adolfo Alberto y Luz Marina
por brindarme su apoyo incondicional,
a mis familiares por su compañía,
y a todos aquellos que fueron
una fuente de motivación y de éxito. Mil gracias.

Agradecimientos

El autor expresa sus agradecimientos al Dr. Ing. Juan Carlos Corrales Muñoz director de la práctica profesional, por su acompañamiento y orientación.

A los Magister Juan David Rincón Patiño (Co-Director), y Juan Fernando Casanova Olaya (Asesor), por su oportuna colaboración, sus consejos, recomendaciones, y aportes realizados durante el desarrollo de la práctica profesional en la empresa Ecotecma S.A.S.

Agradezco al grupo de Ingeniería Telemática, al Departamento de Telemática, y al comité de programa de Ingeniería Electrónica y Telecomunicaciones por sus aportes y recomendaciones.

RESUMEN ESTRUCTURADO

Antecedentes

La empresa Ecotecma S.A.S se dedica al desarrollo de soluciones enmarcadas dentro del ámbito de la agricultura climáticamente inteligente y el soporte en la toma de decisiones. Es por ello, que la empresa se encuentra buscando soluciones para el análisis de mercados agrícolas, de diferentes productos dentro de los cuales se encuentra el aguacate. Debido a la ausencia de datos centralizados y de relevancia con respecto a este mercado se propone el diseño y la implementación de un servicio de captura automatizada de datos cuyo repositorio de datos sea un lago de datos, para así poder obtener información de valor del mercado. Este servicio de captura automatizada de datos es el realizado en la presente práctica profesional.

Objetivos

Objetivo general: construir un repositorio de datos agropecuarios para el análisis del mercado del aguacate.

Objetivos específicos:

- Caracterizar posibles fuentes para la extracción de datos demográficos y económicos del mercado del aguacate.

- Desarrollar un servicio de captura automática de datos demográficos y económicos del mercado de aguacate, para la conformación de un lago de datos.

- Implementar un lago de datos con los registros capturados del mercado del aguacate.

Metodología

Para soportar el análisis del estado del arte se utilizó la herramienta SciMAT y las metodologías Petersen y Kitchenham, para la realización del mapeo sistemático alrededor de temas como los procesos de extracción, carga y transformación de datos, los lagos de datos, los análisis de mercados agrícolas, y la agricultura. Para el desarrollo y validación del servicio de captura automática de datos y del repositorio de datos en el tiempo acordado para la práctica profesional, se usó la metodología Ralph Kimball y para la caracterización de las fuentes de datos se utilizó la metodología CRISP-DM.

Resultados

Como resultado de la práctica profesional, se generó un servicio de captura automática de datos con un lago de datos como repositorio de datos para la empresa Ecotecma S.A.S, logrando de esta manera recopilar una gran cantidad de registros en torno al mercado del aguacate de los Estados Unidos.

Conclusiones

Con la realización de la práctica profesional, se logró obtener una gran cantidad de datos en torno al mercado del aguacate de los Estados Unidos, los cuales podrán ser analizados con mayor detalle para así poder generar información de valor para la empresa Ecotecma S.A.S.

CONTENIDO

1 INTRODUCCIÓN 1

1.1 Planteamiento del problema 1

1.2 Objetivos 3

1.2.1 Objetivo General 3

1.2.2 Objetivos específicos 4

1.2.3 Metodologías para la implementación de lagos de datos 4

2 PLANIFICACIÓN DEL PROYECTO 9

2.1.1 Metodología Petersen 9

2.1.2 Metodología Kitchenham 11

2.1.3 Conceptos relacionados con procesos ETL, ELT, minería de datos, y mercados agrícolas 12

2.1.4 Incidencia de variables en mercados agrícolas 16

2.1.5 Analítica de datos 19

2.1.6 Lagos de datos 22

2.1.7 Resumen del capítulo 26

3 CARACTERIZACIÓN DE LAS FUENTES DE DATOS 27

3.1.1 Metodología CRISP-DM: 27

3.1.2 Fuentes de datos para el análisis de mercado del aguacate en Estados Unidos 29

3.1.3 Resumen del capítulo 38

4 SERVICIO DE CAPTURA AUTOMÁTICA DE DATOS 39

4.1.1 Tecnologías para la extracción, carga y transformación automática de datos 39

4.1.2 Extracción de datos del Departamento de Agricultura de los Estados Unidos (USDA) 48

4.1.3 Manejo y uso del API Quickstats 48

4.1.4 Extracción de datos del sistema de información económica, estadística y de mercado (ESMIS) 52
4.1.5 Extracción de datos de la Organización de la Agricultura y la Alimentación (FAO) 53
4.1.6 Extracción del departamento de agricultura y alimentos de California (CDFA) 54
4.1.7 Extracción de datos del Hass Avocado Board (HAB) 54
4.1.8 Extracción de datos de Kaggle 55
4.1.9 Extracción de datos de la oficina del censo de los Estados Unidos 56
4.1.10 Resumen del capítulo 56
5 IMPLEMENTACIÓN DE LA ARQUITECTURA EN AMAZON WEB SERVICES 57
5.1.1 Modelo de vistas 4+1 del servicio de captura automática de datos 57
5.1.2 Vista lógica 58
5.1.3 Vista de proceso 61
5.1.4 Vista de desarrollo 62
5.1.5 Vista física 64
5.1.6 Despliegue del servicio de captura automática de datos en AWS 66
5.1.7 Resumen del capítulo 90
6 CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS 91
6.1 Conclusiones 91
6.2 Recomendaciones 92
6.3 Trabajos futuros 94
7 REFERENCIAS 95

FIGURAS

Figura 1. Fases de la metodología Ralph Kimball. Fuente [6]. 5
Figura 2. Metodología Kimball adaptada. Fuente propia..... 8
Figura 3. Proceso del mapeo sistemático. Fuente [8] 9
Figura 4. Mapa longitudinal. Fuente propia. 13
Figura 5. Mapa de la ciencia con número de documentos, primer periodo. Fuente propia. 14
Figura 6. Mapa de la ciencia con número de documentos, segundo periodo. Fuente propia. 14
Figura 7. Primer análisis de cluster de Machine Learning. Fuente propia. 15
Figura 8. Diagrama de proceso de la metodología CRISP-DM. Fuente [10]..... 27
Figura 9. Cantidad de citaciones de fuentes de datos relacionados con el mercado del aguacate de los Estados Unidos. Fuente propia……..31
Figura 10. Cantidad de publicaciones por periodo de tiempo. Fuente propia……..32
Figura 11. Periodicidad de fuentes de datos citadas en publicaciones. Fuente propia……..33
Figura 12. Consulta de datos relacionados con el aguacate en USDA. Fuente Propia. 488
Figura 13. Consulta API Quick Stats USDA con CURL. Fuente Propia. 499
Figura 14. Consulta de datos relacionados con el aguacate en USDA sin datos de la categoría ambiental. Fuente Propia……50
Figura 15. Consulta API Quick Stats USDA con CURL sin datos ambientales. Fuente Propia……..50
Figura 16. Consulta de datos relacionados con el aguacate en USDA sin datos de la categoría ambiental, últimos dos años. Fuente Propia……51
Figura 17. Consulta API Quick Stats USDA con CURL sin datos ambientales de la categoría ambiental, últimos dos años. Fuente Propia..... 51
Figura 18. Diagrama de clases de los extractores de datos……..58
Figura 19. Vista lógica, diagrama de clases. Fuente propia……60
Figura 20. Vista de proceso, diagrama de secuencia. Fuente propia……61
Figura 21. Vista de desarrollo, diagrama de componentes. Fuente propia……63
Figura 22. Vista física, diagrama de despliegue. Fuente propia……65
Figura 23. Repositorio del lago de datos en el bucket de S3. Fuente propia……..77
Figura 24. Listado de tablas de datos generadas por el Crawler. Fuente propia……89

ÍNDICE DE TABLAS

Tabla 1. Comparación de lagos de datos y almacenes de datos. Fuente [25] 25
Tabla 2. Fuentes de datos del mercado de aguacate de los Estados Unidos. Fuente propia. 34
Tabla 3. Descripción, ventajas y desventajas de las posibles tecnologías para implementar el proceso ELT. Fuente propia 40
Tabla 4. Criterios para la elección de la tecnología para la extracción de datos. Fuente propia. 44
Tabla 5. Análisis de costos de las diferentes tecnologías para la implementación del proceso ELT. Fuente propia. 45
Tabla 6. Descripción de los registros almacenados de la CDFA en el lago de datos. Fuente propia 69
Tabla 7. Descripción de los registros almacenados de la Oficina del Censo de los Estados Unidos en el lago de datos. Fuente propia 70
Tabla 8. Descripción de los registros almacenados de la FAO en el lago de datos. Fuente propia. 72
Tabla 9. Descripción de los registros almacenados del HAB en el lago de datos. Fuente propia. 72
Tabla 10. Descripción de los registros almacenados de Kaggle en el lago de datos. Fuente propia. 73
Tabla 11. Descripción de los registros almacenados de USDA en el lago de datos. Fuente propia. 73
Tabla 12. Descripción de los registros capturados y almacenados de manera automática en el lago de datos. Fuente propia. 75
Tabla 13. Caracterización de las tablas que conforman el lago de datos. Fuente propia 77

LISTADO DE ABREVIATURAS

ETL: Extract, Transform, and Load. Extracción, Transformación, y Carga.

ELT: Extract, Load, and Transform. Extracción, Carga, y Transformación.

DL: Data Lake. Lagos de datos.

DW: Data Warehouse. Almacenes de datos.

USDA: United States Department of Agriculture. Departamento de Agricultura de los Estados Unidos.

NASS: National Agricultural Statistics Service. Servicio Nacional de Estadística y Agricultura.

ESMIS: USDA Economics, Statistics and Market Information System. Sistema de información económica, estadística y del mercado de USDA.

FAO: Food and Agriculture Organization. Organización de las Naciones Unidas para la Alimentación y la Agricultura.

CDFA: California Department of Food and Agriculture. Departamento de Agricultura y Alimentación de California.

HAB: Hass Avocado Board. Junta del Aguacate Hass.

AWS: Amazon Web Services. Servicios Web de Amazon.

GCP: Google Cloud Platform. Plataforma en la nube de Google.

AWS S3: Amazon Web Services Simple Storage Service. Servicio de almacenamiento simple en los Servicios Web de Amazon.

AWS EC2: Amazon Web Services Elastic Cloud Computing. Computación en la nube elástica en los Servicios Web de Amazon.

AWS IAM: Amazon Web Services Identity and Access Management. Gestión de acceso e identidad de los Servicios Web de Amazon.

API: Application Programming Interface. Interfaz de programación de aplicaciones.

HTTP: HyperText Transfer Protocol. Protocolo de transferencia de hipertexto.

CSV: Comma Separated Values. Valores separados por Comas.

B2B: Business To Business. Empresa a Empresa.

B2C: Business To Consumer: Empresa a Consumidor.

BI: Business Intelligence. Inteligencia de negocio.

AI: Artificial Intelligence. Inteligencia Artificial.

ML: Machine Learning. Aprendizaje Automático.

SDK: Software Development Kit. Kit de desarrollo de software.

IOT: Internet Of Things. Internet De Las Cosas.

PA: Precision Agriculture. Agricultura de Precisión.

CA: Conventional Agriculture. Agricultura Convencional.

WSN: Wireless Sensors Networks. Redes de Sensores Inalámbricos.

SQL: Structured Query Language. Lenguaje de Consulta Estructurado.

HDFS: Hadoop Distributed File System. Sistema de archivos distribuido de Hadoop.

CRISP-DM: CRoss Industry Standard Process for Data Mining. Proceso estándar de minería de datos en toda la industria.

1 INTRODUCCIÓN

1.1 Planteamiento del problema

La agricultura se ha convertido en uno de los principales sectores de crecimiento de la economía colombiana en los últimos años. De acuerdo con el reporte del Producto Interno Bruto (PIB) del cuarto trimestre del 2017 emitido por el Departamento Administrativo Nacional de Estadística (DANE), se tuvo que para el año 2017, el valor agregado del sector de la agricultura creció en un 4,9% en su serie original, respecto al mismo periodo de 2016 [1]. Esta dinámica se explica por los crecimientos de las distintas actividades económicas dentro de las cuales se involucran los cultivos agrícolas transitorios y permanentes. A su vez se han generado 290.000 nuevos puestos de trabajo, los cuales han contribuido con la reducción de la población rural que vive en condiciones de pobreza [2].

En ese sentido, se tiene que la agricultura y los cultivos transitorios y permanentes presentan un rol importante en la dinámica de la economía colombiana. Dentro de los cultivos más importantes en la actualidad se encuentra el aguacate, para el cual la demanda ha estado teniendo un constante crecimiento a nivel mundial. Con una marcada tendencia al alza representada por un aumento de las exportaciones del 37,6%, durante el primer semestre del año 2019 respecto al mismo periodo del 2018, el aguacate se consolida como uno los productos más relevantes del país en materia de exportaciones no mineras. Así mismo, la producción de este cultivo en el país ha tenido un notable crecimiento [3]. En lo que respecta al intervalo de tiempo 2007-2018 se ha dado un crecimiento en su producción del 450%. En este mismo sentido, el departamento del Cauca no es ajeno a dicha tendencia; como muestra se resalta que en el año 2018 se tuvo una producción de 4.827 toneladas, siendo una participación significativa dentro del contexto de producción nacional del aguacate [2].

Sin embargo, los cultivos de frutas y hortalizas, incluyendo el del aguacate, presentan distintas dificultades, como el hecho de que el gremio que trabaja alrededor de estos

se encuentra conformado en gran parte por pequeños productores, los cuales están dispersos en el territorio nacional. Lo anterior genera un esfuerzo institucional mayor para poder agruparlos e incentivar el crecimiento de estos cultivos. Además de ello, existe una problemática con las políticas que se implementan para mitigar los riesgos que enfrentan dichos productores, los cuales se clasifican en 4 tipos: riesgos financieros, riesgos biológicos, riesgos climáticos y riesgos del mercado [5]. En ese sentido, dentro de los riesgos generados por el mercado, se detectan las variaciones de los precios de venta de los productos de acuerdo con los cambios en los mercados nacionales e internacionales, lo cual afecta a los agricultores.

Adicionalmente, los pequeños productores presentan dificultades para vender sus cultivos debido a que poseen pocos canales de venta, incluso únicos, para la comercialización de sus productos. Esto genera una dependencia de los distintos actores dentro de la cadena de valor, trayendo consigo que en muchos casos las ganancias obtenidas por los agricultores sean demasiado bajas. A su vez, los pequeños productores tienen problemas en la toma de decisiones relacionadas con la cuantificación de la producción de sus cultivos, debido a que se presenta una incertidumbre respecto a cuál debe de ser la cantidad óptima de producción. Este fenómeno se genera en parte a que se desconoce el precio al cual podrán vender sus productos y la cantidad de estos, incrementando así los niveles de pobreza en esta población [4].

En ese sentido y para mitigar los diferentes riesgos que enfrentan los agricultores, se han implementado políticas por parte del Ministerio de Agricultura y Desarrollo Rural, dentro de las cuales se encuentra la “estrategia 360”. Esta estrategia tiene como objetivo principal brindar a la población rural, y particularmente a los productores agropecuarios, instrumentos idóneos para gestionar los distintos riesgos asociados con su actividad; permitiendo que ellos se concentren exclusivamente en lo que les compete [5]. Para ello la estrategia define unos objetivos, dentro de los cuales está el brindar a los distintos actores la información adecuada sobre el sector agropecuario para fomentar la inversión del sector privado. En este objetivo se mencionan además las nuevas tecnologías de la información como una herramienta importante para lograr mejorar la calidad de vida de los productores [5].

Dada la problemática existente alrededor de las condiciones de pobreza en las que vive gran parte de la población rural que trabaja como productores en el sector agrícola, estas son generadas en parte por las variaciones existentes en el mercado y la poco efectiva toma de decisiones en lo referente a las cantidades de productos agrícolas generados. Por lo cual, se identifica la necesidad de recopilar nuevos datos provenientes de diferentes fuentes relacionadas con los mercados agrícolas, para así poder generar soluciones basadas en las Tecnologías de la Información y las Comunicaciones (TIC), ya que los datos son un pilar fundamental para este tipo de soluciones; y actualmente existe una alta heterogeneidad y un número limitado de fuentes de información relacionadas con los mercados agrícolas.

Es en este contexto donde Ecotecma S.A.S, una empresa que desarrolla soluciones enmarcadas dentro del ámbito de la agricultura climáticamente inteligente y el soporte en la toma de decisiones, busca contribuir en la solución de la problemática planteada a partir de un sistema que recopile datos demográficos y socioeconómicos relacionados con la dinámica de los mercados agrícolas, para así poder generar información de valor para la organización y los productores de aguacate ubicados en el Departamento del Cauca. Buscando de esta manera contribuir en la eliminación de las brechas existentes entre la implementación de soluciones basadas en las Tecnologías de la Información y las Comunicaciones (TIC) y los procesos productivos y de comercialización del sector agrícola en el departamento.

1.2 Objetivos

1.2.1 Objetivo General

Construir un repositorio de datos agropecuarios para el análisis del mercado del aguacate.

1.2.2 Objetivos específicos

- Caracterizar posibles fuentes para la extracción de datos demográficos y económicos del mercado del aguacate.

- Desarrollar un servicio de captura automática de datos demográficos y económicos del mercado de aguacate, para la conformación de un lago de datos.

- Implementar un lago de datos con los registros capturados del mercado del aguacate.

1.2.3 Metodologías para la implementación de lagos de datos

De acuerdo con la revisión del estado del conocimiento alrededor de las metodologías usadas para la implementación de lagos de datos, se encontró que no se han generado metodologías bien definidas para la construcción de lagos de datos, lo cual genera cierta complejidad para la implementación de este tipo de repositorios. Sin embargo, existen metodologías formales para la implementación de almacenes de datos, algunas de las cuales son: Ralph Kimball, Hefestos y SAS Rapid Data Warehouse Methodology, mencionadas en [6]. A pesar de la existencia de estas metodologías para la construcción de almacenes de datos, no se pudo encontrar alguna adaptación de estas metodologías para la construcción de un lago de datos, por lo tanto, se optó por adaptar una de las metodologías encontradas.
Para ello, partiendo de la revisión bibliográfica realizada se determinó que la metodología más usada para la implementación de almacenes de datos es la de Ralph Kimball. A continuación, se presenta una descripción de la metodología.

1.2.3.1 Metodología Ralph Kimball

La metodología de Kimball, también llamada Modelo Dimensional, se basa en lo que se denomina el Ciclo de Vida Dimensional del Negocio. Este modelo se constituye por modelos de tablas y relaciones con el propósito de optimizar la toma de decisiones. Este modelo es una técnica de diseño lógico que tiene como objetivo presentar los datos dentro de un marco de trabajo estándar e intuitivo, para permitir su acceso con un alto rendimiento [6].

Las fases establecidas por Ralph Kimball han sido diseñadas para que puedan ser desarrolladas en paralelo o en forma secuencial; cada una de las fases planteadas en esta metodología garantiza la calidad en el desarrollo del repositorio de datos.

You

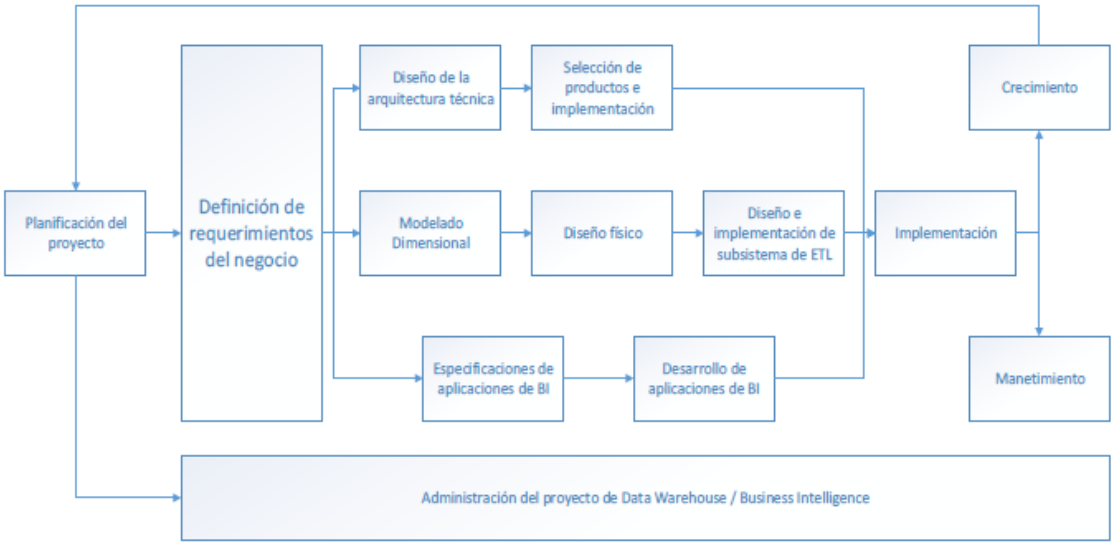

Figura 1. Fases de la metodología Ralph Kimball. Fuente [6].

Definida la figura 1, se procede a describir las diferentes fases que componen esta metodología:

- **Planificación del proyecto**: busca identificar la definición y el alcance del proyecto de almacén de datos. Se focaliza sobre recursos, perfiles, tareas, duraciones y secuencialidad. Es independiente del negocio y sus

requerimientos y busca identificar el escenario del proyecto para saber de dónde surge la necesidad del almacén de datos [7].

- **Definición de los requerimientos del negocio**: los diseñadores de los almacenes de datos deben entender los factores claves que guían al negocio para determinar efectivamente los requerimientos y traducirlos en consideraciones de diseño apropiadas, ya que son esenciales para las tres etapas subsiguientes enfocadas en la tecnología, los datos y las aplicaciones [7].
- **Modelado dimensional**: en esta fase se comienza con una matriz donde se determina la dimensionalidad de cada indicador y después se identifican los diferentes grados de detalle (atributos), dentro de cada concepto del negocio, así como la granularidad de cada indicador y las jerarquías que dan forma al modelo dimensional del negocio o mapa dimensional [7].
- **Diseño físico**: el diseño físico se enfoca sobre la selección de estructuras necesarias para soportar el diseño lógico. Los elementos principales de este proceso son la definición de convenciones estándares de nombres y configuraciones específicas del ambiente de la base de datos [7].
- **Diseño e implementación del subsistema de ETL**: se definen como procesos de extracción aquellos requeridos para obtener los datos permiten efectuar la carga del modelo físico acordado. Estos procesos de carga de datos sirven para incrementar el tamaño del almacén de datos [7].
- **Diseño de la arquitectura técnica**: los ambientes de almacenes de datos requieren la integración de numerosas herramientas. Para la elección de estas herramientas se debe tener en cuenta tres factores: los requerimientos del negocio, los ambientes técnicos actuales, y las directrices técnicas estratégicas futuras planificadas [7].
- **Selección de productos e implementación**: realizado el diseño de la arquitectura técnica se requiere evaluar y seleccionar los componentes específicos de la arquitectura como son la plataforma de software, el motor de base de datos, la herramienta de ETL o el desarrollo necesario, entre otras [7].
- **Especificación de Aplicaciones para usuarios finales**: los diferentes roles o perfiles de los diversos usuarios determinan las interfaces de acceso al almacén de datos. Dentro de estas interfaces se encuentran: herramientas de

diseño de reportes, tableros de control para gerentes, envío de información a usuarios internos o externos, entre otros [7].

- **Desarrollo de aplicaciones para usuarios finales**: de acuerdo con lo especificado en las aplicaciones para los usuarios finales, se procede al desarrollo de las interfaces seleccionadas a partir de las diferentes configuraciones de los metadatos y la generación de los diversos reportes definidos [7].
- **Implementación**: esta fase representa la convergencia de la tecnología, los datos, y las aplicaciones de los usuarios finales. Para ello existen diversos factores adicionales al correcto funcionamiento de las tres etapas mencionadas previamente, dentro de estos factores se tiene: la capacitación, el soporte técnico, y la comunicación entre los diferentes actores [7].
- **Mantenimiento y crecimiento**: el proceso de creación de un almacén de datos es acompañado por la evolución de la organización a lo largo del tiempo, esto significa que se requiere continuar con la actualización y el crecimiento de los diferentes componentes del almacén de datos a medida que las necesidades de la organización van creciendo [7].

De acuerdo con la figura 1 y la descripción de las fases, se puede observar que existe un número significativo de fases de esta metodología que son útiles para la construcción de un lago de datos, sin embargo, la fase de modelado dimensional, diseño físico y el diseño e implementación de un subsistema de ETL, no forman parte del desarrollo del lago de datos, ya que el proceso y la recolección de los datos en un lago de datos se realiza a través de la implementación de procesos ELT, además de que no se requiere hacer una definición en detalle de los diferentes atributos que deben de cumplir los archivos para almacenar y de las clases y relaciones que definen a la base de datos del repositorio de datos. Adicionalmente, para la implementación del Lago de Datos no se incluyó el desarrollo de aplicaciones de inteligencia de negocio (por sus siglas en inglés BI), ya que no está orientado a usuarios operacionales o analistas de negocio, sino a usuarios con un mayor conocimiento de las herramientas analíticas de datos como lo son los científicos de datos. A continuación, se presenta el diagrama que muestra las diferentes fases de la metodología Kimball que se utilizaron en la construcción del lago de datos.

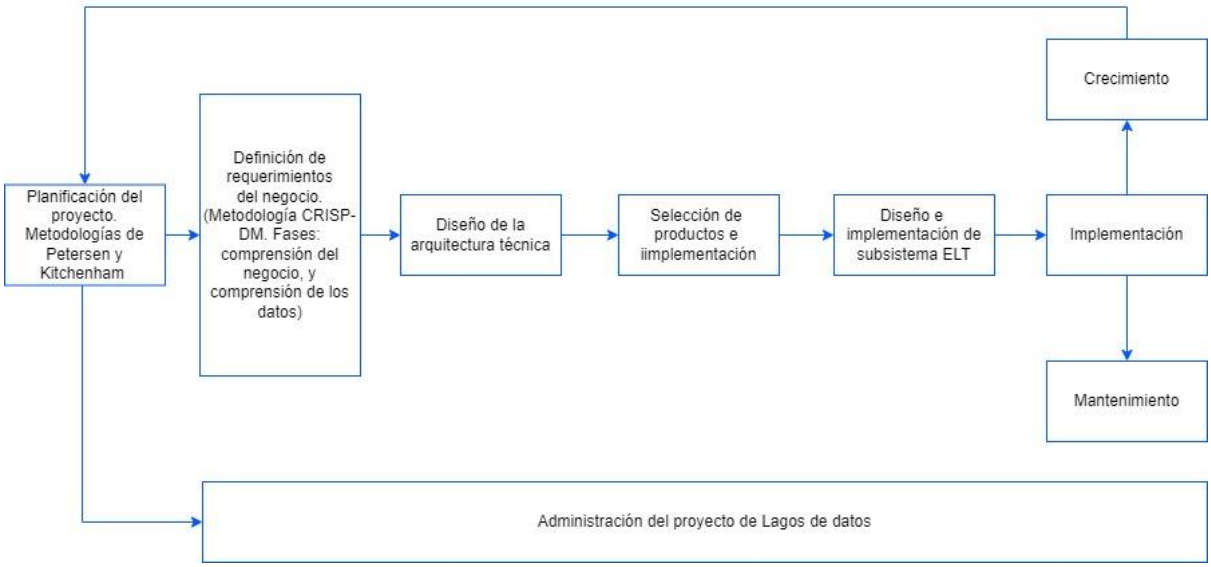

Figura 2. Metodología Kimball adaptada. Fuente propia

Para la planificación del proyecto se utilizaron las metodologías de mapeo y revisión sistemática de Petersen [8] y Kitcherman [9], mediante las cuales se identificaron los temas relevantes para el dominio del mercado de aguacate, como insumo para la adecuada planificación del estudio. Seguidamente, para la definición de requerimientos del negocio, se utilizaron las fases de comprensión del negocio y comprensión de los datos, de la metodología CRISP-DM [10]. Después de ello se incluyen las fases de diseño de la arquitectura técnica, selección de productos e implementación, diseño físico, diseño e implementación de subsistema ELT, como también, la implementación, el crecimiento, el mantenimiento, y la administración del proyecto de lago de datos, fases propias de la metodología de Kimball.

2 PLANIFICACIÓN DEL PROYECTO

En este capítulo se describe el resultado del mapeo sistemático relacionado con la analítica de datos, los procesos ELT y ETL, la extracción de datos, y el análisis de mercados agrícolas, lo cual de acuerdo con la adaptación de la metodología Kimball mostrada anteriormente, corresponde a la fase de planificación del proyecto, la cual se realizó basándose en la metodología de Petersen y de Kitchenham. A su vez se definen las características principales de los procesos de analítica de datos, los lagos de datos y los mercados agrícolas, finalmente este capítulo concluye con la búsqueda de antecedentes de aplicaciones relacionadas con la problemática del proyecto.

2.1.1 Metodología Petersen

La metodología Petersen proporciona un proceso para la realización de mapeos sistemáticos. Dentro de esta metodología, se tienen las siguientes fases: definición de las preguntas de investigación, la búsqueda de los artículos relevantes, la revisión de los artículos, el registro de los resúmenes, la extracción de los datos y el proceso de mapeo. Cada una de estas fases genera un resultado, y el resultado final del proceso es el mapeo sistemático [8]. A continuación, se presenta el diagrama que muestra las fases de la metodología.

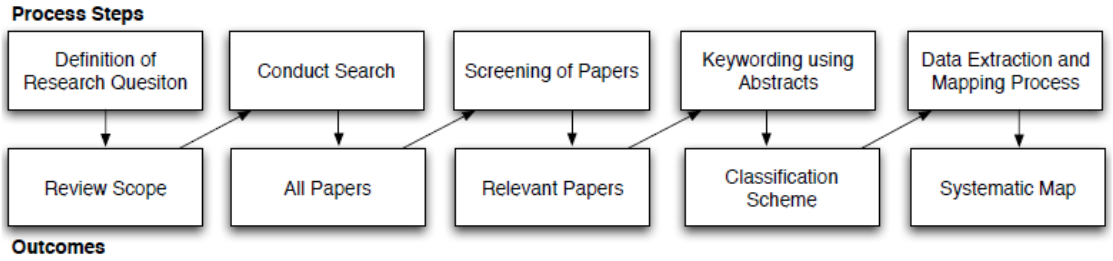

Figura 3. Proceso del mapeo sistemático. Fuente [8]

- **Definición de las preguntas de investigación**: el objetivo principal de un mapeo sistemático es el de proveer una vista general del área de investigación e identificar y cuantificar los diferentes resultados disponibles. Para lograr este objetivo, se deben de generar las preguntas de investigación correctas a las cuales se quieren dar respuesta [8].
- **Búsqueda de artículos relevantes**: los estudios primarios son identificados a partir del uso de las cadenas de búsqueda en las diferentes bases de datos científicas. Una manera adecuada de creación de cadenas de búsqueda se puede generar a partir de términos de población, intervención, comparación y resultados. A su vez la estructura debe de estar en la misma línea que las preguntas de investigación previamente formuladas [8].
- **Revisión de los artículos para inclusión y exclusión**: en esta etapa, se busca excluir los estudios que no son relevantes para dar respuesta a las preguntas de investigación formuladas [8].
- Palabras clave de los resúmenes: en esta fase se involucra el concepto de "Keywording" el cual hace referencia a una manera de reducir el tiempo necesario para el desarrollo del esquema de clasificación y asegurar que el esquema toma los estudios existentes en cuenta. "Keywording", se realiza en dos pasos: primero se leen los resúmenes y se buscan las palabras clave y los conceptos que reflejen la contribución del artículo. Después de realizar este paso con los diferentes artículos se agrupan las palabras clave comunes entre los diferentes artículos, para así entender de una mejor manera la naturaleza y la contribución de la investigación [8].
- **Extracción de los datos y mapeo de los estudios:** realizado el esquema de clasificación, se procede a organizar los artículos dentro del esquema. El análisis de los resultados se enfoca en presentar las frecuencias de las publicaciones en cada categoría, para así poder identificar en cuáles categorías se ha hecho énfasis en el pasado y en cuales hay posibilidades de mayor profundización en el futuro [8].

2.1.2 Metodología Kitchenham

La metodología Kitchenham se utiliza para la generación de una revisión sistemática de la literatura. Los pasos para la generación de esta revisión se presentan a continuación.

- **Preguntas de investigación**: esta fase hace referencia a la formulación de las preguntas de investigación, a las cuales se quiere dar respuesta a través de la revisión de la literatura. Esta fase es la más importante de todo el proceso, ya que define la orientación de la revisión sistemática a realizar [9].
- **Proceso de búsqueda**: hace referencia al proceso de búsqueda específica de diferentes procedimientos y artículos relacionados con el campo de investigación [9].
- **Criterios de inclusión y exclusión**: en esta fase se busca identificar cuáles son los estudios primarios que proveen información directa relacionada con las preguntas de investigación. Estos criterios se definen al principio de la revisión para así no generar un sesgo sobre la selección de los estudios, sin embargo, estos criterios se pueden modificar a lo largo del proceso de búsqueda [9].
- **Evaluación de la calidad**: adicional a los criterios de inclusión y exclusión se requiere realizar una evaluación de la calidad de los estudios primarios para así poder generar unos criterios de inclusión y exclusión más detallados, investigar si las diferencias de calidad explican las diferencias en los resultados de los estudios, orientar la interpretación de los resultados y determinar la fuerza de las inferencias [9].
- **Recolección de datos**: en esta parte se busca recolectar toda la información necesaria para realizar el estudio de calidad. En la mayoría de los casos se definen unos valores numéricos y cualitativos que deben de ser extraídos de cada uno de los estudios, para así poder hacer un metaanálisis sobre los estudios primarios [9].
- **Análisis de datos**: en esta fase, se muestran los resultados obtenidos, mediante tablas y gráficos. Estas tablas y figuras se estructuran para poder presentar las diferentes similitudes y diferencias entre los diversos estudios [9].

2.1.3 Conceptos relacionados con procesos ETL, ELT, minería de datos, y mercados agrícolas

Se realiza el mapeo sistemático basado en las metodologías de Petersen y Kitchenham. Para ello se definió la siguiente pregunta de investigación:

R1. ¿Qué relación existe entre Big Data, procesos ETL y ELT, minería de datos y lagos de datos con la agricultura, los análisis de mercado y las cadenas de suministro?

Para el mapeo sistemático se definieron dos periodos de tiempo: el primero consiste en los documentos publicados antes del año 2018, y el segundo contiene los documentos publicados desde el 2018 hasta el 2021. A continuación, se muestran los mapas longitudinales y los mapas de la ciencia, de los respectivos períodos, obtenidos a partir de la herramienta SciMAT:

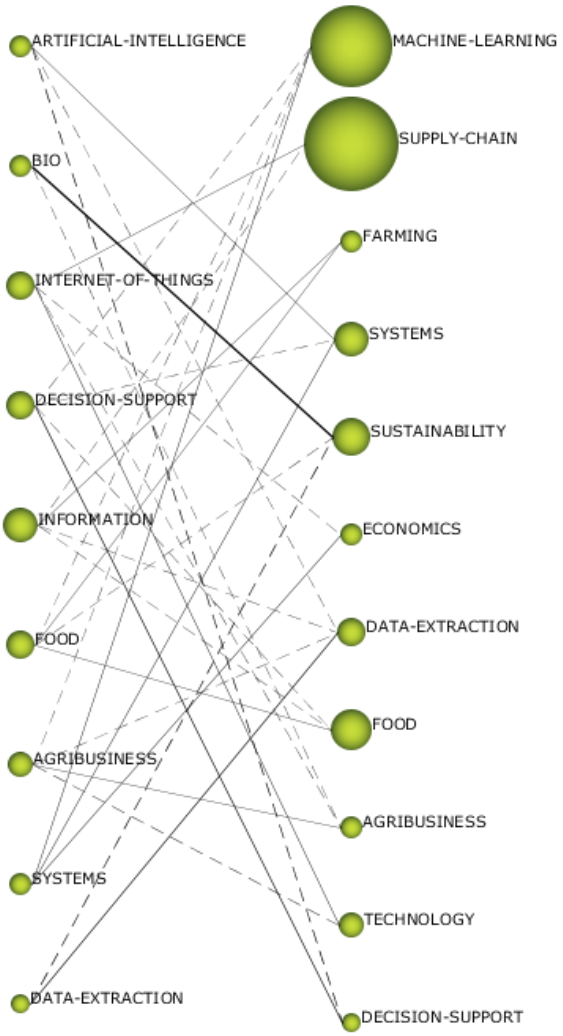

Figura 4. Mapa longitudinal. Fuente propia.

En la figura 4 se observan dos columnas que representan los dos periodos de tiempo considerados, los cuales son: los documentos publicados antes del 2018, y después del 2018. Las principales tendencias que se pueden ver en la figura se relacionan con la transición que se ha generado desde las investigaciones en torno al “Internet Of Things” en el primer periodo de tiempo definido para las publicaciones realizadas hasta antes del año 2018, hacia los temas de “Supply-chain” y de “Machine Learning” en la segunda columna, que representa el segundo periodo de tiempo, comprendido entre el año 2018 y 2021. Además de ello se puede observar que se dio un crecimiento en los temas de “Machine Learning” y de “Supply-chain” en el segundo periodo de tiempo, también se mantuvieron los temas de “data extraction” y de “agribusiness”, y se dio una transición entre los temas relacionados con “bio” hacia “sustainability”.

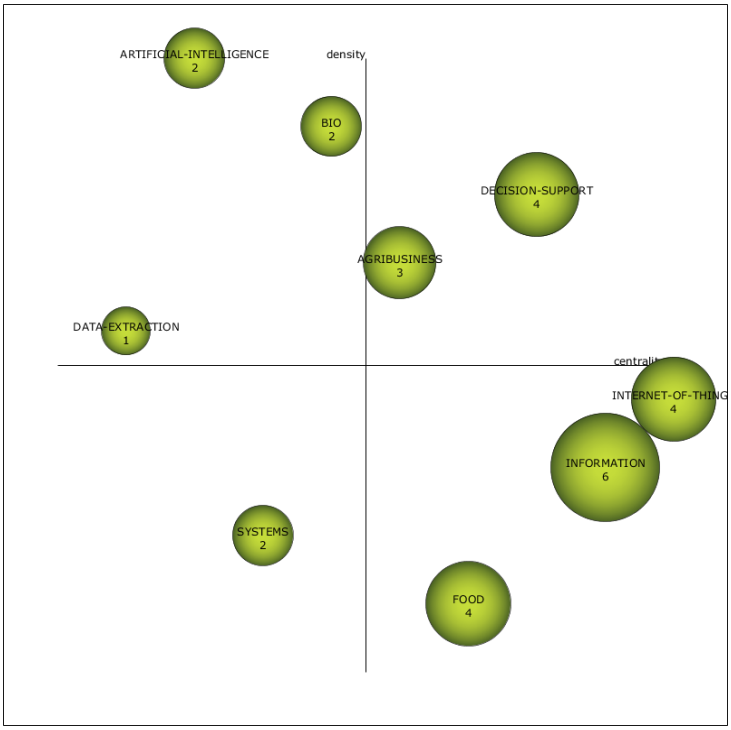

Figura 5. Mapa de la ciencia con número de documentos, primer periodo. Fuente propia.

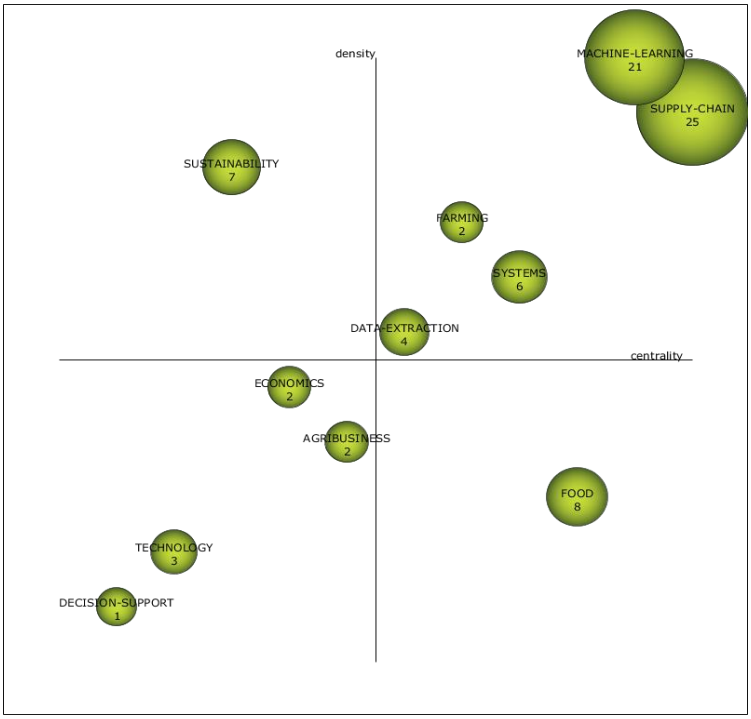

Figura 6. Mapa de la ciencia con número de documentos, segundo periodo. Fuente propia.

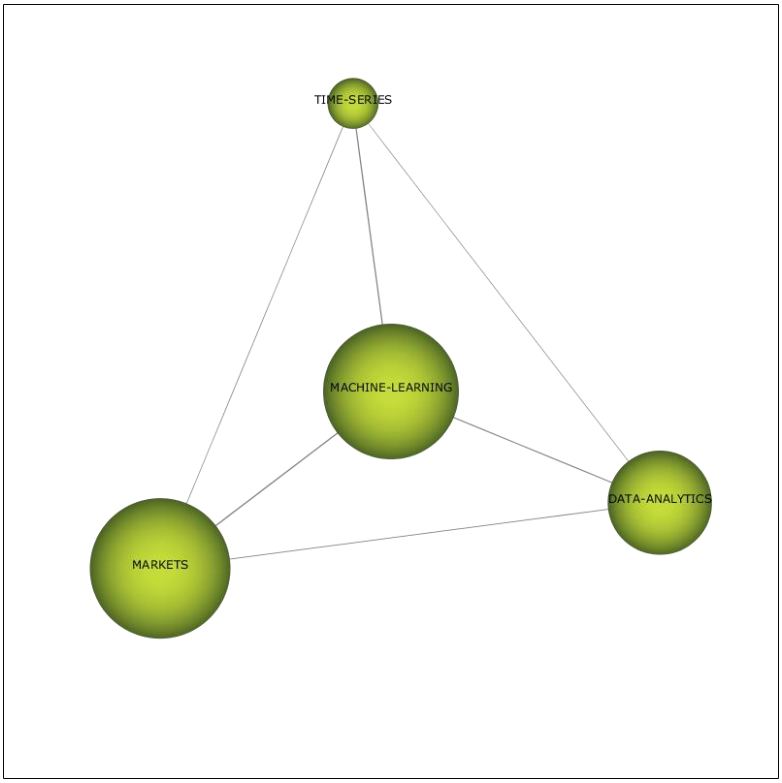

Figura 7. Primer análisis de cluster de Machine Learning. Fuente propia.

De acuerdo con los mapas de la ciencia mostrados en las figuras 5 y 6, los cuales representan la cantidad de documentos principales que se generaron en los dos períodos definidos para la realización del análisis bibliométrico, se visualiza que existen diversos temas con alta densidad y centralidad en el primer periodo de tiempo, dentro de los cuales se encuentran: soporte de decisión, agronegocios e Internet de las Cosas. El tema de soporte de decisión se encuentra sobre el primer cuadrante que corresponde a los temas motor, se puede encontrar también que se tiene el tema de agronegocios el cual no tiene una centralidad muy alta, y el tema de Internet de las Cosas se encuentra en el cuarto cuadrante el cual corresponde a los temas básicos y transversales. En este periodo de tiempo no se encontraron muchos documentos, por lo cual muchos de los temas no tienen más de cinco documentos principales [11].

En el segundo periodo, correspondiente al intervalo del 2018 hasta el 2021, se tuvieron un mayor número de publicaciones, debido a que los temas de procesos ETL, ELT, lagos de datos, y analítica de datos aplicados en el sector agrícola, en especial

en el análisis de mercados, ha sido bastante reciente. En la figura 6 se pueden identificar los temas de Machine Learning, y Supply chain como los temas motores, adicionalmente el mayor número de publicaciones principales se encuentran en estos dos temas. Dentro del cuadrante de temas emergentes y en declive, se encuentra: economía y agronegocios; el tema de sostenibilidad se encuentra como un tema aislado con respecto a la búsqueda realizada. En la figura 7, se muestra el mapa individual del tema de Machine Learning para el segundo periodo de análisis, en el cual se observa que existe una relación directa con los temas de mercados, analítica de datos, y series de tiempo [11].

De acuerdo a la revisión realizada de los diferentes conceptos relacionados con procesos ETL, ELT, minería de datos y mercados agrícolas, se procede a hablar acerca de la incidencia de diferentes variables sobre los mercados agrícolas como tema relevante de acuerdo a la revisión sistemática realizada, luego de ello se habla acerca del concepto y la importancia de la analítica de datos dentro de la agricultura, y del uso del lago de datos en este sector.

2.1.4 Incidencia de variables en mercados agrícolas

Los mercados en economía se definen como un conjunto de transacciones o intercambio de bienes y servicios entre individuos. En el mercado se busca hacer referencia a los acuerdos realizados entre distintos partícipes en un marco de transacciones [5]. Los partícipes se pueden agrupar en dos categorías: ofertantes (productores, vendedores), y demandantes (consumidores, compradores). Las transacciones se suelen clasificar de acuerdo con los participantes, los cuales pueden ser: consumidores, productores, empresas, y gobierno. Dentro de las diferentes combinaciones de participantes, destacan dos principales: empresa-a-consumidor (Business-to-Consumer B2C), y empresa a empresa (Business-to-Business B2B). El mercado es a su vez el ambiente social o virtual que proporciona las condiciones necesarias para realizar las transacciones comerciales de productos y servicios entre los diferentes participantes [12].

Los productos agrícolas por su parte hacen referencia a todos los productos de la agricultura que se han generado a partir de las actividades humanas que se realizan

a través de los cultivos. Estos artículos se suelen clasificar en dos categorías: agrícolas alimentarios, y agrícolas industriales, los cuales se determinan a partir del uso final que se suele dar al producto. Por lo cual los mercados agrícolas, hacen referencia al ambiente social o virtual, que se establece para realizar las diferentes transacciones entre oferentes y demandantes de diferentes tipos, para la comercialización de artículos agrícolas [13].

Los mercados agrícolas a su vez se ven afectados por distintas variables. A continuación, se mencionan algunas de las más relevantes en este trabajo:

- **Densidad poblacional**: la densidad poblacional hace referencia a la distribución de la población que habita en una área urbana o rural de una región determinada. Se mide como número de habitantes por unidad de superficie terrestre; que bien puede ser: kilómetros cuadrados, o millas cuadradas [14].

- **Precipitación**: hace referencia al fenómeno de la condensación del aire, hasta el punto de alcanzar un 100% de humedad, para así generar la precipitación la cual puede darse en forma de: lluvia, nieve, granizo, entre otras. Esta variable por lo regular se mide con un pluviómetro, y su unidad de medida son los milímetros de agua [15].

- **Política monetaria**: la política monetaria o política financiera es una rama de la política económica, que busca mantener la estabilidad económica en una nación o zona económica. Comprende las decisiones de las autoridades monetarias respecto al mercado del dinero, las cuales modifican variables como: la cantidad de dinero, el tipo de interés, y las tasas de cambio de la moneda. En los casos que se aplica para aumentar la cantidad de dinero, se le denomina política monetaria expansiva, y cuando se busca reducir la cantidad de dinero se nombra política monetaria restrictiva [16].

Existen diversos estudios respecto a la influencia de variables meteorológicas, demográficas, poblacionales, y económicas, sobre diferentes tipos de mercados,

entre los cuales están los relacionados con la venta y comercialización de productos agrícolas. Además, se analizan los cambios de diferentes variables y sus efectos sobre el rendimiento de los cultivos, los salarios y el bienestar de los agricultores. A continuación, se presentarán algunos trabajos relacionados con la influencia de este conjunto de variables sobre los cultivos y mercados agrícolas.

Se han realizado estudios con respecto a la relación existente entre la densidad poblacional y los mercados agrícolas. En [14] se analizó como un aumento de la densidad poblacional en una determinada zona rural de Malawi, implica una reducción del tamaño de las tierras de los campesinos. Adicionalmente se disminuían los salarios de los diferentes agricultores que trabajaban en esta zona debido al aumento de mano de obra campesina en la región. Sumado a esto, el incremento de la población en una zona rural permite que se de una mayor participación en el mercado por parte de los agricultores de algunas zonas en Etiopía [16]. En las zonas urbanas donde esta variable es alta o tiende a incrementar, se da un aumento en los precios de los distintos productos agrícolas que son producidos en zonas de baja densidad poblacional debido en parte a los costos de transporte [15].

Se han realizado estudios de la influencia de diferentes variables meteorológicas sobre el rendimiento de diferentes tipos de cultivos. Dentro de las variables estudiadas se encuentra la precipitación y la temperatura. Respecto a la precipitación, se determinó un efecto directo sobre los requisitos hídricos de distintos cultivos que se comercializan. En [15], se identificó el efecto que tienen las precipitaciones en el rendimiento de distintos cultivos de trigo de algunas zonas ubicadas en el centro y en el sur de Canadá. Se menciona una alta correlación entre la precipitación y el rendimiento de los sembrados, para los distintos periodos de crecimiento de los cultivos.

Adicionalmente, el cambio climático ha generado diversas variaciones en las ocurrencias de las precipitaciones, llegando a generar en ciertas regiones de Irán periodos de sequía, los cuales afectan directamente el rendimiento de diferentes cultivos, por lo cual se plantearon diferentes estrategias para afrontar los distintos cambios climáticos en la región [16].

El impacto de la temperatura difiere de acuerdo con el tipo de cultivo que se esté realizando, ya que cada uno tiene un rango de temperatura ideal. En el caso de los cultivos de maíz de diferentes zonas de Estados Unidos, por ejemplo, se tiene que las altas temperaturas generaron una disminución en la producción de estos cultivos. Para el caso de los cultivos de arroz en esas regiones, se tuvo que el rango de temperatura va de: 8°C a 36-40°C, teniendo en cuenta que la temperatura ideal para el crecimiento y reproducción de este cultivo es de 25°C [17].

Sumado a esto el cambio climático, ha generado ciertos aumentos de temperatura que han afectado el rendimiento de distintos cultivos producidos a nivel mundial, como es el caso de algunos cultivos generados en distintas zonas del Norte de Irán. Este aumento en la temperatura junto con la ausencia de precipitaciones en la región ha generado problemáticas en los agricultores, por lo cual se propusieron y desarrollaron algunas metodologías para la mitigación de estos riesgos climáticos [16].

Las políticas monetarias afectan los mercados agrícolas, a través de los cambios que se generan en las tasas de interés de los créditos, los subsidios generados, y las tasas de cambio, y cómo estas variaciones influyen en los precios agrícolas. En [18] se analizó la correlación existente entre los precios agrícolas y las políticas monetarias. En Sudáfrica se observó cómo los cambios en los precios agrícolas afectan los ingresos de los agricultores, sus decisiones de inversión, y la productividad de este sector. También se mencionaba como causa parcial de la volatilidad de los precios agrícolas; el alza que se da en los precios de estos, lo cual afectaba los ingresos y la productividad de los agricultores, además de influir sobre los consumidores finales, sobre todo en los que presentan un menor poder adquisitivo.

2.1.5 Analítica de datos

Se han realizado diferentes investigaciones y aplicaciones de la analítica de datos en diversos campos; con el objetivo de poder contribuir en la toma de decisiones que se realizan en los diferentes procesos de las empresas. Uno de los campos en los cuales la analítica ha comenzado a cobrar mayor relevancia y aplicación es la agricultura.

La analítica de datos en la agricultura ha influido en diferentes aspectos tales como la competencia entre los diferentes componentes de la cadena de valor, y la relación entre los diferentes actores de la cadena de valor agrícola. En [19], se investigó con respecto a la transformación de la agricultura convencional, por sus siglas en inglés (CA), en la agricultura de precisión (PA) en los Estados Unidos. Uno de los principales cambios se dio en la relación existente entre los agricultores o productores y los diferentes proveedores de productos de entrada para los procesos de producción en la agricultura, dentro de los cuales están: productos químicos, semillas, y maquinaria. Su relación cambió en el hecho de que los productores agrícolas comenzaron a suministrar los datos relacionados con el rendimiento y comportamiento de sus cultivos, tales como el comportamiento de variables meteorológicas, la geolocalización y el trayecto recorrido por las maquinarias, a las empresas proveedoras de productos de entrada de los cultivos.

Los datos proporcionados por los productores se utilizaron para mejorar la toma de decisiones por parte de estos, respecto a la producción y el rendimiento en los cultivos. La analítica de estos datos en este contexto se realiza por parte de las empresas proveedoras de productos de entrada agrícolas, como: Monsanto, John Deere, CNH, Bayer, y Agrium. Monsanto por su parte trabajó en el desarrollo de “FieldView”.

FieldView es una plataforma que provee almacenamiento y analíticas de información relacionada con la agricultura, y a su vez realiza recomendaciones a los productores respecto a cómo mejorar el rendimiento de sus cultivos. A su vez Monsanto realizó diferentes convenios con empresas tales como CNH y ACGO, las cuales están involucradas en el mercado de la maquinaria agrícola [19]. Estos convenios se realizaron con el objetivo de proporcionar una integración entre empresas de diferentes sectores de la agricultura tales como: maquinaria agrícola, productos químicos y semillas, para así poder compartir los diferentes datos obtenidos y mejorar el rendimiento de sus plataformas de análisis. Por ende, se ha comenzado a dar una integración entre las diferentes empresas que están involucradas en el mercado de los suministros de productos de entrada en la agricultura, debido en gran parte a la analítica de datos.

Adicional a los estudios realizados respecto a la dinámica de la relación agricultor/proveedor, se han realizado trabajos entorno a la adquisición de datos de los cultivos a partir de redes de sensores inalámbricos (WSN por sus siglas en inglés), datos abiertos, y diversas fuentes de datos; y el procesamiento necesario para la obtención de información de valor para los agricultores. En [20] se habla acerca de la aplicación de la analítica de grandes cantidades de datos en la agricultura en Corea. Se procesan imágenes de cultivos, obtenidas a partir de satélites, y datos obtenidos de las WSN con el objetivo de poder estimar y monitorear el rendimiento de los sembrados. Dentro de los campos de estudio, se encontró también la horticultura, definida como la cultivación de plantas, frutas y vegetales en huertas, diferente a los cultivos. Los datos que se recopilaron a partir de las WSN e imágenes 3D, se procesaron con el objetivo de determinar las condiciones ideales de variables como temperatura, y radiación de energía térmica a partir de luz LED para cultivos hidropónicos de canola.

A su vez en [20], se presenta una revisión literaria alrededor de la analítica de grandes cantidades de datos. En este artículo se detalla con respecto a las herramientas y tecnologías utilizadas en las diferentes fases de la analítica y los diferentes retos que se tenían. El almacenamiento de los datos fue realizado a partir del uso de herramientas tales como: el ecosistema hadoop, plataformas en la nube, y contenedores de datos (Data warehouse). En esta fase se encontraron problemáticas en torno al acceso rápido y seguro de los datos almacenados. La transformación de los datos se realizó mediante algoritmos de aprendizaje automático, normalización, visualización y anonimización de los datos. Las diferentes dificultades que se presentaron en los procesos de esta fase se encontraban en la heterogeneidad de las fuentes de datos, y la preparación y limpieza de los datos. En la analítica de datos, se aplicaron modelos de rendimiento, instrucciones de plantación, evaluación comparativa, ontologías de decisión, y computación cognitiva; algunas de las problemáticas que se identificaban en esta etapa, eran la escalabilidad y las analíticas en tiempo real.

En [21], se propone la aplicación e implementación de un modelo predictivo del rendimiento de un cultivo de algodón en el distrito de Ahmedabad, Gujarat, India. Este desarrollo se realizó a partir del uso de un ambiente de trabajo analítico, en el cual se

definió una cadena de procesamiento de grandes cantidades de datos (Big Data Pipeline). Esta se compone de 5 fases similares a las mencionadas anteriormente en [20], las cuales son: Extracción de datos, almacenamiento de datos en bruto, procesamiento e integración de los datos, implementación del motor analítico (analytical engine), y visualización de los datos.

Mencionada esta línea de trabajo, en [21] se propuso un ambiente de trabajo analítico, en el cual se definió una base de datos NoSQL denominada Cassandra, la cual almacenaba los datos en bruto obtenidos a partir de las diferentes fuentes de información y la información de valor resultante generada por el software Apache Spark. Spark en este entorno está encargado de realizar el procesamiento de los datos en bruto y suministrar la información de valor obtenida a la base de datos. Esto lo realiza a partir de la implementación de trabajos de aprendizaje automático iterativos alrededor de los datos en bruto. Después, los datos resultantes son suministrados por Cassandra a un tablero de visualización de datos implementado en "R", para poder realizar un análisis más profundo respecto a la información obtenida. De acuerdo con las predicciones del rendimiento obtenidas en [21], se pudieron realizar recomendaciones a los agricultores respecto a cómo mejorar el rendimiento de sus cultivos basados en las variables meteorológicas obtenidas.

2.1.6 Lagos de datos

Un lago de datos es un repositorio de datos utilizado para almacenar una masiva cantidad de datos en su formato original. Basado en tecnologías de bajo coste, que buscan mejorar la captura, refinación, almacenamiento, y exploración de datos en bruto de una empresa [22]. Un lago de datos contiene en desorden: datos no-estructurados y multi-estructurados en bruto, los cuales no generan en principio valor para la empresa.

Los lagos de datos se han popularizado como una plataforma de gestión de datos empresariales, la cual analiza diversas fuentes de información en su formato original. Este paradigma busca almacenar los datos en su formato original, en vez de colocar la información en un depósito de datos diseñado para ese tipo de datos en específico,

como es el caso de los contenedores de datos (Data warehouse). Este enfoque elimina ciertos costos del proceso de ingestión de datos, tales como la transformación de los datos. A su vez la información que se encuentra en el lago de datos está disponible para cualquier miembro de la organización o empresa, lo cual incrementa la agilidad en los procesos de análisis de datos y la accesibilidad a los mismos [22].

Alrededor de los lagos de datos, se han realizado investigaciones y aplicaciones en diferentes campos. Sin embargo, este enfoque de recopilación y gestión de los datos no ha sido muy aplicado en los diferentes procesos de la agricultura. A continuación, se darán a conocer algunas investigaciones realizadas respecto a la aplicación de esta tecnología en la agricultura, y adicional a ello se hablará acerca de los requisitos y componentes que debe cumplir un lago de datos para su correcto funcionamiento.

La aplicación e investigación de los lagos de datos en la agricultura de acuerdo con las búsquedas realizadas no ha sido exhaustiva. Sin embargo, en [23] se habla acerca de la aplicación de esta tecnología dentro de la resolución de diferentes problemáticas relacionadas con la exploración de nuevas alternativas tecnológicas que permitan mejorar la extracción de datos provenientes de diferentes fuentes y en distintos formatos.

En [23], se propone un lago de datos el cual está compuesto por un catálogo de metadatos y un componente relacionado con las ontologías; este lago de datos brinda acceso a diferentes actores tales como: el administrador, el científico de datos, y el encargado de la toma de decisiones. Dentro de la arquitectura propuesta, los sistemas de soporte de decisión y los sistemas de información están comunicándose con el lago de datos con el objetivo de tener disponible el acceso a la información de valor y a los datos en bruto para los distintos actores. Este enfoque busca resolver las diferentes cuestiones mencionadas anteriormente, y deja como tema de discusión la implementación de esta arquitectura de los lagos de datos en el campo de la agricultura, y los aportes que esta tecnología podría llegar a generar.

El concepto de lagos de datos suele tener múltiples interpretaciones y relaciones con otros conceptos del campo de los análisis de datos. En [24] se investiga acerca del estado del arte de los lagos de datos, con el fin de identificar los requisitos y

condiciones que debe cumplir un lago de datos para su correcto funcionamiento. Como se ha dicho anteriormente, los lagos de datos buscan almacenar los datos en su formato original, y de manera centralizada, con el objetivo de brindar la accesibilidad de los datos a los diferentes interesados de las empresas. Para ello se requiere de una tecnología de almacenamiento de grandes cantidades de datos en su mayoría no-estructurados, que bien puede ser: Hadoop (HDFS) o MongoDB.

Mencionada esta parte, se procede a hablar sobre las diferentes arquitecturas existentes, las cuales son: la arquitectura por zonas, y la arquitectura tipo estanque (pond architecture) [24]. La arquitectura por zonas hace referencia a cómo los datos en bruto, y los datos preprocesados son almacenados en diferentes zonas de la arquitectura. Una ventaja que brinda esta arquitectura hace referencia a la disponibilidad del acceso a los datos en bruto siempre que se requiera; oportunidad que no se brinda en la arquitectura tipo estanque. En este tipo de arquitectura solo se permite el acceso a la información almacenada en un estanque determinado, esto restringe el acceso a diferentes formatos de la información inicial, dentro de los cuales se encuentra los datos en bruto; esta condición va en oposición al concepto general de los lagos de datos, el cual busca brindar el acceso a los datos en diferentes formatos y de manera centralizada a los diferentes interesados de las empresas o instituciones.

La gestión de los metadatos es a su vez un componente importante dentro de los lagos de datos, ya que permite capturar la información referente a la semántica de los datos, los esquemas utilizados, y la proveniencia de los datos obtenidos. Dentro de las tecnologías que se utilizan para esta gestión, están los catálogos de metadatos, herramientas en las cuales se insertan los metadatos de la información obtenida, para así permitir la consulta de esta información por parte de los usuarios [24]. Esta gestión pertenece al campo del gobierno de los lagos de datos (Data lake governance), el cual involucra todo lo que hace referencia a las políticas y reglas que buscan asegurar la calidad de los datos. Para la realización de este objetivo existe un ambiente de trabajo general el cual sirve como guía dentro de la estructuración de los lagos de datos y su correcto funcionamiento. Existen todavía muchas brechas entre las arquitecturas y el gobierno de los lagos de datos, y su relación, lo cual requiere de mayor investigación y profundización en el futuro [24].

Realizada la descripción del estado del conocimiento en torno a los lagos de datos, se procede a realizar una comparación de los lagos de datos con respecto a los almacenes de datos (en Inglés Data Warehouses-DW).

En la tabla 1, se describen algunas de las principales diferencias y similitudes existentes entre los lagos de datos y los almacenes de datos tales como la estructura de datos que soportan, el procesamiento de los datos, la gestión de los esquemas definidos para el almacenamiento de los registros y el tipo de usuarios a los cuales está orientado cada uno de los repositorios, lo cual es de gran relevancia para este estudio.

Tabla 1. Comparación de lagos de datos y almacenes de datos. Fuente [25]

Características	**Almacenes de Datos**	**Lagos de Datos**
Estructura de datos	Los datos son procesados, sólo la información estructurada es capturada y organizada en esquemas.	Los datos están en bruto, todos los tipos de datos (estructurados, semi-estructurados, y no estructurados) son capturados en su forma original.
Usuarios	Ideal para usuarios operacionales tales como los analistas de negocio dado que los datos son estructurados y fáciles de manipular.	Ideal para usuarios avanzados tales como los científicos de datos, los cuales llevan a cabo análisis profundos con herramientas de analítica avanzadas.
Costos de almacenamiento	El almacenamiento de los datos consume tiempo y es costoso.	El almacenamiento de los datos es relativamente económico.
Accesibilidad	Las actualizaciones son costosas, ya que son complejas de realizar.	Las actualizaciones pueden ser realizadas rápidamente, permitiendo de esta manera una alta accesibilidad.
Posición del esquema	El esquema es definido previo al almacenamiento de los datos, lo cual permite tener rendimiento y seguridad.	El esquema es definido después de que los datos son almacenados, lo cual hace que sea muy ágil y escalable.
Procesamiento de los datos	Usa los procesos de Extracción, Transformación y Carga de datos (por sus siglas en inglés ETL)	Usa los procesos de Extracción, Carga y Transformación de datos (por sus siglas en inglés ELT).

2.1.7 Resumen del capítulo

Se definieron las principales tendencias relacionadas con la analítica de datos, los lagos de datos, los procesos ETL y ELT, y las metodologías usadas para la implementación de lagos de datos y almacenes de datos, haciendo énfasis en las aplicaciones sobre el sector agrícola, en específico sobre el análisis de mercados. Lo que facilitó tener una visión global de las temáticas de dominio del problema y la solución y su estado de desarrollo, siendo la base principal de la planificación del estudio.

3 CARACTERIZACIÓN DE LAS FUENTES DE DATOS

En este capítulo se describe la metodología CRISP-DM, y las fases de la metodología que se utilizaron para la caracterización de las diferentes fuentes de datos para el análisis de mercado del aguacate en los Estados Unidos, lo cual, de acuerdo a la metodología de Kimball adaptada para la implementación de lago de datos, corresponde a la fase de definición de requerimientos del negocio.

3.1.1 Metodología CRISP-DM:

La metodología CRISP-DM es un modelo del proceso con seis fases que describen el ciclo de vida de los proyectos de minería de datos. Las seis fases que componen este proceso son las siguientes: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación, y despliegue, las cuales se pueden observar en el siguiente diagrama [10]:

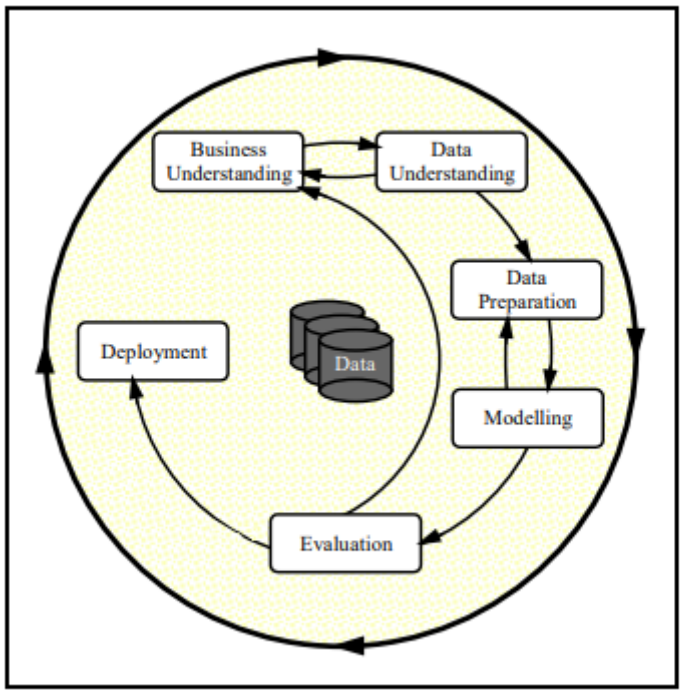

Figura 8. Diagrama de proceso de la metodología CRISP-DM. Fuente [10]

Ahora se procede a describir las fases mencionadas anteriormente:

- **Comprensión del negocio**: esta fase inicial se enfoca en el entendimiento de los objetivos y requisitos del proyecto desde una perspectiva de negocio, para así poder convertir este conocimiento en la definición de un problema de minería de datos [10].
- **Comprensión de los datos**: esta fase comprende la recopilación inicial de los datos y las diferentes actividades que se realizan para generar una primera aproximación sobre los datos, con el objetivo de poder descubrir algunas hipótesis iniciales, y obtener algunas conclusiones relevantes de los datos [10].
- **Preparación de los datos**: la preparación de los datos abarca todas las actividades relacionadas con la generación del conjunto de datos final que pasará a la fase de modelado. Dentro de las actividades que se encuentran en esta fase, se pueden mencionar: la limpieza de los datos, la ingeniería de requisitos, y la selección de requisitos [10].
- **Modelado**: en esta fase varias técnicas de modelado son seleccionadas y aplicadas, además de que sus parámetros se definen para obtener resultados óptimos. Además, existe una fuerte relación entre la fase de modelado y de preparación de los datos, ya que en la mayoría de los casos el correcto desempeño de los modelos depende de los conjuntos de datos obtenidos en la anterior fase [10].
- **Evaluación**: en esta etapa se han construido diferentes modelos de alta calidad desde una perspectiva de análisis de los datos. Sin embargo, antes de proceder al despliegue final del modelo es necesario cerciorarse de que los pasos para la construcción de este modelo fueron los correctos, y que se cumplieron los objetivos del negocio [10].
- **Despliegue**: en la fase de despliegue se procede a organizar y presentar la información obtenida a partir de todo el proceso realizado, de una manera en la que el cliente final la pueda utilizar. La complejidad de esta fase, depende en gran medida de lo que el cliente final quiera realizar, ya que el producto final bien puede ser un reporte o una aplicación que utilice los diferentes modelos y datos obtenidos [10].

Descritas las diferentes fases que componen a la metodología CRISP-DM, se define que para la caracterización de las fuentes de datos en torno al mercado del aguacate, se realizarán las fases de comprensión del negocio y de comprensión de los datos, ya que de acuerdo a la descripción de las diferentes etapas, se llegó a la conclusión de que en si se debe de realizar una descripción de la problemática de escasez de datos, entender que tipos de datos se quieren obtener y de que mercados en específico a partir de la fase de comprensión del negocio de la metodología, luego de ello ya se procede a analizar las diferentes fuentes de datos disponibles, describir las diferentes variables de las que se tienen registro en las fuentes, los métodos de acceso, y la periodicidad de los archivos, esto ya se realiza a partir del desarrollo de la fase de comprensión de los datos, que también tiene una relación bidireccional fuerte con la fase de comprensión del negocio.

3.1.2 Fuentes de datos para el análisis de mercado del aguacate en Estados Unidos

Para la identificación y búsqueda de las diferentes fuentes de datos relacionados con el mercado del aguacate, se utilizó la metodología de Kitchenham, la cual fue usada previamente para la realización del mapeo sistemático. Para ello, se definió la siguiente pregunta de investigación relacionada con las fuentes de datos:

R1: ¿Cuáles son las fuentes de datos existentes en torno al mercado del aguacate de los Estados Unidos?

La pregunta de investigación busca encontrar resultados de diferentes fuentes de datos del mercado del aguacate en los Estados Unidos, debido a que en la caracterización inicial se buscaron fuentes de datos en Colombia relacionadas con el mercado del aguacate, sin embargo esta búsqueda no retorno un número significativo de fuentes de datos, además del hecho de que la documentación para acceder a los registros de estas fuentes no era muy ordenada, los registros no eran demasiados, y a la vez la periodicidad de actualización de estos archivos era bastante amplia. Por ello, de acuerdo a los resultados obtenidos y a lo hablado con la empresa, se definió

que se iba a realizar la caracterización de las fuentes de datos del mercado del aguacate de los Estados Unidos.

De acuerdo, a la pregunta de investigación planteada se generaron las diferentes palabras clave, para la definición de las siguientes cadenas de búsqueda para Web Of Science (WOS) y Scopus.

Cadena de busqueda WOS:

TS=(avocado) AND TS = (repository Or "data sources" OR "statisti* service" OR "database") AND CU=(USA OR "united states" OR "united states of america")

Cadena de busqueda Scopus:
TITLE-ABS-KEY (avocado) AND TITLE-ABS-KEY (repository OR "data sources" OR "statisti* service" OR "database") AND TITLE-ABS-KEY (usa OR "united states" OR "united states of america")

Dentro de los resultados que se obtuvieron de las cadenas de búsqueda, se puede mencionar que dentro de las publicaciones [26, 27, 28, 29] se utilizaron diversas fuentes de datos relacionadas con el mercado del aguacate, dentro de las que se mencionaban se destacaron las siguientes: USDA (United States Department of Agriculture), FAO (Food and Agriculture Organization), y el CDFA (California Department of Food and Agriculture). A continuación, se presenta una gráfica que muestra la cantidad de citaciones que tuvo cada una de estas tres fuentes.

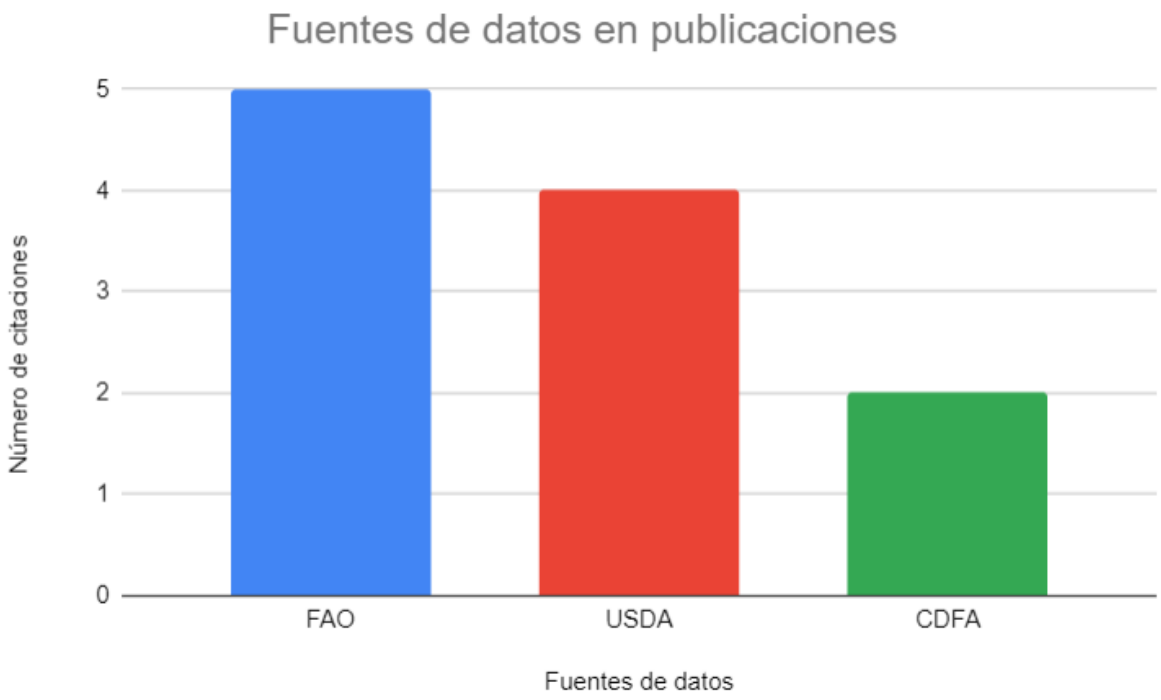

Figura 9. Cantidad de citaciones de fuentes de datos relacionados con el mercado del aguacate de los Estados Unidos. Fuente propia.

A partir de la figura 9, se puede ver que la FAO es la fuente de datos que más fue citada en las publicaciones halladas a partir de las cadenas de búsqueda, después de ello se tiene a USDA y luego al CDFA. Cabe mencionar que se obtuvieron 15 publicaciones a partir de la cadena de búsqueda en el caso de WOS y 13 publicaciones en Scopus, a su vez se repitieron cinco publicaciones en las dos bases de datos científicas.

Para el análisis de los años en los que fueron publicados los diferentes artículos, se definen dos periodos de tiempo, los cuales corresponden a: antes del 2012, y después del 2012, para así comprender un poco acerca de que tan reciente han sido las publicaciones sobre el mercado del aguacate en los Estados Unidos y sus diferentes dinámicas, para ello se presenta la siguiente figura.

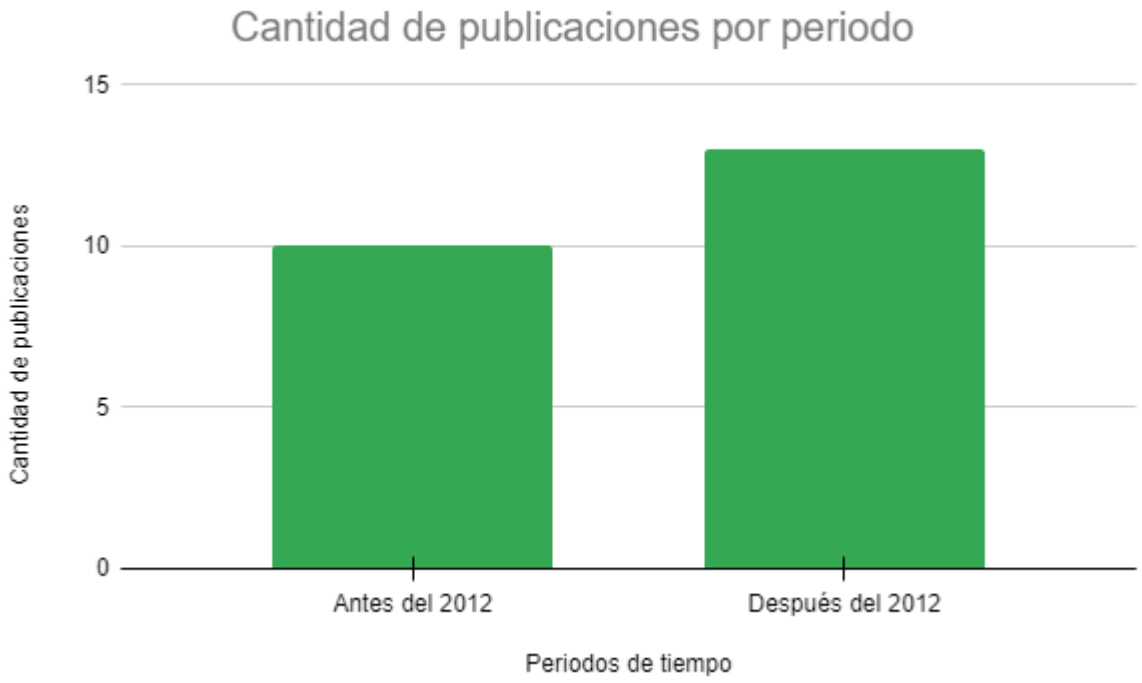

Figura 10. Cantidad de publicaciones por periodo de tiempo. Fuente propia

En la figura 10, se representan los dos periodos de tiempo previamente definidos, y en ellos se tiene que para el periodo antes del 2012, se generaron 10 publicaciones alrededor del tema de búsqueda, y para el periodo que incluye el 2012 y sus años posteriores se publicaron 13 investigaciones, esto indica que el número de publicaciones en torno a los mercados del aguacate no es muy alto ya que se identificaron 23 publicaciones en las dos bases de datos científicas, además se pudo identificar que en estas publicaciones no existía una gran cantidad de fuentes de datos definidas en torno al mercado del aguacate de los Estados Unidos, por lo cual se realizaron búsquedas adicionales de otras fuentes de información.

A continuación, se presenta la siguiente figura, que muestra los diferentes periodos de tiempo de los registros de datos disponibles que brinda las fuentes de la FAO, USDA, y el CDFA:

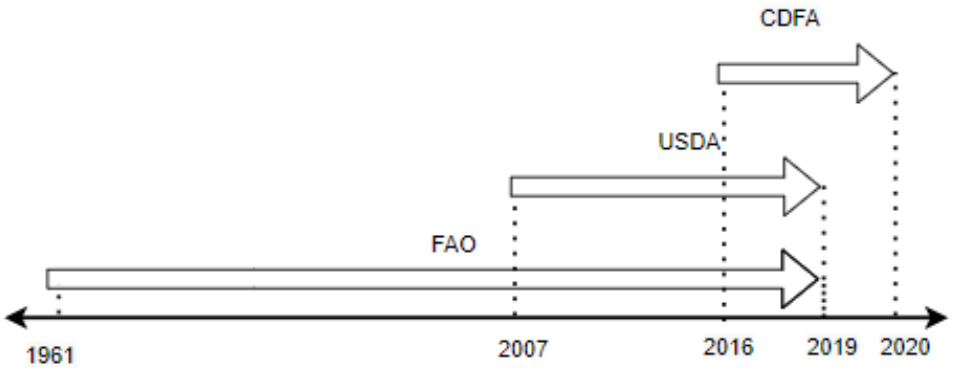

Figura 11. Periodicidad de fuentes de datos citadas en publicaciones. Fuente propia.

En la figura 11, se puede observar que la FAO es la fuente de datos que tiene un mayor registro de datos histórico ya que comprende desde el año 1961 hasta el 2019, luego se tiene a la USDA, la cual tiene registros desde el año 2007 hasta el año 2019, y por último se tiene al CDFA con un margen de tiempo que va desde el 2016 hasta el 2020.

Además de las fuentes de datos mencionadas, también se encontraron otras fuentes de información adicionales, dentro de las cuales esta: HAB (Hass Avocado Board), la cual es una organización dentro de los Estados Unidos que busca incrementar el consumo y la comercialización del aguacate en este pais, también se tiene a la oficina de censo de los Estados Unidos (United States Census Bureau), la cual es una fuente de datos importante en lo relacionado con la demografía del país y en el sector agrícola, y por último se tiene a Kaggle, la cual es una plataforma de Google que tiene una gran cantidad de conjuntos de datos de diferentes campos, temáticas y ramas de investigación dentro de las cuales se encuentra la agricultura y la producción y comercialización del aguacate.

De acuerdo a la investigación y búsqueda realizada de las diferentes fuentes de datos relacionadas con el mercado del aguacate en los Estados Unidos enmarcado en la fase de comprensión del negocio y de los datos de la metodología CRISP-DM, se caracterizaron todas las fuentes de datos que proporcionaron datos relacionados con la comercialización y producción del aguacate en los Estados Unidos, de acuerdo a los diferentes registros que proporcionaban y a lo hablado con la empresa Ecotecma

S.A.S. Los resultados de las fuentes de datos encontrados se muestran en la siguiente tabla:

Tabla 2. Fuentes de datos del mercado de aguacate de los Estados Unidos. Fuente propia.

Nombre fuente de datos	**Datos y variables**	**Descripción**	**Formatos de los datos**	**Periodicidad**
United States Department of Agriculture (USDA), National Agricultural Statistics Service (NASS)	Producción de aguacates a nivel nacional. Datos de consumo del aguacate producido. **Variables de producción** Aguacates Acres Bearing. Producción de aguacates medida en dólares y toneladas. Rendimiento de la producción. Producción del mercado fresco, en proceso, no vendido y utilizado de aguacate en dólares y toneladas.	Se tienen datos de la producción de aguacate medidos en dólares y toneladas. Además, se tienen datos de la producción total, de rendimiento, del mercado fresco, no vendidos y en proceso. Se puede acceder a la información de cada estado y ciudad en específico en un año determinado. Existen dos tipos de fuentes: encuesta y censo.	EXCEL, CSV	Estadísticas del aguacate desde el 2007 hasta el 2019, anuales.
United States Department of Agriculture (USDA), National Agricultural Statistics Service (NASS)	Envíos semanales de aguacate.	Este informe se refiere a las cantidades de envíos semanales por peso de aguacates. Los datos presentados están organizados por variedad de producto, modo de transporte y origen del envío.	TXT	Semanal. Registros desde el 27 de noviembre del 2018, hasta la actualidad.
FAOSTAT	Área cosechada, producción, y rendimiento de los cultivos de aguacate.	Se pueden acceder a los datos de producción, rendimiento y área cosechada de diferentes países dentro de los cuales se encuentra	CSV, XLS (EXCEL)	Registros anuales desde el año 1961 hasta el 2019. Se tienen

	Cantidad de importaciones y exportaciones en dólares y toneladas de aguacate. Estimaciones de la población. **Variables de producción** Área cosechada Rendimiento Producción cantidad. Importaciones y Exportaciones. Cantidad y valor.	Estados Unidos. Se puede obtener la cantidad de importaciones y exportaciones del aguacate a partir de la definición de los países socios: Estados Unidos, y declarantes: Colombia, México, Chile.		registros en la matriz de comercio hasta el 2019, sin embargo, el año de registros iniciales varía de acuerdo con el país.
FAOSTAT	Se tiene acceso a diferentes datos demográficos de los Estados Unidos y diferentes países. Las variables que se encuentran son las siguientes: **Variables demográficas** Población total Hombres Mujeres Población rural Población urbana	Se pueden acceder a diferentes valores correspondientes a diferentes estimaciones de diversos grupos de población: rural, urbana, hombres y mujeres.	CSC, XLS(EXCEL)	Registros anuales desde el año 1950 hasta el 2018.
United States	Cifras acerca	Archivos que contienen	EXCEL,	Los reportes de

Census Bureau	de la producción del aguacate. Cifras y estadísticas acerca de la población de los Estados Unidos. **Variables de producción.** Producción de aguacates utilizada. Valor de la producción de Estados líderes en producción	datos de importación y exportación de productos agrícolas de 1990 hasta 2010, existen categorías de frutas y vegetales. Se tienen archivos referentes a la producción, el suministro y el valor de diferentes frutas dentro de las cuales se encuentra el aguacate. Se tiene acceso a los datos que describen la población de los Estados Unidos. En los conjuntos de datos se pueden encontrar diferentes categorías como: economía, empleo, gobierno, salud, población, ingresos y pobreza.	PDF	agricultura se realizaron en el año 2012, no existen reportes de otros años.
California Department of Agriculture and Food	Producción de aguacates del estado de California. **Variables de producción:** Área cultivada Producción en toneladas. Producción utilizada. Valor por unidad. Rendimiento por Acre Valor total.	Se tiene acceso a los reportes agrícolas para el rango de años 2016-2020. En los archivos de reporte se pueden encontrar datos de producción de diferentes cultivos, entre los cuales se encuentran los del aguacate. En estos informes se encuentra información de la producción, del rendimiento de los cultivos, de las unidades vendidas y los respectivos precios.	PDF	Reportes desde el año 2016 hasta el 2020.
Hass Avocado Board	Datos relacionados con la producción y	Se encuentran los datos de producción e importación de aguacate Hass en los Estados	CSV	Se tienen reportes actualizados semanalmente

	comercialización de aguacate. **Variables de producción** Volumen y proyección del aguacate en Estados Unidos. Ventas de aguacate en unidades y dólares en Estados Unidos y por estados. Volumen de ventas minoristas. Precios medios de venta entregados trimestralmente	Unidos. Se tienen datos de países como: México, Chile, Colombia y Perú. Se tienen muchas gráficas que muestran la dinámica del mercado del aguacate en Estados Unidos acerca de la producción, importación y exportación. También se encuentra información acerca de las ventas del aguacate Hass realizadas en los Estados Unidos, en especial gráficas.		desde el 2018 hasta la actualidad, acerca de la producción de aguacate. Adicionalmente también se definen unas proyecciones de la producción.
Kaggle	Existen diferentes conjuntos de datos relacionados con la producción y comercialización del aguacate en el mercado de Estados Unidos y de diferentes países. **Variables económicas** Precio promedio de la unidad de aguacate. Número de aguacates vendidos. Volumen total.	Kaggle es una plataforma que brinda acceso gratuito a diferentes conjuntos de datos, de diversas áreas.	CSV	El dataset de precios tiene registros desde el 4 de enero del 2015 hasta el 25 de Marzo del 2018. Existe también un conjunto de datos actualizado con datos hasta el 29 de noviembre del 2020.

3.1.3 Resumen del capítulo

De acuerdo con la caracterización de las fuentes de datos del mercado del aguacate realizadas, se puede observar que las diferentes fuentes poseen registros en diferentes formatos como: CSV, EXCEL, PDF Y TXT; adicionalmente todas estas fuentes proveen herramientas para la extracción de datos como las páginas web de estas y algunas APIs que proporcionan fuentes como USDA y Kaggle.

4 SERVICIO DE CAPTURA AUTOMÁTICA DE DATOS

En este capítulo se describe el proceso realizado para la implementación del servicio de captura automática de datos de las diferentes fuentes de datos relacionadas con el mercado del aguacate en Estados Unidos, lo cual de acuerdo con la metodología Kimball corresponde a las fases de: diseño e implementación del proceso ELT, y el diseño físico. Cabe mencionar que las fases de planificación del proyecto y definición de requerimientos del negocio se abordaron en el capítulo uno de introducción en el cual se realiza el planteamiento del problema y se definen los requerimientos del servicio a realizar. Se presentará el entorno de desarrollo utilizado, las técnicas de scraping usadas y el consumo de APIs realizado sobre las diferentes fuentes de datos.

4.1.1 Tecnologías para la extracción, carga y transformación automática de datos

Se realizó la investigación de diferentes posibles tecnologías para la implementación del proceso ELT. Para la elección y evaluación de las diferentes tecnologías se elaboraron las siguientes tablas:

En la tabla se describen las diferentes posibles tecnologías para la implementación del proceso ELT, se menciona el hecho de que existen herramientas de código abierto, como lo son Selenium y Scrapy, que proporcionan las facilidades necesarias para la extracción de los datos de las fuentes caracterizadas previamente, a su vez existen herramientas como Automation Anywhere, y Mozenda las cuales proveen mayores capacidades para la extracción de datos, pero al mismo tiempo tienen un costo significativo, variable a tener en cuenta. Luego de ello, se realiza la descripción de diferentes servicios proporcionados por AWS y GCP, los cuales proveen herramientas necesarias para la generación de la carga y el almacenamiento de los datos. La descripción de los costos de las herramientas se realizará en la tabla 5 con mayor detalle. A continuación, se presentan los diferentes criterios de selección para la tecnología encargada de realizar la extracción de datos.

Tabla 3. Descripción, ventajas y desventajas de las posibles tecnologías para implementar el proceso ELT. Fuente propia

Nombre de la herramienta	**Descripción**	**Ventajas**	**Desventajas**
Scrapy	Scrapy es un ambiente de trabajo usado para la realización de Web Crawling, esta tecnología puede tener múltiples aplicaciones como minería de datos, procesamiento de información e implementación de bots [30].	Es de código abierto. Es multiplataforma y puede trabajar en Windows y Linux. Scrapy proporciona un ambiente de trabajo asíncrono.	Extrae información sólo de datos estructurados.
Selenium	Selenium es un conjunto de herramientas de software que realizan principalmente la automatización de procesos en las aplicaciones web, sobre todo en la parte de testing. Existen tres herramientas que componen el ambiente de trabajo de Selenium, las cuales son: Selenium IDE, Selenium Grid, y Selenium WebDriver. Se puede utilizar para automatización de procesos en navegadores [31]	Es de código abierto. Es un conjunto de herramientas para hacer web mining. Permite una mayor interacción y configuración de las acciones en el navegador. Soporta múltiples lenguajes como: Java, Python, C y Ruby.	Se utiliza solo para automatizar procesos en los navegadores.
Mozenda	Es una herramienta que permite la extracción y gestión de datos de la web. Los usuarios pueden extraer, almacenar y compartir información extraída de la web. Existen dos partes que conforman la herramienta: la consola web y el constructor de agentes [32].	Extrae información de datos estructurados y no estructurados. Interfaz de usuario amigable.	Pago mínimo de 250 USD por mes. Incluye: 1 usuario 20k procesamiento de créditos/month 10 agentes de web scraping. Solo disponible en Windows.
Automation Anywhere	Es una empresa que desarrolla software encargado de realizar la automatización de	Puede automatizar procesos de End to End (E2E) en las organizaciones	AA no extrae información de imágenes.

	procesos robóticos (RPA), los cuales usan bots para completar las diferentes tareas en los negocios [33].	Permite generar bots sin necesidad de código. También permite la personalización de los bots a partir de múltiples lenguajes de backend como .NET y C#. Puede convertir datos no estructurados, en estructurados a partir del servicio de IQ-Bot. Bot-Insight a su vez analiza el comportamiento y desempeño de los bots.	Puede llegar a ser costoso. Ofrece una versión gratuita, y además tiene una versión enterprise cuyo costo es de 995 USD. Este valor solo se paga una vez. La opción de descargar un cliente de escritorio sólo está disponible para la versión Enterprise. Proporcionan 30 días de free-trial para la versión enterprise. La versión web del "cuarto de control", no brinda la opción de generar el web recording. Este servicio es el que permite obtener información de las páginas web.
Web Content Extractor	Es una herramienta de web scraping y extracción de datos de Internet. Esta herramienta ofrece una interfaz para la extracción de datos, y la generación de patrones para los diferentes webs crawlers [34].	Permite exportar los archivos en diferentes formatos, como: CSV, TXT, XML Y SQL. Permite la extracción de datos en tiempo real. Posee una versión web y una versión de escritorio.	Solo tiene soporte para Windows. Solo tiene una versión de prueba de 14 días de duración.
AWS Glue	Es un servicio de integración de datos sin servidores totalmente administrado que facilita descubrir, preparar y combinar datos para análisis, aprendizaje automático y desarrollo de aplicaciones, a partir de la implementación de procesos ETL y ELT [35].	Provee un contexto para trabajos basados en Python, Spark y Scala. Provee acceso para que otros servicios como: Amazon Redshift y Amazon Atena, puedan consultar la información. Se pueden generar ETLs o ELTs anidados. Puede generar	Glue demanda altos recursos computacionales. Glue utiliza unidades de procesamiento llamadas DPU equivalentes a 4vCPU y 16GB RAM.

		catálogos de metadatos para que otros servicios puedan consultar la información.	
AWS Lambda	AWS Lambda es un servicio informático que permite ejecutar código sin administrar servidores. Lambda ejecuta el código sólo cuando es necesario y se escala de manera automática, pasando de pocas solicitudes al día a miles por segundo [36].	Tiene soporte para diferentes lenguajes de programación como: Python, Java y NodeJs. Solo se pagará por el tiempo informático que se consuma, no se cobra cuando el código no se está ejecutando. Se pueden generar aplicaciones compuestas por funciones que se activan de acuerdo con el suceso de eventos.	No permite la administración de los recursos informáticos necesarios para ejecutar nuestro código.
AWS Redshift	Es un servicio que administra todo el trabajo necesario para la configuración, el uso y el escalado de un data warehouse. Entre estas tareas se incluyen el aprovisionamiento de capacidad, la monitorización y la realización de copias de seguridad del clúster, y la aplicación de parches y actualizaciones al motor de Amazon Redshift [37].	Este servicio permite trabajar de una manera más sencilla con los datos en formatos abiertos. Además, permite consultar y exportar los datos hacia sobre un lago de datos. AWS Redshift permite analizar, crear e implementar modelos de aprendizaje automático en Amazon SageMaker a través de SQL.	Es un servicio totalmente administrado que hasta cierto punto no permite realizar gestión sobre los diferentes recursos informáticos que utiliza para la ejecución de los diversos procesos.
AWS Simple Storage Service (S3)	Es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento. Amazon S3, brinda la posibilidad de almacenar y proteger cualquier volumen de datos para los más	Permite la creación de un lago de datos de una manera rápida y sencilla a partir del uso de AWS Lake Formation y del almacenamiento para lago de datos [38].	Si comienza a crecer demasiado el número de solicitudes sobre los diferentes buckets o servicios de almacenamiento de S3 los costos pueden incrementar de manera significativa. Además, las configuraciones de seguridad y control de

	variados fines, como usarlos en lagos de datos, sitios web, aplicaciones móviles, procesos de copia de seguridad y restauración, operaciones de archivado, aplicaciones empresariales, dispositivos IoT y análisis de big data [38].		acceso no son tan sencillas de realizar.
Google Cloud Dataflow	Dataflow es un servicio administrado que ejecuta una amplia variedad de patrones de procesamiento de datos. La documentación que se ofrece en este sitio muestra cómo implementar las canalizaciones de procesamiento de datos por lotes y de transmisión [39].	Servicio totalmente gestionado de procesamiento de datos Aprovisionamiento y gestión automática de recursos de procesamiento Innovación en software libre motivada por la comunidad mediante el SDK de Apache Beam	Requiere de una doble actualización sobre los datos procesados que a su vez no se puede automatizar.
Google Cloud Functions	Cloud Functions es una solución de procesamiento ligera que les permite crear funciones independientes y de un solo propósito que respondan a eventos de Cloud sin tener que administrar un servidor o un entorno de ejecución [40].	No se requiere aprovisionar, administrar y actualizar servidores. Se escala automáticamente según la carga Funciones integradas de supervisión, registro y depuración. Seguridad integrada a nivel de funciones y por función que se basa en el principio de mínimo privilegio.	No permite la administración de los recursos informáticos necesarios para ejecutar nuestro código.
Google Cloud BigQuery	Es un almacén de datos que permite realizar consultas de SQL de alta velocidad mediante el poder de procesamiento de la infraestructura de	Permite acceder y compartir los datos y la información de valor de una forma sencilla. Se puede configurar como un repositorio de	No permite la administración de los recursos informáticos necesarios para ejecutar nuestras consultas SQL.

	Google [41].	datos tipo almacén de datos o lago de datos.	

En la tabla 4, se muestran los diferentes criterios utilizados para la selección de la tecnología para la extracción de datos, dentro de ellos se analiza si la herramienta permite extraer información de datos estructurados y no estructurados, para la evaluación de este criterio, se tiene que Scrapy y Selenium permiten la extracción de datos estructurados más no de no estructurados, por otro lado, las herramientas de Automation Anywhere, Mozenda y Web Content Extractor manejan datos estructurados y no estructurados, sin embargo estas herramientas tienen un costo significativo y a su vez Mozenda y Web Content Extractor sólo pueden ejecutarse en el sistema operativo de Windows. Debido a que en la ejecución de la práctica se busca realizar una primera aproximación en la extracción de datos de las fuentes del mercado del aguacate y no se requiere una herramienta que extraiga datos de archivos no estructurados, es por ello por lo que se escoge al ambiente de trabajo de Selenium como herramienta para la extracción de los registros, ya que se puede ejecutar en diferentes sistemas operativos, permite interactuar con el navegador web y es de código abierto.

Tabla 4. Criterios para la elección de la tecnología para la extracción de datos. Fuente propia.

Nombre de la herramienta	**Criterios extractores de datos**			
	¿Es de código abierto?	**¿Maneja datos estructurados?**	**¿Maneja datos no estructurados?**	**Sistemas Operativos**
Scrapy	Si	Si	No	Windows, macOS, Linux
Selenium	Si	Si	No	Windows, macOS, Linux
Mozenda	No	Si	Si	Solo windows
Automation Anywhere	No	Si	Si	Windows, Linux
Web Content Extractor	No	Si	Si	Solo windows

A continuación, se presenta la tabla de costos, de las tecnologías para la extracción de datos:

Tabla 5. Análisis de costos de las diferentes tecnologías para la implementación del proceso ELT. Fuente propia.

Nombre de la herramienta	Costos	Características
Scrapy	No tiene costo	No tiene costo
Selenium	No tiene costo	No tiene costo
Mozenda	Requiere un pago mensual de 250 USD por mes [32]	El pago incluye un usuario, 20k de créditos de procesamiento. 10 agentes de web scraping [32]
Automation Anywhere	La versión Enterprise se puede adquirir a partir de un pago único de 995 USD [33].	La descarga del cliente de escritorio solo se puede realizar adquiriendo la versión enterprise. Sin el cliente de escritorio no se puede realizar web scraping [33].
Web Content Extractor	Se puede adquirir, a partir de un pago único de 59 USD [34].	Tiene una versión de prueba de 14 días [34].
AWS Glue	Si se sobrepasa la capa gratuita, los precios de cobro serían de: 0,44 USD por hora de DPU. Se factura por segundo, con un mínimo de 1 minuto (Glue versión 2.0) o de 10 minutos (Glue versión 0.9/1.0) para cada trabajo ETL o ELT de tipo Apache Spark [42]	Capa gratuita por siempre.1 millón de objetos almacenados en el catálogo de datos de AWS Glue 1 millón de solicitudes realizadas por mes al catálogo de datos de AWS Glue [43]
AWS Lambda	0,20 USD por un millón de solicitudes 0,0000166667 USD por cada GB/segundo [44]	Capa gratuita por siempre, un millón de solicitudes gratuitas por mes. Hasta 3,2 millones de segundos de tiempo de informática por mes [43].
AWS Redshift	AWS Redshift maneja un	AWS Redshift tiene una

	esquema de precios bajo demanda, entre los cuales se incluye una instancia tipo dc2.large con 2 CPU Virtual, 15 GiB, 0.16 TB SSD de almacenamiento, a un precio de 0,25 USD por hora [45].	versión de prueba gratuita que dura dos meses, en ella se incluyen 750 horas gratis al mes, lo cual permite ejecutar un nodo DC2.Large ininterrumpidamente con 160 GB de almacenamiento SSD comprimido [43].
AWS Simple Storage Service (S3)	S3 Estándar siendo el servicio de almacenamiento de propósito general, tiene un costo de 0,023 USD por GB en los primeros 50 TB por mes [46]	AWS brinda un periodo de 1 año gratuito de 5 GB de almacenamiento estándar 20.000 solicitudes Get y 2.000 solicitudes Put mensuales [43].
Google Cloud Dataflow	Los precios de Cloud Dataflow, varían de acuerdo con el tipo de trabajador: Por lotes, se tiene que: un vCPU (por hora) cuesta 0,056 USD, Memoria (por GB y hora) cuesta 0,003557 USD, Datos procesados (por GB) cuesta 0,011 USD. Por FlexRS, se tiene que: vCPU(por hora) cuesta 0,0336 USD, Memoria (por GB y hora) cuesta 0,0021342 USD, Datos procesados (por GB) cuesta 0,011 USD. En streaming, se tiene que: vCPU(por hora) cuesta 0,069 USD, Memoria (por GB y hora) cuesta 0,003557 USD, Datos procesados (por GB) cuesta 0,018 USD	Cloud Dataflow, no se encuentra dentro de los servicios de capa gratuita [48].

	[47].	
Google Cloud Functions	Al sobrepasar los 2 millones de solicitudes, el precio por un millón de solicitudes es de 0,4 USD [48].	Google Cloud provee una capa gratuita, dentro de la que se encuentra el servicio de Google Cloud Functions con: 2 millones de invocaciones al mes (incluye las invocaciones HTTP y en segundo plano) 400.000 GB por segundo y 200.000 GHz por segundo de tiempo de procesamiento 5 GB de salida de red al mes [48].
Google Cloud BigQuery	Google BigQuery maneja los siguientes precios de almacenamiento. Almacenamiento activo por 0,020 por GB y almacenamiento a largo plazo 0,010 por GB [50]	Google Cloud provee una capa gratuita, dentro de la que se encuentra BigQuery, en ella se brindan 10GB de almacenamiento de manera gratuita por mes [48].

De acuerdo con la tabla 5, se puede concluir que los servicios proporcionados por AWS y GCP manejan en su mayoría unos costos no tan elevados y cobran por el tiempo de uso del servicio, además estas plataformas proporcionan una capa gratuita en la cual se encuentran la mayoría de los servicios que se requieren para la conformación del lago de datos. Debido a que la plataforma de AWS, incluye en su capa gratuita los servicios que se requieren para la implementación del proceso ELT, y a su vez es la que más lagos de datos implementados tiene en su infraestructura, será la seleccionada para el despliegue del servicio.

A continuación, se presentan los procesos de extracción de registros de las diferentes fuentes de datos, a partir del uso de web scraping con el ambiente de trabajo de Selenium, y el consumo de las diferentes APIs que proporcionan algunas de las fuentes caracterizadas.

4.1.2 Extracción de datos del Departamento de Agricultura de los Estados Unidos (USDA)

El departamento de agricultura de los Estados Unidos (USDA), es un departamento ejecutivo del Gobierno Federal de los Estados Unidos, cuyo propósito es desarrollar y ejecutar políticas de ganadería, agricultura, y alimentación. Uno de sus objetivos es entender las necesidades de los productores, incentivando el comercio agrícola y la producción, y protegiendo los recursos naturales. USDA proporciona acceso a diferentes datos y archivos, relacionados con la producción, y comercialización de diferentes productos agrícolas, y ganaderos; dentro de la información de productos agrícolas se encuentran datos relacionados con el aguacate. A continuación, se hablará con respecto al uso de una de las Application Programming Interface (API) que proporciona USDA, la cual es Quickstats, para la obtención de los datos requeridos.

4.1.3 Manejo y uso del API Quickstats

Se requiere solicitar un API KEY, para poder realizar el consumo de los servicios a través de las APIs. De acuerdo con la documentación de la API de Quick Stats de USDA [51], se permite la realización de solicitudes de tipo HTTP GET, en las cuales se pueden definir diferentes parámetros de búsqueda y cabeceras, de acuerdo con la consulta que se requiera realizar. En este caso se realizaron las siguientes consultas, a partir del uso de las herramientas de CURL y POSTMAN.

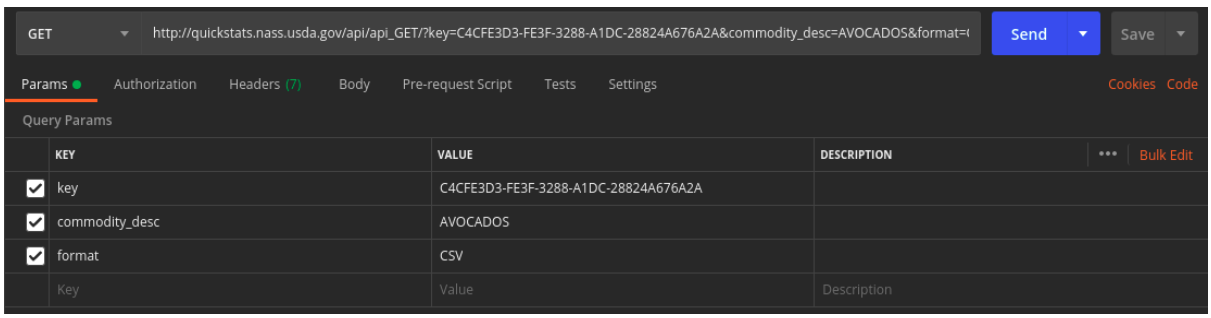

Figura 12. Consulta de datos relacionados con el aguacate en USDA. Fuente Propia.

A continuación, se presenta el equivalente de la solicitud en Postman, con la herramienta CURL; además en el comando se define el archivo en formato CSV, en el cual se guardará la respuesta de la API de Quick Stats de USDA.

“curl -o FINAL_CURL_AVOCADO_USDA.csv --location --request GET 'http://quickstats.nass.usda.gov/api/api_GET/?key=C4CFE3D3-FE3F-3288-A1DC-28824A676A2A&commodity_desc=AVOCADOS&format=CSV' ”

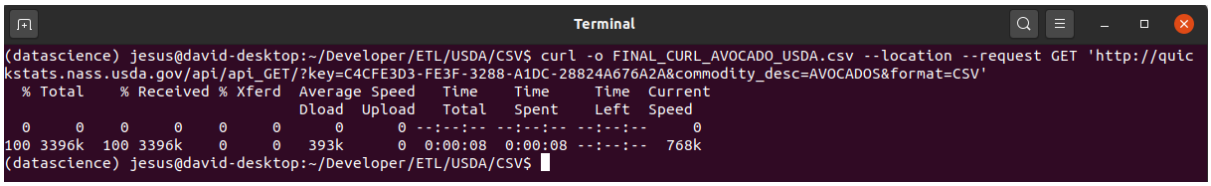

Figura 13. Consulta API Quick Stats USDA con CURL. Fuente Propia.

En las figuras, se definen los siguientes parámetros:

- Key: En este parámetro se debe definir el API Key que se solicitó previamente, con esta llave se pueden realizar las respectivas solicitudes.
- commodity_desc: En este parámetro se define el tema principal de interés, el cual en este caso es el aguacate, en inglés: Avocado.
- format: En este parámetro se debe definir el tipo de formato de salida que se requiere, se puede generar en: XML, JSON, o CSV. En este caso se definió la salida en formato CSV.

En el archivo CSV resultante, se obtuvieron 7814 filas con 39 columnas, en las cuales se encuentra información de producción, comercial, ambiental, y económica relacionada con el mercado de aguacate en los Estados Unidos. Sin embargo, se requirió generar otra consulta para así solo tener datos relacionados con la producción, y la comercialización del aguacate, excluyendo la parte ambiental. La solicitud fue la siguiente, se muestra su ejecución en Postman y CURL:

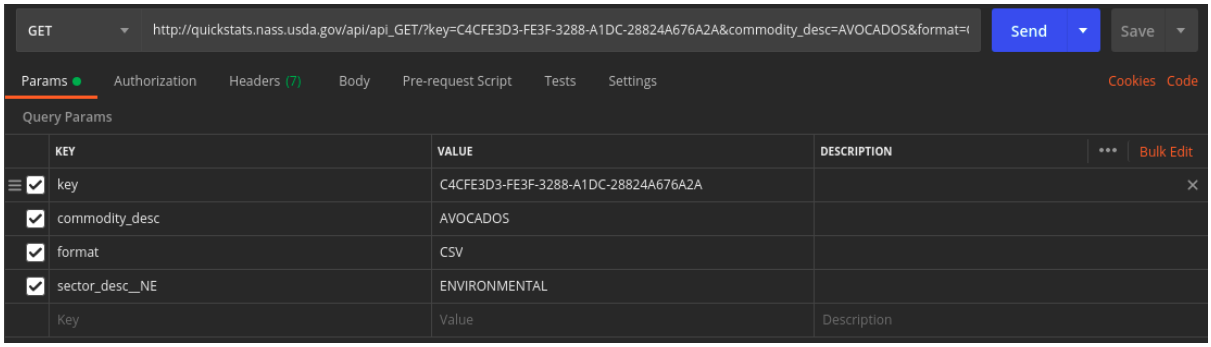

Figura 14. Consulta de datos relacionados con el aguacate en USDA sin datos de la categoría ambiental. Fuente Propia.

“curl -o AVOCADO_NOT_ENVIRONMENTAL_FINAL.csv --location --request GET 'http://quickstats.nass.usda.gov/api/api_GET/?key=C4CFE3D3-FE3F-3288-A1DC-28824A676A2A&commodity_desc=AVOCADOS&format=CSV&sector_desc__NE=ENVIRONMENTAL'”

```
(datascience) jesus@david-desktop:~/Developer/ETL/USDA/CSV$ curl -o AVOCADO_NOT_ENVIRONMENTAL_FINAL.csv  --location --request GET 'ht
tp://quickstats.nass.usda.gov/api/api_GET/?key=C4CFE3D3-FE3F-3288-A1DC-28824A676A2A&commodity_desc=AVOCADOS&format=CSV&sector_desc__N
E=ENVIRONMENTAL'
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100 2047k  100 2047k    0     0   313k      0  0:00:06  0:00:06 --:--:--  572k
(datascience) jesus@david-desktop:~/Developer/ETL/USDA/CSV$
```

Figura 15. Consulta API Quick Stats USDA con CURL sin datos ambientales. Fuente Propia.

En la solicitud HTTP GET realizada se añadió un nuevo parámetro el cual es “sector_desc”, este parámetro define las cinco categorías en las cuales se encuentran los datos en la base de datos de USDA, estas categorías son: “CROPS, ANIMALS & PRODUCTS, ECONOMICS, DEMOGRAPHICS, and ENVIRONMENTAL” [46]. En el archivo anterior se encontró que existían tres tipos de categorías: CROPS, DEMOGRAPHICS, y ENVIRONMENTAL, en este caso solo requerimos excluir la categoría ENVIRONMENTAL debido a que esta categoría se relaciona con la caracterización del uso de productos químicos en los diferentes cultivos, por lo cual se definió el parámetro de la siguiente manera: “sector_desc__NE = ENVIRONMENTAL”, excluyendo de esta manera los datos de categoría ENVIRONMENTAL, a partir del operador excluyente: “__NE” definido en [51].

Dentro de los datos obtenidos relacionados con el aguacate se obtuvieron las siguientes variables de producción y comercialización del aguacate en los Estados Unidos: acres de producción de aguacate, acres de no producción de aguacate,

superficie con y sin frutos de aguacate medido en árboles, cuerdas con y sin carga, operaciones con superficie portante, operaciones con producción, precio recibido, base ajustada, medido en $USD/tonelada, precio recibido, medido en $USD/tonelada, producción, medida en $ USD/ tonelada, producción medida en toneladas, producción medida en $USD, ventas medidas en $USD, ventas medidas en libras, rendimiento medido en toneladas/acre, producción medida en toneladas y $USD del mercado fresco de aguacate, consumo en el hogar uso en la granja medido en toneladas y $USD, producción no vendida medida en toneladas, acres cosechados para aguacates orgánicos, operaciones con ventas de aguacates orgánicos vendidos en mercados convencionales y orgánicos, ventas de aguacates orgánicos medidas en toneladas y $USD.

Dentro de los datos obtenidos, se encontraron registros desde el año 1924 hasta la actualidad, en diferentes cantidades de acuerdo con el año. Con todos los datos obtenidos a partir de la API de Quick Stats del departamento de agricultura de los Estados Unidos, por sus siglas en inglés (USDA), se prosigue a la realización de la actualización de los datos, para ello se realizarán las consultas con CURL para los dos años más recientes, los cuales en este caso son 2020 y 2021, ya que, en el periodo de los dos últimos años, se suelen realizar actualizaciones o subidas de nuevos datos. A continuación, se presenta la solicitud realizada en Postman y CURL:

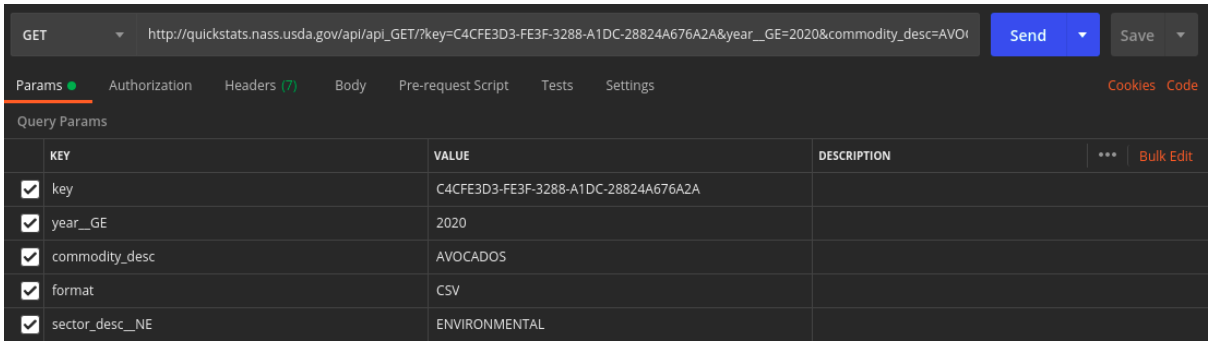

Figura 16. Consulta de datos relacionados con el aguacate en USDA sin datos de la categoría ambiental, últimos dos años. Fuente Propia.

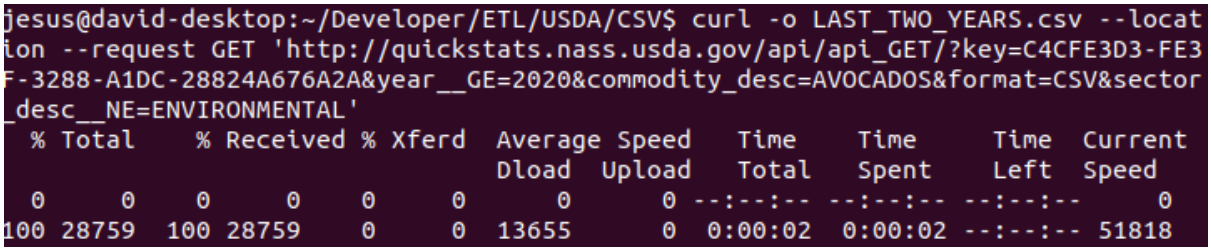

Figura 17. Consulta API Quick Stats USDA con CURL sin datos ambientales de la categoría ambiental, últimos dos años. Fuente Propia.

“curl -o LAST_TWO_YEARS.csv --location --request GET 'http://quickstats.nass.usda.gov/api/api_GET/?key=C4CFE3D3-FE3F-3288-A1DC-28824A676A2A&year__GE=2020&commodity_desc=AVOCADOS&format=CSV&sector_desc__NE=ENVIRONMENTAL'”

En las solicitudes, se definió el parámetro “sector_desc__NE” para excluir los datos relacionados con la categoría: “ENVIRONMENTAL”, ya que estos datos se relacionan con el uso de los productos químicos sobre los diferentes cultivos agrícolas, y no con la cantidad de producción y la comercialización del aguacate en los Estados Unidos. Además, también se incluyó el parámetro: “year__GE” con valor igual a 2020, para así poder incluir los datos cuyos años sean del 2020 y 2021, este parámetro se definió de acuerdo con la documentación [51], se definió este parámetro para la obtención de los datos registrados después del 2020, ya que los datos de años anteriores se obtuvieron mediante el uso de las solicitudes a la API previamente definidas.

4.1.4 Extracción de datos del sistema de información económica, estadística y de mercado (ESMIS)

USDA tiene una plataforma adicional denominada USDA Economics, Statistics and Market Information System (ESMIS) [52], en la cual se puede encontrar información de relevancia con respecto a la agricultura, y sus diferentes mercados, dentro de los cuales se encuentra el mercado del aguacate. En esta plataforma se encuentran archivos relacionados con las importaciones y exportaciones semanales de aguacate en los Estados Unidos. A continuación, se presenta el proceso realizado para la obtención de estos datos.

En principio, en la fuente de datos USDA-ESMIS, solo se brinda una publicación de datos relacionados con el aguacate, esta publicación contiene los datos de importación y exportación de aguacate semanales desde noviembre 27 del 2018, hasta la actualidad. La actualización de estos archivos se realiza normalmente los martes de cada semana, y además las publicaciones se encuentran en formato TXT, la recolección de los datos se realiza con web scraping a partir del uso del ambiente de trabajo de Selenium, y en específico con el uso de Selenium Webdriver con Python, para la descarga de los archivos TXT.

4.1.5 Extracción de datos de la Organización de la Agricultura y la Alimentación (FAO)

La Organización de la Agricultura y la Alimentación por sus siglas en inglés (FAO) es la agencia de las Naciones Unidas para la agricultura y la alimentación que lidera el esfuerzo internacional para poner fin al hambre. Esta organización contribuye como una fuente de conocimiento e información que busca ayudar a los diferentes países en vías de desarrollo para la modernización y mejora de sus actividades agrícolas, forestales y pesqueras. Siendo la FAO una fuente de conocimiento, proporciona acceso a información relacionada con los diferentes procesos de producción y comercialización de diferentes productos agrícolas en diversos países, dentro de los cuales se encuentran Estados Unidos y Colombia.

La FAO proporciona acceso a diferentes fuentes de datos relacionadas con la producción, y comercialización del aguacate en diferentes países. Para la obtención de estos datos se utilizó el ambiente de trabajo de Selenium con Python como lenguaje de programación para la descarga de archivos que proveen información acerca del área cosechada, la producción, y el rendimiento del cultivo de aguacate en Estados Unidos, además también se tiene acceso a la cantidad de importaciones y exportaciones en $USD y en toneladas de aguacate. Con respecto a la información demográfica de los Estados Unidos, se tienen datos relacionados a la estimación de la población, la población total, la población urbana y rural, y la distribución del género en la población [53, 54, 55].

4.1.6 Extracción del departamento de agricultura y alimentos de California (CDFA)

El CDFA es un gabinete a nivel de agencia del gobierno de California, es el responsable de la seguridad de los alimentos del estado, y de promover el desarrollo de la industria de la agricultura de California [56]. El CDFA provee acceso a los reportes agrícolas para el rango de años 2016-2020. En los reportes, se pueden encontrar datos de producción de diferentes cultivos, entre los cuales se encuentran los del aguacate. En estos informes se encuentra información de la producción, el rendimiento de los cultivos, las unidades vendidas y los respectivos precios [57]. Para la obtención de estos reportes se implementó un scraper a partir del uso de Selenium Webdriver con Python.

4.1.7 Extracción de datos del Hass Avocado Board (HAB)

El Hass Avocado Board (HAB) es una organización que busca recaudar y enfocar los recursos con el objetivo de expandir la demanda de aguacate Hass en Estados Unidos, para así convertirla en la fruta preferida en este país. HAB brinda información actualizada sobre la oferta y la demanda de aguacate en el país, a su vez tiene relaciones con diferentes organizaciones de países productores de aguacate como: México, Colombia y Chile, de los cuales se obtienen diferentes variables de producción y de comercialización en torno a las importaciones y exportaciones de esta fruta [58]. Dentro de las variables del mercado del aguacate se encuentra: el volumen y proyección del aguacate en los Estados Unidos, ventas de aguacate en unidades y dólares en Estados Unidos y por estados, volumen de ventas minoristas, y precios medios de venta entregados trimestralmente. Los reportes generados se encuentran actualizados desde el 2018 hasta el día de hoy en el año 2021, además de ello también se definen diferentes proyecciones de producción. Para la descarga de estos reportes, se requiere registrarse en el sitio web, por ello el inicio de sesión y la obtención de los diferentes reportes se realizará a través del uso de Selenium Webdriver con Python.

4.1.8 Extracción de datos de Kaggle

Kaggle es una plataforma subsidiaria de Google, la cual es a su vez una comunidad de científicos de datos y profesionales en las áreas de análisis de datos y aprendizaje automático. Esta plataforma brinda la posibilidad de poder encontrar y publicar conjuntos de datos relacionados con múltiples campos de investigación y de la industria, además de ello Kaggle brinda la opción de crear y participar en competencias de análisis de datos [59]. Siendo Kaggle una plataforma con una gran cantidad de conjuntos de datos, es una de las fuentes de datos que se eligió para la obtención de datos relacionados con el mercado del aguacate, en Kaggle se encontraron diferentes conjuntos de datos que tienen diversas variables del mercado de aguacate en Estados Unidos, como el precio promedio de la unidad de aguacate, número de aguacates vendidos, y el volumen total en diferentes periodos de tiempo. Kaggle tiene una API, la cual busca facilitar la subida y descarga de los diferentes conjuntos de datos, para poder utilizar la API, se requiere generar una cuenta en la plataforma, después de ello se solicita el API Key, necesario para la descarga de los archivos. Con el API Key, se procede a descargar el archivo JSON con las credenciales necesarias de la API, este archivo debe de copiarse en la siguiente ruta: "~/.kaggle", en caso de no existir el directorio oculto, se genera uno nuevo. Estando esta parte realizada, se procede a ejecutar los siguientes comandos para la obtención de los conjuntos de datos relacionados con el aguacate:

"kaggle datasets list -s 'avocado' "

"kaggle datasets download -d <filename>"

El primer comando permite listar los diferentes conjuntos de datos relacionados con el aguacate, y el segundo comando se utiliza para la descarga del respectivo conjunto de datos a partir del nombre del archivo.

4.1.9 Extracción de datos de la oficina del censo de los Estados Unidos

La oficina del censo de los Estados Unidos es la principal agencia del Sistema Estadístico Federal de los Estados Unidos, encargada de la generación de datos relacionados con los habitantes y la economía del país, una de sus tareas principales es la de llevar a cabo el censo de los Estados Unidos cada diez años [60]. La oficina del censo brinda acceso a diferentes archivos relacionados con la importación y exportación de diferentes productos agrícolas entre los cuales se encuentra el aguacate, además de ello también proporciona múltiples datos y estadísticas que buscan caracterizar la población de los Estados Unidos. Dentro de las variables relacionadas con el mercado del aguacate se encuentran: la producción de aguacates utilizada, el valor de la producción, y los estados líderes en producción. Para la obtención de los respectivos archivos Excel, se utilizó Selenium Webdriver con Python como lenguaje de programación.

4.1.10 Resumen del capítulo

En el presente capítulo, se caracterizaron las diferentes posibles tecnologías para la implementación del proceso ELT, además de ello se escogió la herramienta Selenium y la plataforma cloud de AWS para la implementación y despliegue del proceso ELT, de acuerdo con la tabla 3 y 4 realizada con los diferentes criterios y la estimación de costos mostrados previamente. Adicionalmente, se realizó la extracción de datos de las diferentes fuentes a partir del uso de Selenium y el consumo de APIs.

5 IMPLEMENTACIÓN DE LA ARQUITECTURA EN AMAZON WEB SERVICES

En este capítulo se describe el proceso realizado en la fase de diseño de la arquitectura técnica, selección de los productos para la implementación y la construcción del servicio de captura automática de datos junto con el lago de datos de la metodología Kimball. Para ello se describe la implementación del despliegue de los extractores de datos y el lago de datos en Amazon Web Services (AWS), a su vez se muestra el proceso de configuración para la automatización y generación del servicio de extracción y carga de datos en el lago de datos y también se describen los diferentes registros almacenados en el lago de datos, y las respectivas tablas que conforman el catálogo de datos del lago de datos.

5.1.1 Modelo de vistas 4+1 del servicio de captura automática de datos

Para la implementación y despliegue del servicio de captura automática de datos en AWS se utilizó el modelo de vistas 4+1 el cual permite realizar la descripción de la arquitectura de software basado en cinco múltiples vistas concurrentes, lo cual corresponde a la fase diseño de la arquitectura técnica de la metodología Kimball. Dentro de las cinco vistas se encuentran: la vista lógica, la cual representa la funcionalidad que el sistema genera a los usuarios finales, la vista del proceso la cual representa la concurrencia y sincronización de la arquitectura; explica los aspectos dinámicos del sistema tales como los procesos de comunicación, integración y distribución de los diferentes procesos y tareas, la vista de desarrollo o implementación representa el sistema desde la perspectiva de un programador y se relaciona con la gestión del software, la vista física muestra el sistema desde un punto de vista de ingeniería, describe los diferentes componentes de la capa física del sistema y cómo se relacionan entre sí. La descripción de cómo funcionan las cuatro vistas en conjunto se realiza mediante la definición de los escenarios que se encuentran representados con los diagramas de casos de uso [61].

5.1.2 Vista lógica

De acuerdo con lo mencionado anteriormente, se presentará la arquitectura del sistema, a partir del uso del modelo de vistas de 4+1. A continuación se presenta la vista lógica, la cual para este sistema se representará mediante dos diagramas de clases:

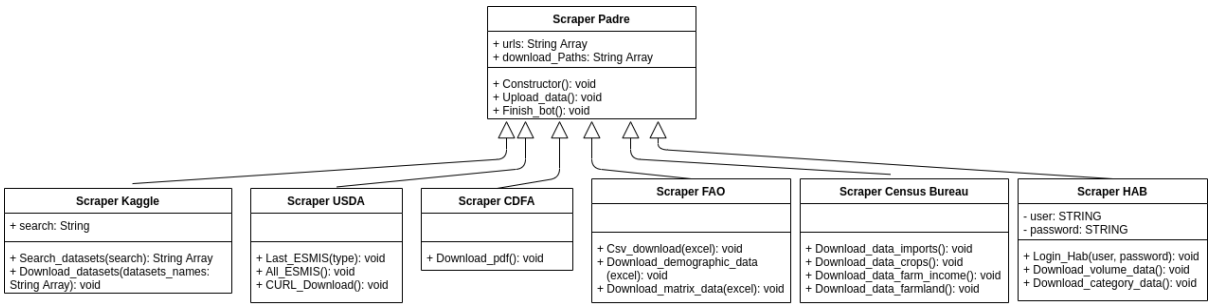

Figura 18. Diagrama de clases de los extractores de datos.

En el diagrama de clases de los extractores de datos, se presenta un Scraper Padre el cual posee dos atributos, los cuales son: “urls”, el cual corresponde a las direcciones de las páginas web de las cuales se quiere obtener los datos, luego se encuentra el atributo “download_paths” el cual contiene las rutas para el almacenamiento de los diferentes archivos que se capturen de las fuentes de datos. Dentro de los métodos, se encuentra el constructor, en el cual se incluye la declaración del controlador headless, el cual será el encargado de automatizar las acciones sobre el navegador sin interfaz gráfica. Luego, se define un método “Upload_data” el cual se encarga de subir los datos obtenidos al lago de datos, y por último se tiene el método “Finish_bot” el cual finaliza el scraper.

En el diagrama, se tienen seis clases las cuales heredan de la clase Scraper Padre. Estas clases corresponden a los Scrapers que obtienen datos, de cada una de las seis fuentes de datos principales del mercado del aguacate de los Estados Unidos caracterizadas anteriormente. La clase “Scraper Kaggle”, tiene definido un atributo “search”, el cual corresponde al parámetro de búsqueda dentro de los conjuntos de datos existentes en kaggle. Luego, se tienen los métodos dentro de los cuales está el

de “Search Datasets”, el cual se encarga de buscar los conjuntos de datos que se relacionen con el parámetro de búsqueda. Adicionalmente se tiene el método de “Download_datasets”, para la descarga de todos los conjuntos de datos encontrados a partir del método “search_datasets”, el cual retorna un arreglo que contiene los nombres de los conjuntos de datos relacionados con el tema. En la clase “Scraper USDA” no se tienen atributos diferentes a los que se heredan de la clase padre, con respecto a los métodos se tienen tres. “Last_ESMIS”, accede a una página de ESMIS de USDA, con el objetivo de poder descargar el archivo más reciente de importaciones y exportaciones de aguacate en USDA, “All_ESMIS” obtiene todos los archivos históricos correspondientes a las importaciones y exportaciones de aguacate en la página de ESMIS, por último “CURL_download” hace referencia al método encargado de descargar la información histórica relacionada con el aguacate a traves de la API de quickstats.

En la clase “Scraper CDFA” solo se tiene un método definido, el cual es el encargado de realizar la descarga de los archivos en formato PDF, de la página de CDFA, los cuales contienen información relacionada con el sector agrícola, y la producción de aguacates en California. La clase “Scraper FAO” posee tres métodos, los cuales se encargan de obtener los archivos en formato CSV relacionados con la producción, y comercialización del aguacate, y además también se capturan archivos que proporcionan datos demográficos históricos de los Estados Unidos provenientes de la página oficial de la FAO. En la clase “Scraper Census Bureau”, se tienen cuatro métodos, los cuales obtienen datos relacionados con el mercado del aguacate de los Estados Unidos de la página oficial de la oficina del censo de los Estados Unidos. Finalmente, se tiene la clase “Scraper HAB”, la cual tiene dos atributos, los cuales corresponden a las credenciales necesarias para el inicio de sesión en la página oficial del Hass Avocado Board, estas credenciales son las utilizadas por el método “login_hab”, el cual inicia la sesión, para así ya poder tener acceso a los datos de producción y de comercialización proporcionados por esta fuente, los cuales se obtienen a partir de los otros dos métodos de esta clase.

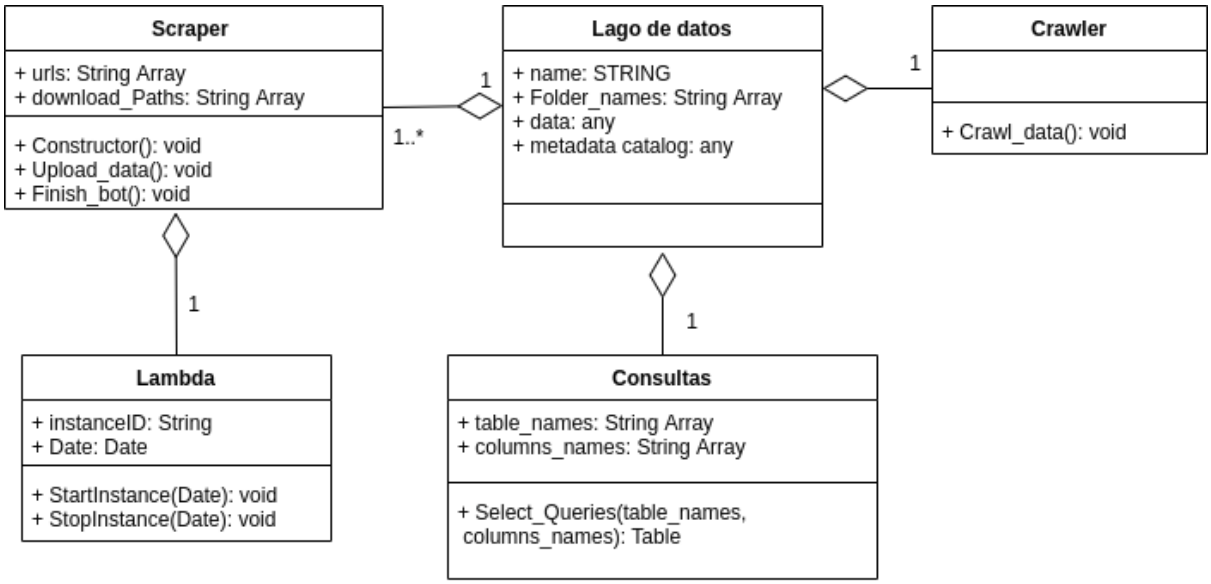

Figura 19. Vista lógica, diagrama de clases. Fuente propia.

En este segundo diagrama de clases, correspondiente a la vista lógica del modelo de vistas 4+1, se muestra de una manera más general, las diferentes clases que conforman el sistema de captura automática de datos y el lago de datos. En el diagrama, se puede identificar, la clase “Lambda”, la cual es la encargada de gestionar la activación del Scraper en unos horarios específicos de manera semanal, por ello se definen dos atributos, los cuales son el identificador de la instancia “instanceID”, donde se va a ejecutar el Scraper, y la fecha y hora en la cual se activará la respectiva instancia, a su vez se tienen los dos métodos de iniciar y de parar el Scraper. La clase “Scraper” se ha descrito previamente en el anterior diagrama en el cual se mostraba el scraper padre, y los diferentes tipos de Scraper, a su vez existe una relación de agregación de multiplicidad uno entre la clase Scraper, y la clase Lambda. Después, se tiene la clase “Lago de Datos”, la cual está compuesta por los atributos de: “name”, “Folder_names”, “data”, y “metadata catalog”. El atributo data, hace referencia a todos los datos en bruto almacenados dentro del lago de datos, a partir de estos datos se genera un catálogo de metadatos a partir del método “Crawl_data” de la clase “Crawler”, el cual se ejecuta cada vez que haya una actualización dentro del atributo de datos del lago de datos. Por último se tiene la clase “Consultas”, la cual posee los atributos “table_names” y “column_names”, los cuales serán necesarios para el método de “Select_queries” el cual hará consultas en el catálogo de metadatos y en

los datos que componen el lago de datos; para ello se le pasa el nombre de la base de datos, de las tablas sobre las que se quiere consultar y los nombres de las respectivas columnas.

5.1.3 Vista de proceso

Para la representación de la vista de proceso se muestra el siguiente diagrama de secuencia:

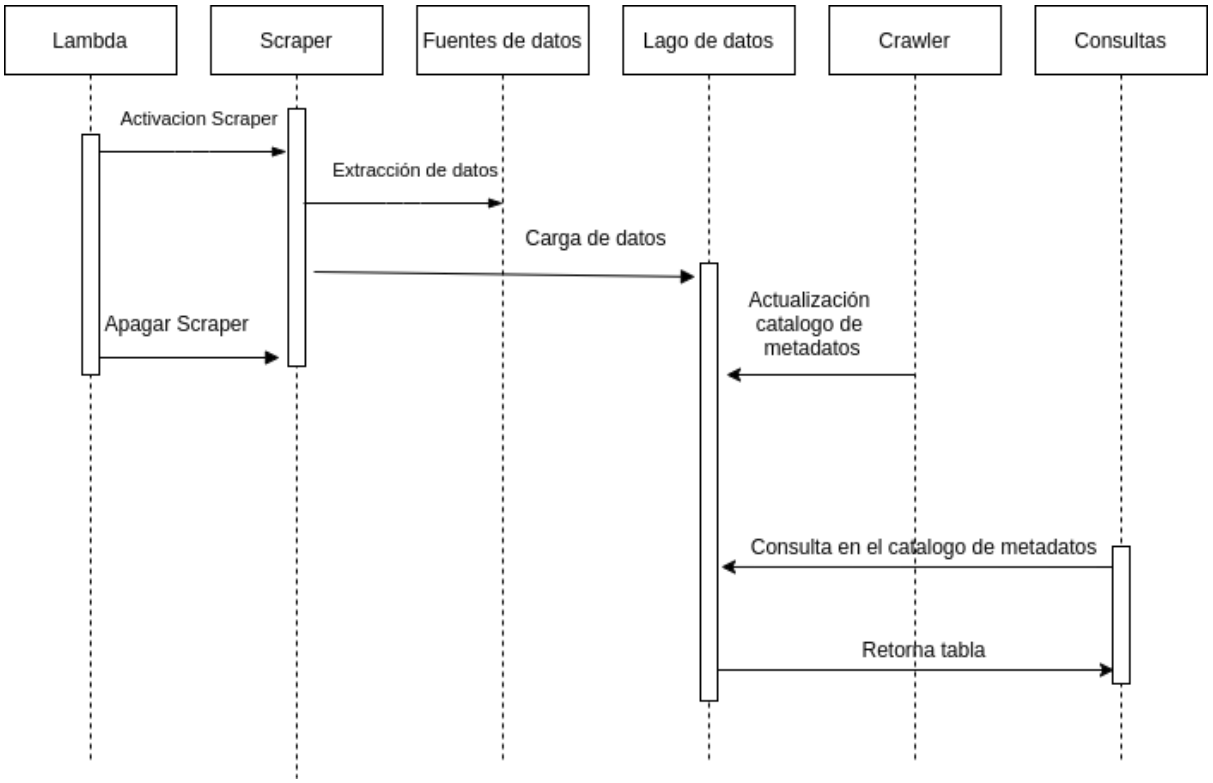

Figura 20. Vista de proceso, diagrama de secuencia. Fuente propia.

En la figura 20 se presenta el diagrama de secuencia del servicio de captura automática de datos. En principio, el flujo del diagrama será el siguiente, comienza con la línea de vida que representa a “Lambda”, en ella se genera la activación del Scraper, después de ello el Scraper realiza la respectiva extracción de datos de las diferentes fuentes, y ya estos datos recopilados se cargan al lago de datos, realizada la actualización de los datos en el repositorio, el Crawler realiza la correspondiente actualización del catálogo de metadatos que describe los datos contenidos en el lago de datos. Realizado esto, se procede a generar una consulta en el catálogo de metadatos, sobre los datos actualizados; la respuesta a esta solicitud es una tabla con los respectivos datos solicitados, y ahí ya finaliza la ejecución del servicio de captura automática de datos.

5.1.4 Vista de desarrollo

Para la representación de la vista de desarrollo se muestra el siguiente diagrama de componentes:

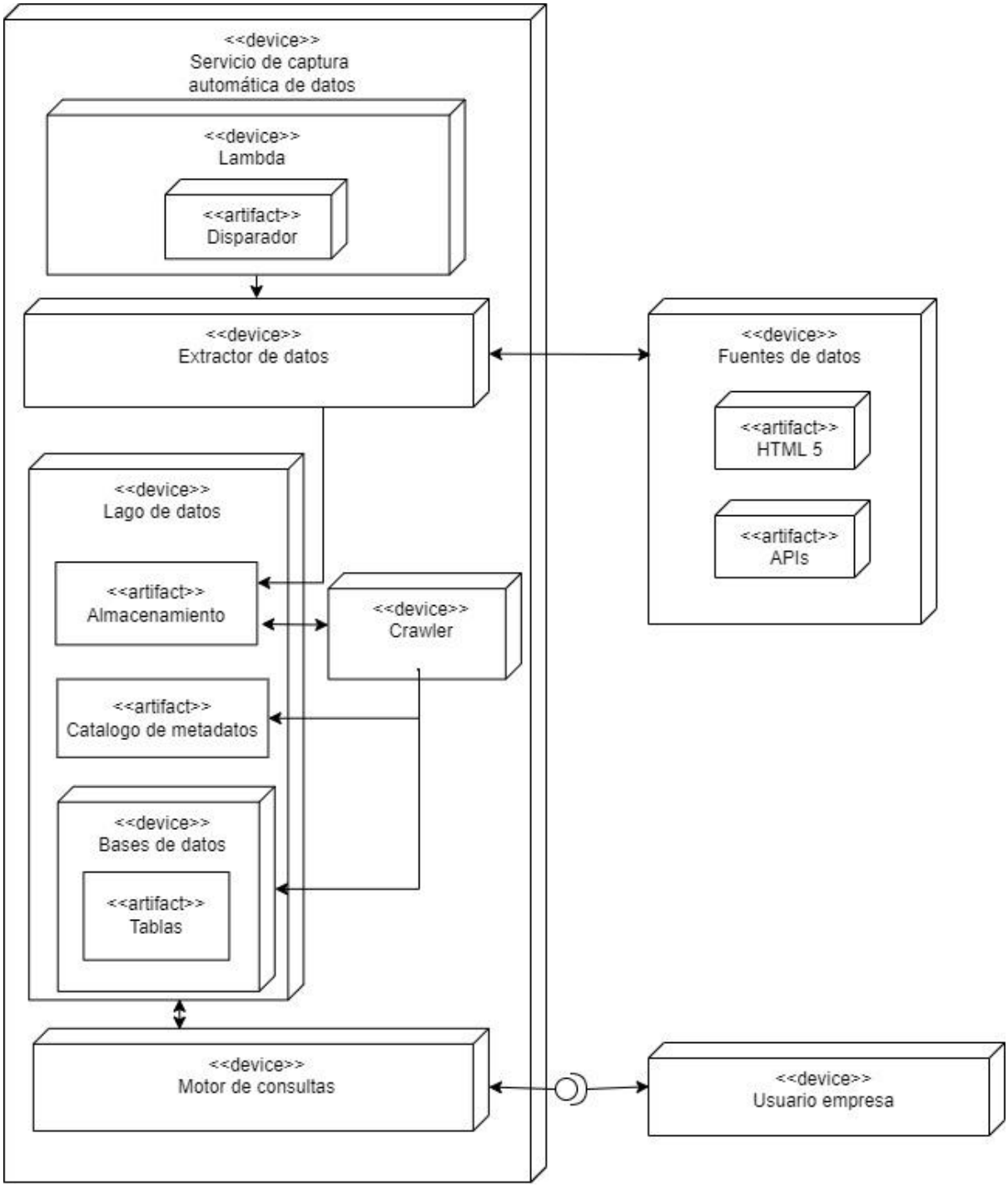

Figura 21. Vista de desarrollo, diagrama de componentes. Fuente propia.

En la figura 21, se presenta el diagrama de componentes, el cual se puede dividir en tres partes principales: por un lado, se tienen las diferentes páginas web de las fuentes de datos; en la segunda parte se tienen todos los componentes y servicios que se encuentran dentro del servicio de captura automática de datos, y en la tercera parte se puede identificar el usuario de la empresa. En la primera parte se tienen las diferentes páginas web las cuales están compuestas por un código HTML, la

respectiva extracción de los datos se realiza a partir del análisis de la estructura y la manipulación de las diferentes etiquetas que componen la página. Para hacer las respectivas solicitudes a estas páginas desde el Scraper se utiliza el protocolo HyperText Transfer Protocol (HTTP).

En el segundo componente, se tienen los diferentes servicios que se encuentran alojados en el servicio de captura automática de datos. En la parte superior de este componente se encuentra el componente “lambda” el cual se encarga de gestionar el encendido y apagado de los extractores de datos, a partir del disparador que se activa de manera semanal, luego está el “Extractor de datos”, en él se ejecutan los diferentes scrapers encargados de obtener los datos de las diversas fuentes que son representadas por el componente “fuente de datos”. Después de la obtención de los datos, se realiza la respectiva carga de estos al lago de datos, en especifico el componente de almacenamiento, el cual es el repositorio del lago de datos.. Con los datos actualizados y cargados en el lago de datos el componente “Crawler” se dirige al “almacenamiento” para identificar nuevos archivos, después de eso el “Crawler” prosigue a realizar la actualización del catálogo de metadatos basado en los nuevos archivos cargados. Esta actualización también se realiza en la base de datos del lago de datos y sus respectivas tablas por parte del Crawler, después de ello ya el device de “Motor de consultas” realiza la consulta sobre la base de datos y el catálogo de metadatos a través del lenguaje SQL, a partir de la solicitud que genera el componente “Usuario empresa” mediante una interfaz gráfica sobre el motor de consultas” y se obtienen los resultados de la consulta ejecutada.

5.1.5 Vista física

Para la representación de la vista de despliegue se muestra el siguiente diagrama de despliegue:

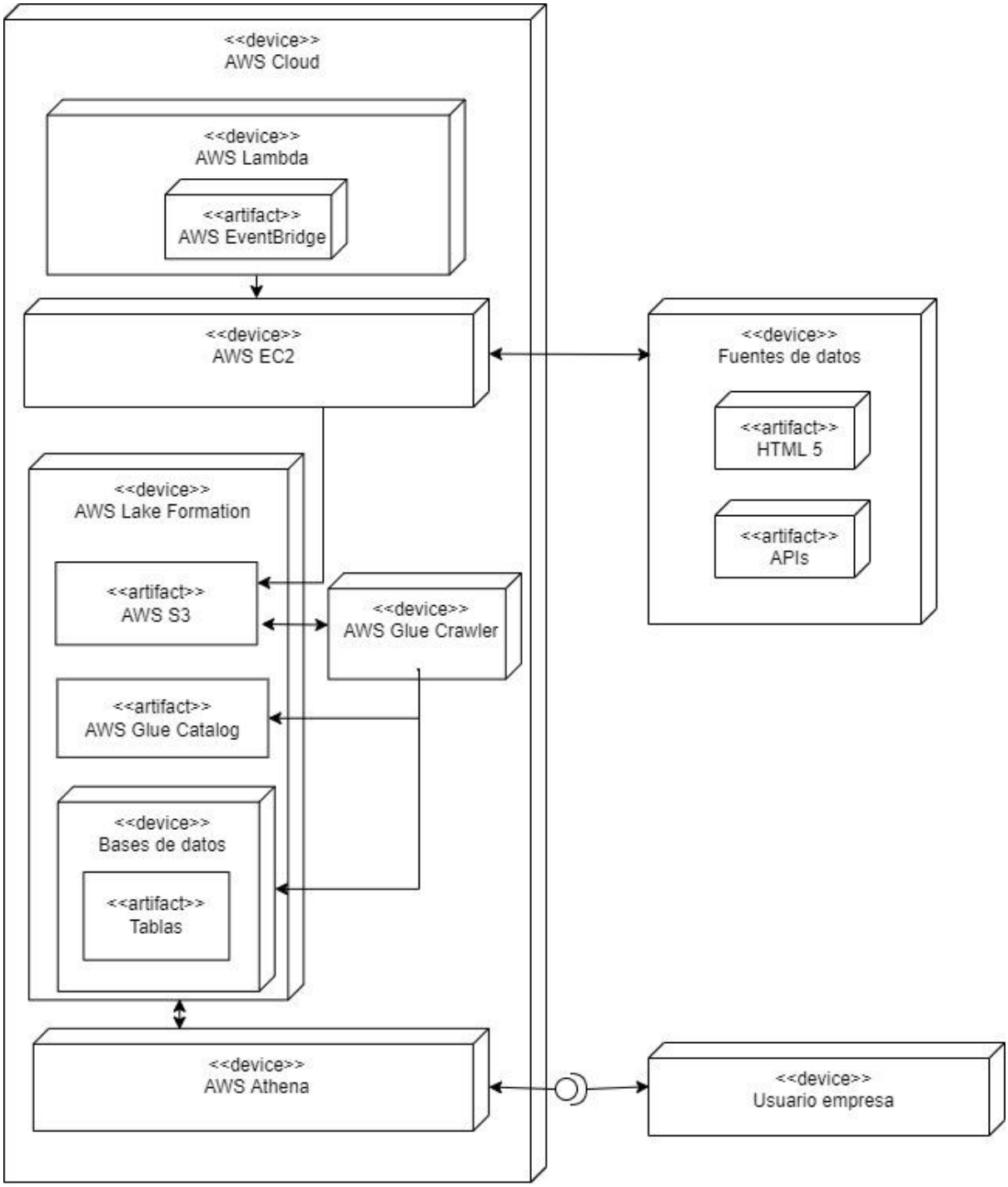

Figura 22. Vista física, diagrama de despliegue. Fuente propia.

En el diagrama de despliegue se define prácticamente toda la arquitectura sobre los servicios de cloud de AWS. La arquitectura, funciona de la siguiente manera: en principio AWS EventBridge es un servicio de AWS que permite generar eventos para activar diferentes servicios, entre los cuales se encuentran las funciones lambda; en este caso AWS EventBridge se utiliza para generar una alarma para ejecutar la función lambda los días Jueves a las nueve de la mañana; esta función lambda se

encarga de iniciar una instancia en AWS Elastic Cloud Computing (EC2), la cual tiene configurada la ejecución de un script en bash a las nueve y veinte del Jueves de cada semana; este script se encarga de obtener los diferentes archivos periódicos de las diferentes fuentes de datos, para luego subirlos a un bucket que se encuentra alojado en AWS Simple Storage Service (S3 por sus siglas en inglés).

En el bucket que se encuentra en S3, se tienen almacenados todos los datos recopilados de las diferentes fuentes, este bucket será el servicio que almacenará los datos del lago de datos. A partir de los datos que se encuentran en el bucket se genera el catálogo de metadatos de AWS Glue; este catálogo se produce por parte de un Crawler que será implementado sobre el servicio de AWS Glue, adicional a ello el Crawler se encarga de actualizar la base de datos y sus respectivas tablas a partir de los nuevos registros capturados. Generado el catálogo de metadatos, se realiza una consulta SQL a través del servicio de AWS Athena al catálogo por parte del Usuario empresa, para poder así visualizar los diferentes resultados de las consultas sobre los datos generados. También cabe mencionar el servicio de Data Lake Formation, el cual es el encargado de la gestión del lago de datos, en él se define la ubicación del almacenamiento de los datos, las tablas y los permisos de acceso a los registros alojados en el lago datos.

Descritas las cuatro vistas del modelo de vista 4 +1, queda restante la descripción de los escenarios a partir de un conjunto de diagramas de caso de uso del servicio. Sin embargo, en este caso, cabe mencionar que no es necesaria la definición de la parte de escenarios, ya que este servicio no está orientado hacia usuarios finales, sino que su propósito es en sí la recopilación y persistencia automática de registros en el lago de datos, y no requiere de la definición de una situación diferente a esta.

5.1.6 Despliegue del servicio de captura automática de datos en AWS

En esta sección del capítulo, se hablará con respecto al despliegue del proceso ELT en la plataforma cloud de Amazon Web Services (AWS). Como se mencionó en el capítulo anterior, se utilizarán algunos servicios de AWS para la conformación del proceso ELT y el lago de datos, dentro de los servicios de Amazon que se utilizaron

en el desarrollo del sistema, se encuentran: AWS Lambda, AWS Simple Storage Service (S3), AWS Glue, AWS Data Lake Formation, AWS Athena, AWS CloudWatch Logs, y AWS IAM; posteriormente se hablará más detalladamente con respecto a la configuración de cada uno de estos servicios, y la función que cumplen en la arquitectura del sistema.

Con los diferentes conjuntos de datos obtenidos a partir de Selenium Webdriver y Python, se prosigue a la conformación del lago de datos, el cual cumple con la función de repositorio de datos en la arquitectura. Para ello, se requiere generar un Bucket en el servicio de AWS S3; este Bucket denominado “avocado-datalake” será el repositorio donde se almacenarán los datos obtenidos de las diferentes fuentes. En este bucket se definen dos carpetas: “DATA”, y “SCRIPTS”. En “DATA” se tendrán todos los datos obtenidos a partir del algoritmo de captura de datos, y en “SCRIPTS” se almacenan las diferentes consultas y pruebas que se realicen sobre el lago de datos.

Definido el bucket, se procede a la creación del usuario que tendrá acceso al lago de datos, para realizar las diferentes consultas, este usuario se define a través del servicio de AWS Identity and Access Management (IAM por sus siglas en inglés) con el nombre de “avocadouser”, para este usuario, se definen los siguientes permisos a través de la siguientes políticas: AmazonS3FullAccess, AmazonAthenaFullAccess, CloudWatchLogsReadOnlyAccess, AWSGlueConsoleFullAccess, y AWSCloudFormationReadOnlyAccess; con los permisos definidos ya se crea el usuario. Después de ello, se definirá el rol en IAM el cual será usado posteriormente por el servicio de AWS Glue para catalogar los datos del lago de datos que se encuentran almacenados en el bucket “avocado-datalake”. Realizada la configuración del Rol y del S3 Bucket, se procede a la configuración del lago de datos en AWS a partir del servicio de AWS Lake Formation. Al principio se requiere la configuración de un administrador del lago de datos a través del uso de IAM, después de ello se crea una nueva base de datos en el servicio de AWS Lake Formation denominada “avocadodb”, y se define la ubicación de los datos, la cual en esta implementación hace referencia a la carpeta “data” que se encuentra en el bucket de “avocado-datalake”. Realizada esta parte se configura la localización del bucket generado en S3 como el almacenamiento del lago de datos; para ello se añadió un nuevo registro

en la parte de ubicaciones de lago de datos, y se definió la ruta del bucket “avocado-datalake” dentro de la carpeta “DATA”.

Definida la base de datos y la ubicación del almacenamiento para el lago de datos, se procede a la configuración del crawler a partir del uso de AWS Glue. Uno de los principios fundamentales para la generación del lago de datos es el de catalogar todos los datos que se encuentran en el lago de datos; para la generación automática de este catálogo de datos se configura el Crawler. Para ello se accede a la base de datos de “avocadodb” y se accede al permiso de grant ubicado en el menú desplegable de acciones de la base de datos; esto se realiza para definir que el rol “avocadocrawlerrole” puede crear y cambiar las tablas que se encuentren en la base de datos. Después se crea el Crawler en la consola de AWS Glue, se define el nombre, el tipo de fuente que en este caso es almacenes de datos, se define la ubicación del almacén de datos en el bucket de “avocado-datalake”, se especifica que el rol será el de “avocadocrawlerrole”, que se ejecutará en demanda, y se selecciona la base de datos “avocadodb”, después de ello se finaliza la creación del crawler, y se ejecuta. Después de ello se podrá observar que se generaron diferentes tablas por parte del Crawler en la base de datos del lago de datos.

Terminada esta parte, se definen los permisos de acceso del usuario “avocadouser” para la realización de las diferentes consultas en las tablas generadas por el Crawler. Para ello se accede a las diferentes tablas y se configuran los permisos a partir de la acción Grant; dentro de esta nueva ventana se selecciona el usuario “avocadouser” y en los permisos de la tabla solo se selecciona la opción “Select”, y se guarda. Finalizada esta parte, se procede a probar el acceso a las diferentes tablas, a partir de la ejecución de una consulta SQL, a partir del servicio de Athena; para ello se requiere acceder como el usuario “avocadouser” en la consola de AWS; luego se accede al servicio de Athena y se selecciona “avocadodb” como la base de datos sobre la cual se ejecutarán las consultas, en el menú lateral se podrán observar las diferentes tablas que se encuentran dentro de la base de datos, del lago de datos. Antes de ejecutar la consulta, se requiere definir la ubicación de almacenamiento de los resultados obtenidos a partir de la consulta, la cual en este caso será el bucket “avocado-datalake” en la carpeta “SCRIPTS”, el cual se configura en la sección de configuración de Amazon Athena.

Definida la parte de cómo se realizó la implementación del lago de datos, se procede a caracterizar los diferentes archivos y registros recopilados por parte del servicio de captura automática de datos. Como se había mencionado anteriormente, el repositorio del lago de datos se encuentra en el bucket de S3 denominado “avocado-datalake”, y en él se tienen dos carpetas: “DATA” la cual almacena todos los datos recopilados y las tablas generadas por el crawler, y la carpeta “SCRIPTS” donde se almacenan los resultados de las consultas realizadas por el servicio de Amazon Athena. Dentro de la carpeta de “DATA” se tienen las siguientes subcarpetas: “CDFA”, “CENSUS_BUREAU”, “FAO”, “HAB”, “KAGGLE”, “LAST_FILES”, y “USDA”, que contienen los datos previamente recopilados. Dentro de la carpeta “LAST_FILES” se cargan los archivos que se extraen semanalmente de manera automática. Dicho esto, se prosigue a realizar la respectiva descripción de los archivos almacenados en el bucket de S3.

Tabla 6. Descripción de los registros almacenados de la CDFA en el lago de datos. Fuente propia

Carpeta fuente: subcarpeta	Nombre archivo	Tamaño
CDFA: PDF	2016-17AgReport.pdf	27.6 MB
	2016Report.pdf	26.9 MB
	2017-18AgReport.pdf	23.1 MB
	2020_Ag_Stats_Review.pdf	23.1 MB
	2018-2019AgReportnass.pdf	22.0 MB
	2020_Organics_Publication.pdf	6.7 MB
	2017-18AgExports.pdf	6.6 MB
	AgExports2015-2016.pdf	5.4 MB
	2020_Exports_Publication.pdf	5.4 MB
	AgExports2018-2019.pdf	5.3 MB

	2017AgExports.pdf	4.7 MB
	2018-19CaliforniaAgriculturalOrganicReport.pdf	3.4 MB

Tabla 7. Descripción de los registros almacenados de la Oficina del Censo de los Estados Unidos en el lago de datos. Fuente propia

Carpeta fuente: subcarpeta	**Nombre archivo**	**Tamaño**
CENSUS BUREAU: CROPS	12s0858.xls	54.5 KB
	12s0863.xls	45.5 KB
	12s0864.xls	41.5 KB
	12s0862.xls	40.5 KB
	12s0868.xls	33.0 KB
	12s0857.xls	31.0 KB
CENSUS BUREAU: EXPORTS-IMPORTS	12s0848.xls	43.5 KB
	12s0851.xls	42.0 KB
	12s0856.xls 12s0855.xls	40.5 KB 37.5 KB
	12s0850.xls	37.0 KB
	12s0853.xls	34.5 KB

	12s0852.xls	32.5 KB
	12s0849.xls	31.5 KB
	12s0854.xls	28.5 KB
CENSUS BUREAU: FARMLAND	12s0825.xls	130.5 KB
	12s0838.xls	42.5 KB
	12s0823.xls	37.0 KB
	12s0828.xls	34.5 KB
	12s0831.xls	33.5 KB
	12s0833.xls	33.5 KB
	12s0868.xls	33.0 KB
	12s0829.xls	32.0 KB
	12s0826.xls	31.5 KB
	12s0827.xls	31.0 KB
	12s0835.xls	30.5 KB
	12s0837.xls	30.0 KB
	12s0836.xls	29.0 KB
	12s0824.xls	28.0 KB
	12s0832.xls	27.5 KB
CENSUS BUREAU:	12s0842.xls	113.5 KB
	12s0845.xls	112.5 KB
	12s0841.xls	69.5 KB

FARM_INCOME	12s0840.xls	69.0 KB
	12s0843.xls	67.5 KB
	12s0844.xls	67.0 KB
	12s0839.xls	47.0 KB
	12s0830.xls	37.5 KB
	12s0846.xls	37.5 KB

Tabla 8. Descripción de los registros almacenados de la FAO en el lago de datos. Fuente propia.

Carpeta fuente: subcarpeta	**Nombre archivo**	**Tamaño**
FAO: CSV	FAOSTAT_data_5-24-2021 (1).csv	76.3 KB
	FAOSTAT_data_5-24-2021 (2).csv	24.8 KB
	FAOSTAT_data_5-24-2021 (3).csv	25.0 KB
FAO: EXCEL	FAOSTAT_data_5-24-2021.xls	179.0 KB
	FAOSTAT_data_5-24-2021 (1).xls	66.5 KB
	FAOSTAT_data_5-24-2021 (2).xls	64.0 KB

Tabla 9. Descripción de los registros almacenados del HAB en el lago de datos. Fuente propia.

Carpeta fuente: subcarpeta	**Nombre archivo**	**Tamaño**
	2018-plu-total-hab-data.csv	680.0 KB
	2019-plu-total-hab-data.csv	728.7 KB

HAB: CATEGORY_DATA	2020-plu-total-hab-data.csv	674.4KB
	2021-plu-total-hab-data.csv	253.6 KB
	2018-market-region-hab-data.csv	116.6 KB
	2019-market-region-hab-data.csv	116.7 KB
	2020-market-region-hab-data.csv	116.8 KB
	2021-market-region-hab-data.csv	39.4 KB
HAB: VOLUME_DATA	Volume Data Projections - Hass Avocado Board.csv	4.5 KB
	Volume Data Projections - Hass Avocado Board (1).csv	4.6 KB
	Volume Data Projections - Hass Avocado Board (2).csv	4.3 KB
	Volume Data Projections - Hass Avocado Board (3).csv	226.0 B

Tabla 10. Descripción de los registros almacenados de Kaggle en el lago de datos. Fuente propia.

Carpeta fuente	Nombre archivo	Tamaño
KAGGLE	Augmented_avocado.csv	226.5 MB
	avocado-updated-2020.csv	3.4 MB
	Avocado.csv	2.7 MB
	avocado.csv	1.9 MB
	avocado_data.csv	1.8 MB
	2019-plu-total-hab-data.csv	608.8 KB
	Avocado2016-2019.csv	4.4 KB

	List_of_countries_by_avocado_production.csv	2.0 KB

Tabla 11. Descripción de los registros almacenados de USDA en el lago de datos. Fuente propia.

Carpeta fuente: subcarpeta	**Nombre archivo**	**Tamaño**
USDA: CSV	FINAL_AVOCADO_USDA.csv	3.3 MB
	FINAL_CURL_AVOCADO_USDA.csv	3.3 MB
	AVOCADO_SINCE_2007.csv	2.6 MB
	AVOCADO_NOT_ENVIRONMENTAL_FINAL.csv	2.0 MB
	AVOCADO_BEFORE_2007.csv	686.6 KB
	LAST_TWO_YEARS.csv	28.1 KB
	AVOCADO_2020.csv	26.9 KB
	AVOCADO_2021_CURL.csv	1.7 KB
USDA: ESMIS_TXT (139 Archivos)	WA_FV404.TXT	4.9 KB
	WA_FV404 (1).TXT	4.9 KB
	WA_FV404 (2).TXT	4.9 KB
	WA_FV404 (3).TXT	4.9 KB
	WA_FV404 (4).TXT	4.9 KB
	WA_FV404 (5).TXT	4.9 KB
	WA_FV404 (6).TXT	4.9 KB
	WA_FV404 (7).TXT	4.9 KB
	WA_FV404 (8).TXT	4.9 KB

	WA_FV404 (9).TXT	4.9 KB
	WA_FV404 (10).TXT	4.9 KB
	WA_FV404 (11).TXT	4.9 KB
	WA_FV404 (12).TXT	4.9 KB
	WA_FV404 (13).TXT	4.9 KB
	WA_FV404 (14).TXT	4.9 KB
	WA_FV404 (15).TXT	4.9 KB
	WA_FV404 (135).TXT	4.9 KB
	WA_FV404 (136).TXT	4.9 KB
	WA_FV404 (137).TXT	4.9 KB
	WA_FV404 (138).TXT	4.9 KB

En la siguiente tabla se describen los archivos almacenados en la carpeta LAST_FILES donde se recopilan los archivos obtenidos de manera semanal y automatizada:

Tabla 12. Descripción de los registros capturados y almacenados de manera automática en el lago de datos. Fuente propia.

Nombre Archivo	Tamaño
AVOCADO_NOT_ENVIRONMENTAL-01-07-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-08-07-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-09-06-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-10-06-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-15-07-	2.0 MB

21.csv	
AVOCADO_NOT_ENVIRONMENTAL-17-06-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-22-06-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-22-07-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-23-06-21.csv	2.0 MB
AVOCADO_NOT_ENVIRONMENTAL-24-06-21.csv	2.0 MB
Volume Data Projections - Hass Avocado Board-01-07-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-08-07-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-10-06-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-15-07-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-17-06-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-22-06-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-22-07-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-23-06-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-24-06-21.csv	4.4 KB
Volume Data Projections - Hass Avocado Board-01-07-21.csv	4.4 KB
WA_FV404 (1).TXT	4.4 KB
WA_FV404 (1).TXT	4.4 KB
WA_FV404-01-07-21.TXT	4.4 KB
WA_FV404-08-07-21.TXT	4.4 KB
WA_FV404-10-06-21.TXT	4.4 KB
WA_FV404-15-07-21.TXT	4.4 KB

WA_FV404-17-06-21.TXT	4.4 KB
WA_FV404-22-07-21.TXT	4.4 KB
WA_FV404-22-06-21.TXT	4.4 KB
WA_FV404-23-07-21.TXT	4.4 KB
WA_FV404-24-06-21.TXT	4.4 KB

En la siguiente imagen se presentan las diferentes subcarpetas que existen dentro de la carpeta “DATA” del bucket “avocado-datalake”:

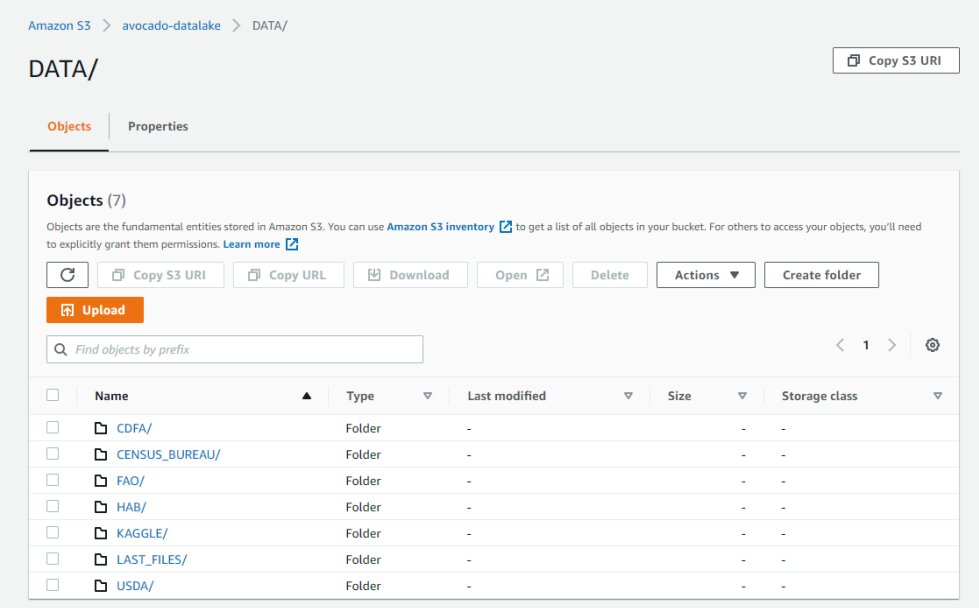

Figura 23. Repositorio del lago de datos en el bucket de S3. Fuente propia.

Además de ello, se generaron las siguientes tablas a partir del crawler de AWS Glue, las cuales conforman el catálogo de datos del lago de datos. A continuación, se muestra el listado de tablas:

Tabla 13. Caracterización de las tablas que conforman el lago de datos. Fuente propia

Nombre tabla	Columnas y tipos	Descripción tabla
2020_plu_total_hab_data_csv	1. geography:	En esta tabla se pueden encontrar los

	string 2. timeframe: string 3. current year week ending: string 4. type: string 5. asp current year: double 6. total bulk and bags units: double 7. 4046 units: double 8. 4225 units: double 9. 4770 units: double 10. totalbagged units: double 11. smlbagged units: double 12. lrgbagged units: double 13. x-lrgbagged units: double 14. bulk gtin: string	datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2020.
avocado2016_2019_csv	1. date: date 2. averageprice: double 3. total volume: double 4. plu-4046: double 5. plu-4225: double	Esta tabla se generó a partir de uno de los datasets de Kaggle relacionados con el consumo y la producción de aguacate en los Estados Unidos.

	6. plu-4770: double 7. total bags: double 8. small bags: double 9. large bags: double 10. xlarge bags: double	
avocado_data_csv	1. date: date 2. type: string 3. region: string 4. year: bigint 5. total volume: double 6. plu4046: double 7. plu4225: double 8. plu4770: double 9. total bags: double 10. small bags: double 11. large bags: double 12. xlarge bags: double 13. averageprice: double	Esta tabla se generó a partir de uno de los datasets de Kaggle relacionados con el consumo y la producción de aguacate en los Estados Unidos. En este dataset se tiene un mayor número de columnas, entre las cuales se encuentran: type, region, year.
2021_plu_total_hab_data_csv	1. geography: string 2. timeframe: string 3. current year week	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2021.

	ending: string 4. type: string 5. asp current year: double 6. total bulk and bags units: double 7. 4046 units: double 8. 4225 units: double 9. 4770 units: double 10. totalbagged units: double 11. smlbagged units: double 12. lrgbagged units: double 13. x-lrgbagged units: double 14. bulk gtin: string	
fao	1. código ámbito: string 2. ámbito: string 3. código área: bigint 4. área: string 5. código elemento: bigint 6. elemento: string 7. código producto: bigint 8. producto: string 9. código	Esta tabla se generó a partir de los diferentes datos relacionados con la comercialización y producción de aguacate que se encontraron en la fao.

	año: bigint 10. año: bigint 11. unidad: string 12. valor: double 13. símbolo: string 14. descripción del símbolo: string 15. nota: string 16. código país declarante: bigint 17. países declarantes: string 18. código país socio: bigint 19. países socios: string	
avocado_updated_2020_csv	1. date: date 2. average_price: string 3. total_volume: string 4. 4046: double 5. 4225: double 6. 4770: double 7. total_bags: string 8. small_bags: string 9. large_bags: string 10. xlarge_bags: string 11. type: string 12. year: string	En esta tabla se encuentran los datos recolectados del dataset de avocado_updated_2020_csv obtenidos de la plataforma Kaggle

	13. geography: string	
2018_market_region_hab_data_csv	1. geography: string 2. segment: string 3 variety: string 4. timeframe: string 5. period: bigint 6. current year week ending: string 7. units prior year: double 8. units current year: double 9. unit variance: double 10. dollars prior year: double 11. dollars current year: double 12. dollar variance: double 13. asp prior year: double 14. asp current year: double 15. asp variance: double	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2018.
2019_plu_total_hab_data_csv	1. geography: string 2. timeframe: string 3. current year week ending: string	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2019.

	4. type: string 5. asp current year: double 6. total bulk and bags units: double 7. 4046 units: double 8. 4225 units: double 9. 4770 units: double 10. totalbagged units: double 11. smlbagged units: double 12. lrgbagged units: double 13. x-lrgbagged units: double	
avocado_csv	1. date: string 2. average_price: string 3. total_volume: string 4. 4046: double 5. 4225: double 6. 4770: double 7. total_bags: string 8. small_bags: string 9. large_bags: string 10.	Esta tabla se generó a partir de uno de los datasets de Kaggle relacionados con el consumo y la producción de aguacate en los Estados Unidos. En este dataset se tiene un mayor número de columnas, entre las cuales se encuentran: type, region, year.

	xlarge_bags : string 11. type: string 12. year: string 13. region: string	
augmented_avocado_csv	1. id: bigint 2. date: date 3. average_price: double 4. total_volume: double 5. 4046: double 6. 4225: double 7. 4770: double 8. total_bags: string 9. small_bags: string 10. large_bags: string 11. xlarge_bags : string 12. type: string 13. year: int 14. region: string	Esta tabla se generó a partir de uno de los datasets de Kaggle relacionados con el consumo y la producción de aguacate en los Estados Unidos. En este dataset se tiene un mayor número de columnas, entre las cuales se encuentran: type, region, year. Adicionalmente este es uno de los datasets más grandes que se recopilaron a partir del servicio de captura automática de datos
2021_market_region_hab_data_csv	1. geography: string 2. segment: string 3 variety: string 4. timeframe: string 5. period: bigint 6. current year week ending: string	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2021.

	7. units prior year: double 8. units current year: double 9. unit variance: double 10. dollars prior year: double 11. dollars current year: double 12. dollar variance: double 13. asp prior year: double 14. asp current year: double 15. asp variance: double	
list_of_countries_by_avocado_production_csv	1. rank: string 2. country/region: string 3. 2018: bigint 4. 2017: bigint 5. 2016: bigint	Esta tabla se generó a partir de uno de los datasets de Kaggle relacionados con la producción de aguacate a nivel mundial. En este conjunto de datos, se listan los diferentes países productores de aguacate y su respectiva producción de aguacate para los años 2016, 2017 y 2018.
avocado_csv_18f55474801eb0b7dc67881e8eaa73e0	1. id: bigint 2. date: date 3. average_price: double 4. total_volume: double 5. 4046: double 6. 4225: double 7. 4770:	Esta tabla, se generó a partir de uno de los datasets de Kaggle relacionados con el consumo y la producción de aguacate en los Estados Unidos. En este dataset se tiene un mayor número de columnas, entre las cuales se encuentran: type, region, year.

	double 8. total_bags: string 9. small_bags: string 10. large_bags: string 11. xlarge_bags : string 12. type: string 13. year: int 14. region: string	
2018_plu_total_hab_data_csv	1. geography: string 2. timeframe: string 3. current year week ending: string 4. type: string 5. asp current year: double 6. total bulk and bags units: double 7. 4046 units: double 8. 4225 units: double 9. 4770 units: double 10. totalbagged units: double 11. smlbagged units: double 12.	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2018.

	lrgbagged units: double 13. x-lrgbagged units: double 14. bulk gtin: string	
2020_market_region_hab_data_csv	1. geography: string 2. segment: string 3 variety: string 4. timeframe: string 5. period: bigint 6. current year week ending: string 7. units prior year: double 8. units current year: double 9. unit variance: double 10. dollars prior year: double 11. dollars current year: double 12. dollar variance: double 13. asp prior year: double 14. asp current year: double 15. asp variance: double	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2020.
2019_market_region_hab_data_csv	1. geography: string 2. segment: string 3 variety:	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos,

	string 4. timeframe: string 5. period: bigint 6. current year week ending: string 7. units prior year: double 8. units current year: double 9. unit variance: double 10. dollars prior year: double 11. dollars current year: double 12. dollar variance: double 13. asp prior year: double 14. asp current year: double 15. asp variance: double	proporcionados por el Hass Avocado Board para el año 2019.
2019_plu_total_hab_data_csv_b3703c1137fc3aa4a6f4d23ccbd87d13	1. geography: string 2. timeframe: string 3. current year week ending: string 4. type: string 5. asp current year: double 6. total bulk and bags units: double 7. 4046	En esta tabla se pueden encontrar los datos relacionados con la comercialización del aguacate en los Estados Unidos, proporcionados por el Hass Avocado Board para el año 2019.

	units: double 8. 4225 units: double 9. 4770 units: double 10. totalbagged units: double 11. smlbagged units: double 12. lrgbagged units: double 13. x-lrgbagged units: double 14. bulk gtin: string	

En la siguiente figura, se muestran algunas de las tablas previamente caracterizadas:

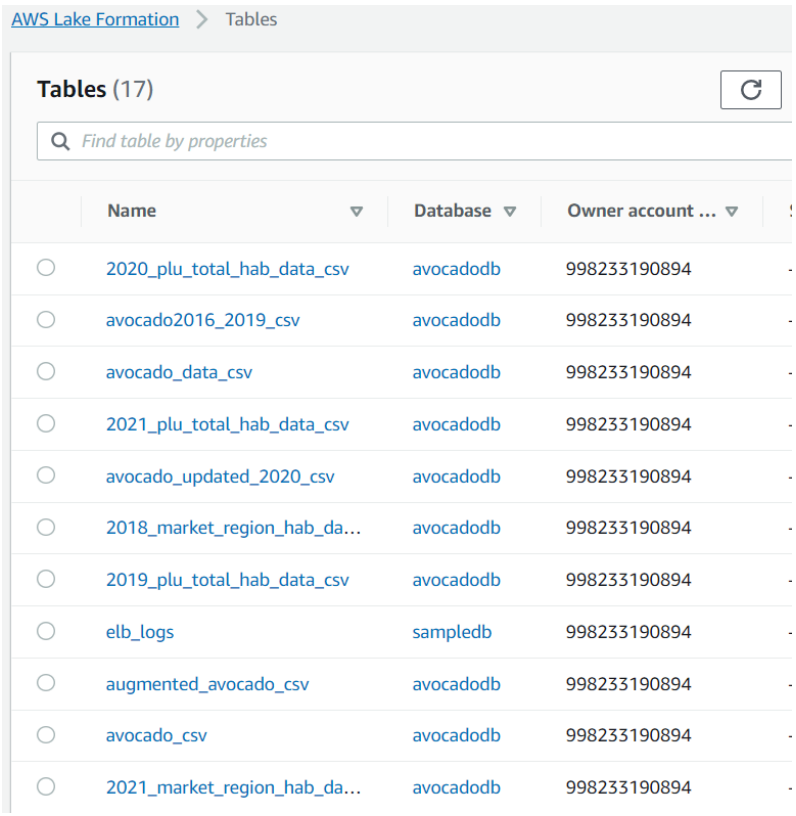

Figura 24. Listado de tablas de datos generadas por el Crawler. Fuente propia.

Como se pudo observar en las tablas anteriores numeradas del 7 al 13, se pudieron obtener múltiples registros y archivos a partir del servicio de captura automática de datos, de diferentes tamaños, formatos y estructuras. A partir de estos registros se conformó el catálogo de datos gestionado por AWS Glue a partir del Crawler, el cual generó las tablas que se caracterizaron en la tabla 13, donde se incluían las diferentes columnas, tipos de datos y descripción de cada una de las mismas.

5.1.7 Resumen del capítulo

En el presente capítulo se mostró el modelo de vistas 4+1 de la arquitectura del proceso ELT con sus diferentes vistas, adicionalmente se explicó cómo fue el desarrollo y despliegue del servicio de captura automática de datos sobre la plataforma de AWS y finalmente se mostraron los resultados obtenidos a partir de la implementación del servicio y del lago de datos.

6 CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

En este capítulo se presentan conclusiones, recomendaciones y trabajos futuros relacionados con el desarrollo de la práctica profesional, con respecto a los aspectos más importantes en la realización de los objetivos propuestos.

6.1 Conclusiones

- La caracterización de las fuentes de datos como primer paso, permitió comprender la estructura de las diferentes páginas de las fuentes, y los diferentes tipos de datos que proporcionaban en torno el mercado del aguacate. Esta caracterización, junto con la exploración de las diferentes tecnologías para el proceso de captura automática de datos permitió la formulación de un servicio que cumpliera con los objetivos de la práctica establecidos.

- La conformación del lago de datos, se pudo realizar de una manera efectiva gracias al servicio de AWS Lake Formation, el cual permitió la implementación del lago de datos de una manera más sencilla, y rápida que se integraba con el servicio de AWS Glue, para la generación y actualización del catálogo de datos del lago de datos a partir del Crawler.

- A partir de los datos obtenidos por parte del servicio de captura automática de datos, se puede llegar a realizar análisis importantes acerca de las diferentes dinámicas del mercado del aguacate, que involucrarían muchos factores, dentro de los cuales esta la cantidad de importaciones y exportaciones sobre diferentes paises, dentro de los cuales esta Colombia, buscando identificar el crecimiento que este teniendo este producto sobre el mercado de este país.

- Al estar el servicio de captura automática de datos alojado totalmente en la nube de AWS se facilitó la integración de los diferentes componentes que conforman el servicio de captura automática de datos, además de que los costos de operación fueron prácticamente nulos, y las tareas de gestión y revisión del servicio se pudieron realizar de una manera más sencilla.

- Al estar el servicio completamente alojado en la nube, se puede escalar de una manera más sencilla el alcance de este servicio de captura automática de datos, a partir de la extracción de datos provenientes de fuentes adicionales, y también se puede gestionar el acceso al lago de datos, para que diferentes usuarios puedan generar información de valor para la empresa, a partir del uso de técnicas de análisis de datos y de aprendizaje automático.

- Se obtuvieron todo tipo de archivos: estructurados, semiestructurados y no estructurados, de los cuales se puede obtener registros del crecimiento en la producción del aguacate en este país, los diferentes precios de venta de este producto, las cantidades de aguacate vendidas, las ganancias obtenidas, y los diferentes canales de distribución de este producto.

- Dentro de los registros obtenidos, existen datos que describen la demografía de los Estados Unidos, basados en estos registros se puede analizar el crecimiento de la población de este país, los diferentes ingresos económicos de los diversos grupos demográficos y las variaciones de las tasas de empleo, para identificar si puede llegar a existir una relación entre la variación de las variables demográficas, sobre el consumo del aguacate en los diferentes estados.

6.2 Recomendaciones

Para el desarrollo de un servicio de captura automática de datos, es necesario considerar las siguientes recomendaciones las cuales permitieron obtener los resultados esperados en el diseño, y la implementación del servicio.

- Caracterizar las fuentes de datos, sobre las cuales se quiere realizar la extracción de los datos, es una tarea muy importante, ya que, en el entorno de los análisis de datos, las fuentes de datos y los datos en bruto son el ítem principal para el desarrollo de los proyectos.

- Definir una arquitectura del sistema fue algo crucial en el desarrollo del servicio de captura automática de datos, ya que gracias al modelo de vistas 4+1 usado en el proyecto, se pudo abstraer y entender de una mejor manera como debía funcionar el servicio y sus diferentes componentes.

- Investigar acerca de los diferentes métodos y tecnologías para la extracción de datos, fue algo muy importante, ya que permitió comprender los diferentes tipos de datos: estructurados, semi estructurados y no estructurados, además de los diferentes enfoques y tecnologías para la extracción de datos e información relevante de los diferentes archivos y fuentes de datos.

- Definir la arquitectura del sistema completamente en la nube fue una decisión que brindo una mejor gestión y control de los diferentes procesos que se ejecutan en el servicio de captura automática de datos, además de que facilitó la creación del lago de datos y sus diferentes componentes.

- Estudiar y comprender las buenas prácticas de programación, ya que esto permitió desarrollar e implementar el servicio de una manera efectiva, en la cual se abstrajeron y definieron las diferentes clases en el servicio de acuerdo al paradigma de Programación Orientada a Objetos (POO).

- En el presente estudio realizado, no se desarrolló una transformación sobre los datos obtenidos, ya que se buscaba recopilar la mayor cantidad de registros en su formato original, adicional a ello en este trabajo no se procedió a desarrollar análisis sobre los datos, limpieza de los datos, y la continuación de las fases del proceso de CRISP-DM en torno a los datos capturados, de acuerdo a lo establecido por la empresa. Estos desarrollos ya quedan como trabajos futuros a realizar.

6.3 Trabajos futuros

Con el desarrollo del servicio de captura automática de datos, se identificaron nuevas propuestas de trabajo, tales como:

- Implementar extractores de datos orientados hacia los datos semi estructurados y no estructurados, principalmente texto. Para ello se requiere trabajar con herramientas y tecnologías orientadas hacia el análisis de texto y una mayor comprensión de los diferentes tipos de archivos.

- Realizar una caracterización de las fuentes de datos de los diferentes mercados del aguacate alrededor del mundo, principalmente en Europa, ya que en Colombia apenas se está comenzando a cultivar la cultura orientada a los datos, por la cual existen pocas fuentes de datos y además su documentación todavía es en parte insuficiente.

- Implementar métodos de análisis de datos, visualización de datos, exploración de datos, y aprendizaje automático para realizar análisis descriptivos, predictivos y prescriptivos con respecto al comportamiento del mercado del aguacate en los Estados Unidos.

7 REFERENCIAS

[1] Departamento Administrativo Nacional de Estadística, “Boletín técnico - Producto Interno Bruto PIB -Cuarto trimestre de 2017,” *Dane*, p. 28, 2018.

[2] DANE, *3er Censo Nacional Agropecuario: Resultados*, vol. 2. 2016.

[3] Procolombia. (2019). *Exportaciones de aguacate crecen 37,6% con relación a 2018 | Sala de Prensa | PROCOLOMBIA*. Procolombia. https://procolombia.co/noticias/exportaciones-de-aguacate-crecen-376-con-relacion-2018

[4] A. V. Pinzón, “Estrategia 360 : Cobertura total de riesgos y financiamiento,” 2019.

[5] K. J. Ferreira, J. Goh, and E. Valavi, “Intermediation in the Supply of Agricultural Products in Developing Economies,” *SSRN Electron. J.*, 2017, doi: 10.2139/ssrn.3047520.

[6] G. E. Silva Peñafiel, V. M. Zapata Yánez, K. P. Morales Guamán, and L. M. Toaquiza Padilla, “Análisis de metodologías para desarrollar Data Warehouse aplicado a la toma de decisiones,” Cienc. Digit., vol. 3, no. 3.4., pp. 397–418, 2019, doi: 10.33262/cienciadigital.v3i3.4..922.

[7] “Inteligencia de Negocio: Ciclo de vida de Ralph Kimball.” [Online]. Available: http://luisleonin.blogspot.com/2014/02/ciclo-de-vida-de-ralph-kimball.html. [Accessed: 09-Sep-2021].

[8] T. Marew, J. Kim, and D. H. Bae, “Systematic functional decomposition in a product line using aspect-oriented software development: A case study,” Int. J. Softw. Eng. Knowl. Eng., vol. 17, no. 1, pp. 33–55, 2007, doi: 10.1142/S0218194007003112.

[9] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering - A systematic literature review,” Inf. Softw. Technol., vol. 51, no. 1, pp. 7–15, 2009, doi: 10.1016/j.infsof.2008.09.009.

[10] R. Wirth and J. Hipp, “CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39,” Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min., no. 24959, pp. 29–39, 2000.

[11] M. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, “SciMAT: A new science mapping analysis software tool,” J. Am. Soc. Inf. Sci. Technol., vol. 63, pp. 1609–1630, 2012, doi: 10.1002/asi.22688.

[12] P. Conforti, “Price Transmission in Selected Agricultural Markets. FAO Commodity and Trade Policy Research Paper No. 7,” no. 7, 2004.

[13] R. A. E. Mueller, “E-commerce and entrepreneurship in agricultural markets,” Am. J. Agric. Econ., vol. 83, no. 5, pp. 1243–1249, 2001, doi: 10.1111/0002-9092.00274.

[14] J. Ricker-Gilbert, C. Jumbe, and J. Chamberlin, “How does population density influence agricultural intensification and productivity? Evidence from Malawi,” Food Policy, vol. 48, pp. 114–128, 2014, doi: 10.1016/j.foodpol.2014.02.006.

[15] J. W. Hopkins, “WEATHER AND WHEAT YIELD IN WESTERN CANADA: II. INFLUENCE OF PRE-SEASONAL PRECIPITATION ON PLOT YIELDS III. RELATION BETWEEN PRECIPITATION AND AGRICULTURAL YIELD,” Can. J. Res., vol. 14c, no. 6, pp. 229–244, Jun. 1936, doi: 10.1139/cjr36c-020.

[16] M. Allahyari, S. Ghavami, Z. Daghighi Masuleh, A. Michailidis, and S. Nastis, “Understanding Farmers’ Perceptions and Adaptations to Precipitation and Temperature Variability: Evidence from Northern Iran,” Climate, vol. 4, no. 4, p. 58, Dec. 2016, doi: 10.3390/cli4040058.

[17] J. L. Hatfield et al., “Climate impacts on agriculture: Implications for crop production,” Agron. J., vol. 103, no. 2, pp. 351–370, 2011, doi: 10.2134/agronj2010.0303.

[18] T. A. Asfaha and A. Jooste, “The effect of monetary changes on relative agricultural prices,” Agrekon, vol. 46, no. 4, pp. 460–474, Dec. 2007, doi: 10.1080/03031853.2007.9523781.

[19] X. Pham and M. Stack, “How data analytics is transforming agriculture,” Bus. Horiz., vol. 61, no. 1, pp. 125–133, 2018, doi: 10.1016/j.bushor.2017.09.011.

[20] B. Basnet and J. Bang, “The state-of-the-art of knowledge-intensive agriculture: A review on applied sensing systems and data analytics,” J. Sensors, vol. 2018, 2018, doi: 10.1155/2018/7425720.

[21] P. Shah, D. Hiremath, and S. Chaudhary, “Big data analytics architecture for agro advisory system,” Proc. - 23rd IEEE Int. Conf. High Perform. Comput. Work. HiPCW 2016, pp. 43–49, 2017, doi: 10.1109/HiPCW.2016.11.

[22] H. Fang, “Managing data lakes in big data era: What’s a data lake and why has it became popular in data management ecosystem,” 2015 IEEE Int. Conf. Cyber Technol. Autom. Control Intell. Syst. IEEE-CYBER 2015, pp. 820–824, 2015, doi: 10.1109/CYBER.2015.7288049.

[23] C. Madera, A. Laurent, T. Libourel, and A. Miralles, “How can the data lake concept influence information system design for agriculture?,” Efita Congr., 2017.

[24] C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang, “Leveraging the Data Lake: Current State and Challenges,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11708 LNCS, no. DaWaK, pp. 179–188, 2019, doi: 10.1007/978-3-030-27520-4_13.

[25] R. Liu, H. Isah, and F. Zulkernine, “A Big Data Lake for Multilevel Streaming Analytics,” 2020 1st Int. Conf. Big Data Anal. Pract. IBDAP 2020, 2020, doi: 10.1109/IBDAP50342.2020.9245460.

[26] R. Ploetz, R. J. Schnell, and J. Haynes, “Variable response of open-pollinated seedling progeny of avocado to Phytophthora root rot,” Phytoparasit. 2002 303, vol. 30, no. 3, pp. 262–268, 2002, doi: 10.1007/BF03039994.

[27] D. N. Kuhn et al., “Creation of an avocado unambiguous genotype SNP database for germplasm curation and as an aid to breeders,” Tree Genet. Genomes 2019 155, vol. 15, no. 5, pp. 1–12, Aug. 2019, doi: 10.1007/S11295-019-1374-1.

[28] Y. Qin and A. Horvath, “Contribution of food loss to greenhouse gas assessment of high-value agricultural produce: California production, U.S. consumption,” Environ. Res. Lett., vol. 16, no. 1, p. 014024, Dec. 2020, doi: 10.1088/1748-9326/ABCFDF.

[29] V. L. Fulgoni, M. Dreher, and A. J. Davenport, “Avocado consumption is associated with better diet quality and nutrient intake, and lower metabolic syndrome risk in US adults: results from the National Health and Nutrition Examination Survey (NHANES) 2001–2008,” Nutr. J. 2013 121, vol. 12, no. 1, pp. 1–6, Jan. 2013, doi: 10.1186/1475-2891-12-1.

[30] “Scrapy | A Fast and Powerful Scraping and Web Crawling Framework.” [Online]. Available: https://scrapy.org/. [Accessed: 13-May-2021].

[31] “SeleniumHQ Browser Automation.” [Online]. Available: https://www.selenium.dev/. [Accessed: 13-May-2021].

[32] “Mozenda - Scalable Web Data Extraction Software & Services.” [Online]. Available: https://www.mozenda.com/. [Accessed: 13-May-2021].

[33] “A Global Leader in Intelligent Automation & RPA | Automation Anywhere.” [Online]. Available: https://www.automationanywhere.com/. [Accessed: 13-May-2021].

[34] “Web Content Extractor - Cloud version.” [Online]. Available: https://www.webcontentextractor.com/. [Accessed: 13-May-2021].

[35] “AWS Glue Documentation.” [Online]. Available: https://docs.aws.amazon.com/glue/. [Accessed: 23-May-2021].

[36] “AWS Lambda Documentation.” [Online]. Available: https://docs.aws.amazon.com/lambda/. [Accessed: 23-May-2021].

[37] “Amazon Redshift Documentation.” [Online]. Available: https://docs.aws.amazon.com/redshift/. [Accessed: 23-May-2021].

[38] “AWS | Almacenamiento de datos seguro en la nube (S3).” [Online]. Available: https://aws.amazon.com/es/s3/. [Accessed: 23-May-2021].

[39] “Dataflow | Google Cloud.” [Online]. Available: https://cloud.google.com/dataflow. [Accessed: 23-May-2021].

[40] “Cloud Functions | Google Cloud.” [Online]. Available: https://cloud.google.com/functions. [Accessed: 23-May-2021].

[41] “BigQuery: almacén de datos en la nube.” [Online]. Available: https://cloud.google.com/bigquery. [Accessed: 23-May-2021].

[42] “Precios AWS Glue: servicio ETL administrado (Amazon Web Services).” [Online]. Available: https://aws.amazon.com/es/glue/pricing/. [Accessed: 23-May-2021].

[43] “Capa gratuita de AWS | Cloud computing gratis |AWS.” [Online]. Available: https://aws.amazon.com/es/free/?all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc&awsf.Free Tier Types=*all&awsf.Free Tier Categories=*all. [Accessed: 23-May-2021].

[44] “AWS Lambda – Precios.” [Online]. Available: https://aws.amazon.com/es/lambda/pricing/. [Accessed: 23-May-2021].

[45] “Precios de Amazon Redshift - Almacén de datos en la nube - Amazon Web Services.” [Online]. Available: https://aws.amazon.com/es/redshift/pricing/. [Accessed: 23-May-2021].

[46] “Precios Amazon Web Service S3 | Amazon Simple Storage Service.” [Online]. Available: https://aws.amazon.com/es/s3/pricing/?nc=sn&loc=4. [Accessed: 23-May-2021].

[47] “Precios de Dataflow | Cloud Dataflow | Google Cloud.” [Online]. Available: https://cloud.google.com/dataflow/pricing. [Accessed: 23-May-2021].

[48] “Programa gratuito de Google Cloud | Programa gratuito de Google Cloud.” [Online]. Available: https://cloud.google.com/free/docs/gcp-free-tier#free-tier. [Accessed: 23-May-2021].

[49] “Precios | Documentación de Cloud Functions | Google Cloud.” [Online]. Available: https://cloud.google.com/functions/pricing. [Accessed: 23-May-2021].

[50] “Precios | BigQuery | Google Cloud.” [Online]. Available: https://cloud.google.com/bigquery/pricing. [Accessed: 23-May-2021].

[51] “USDA/NASS QuickStats Ad-hoc Query Tool.” [Online]. Available: https://quickstats.nass.usda.gov/api. [Accessed: 23-May-2021].

[52] “USDA Economics, Statistics and Market Information System.” [Online]. Available: https://usda.library.cornell.edu/. [Accessed: 23-May-2021].

[53] “FAOSTAT.” [Online]. Available: http://www.fao.org/faostat/es/#data/OA. [Accessed: 23-May-2021].

[54] “FAOSTAT.” [Online]. Available: http://www.fao.org/faostat/es/#data/TM. [Accessed: 23-May-2021].

[55] “FAOSTAT.” [Online]. Available: http://www.fao.org/faostat/es/#data/QC. [Accessed: 23-May-2021].

[56] “California Department of Food and Agriculture.” [Online]. Available: https://www.cdfa.ca.gov/. [Accessed: 23-May-2021].

[57] “CDFA - Statistics.” [Online]. Available: https://www.cdfa.ca.gov/Statistics/. [Accessed: 13-May-2021].

[58] “Home - Hass Avocado Board.” [Online]. Available: https://hassavocadoboard.com/. [Accessed: 23-May-2021].

[59] “Kaggle: Your Machine Learning and Data Science Community.” [Online]. Available: https://www.kaggle.com/. [Accessed: 13-May-2021].

[60] “Census.gov.” [Online]. Available: https://www.census.gov/. [Accessed: 13-May-2021].

[61] P. B. Kruchten, "The 4+1 View Model of architecture," in IEEE Software, vol. 12, no. 6, pp. 42-50, Nov. 1995, doi: 10.1109/52.469759.