UNIVERSIDAD PRIVADA ANTENOR ORREGO

FACULTAD DE INGENIERÍA

PROGRAMA DE ESTUDIO DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

"SOLUCIÓN DE BIG DATA PARA EL ÁREA DE COMERCIALIZACIÓN DE LA EMPRESA INVERSIONES SANTA MARÍA EN EL PERÍODO 2021 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE"

Área de Investigación:

Gestión de Datos e Información

Autor(es):

Br. Chávez Rengifo, Keilita

Br. Amaya Pacheco, Orbit Jhampool

Jurado Evaluador:

Presidente: Dr. Urrelo Huiman, Luis Vladimir

Secretario: Ms. Meléndez Revilla, Karla Vanessa

Vocal: Ms. Abanto Cabrera, Heber Gerson

Asesor:

Ing. Ullón Ramírez, Agustin Eduardo

Código Orcid: https://orcid.org/0000-0003-1198-1855

TRUJILLO – PERÚ

2022

Fecha de sustentación: 2022/11/26

“SOLUCIÓN DE BIG DATA PARA EL ÁREA DE COMERCIALIZACIÓN DE LA EMPRESA INVERSIONES SANTA MARÍA EN EL PERÍODO 2021 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE”

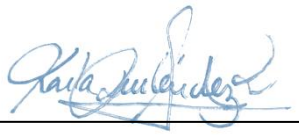
Elaborado por:

Br. Chávez Rengifo, Keilita

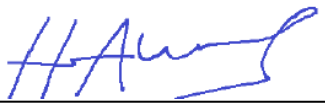
Br. Amaya Pacheco, Orbit Jhampool

Aprobada por:

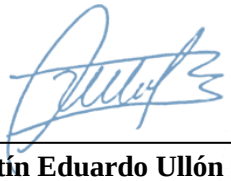
Dr. Urrelo Huiman, Luis Vladimir
Presidente
CIP: 88212



Ms. Karla Vanessa Meléndez Revilla
Secretario
CIP: 120097



Ms. Abanto Cabrera, Heber Gerson
Vocal
CIP: 48234



Ing. Agustín Eduardo Ullón Ramírez
Asesor
CIP: 137602

PRESENTACIÓN

Señores Miembros del Jurado:

De conformidad y en cumplimiento de los requisitos estipulados en el reglamento de grados y Títulos de la Universidad Privada Antenor Orrego y el Reglamento Interno de la Escuela Profesional de Ingeniería de Computación y Sistemas, ponemos a vuestra disposición el presente Trabajo de Tesis: "**SOLUCIÓN DE BIG DATA PARA EL ÁREA DE COMERCIALIZACIÓN DE LA EMPRESA INVERSIONES SANTA MARÍA EN EL PERÍODO 2021 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE"** para obtener el Título Profesional de Ingeniero de Computación y Sistemas

Los Autores.

DEDICATORIA

A Dios, quien siempre me ha dado mucha fuerza y me ha concedido llegar a este momento único de mi vida.

De igual manera, a mis queridos padres que siempre estuvieron junto a mí brindándome su apoyo incondicional en mi trayecto universitario para poder convertirme en una profesional.

Y gracias a todos los que nos brindaron su ayuda en este proyecto.

Br. Chávez Rengifo, Keilita

A Dios, porque es la razón de nuestra existencia, el que me ha dado la capacidad, valentía, fortaleza y perseverancia para poder alcanzar esta meta.

A mis padres Francisco Amaya y Rosa Pacheco, por su amor, consejos y apoyo incondicional, por darme ejemplo de superación y lucha constante, por cada esfuerzo y sacrificio que han hecho por mí, este título también es por y para ustedes.

Br. Amaya Pacheco, Orbit Jhampool

AGRADECIMIENTO

A nuestros profesores porque durante toda la carrera profesional nos han aportado con un granito de arena a nuestra formación, asimismo por sus consejos, enseñanzas y más que todo por su amistad.

Al Ing. Agustín Eduardo Ullón Ramírez, por su apoyo y asesoramiento en el desarrollo de la presente Tesis.

Muchas Gracias.

Los autores.

RESUMEN

“SOLUCIÓN DE BIG DATA PARA EL ÁREA DE COMERCIALIZACIÓN DE LA EMPRESA INVERSIONES SANTA MARIA EN EL PERÍODO 2021 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE”

Por:

Br. Chávez Rengifo, Keilita

Br. Amaya Pacheco, Orbit Jhampool

En la actualidad el uso de soluciones de big data es un criterio muy importante en la estrategia de una empresa u organización generando una ventaja potencial en su competitiva, por lo que proporcionar información valiosa frente a los problemas de negocio y de esta manera puede obtener entradas a nuevos mercados, un mejor control financiero, promociones, ofertas, eliminación de información irrelevante y una mejor planificación de la producción.

Inversiones Santa María es actualmente una empresa con un sistema transaccional con datos almacenados en SQL Server y otros almacenados en Excel. La empresa tiene previsto formular nuevos objetivos con miras de expansión de mercado y gestionar riesgos, por ejemplo, el crecimiento exponencial de los datos puede causar inconvenientes en el procesamiento y almacenamiento de datos en el futuro, lo que se traduce en un aumento de los costos, por lo que es necesario confiar en el apoyo de la plataforma tecnológica para el procesamiento en un ecosistema escalable y de alta disponibilidad, que beneficia a las empresas para una adecuada gestión de sus procesos. Por lo tanto, el objetivo de este trabajo es implementar una solución de big data mejorando de esta manera la gestión de la información en el proceso de ventas de Inversiones Santa María usando el ecosistema de Apache Hadoop y MS Azure.

La solución de big data permite centralizar la información de manera eficiente optimizando el tiempo, optimizando hardware y software, optimizando el tamaño de datos y de esta manera contar con información oportuna y valida.

ABSTRACT

“BIG DATA SOLUTION FOR THE MARKETING AREA OF THE INVERSIONES SANTA MARIA COMPANY IN THE 2021 PERIOD UNDER THE APACHE HADOOP AND MICROSOFT AZURE ECOSYSTEM”

By:

Br. Chavez Rengifo, Keilita

Br. Amaya Pacheco, Orbit Jhampool

At present, the use of big data solutions is a very important criterion in the strategy of a company or organization, it will reveal a potential advantage in its competition, so it will provide valuable information regarding business problems and in this way you can obtain inputs. to new markets, better financial control, promotions, offers, elimination of irrelevant information and better production planning.

Inversiones Santa María is currently a company with a transactional system with data stored in a SQL Server database and some other data stored in MS Excel sheets. The company plans to formulate strategies to achieve further market expansion, avoid and manage potential risks, for example, exponential growth of data may cause inconvenience in data processing and storage in the future, resulting in a increase in costs, so it is necessary to rely on the support of the technological platform for processing in a scalable and highly available ecosystem, which benefits companies to achieve better performance of their processes. Therefore, the objective of this work is to implement a big data solution to improve the analysis of information in the sales process of the company Inversiones Santa María in the Apache Hadoop and MS Azure ecosystems.

The big data solution allows you to centralize information efficiently, optimizing time, optimizing hardware and software, optimizing data size and thus have timely and valid information.

ÍNDICE DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Planteamiento del problema

Las soluciones basadas en big data se están volviendo tan importantes en las organizaciones actuales respaldando negocios en todas sus áreas o procesos. Por ejemplo, algunas de ellas confían en los componentes que conforman las soluciones de análisis de datos para realizar sus procesos de cierre financiero y generar informes normativos.

Hay muchas empresas involucradas en el suministro de soluciones de análisis de datos, grandes, pequeñas, específicas de la industria, gratuitas, etc. El uso de big data es un criterio muy importante en la estrategia de una empresa u organización, que puede generar ventajas competitivas potenciales, brindar información valiosa al momento de enfrentar los problemas del negocio, y así permitir una mejor entrada a nuevos mercados. Control financiero, promociones, ofertas, eliminación de información inapropiada y mejor planificación de la producción para una mayor rentabilidad de productos específicos.

Inversiones Santa María fue creada en 2012 para encarnar el ideal de sus socios, un verdadero servicio de ingeniería en la venta al por mayor y menor de materiales de construcción, grifería, equipos y materiales de plomería y calefacción.

Inversiones Santa María hoy en día cuenta con nuevos servicios los cuales son:

- Ventas de materiales para construcción
- Materiales de Mantenimiento.
- Servicio en arenado y pinturas especiales.

Inversiones Santa María en la actualidad tiene su sistema operacional trabajando con una base de datos en SQL y hojas de Excel. La empresa tiene problemas al momento de preparar reportes que son requeridas por parte de la Administración y el jefe de ventas, los cuales no están disponibles en el momento oportuno o de alguna manera se tienen que volver a procesar la información para que sea más entendible o manejable para los tomadores de decisiones, por lo que la solución de problemas se vuelve lenta y con problemas de análisis y obtención de datos.

Se espera que la empresa pueda crecer en base a mejores decisiones tomadas en base a la información consultada acerca de las ventas y de las tendencias de productos para sus clientes en base a sectores, asimismo se espera poder reducir las probabilidades de que en un futuro se haga una mal inversión. La prioridad de esta organización es la satisfacción de sus clientes brindándoles un mejor servicio.

La empresa pretende desarrollar estrategias de cómo lograr una mayor expansión en el mercado, evitando y manejando posibles riesgos como son el crecimiento exponencial de los datos que en un futuro podria presentar inconvenientes para la gestión y almacenamiento de datos generando un incremento en costos siendo necesario contar con una plataforma tecnológica que pueda soportar los procesamientos en un ecosistema escalable y de alta disponibilidad, beneficiando a la empresa y así lograr un mejor desempeño de sus procesos.

1.2. Delimitación del problema

El siguiente proyecto se realizará analizando la realidad en que se encuentran el área de comercialización en la empresa Inversiones Santa María utilizando Big data en un ecosistema Hadoop y MS Azure, para luego visualizar la información en PowerBI.

1.3. Características del problema

- Procesos con deficiencias.
- Altos tiempos de ejecución de consultas.
- Costos excedidos en tomar malas decisiones por no contar con la información adecuada en el momento oportuno.
- Falta de personas con capacidad en análisis de datos.

1.4. Análisis de características del problema

- **Procesos con deficiencias.**

La empresa, especialmente en el proceso de ventas, muestra deficiencias en el tratamiento de la información y esto con lleva a no brindar un adecuado soporte a la toma de decisiones para esta área, siendo esta el área critica del negocio.

- **Altos tiempos de ejecución de consultas.**

  Actualmente, las áreas administrativas de la empresa deben generar informes que son lentos y analíticamente problemáticos, y que no miden adecuadamente el desempeño del negocio. Este trabajo se realiza de acuerdo a estos pasos:

  - ✓ Los datos requeridos del área de ventas, que se ingresaron en Excel como gráficos, se realizaron con el trabajo extra de trabajadores de sistemas y, por lo general, tardaron hasta 3 días en llegar.
  - ✓ Esto pasa al administrador o al jefe de área, quien procede al análisis de la información, pero es muy posible que se necesite contar con más información de meses anteriores y poder realizar comparaciones, lo cual significa repetir el ciclo mencionado.
  - ✓ Esto genera:
    - Pérdida de tiempo y esfuerzo por parte del personal del área de sistemas
    - Demora en la entrega de la información
    - Generación de reportes ineficaces.
    - Dificultad para acceder a informes históricos.

- **Costos excedidos en tomar malas decisiones por no contar con la información adecuada en el momento oportuno.**

  Los indicadores y los resultados que se pueden obtener mejorará el análisis de información, obteniendo datos más procesados, que van a dar una mejor visión estado actual del negocio. Este es el problema principal de la empresa, teniendo un solo reportes estáticos sin indicadores trayendo consigo problemas de costos, procesos lentos y por consecuencia demora en toma de decisiones.

- **Falta de personas con capacidad en análisis de datos.**

  Las soluciones de análisis de datos apoyan la toma de decisiones y son útiles en la elaboración de informes, simplificando así el proceso y actuando como una herramienta para que los responsables de obtener informes entreguen resultados en un mejor tiempo. e información relacionada para definir estrategias clave.

  Esta problemática se debe a que los sistemas con los que cuenta la empresa no fueron desarrollados con el fin de brindar síntesis, análisis, consolidación, búsquedas de datos.

1.5. Definición del problema

La falta de un análisis de la información en los procesos de comercialización no le permite generar valor a la empresa y tomar mejores decisiones.

1.6. Formulación del problema

¿Cómo brindar un mejor análisis de información al proceso de ventas en Inversiones Santa María?

1.7. Formulación de la hipótesis

Una Solución de Big data permite mejorar análisis de la información de los procesos de ventas de la empresa Inversiones Santa María utilizando el ecosistema de Apache Hadoop y MS Azure.

1.8. Objetivos del estudio

Objetivo general:

Mejorar el análisis de los datos en el proceso de venta en la empresa Inversiones Santa María bajo el ecosistema de Apache Hadoop y MS Azure.

Objetivos específicos:

1. Analizar la situación actual y describir el proceso en el área de ventas de la empresa.
2. Realizar un análisis de requerimientos de las áreas críticas de la empresa.
3. Utilizar Apache Hadoop en un clúster de Azure HDInsight para la administración de clusteres.
4. Implementar el proceso ETL usando el servicio de Data Factory y Synapse Analytics.
5. Crear y diseñar un Dashboard usando PowerBI.

1.9. Justificación de la investigación

1.9.1. Importancia del trabajo

- **Académica o científica:** Big Data es una tendencia emergente dado a su uso y análisis de datos masivos mejorando el análisis de datos, apoyando al aprendizaje sobre nuevas soluciones en gestión de la información de esta manera fomenta o apoya a las futuras investigaciones dentro de la comunidad universitaria.

- **Social:** Con la solución de big data se pretende agilizar la obtención de información sirviendo para la captura de datos desde una base de datos relacional obteniendo conocimiento del volumen de información mejorando el procesamiento de la información y haciendo de este modo más eficiente la atención a las personas.

- **Empresa:** La solución de big data permitirá centralizar la información de manera eficiente optimizando el tiempo, optimizando hardware y software, optimizando el tamaño de datos y de esta manera contar con información oportuna y valida. Las soluciones de big data permitirán el desarrollo de informes prediseñados para procesos comerciales importantes.

1.9.2. Viabilidad de la investigación

- ✓ Esto es posible dado que los investigadores conocen de herramientas indispensables para el trabajo en este tipo de proyecto, basado en la cantidad de información en la empresa, el nivel de dificultad y la capacidad de rápido aprendizaje de los usuarios de la solución.

- ✓ Es viable porque la información de la empresa se encuentra a la mano de los investigadores. Los responsables de las áreas de la empresa en investigación muestran el total apoyo a la realización de la investigación.

2. MARCO TEÓRICO

2.1. ANTECEDENTES

- ✓ **Autor:** Carlos Linares Berrocal

  Título de Investigación: "Implementación de un sistema de big data aplicado a la migración de datos bajo la distribución cloudera con apache hadoop, en el banco Interbank", UTP , Lima 2019

  Descripción:

  En el presente trabajo tiene como objetivo "Implementar un sistema de migración de datos que permita la optimización de tiempos en el procesamiento de los datos, un ahorro en costos a nivel de infraestructura, además de gestionar el crecimiento exponencial de los datos con el fin de garantizar la escalabilidad y alta disponibilidad del sistema". Para lograr dicho objetivo se apoyo en 3 objetivos específicos: "Realizar la ingesta de datos a Hadoop con Apache Sqoop, Realizar el procesamiento de datos en Hadoop con Apache Hive, Evidenciar tiempos de ejecución entre las 2 soluciones, Apache Hadoop vs ETL Tradicional" , el trabajo concluye que: "Al realizar la ingesta de datos con Apache Sqoop a Hadoop, se pudo centralizar todos los datos provenientes de la fuente de datos ORACLE depositados en nuestro DATA LAKE, lo que permitirá no solo trabajar con información estructurada si no también semiestructurados y no estructurada con el fin de potenciar la toma de decisiones en la compañía".

- ✓ **Autor:** Rojas García, José Antonio

  Título de Investigación: "Propuesta de un modelo de negocio basado en big data que facilite la integración de los datos de las personas naturales y de soporte a las políticas de e-government en el Perú, apoyado en una empresa de logística integral" UPC – Lima 2018

  Descripción:

  El objetivo general del presente trabajos es "atender las necesidades de los consumidores futuros de la Empresa de Logística Ligera, así como que esta pueda servir de soporte para varias de las políticas implementadas en la actualidad para el desarrollo de un E- government en el Perú". Tiene como resultado "la

generación de beneficios adicionales que podrán ser complemento y soporte de algunas de las políticas del Estado Peruano para los próximos años como lo es la planificación de los servicios para la sociedad mediante la identificación real del número de ciudadanos en cada ámbito geográfico, así como la reducción de los costos de las organizaciones y de los ciudadanos para actualizar los datos de contractibilidad, generando de esta manera la transformación de la sociedad actual en una en una nueva Sociedad Digital".

- ✓ **Autor:** Guillermo Magaña Bou

  Título de Investigación: El big data y la convergencia con los insigths sociales . México, D.F. 2019

  Descripción:

  El presente trabajo tiene como objetivo conocer cómo las nuevas tendencias de publicidad en guías de programación pretenden alcanzar los Insigths Sociales. Los objetivos de la investigación "se planean acotar en audiencias e insigths, por lo que es de primordial importancia entender la ecología mediática que se propone en un mundo futuro. Concluyendo en el combinar el poder del Big-Data para la creación de insights accionables ayuda a los canales de televisión y proveedores de servicios, con el fin de alcanzar publicidad a segmentos y con un alto poder de selección de target o público objetivo. Tienen un gran potencial para aprovechar el Big Data ya que tienen un nivel transaccional y de participación de usuarios muy alto. Sin embargo la resistencia a utilizar analíticas de Big Data es un factor especialmente importante en este sector, ya que gran parte de la información que generan y gestionan es muy sensible".

- ✓ **Autores:** Rodriguez Torres, Eduardo y Pereda Morales, Piero Armando

  Título de Investigación: "Implementación de un Dashboard para la toma de decisiones estratégicas en la unidad de negocio de producción de huevo incubable de la empresa Avícola Santa Fe S.A.C. Usando Tecnologías Oracle Business Intelligence", Trujillo 2017

  Descripción:

En el presente trabajo tiene como objetivo “la implementación de Dashboards (Reportes Estratégicos), que será usados en la Unidad de Negocio de Producción de Huevo Incubable de la Empresa Avícola Santa Fe S.A.C”. Para lograr dicho objetivo, se usó la herramienta Oracle Business Intelligence. Para el desarrollo del trabajo “se utilizó la metodología de Ralph Kimball con la herramienta Business Intelligence de Oracle para implementar los Dashboards, que permitirán a las gerencias tener un espacio de trabajo adecuado donde puedan consultar los indicadores a través de estos”.

- ✓ **Autor:** Sonia Carolina Guama Morales

  Título de Investigación: “Estudio comparativo de métodos existentes para integrar la información estructurada y no estructurada de una industria enfocado en la generación de conocimiento, desde la perspectiva de una solución integral de big data.” 2018

  Descripción:

  La investigación del presente estudio se “desarrolló en base a la exploración de diversas fuentes entre los cuales se destacan textos de autores especializados en Big Data y consultas en la web de sitios certificados como fuente de apoyo”. Esta investigación ha permitido desarrollar “una guía para la introducción de Big Data en una organización, independientemente de su vertical, para generar conocimiento que les permita innovar, renovar o mejorar la visión de negocio que desean alcanzar. Un estudio comparativo de las principales técnicas de análisis e integración de los datos, consideraciones claves para la definición de una estrategia de implementación, generalidades de la metodología de implementación y arquitectura de Big Data. Esta investigación está dirigida a empresas que deseen innovar y reemplazar una cultura basada en los datos estructurados y desee aprovechar la gran cantidad de información disponible para optimizar las oportunidades de mercado basados en el análisis de los datos”.

2.2. DEFINICIONES

2.2.1. BIG DATA

Big data es un término que describe el gran volumen de datos – estructurados y no estructurados – que inundan una empresa todos los días. Pero no es la cantidad de datos lo importante. Lo que importa es lo que las organizaciones hacen con los datos. El big data puede ser analizado para obtener insights que conlleven a mejores decisiones y acciones de negocios estratégicas. (SAS, 2022)

Big data "es un término en desarrollo que describe un gran volumen de datos estructurados, semiestructurados y no estructurados que tienen el potencial de ser extraídos para obtener información y usarse en proyectos de aprendizaje automático y otras aplicaciones de análisis avanzado". (Iebschool, 2019)

CARACTERÍSTICAS PRINCIPALES:

Sus características a menudo se denominan la "V" de big data porque todos estos atributos comienzan con una letra del alfabeto. No hay consenso sobre cuántas "V" considerar, de hecho, la lista de estas "V" se amplía todo el tiempo, pero podemos estar seguros de que las 7 "V" más comunes en big data son:

- ✓ **Velocidad:** Nuestro concepto de inmediatez ha cambiado recientemente ya que la gente busca información que llega casi de inmediato. Menos de un día de noticias pueden haber perdido interés en horas o incluso minutos. Por lo tanto, la velocidad de análisis requerida en la sociedad actual es una de las características fundamentales de los datos a gran escala, donde en el centro se mueven constantemente datos procesados en tiempo real y se ejecutan algoritmos cada vez más complejos en un menor tiempo.
- ✓ **Variedad:** los datos sobre los que trabajan las técnicas de macrodatos son diversos pues, como hemos visto, proceden de numerosas fuentes y

se encuentran en distintos formatos. Además, continúa en aumento la cuantía de datos no estructurados en proporción a los tradicionales. Igual que pasaba con el volumen, esta entrada en escena con fuerza de los datos no estructurados requiere nuevos tratamientos de la información, necesitando de nuevas metodologías y tecnologías para poder ser analizadas.

- ✓ **Valor:** una gran cuantía de datos frecuentemente extrae pequeñas informaciones de valor. Cómo conseguir dicha información de manera eficiente es uno de los retos que afronta día a día el área de la inteligencia de datos. El valor es sin duda una cualidad fundamental en el análisis.

- ✓ **Variabilidad:** en un entorno tan cambiante como el de los macrodatos, la información varía mucho. Y también han de hacerlo los modelos o tratamientos que se aplican en torno a esta, pues no son fijos en el tiempo y requieren de un control periódico.

- ✓ **Volumen:** como hemos comentado, la cantidad de datos generados está aumentando. Según crecen las bases de datos, también lo han de hacer las aplicaciones y arquitectura construida para soportar la recogida y almacenamiento de datos cada vez más variados. Además, se han reducido los costes de almacenamiento propiciando almacenar grandes cuantías de información a un precio mucho más reducido que antiguamente.

- ✓ **Veracidad:** saber la fiabilidad de la información recogida es importante para obtener unos datos de calidad e, incluso, dependiendo de las aplicaciones que se le vaya a dar a misma, se convierte en fundamental. Es un factor que puede influir mucho en conseguir una ventaja competitiva en la explotación del Big Data.

- ✓ **Visualización:** Convertir cientos de tablas de datos en un solo gráfico que muestra claramente una perspectiva predecible es un ejemplo de

cómo presentar los resultados de forma clara y sencilla en lo que parece ser un ejercicio de síntesis..

2.2.2. DASHBOARD

Un dashboard es una interfaz de usuario, que puede presentar algo de semejanza con el panel de control de un coche, donde se organiza y se presenta la información de una manera que es fácil de leer. Este panel de control es más interactivo que el que nos puede presentar un coche, a menos que sea más moderno y esté basado en una pantalla de ordenador. (Arimetrics, 2022)

"Un dashboard o tablero de operaciones es una herramienta que sirve para visualizar y dar seguimiento a determinados indicadores de desempeño o estado. Condensa en un solo lugar la información crítica de una máquina, una empresa, una estrategia, etc". (Workana, 2020)

Dashboard, término inglés que se traduce normalmente como Panel o Tablero de Control. Es una representación gráfica de los Indicadores Clave de Performance (KPI), que intervienen en la medición de los objetivos del negocio y está orientada a la toma de decisiones para optimizar la estrategia en una empresa, mostrando datos relevantes. (TableauPeru, 2022)

2.2.3. APACHE HADOOP

"Apache Hadoop es una estructura para componentes de software diversos basada en Java, que permite fragmentar tareas de cálculo (jobs) en diferentes procesos y distribuirlos en los nodos de un clúster de ordenadores, de forma que puedan trabajar en paralelo. En las arquitecturas Hadoop más grandes pueden usarse incluso varios miles de ordenadores. La ventaja de este concepto es que a cada ordenador del clúster solo se le ha de proporcionar una fracción de los recursos de hardware necesarios. De esta manera, el trabajo con grandes volúmenes de datos no presupone ninguna máquina de última generación, sino

que se puede llevar a cabo de forma más rentable con varios servidores estándar". (Ionos, 2019)

Componentes básicos de la arquitectura Hadoop:

El fundamento del ecosistema Hadoop lo constituye el Core Hadoop. Sus componentes en la primera versión son el módulo básico Hadoop Common, el Hadoop Distributed File System (HDFS) y un motor MapReduce. A partir de la versión 2.3 este último fue sustituido por la tecnología de gestión de clústers YARN, también denominada MapReduce 2.0. Esta técnica excluye el algoritmo MapReduce del sistema de gestión en sí, de forma que a partir de este momento se convierte en un plugin basado en YARN. (Ionos, 2019)

2.2.4. HDINSIGHT

Azure HDInsight es un servicio de análisis, de código abierto, espectro completo y totalmente administrado en la nube para empresas. Con HDInsight, puede usar plataformas de código abierto, como Apache Spark, Apache Hive, LLAP, Apache Kafka, Hadoop, etc., en el entorno de Azure. (Azure HDInsight, 2019)

Azure HDInsight permite crear clústeres optimizados para Spark, Interactive Query (LLAP), Kafka, HBase y Hadoop en Azure. HDInsight también proporciona un Acuerdo de Nivel de Servicio de un extremo a otro en las cargas de trabajo de producción. (Microsoft, 2022)

HDInsight le permite escalar o reducir verticalmente las cargas de trabajo. Puede reducir el costo mediante la creación de clústeres a petición y pagar solo por lo que se utiliza. También puede compilar canalizaciones de datos para poner en marcha los trabajos. El procesamiento y el almacenamiento desacoplados ofrecen un mejor rendimiento y flexibilidad. (Microsoft, 2022)

Microsoft Azure es conjunto en constante expansión de servicios en la nube para ayudar a las organizaciones a satisfacer sus necesidades comerciales.

Otorga la libertad de crear, administrar e implementar aplicaciones en una red mundial con sus herramientas y marcos favoritos. (Microsoft, 2020)

Azure “es una nube pública de pago por uso que te permite compilar, implementar y administrar rápidamente aplicaciones en una red global de datacenters (centros de datos) de Microsoft”. (Tecon, 2019)

El concepto de Azure surgió como una plataforma de cloud computing “diseñada para crear, desarrollar y administrar aplicaciones, software y servicios a través de una red global de centros de datos administrados por Microsoft. Estos centros de datos están repartidos por todo el mundo, lo que también impulsa la creación de redes de trabajo internacionales para compañías con sedes en varios países”. (Ticportal, 2019)

Si entramos en concreto en el posicionamiento de proveedores Big Data (Data & Analytics) Microsoft está posicionado como líder destacado gracias a sus diferentes servicios y productos como Power BI Suite, SQL Azure, HDInsight, Data Lake, Stream Analytics, Machine Learning, etc. (RAYO, 2017)

2.2.5. MS Azure

Microsoft Azure es conjunto en constante expansión de servicios en la nube para ayudar a las organizaciones a satisfacer sus necesidades comerciales. Otorga la libertad de crear, administrar e implementar aplicaciones en una red mundial con sus herramientas y marcos favoritos. (Microsoft, 2020)

Azure “es una nube pública de pago por uso que te permite compilar, implementar y administrar rápidamente aplicaciones en una red global de datacenters (centros de datos) de Microsoft”. (Tecon, 2019)

El concepto de Azure surgió como una plataforma de cloud computing “diseñada para crear, desarrollar y administrar aplicaciones, software y servicios a través de una red global de centros de datos administrados por Microsoft. Estos centros de datos están repartidos por todo el mundo, lo que también impulsa la creación de

redes de trabajo internacionales para compañías con sedes en varios países". (Ticportal, 2019)

Si entramos en concreto en el posicionamiento de proveedores Big Data (Data & Analytics) Microsoft está posicionado como líder destacado gracias a sus diferentes servicios y productos como Power BI Suite, SQL Azure, HDInsight, Data Lake, Stream Analytics, Machine Learning, etc. (RAYO, 2017)

2.2.6. ANALÍTICA DE DATOS:

La gestión de datos tiene como objetivo último "dotar a las organizaciones de conocimiento y esto no es posible sin la Analítica de datos (Data Analytics). Significa traducir la información en oportunidades para el desarrollo de negocio y mejorar el rendimiento de la organización. En definitiva: se trata de sacar conclusiones de la información. En general de nada sirve tener datos, si luego no hacemos nada con ellos o más concreto: aprendemos de ellos, por lo que hoy tanto los datos, como el análisis, tenemos que comprender que van de la mano. La analítica de datos implica un proceso de limpieza y transformación cuyo objetivo es descubrir cuál es la información que nos ayudará a la mejor toma de decisiones y a extraer conclusiones que mejoren la competitividad de las compañías". (Prometeusgs, 2019)

Es muy importante conocer los diferentes tipos de analítica de datos y estos están divididos en 4 categorías:

a. **Descriptivos ¿Qué está pasando?**

   "Este es el formato más común. En un negocio te permite ver las métricas principales dentro del negocio. Por ejemplo, ganancias y pérdidas en el mes, ventas realizadas, etc". (Executrain, 2017)

b. Diagnóstico ¿Por qué está pasando?

"Este es el siguiente paso de complejidad del análisis de datos. Se debe contar con las herramientas necesarias para que el analista pueda profundizar en los datos y aislar la causa raíz de un problema". (Executrain, 2017)

c. Predictiva ¿Qué es lo más probable que pueda pasar?

"El análisis predictivo tiene que ver con la predicción. Ya sea la probabilidad de que ocurra un evento en el futuro, la previsión de una cantidad cuantificable o la estimación de un punto en el tiempo en el que algo podría suceder - todos ellos se hacen a través de modelos predictivos". (Executrain, 2017)

"Los modelos predictivos suelen utilizar una variedad de datos variables para hacer la predicción. La variabilidad de los datos de los componentes tendrá una relación con lo que es probable predecir (por ejemplo, cuanto más vieja sea una persona, más susceptible será un ataque al corazón - diríamos que la edad tiene una correlación lineal con el riesgo de ataque cardíaco)." (Executrain, 2017)

"En un mundo de gran incertidumbre, ser capaz de predecir permite tomar mejores decisiones. Los modelos predictivos son algunos de los más importantes utilizados en una serie de campos." (Executrain, 2017)

d. Prescriptivos ¿Qué necesito hacer?

"El siguiente paso en términos de valor y complejidad es el modelo prescriptivo. El modelo prescriptivo utiliza un entendimiento de lo que ha sucedido, por qué ha sucedido y una variedad de análisis de lo que podría suceder para ayudar al usuario a determinar el mejor curso de acción a tomar. El análisis prescriptivo no suele ser sólo con una acción individual, sino que de hecho es una serie de otras acciones. Un buen ejemplo de esto es una aplicación de tráfico que le ayuda a elegir la mejor ruta a casa y teniendo en cuenta la distancia de cada ruta, la velocidad a la que uno

puede viajar en cada carretera y, crucialmente, las restricciones de tráfico actuales". (Executrain, 2017)

2.2.7. ETL

ETL es un tipo de integración de datos que hace referencia a los tres pasos (extraer, transformar, cargar) que se utilizan para mezclar datos de múltiples fuentes. Se utiliza a menudo para construir un almacén de datos. Durante este proceso, los datos se toman (extraen) de un sistema de origen, se convierten (transforman) en un formato que se puede almacenar y se almacenan (cargan) en un data warehouse u otro sistema. Extraer, cargar, transformar (ELT) es un enfoque alterno pero relacionado diseñado para canalizar el procesamiento a la base de datos para mejorar el desempeño. (ETL, 2022)

2.2.8. INTEGRACIÓN DE LOS DATOS

"La integración de los datos es el proceso que implica combinar datos desde distintas fuentes en una única visión unificada: empezando por la ingesta, la limpieza, el mapeo hasta la transformación en un colector determinado y, por último, convertir los datos en elementos más explotables y valiosos para aquellos que acceden a ellos. Actualmente las empresas llevan a cabo iniciativas de integración de datos para analizar y tomar decisiones a partir de sus datos de forma más eficaz, en especial dada la explosión de datos y de nuevas tecnologías cloud y de big data. La integración de datos es una obligación, puesto que permite a las empresas modernas mejorar la toma de decisiones estratégica y aumentar su ventaja competitiva". (Talend, 2019)

La importancia de la integración de datos pone de manifiesto las importantes ventajas que supone disponer de un enfoque meditado para la integración de datos, como son:

- ✓ La integración de datos mejora la colaboración y unificación de sistemas y ahorro de tiempo, reduciendo errores (y modificaciones posteriores)
- ✓ La integración de datos suministra datos más valiosos (Talend, 2019)

"Integrar significa combinar datos que se encuentran en diferentes fuentes para permitirle al usuario final tener una vista unificada de los mismos para una accesibilidad idónea, que sirva a las necesidades de negocio". (PowerData, 2019)

2.2.9. MS POWER BI

Power BI es un servicio de análisis empresarial que proporciona información detallada para permitir la toma de decisiones rápidas e informadas. (Power BI, 2019)

Power BI "es un conjunto de aplicaciones de análisis de negocios que permite analizar datos y compartir información de forma rápida y muy intuitiva. Sabrás en todo momento cual es el estado de tu empresa y si necesita tu atención, podrás resolver cualquier problema al momento gracias a los paneles intuitivos en tiempo real." (Stratebi, 2019)

Power BI Embedded está concebido para simplificar la manera en que los ISV y los desarrolladores usan las funcionalidades de Power BI con análisis integrado. Power BI Embedded simplifica las capacidades de Power BI, al ayudar a agregar rápidamente objetos visuales, informes y paneles impactantes a las aplicaciones. De forma similar a las aplicaciones basadas en Microsoft Azure, usa servicios como Machine Learning e IoT. Al habilitar en sus aplicaciones la exploración de datos por los que es fácil navegar, los ISV permiten que sus clientes tomen decisiones rápidas y fundamentadas dentro de un contexto. (Microsoft, 2019)

2.3. METODOLOGÍA DEL PROYECTO

Se basa en el desarrollo del Moldeo propuesto por Ángelo Martínez en su tesis de Maestría en la Universidad de Palermo de Argentina, la cual se basa en las siguientes fases:

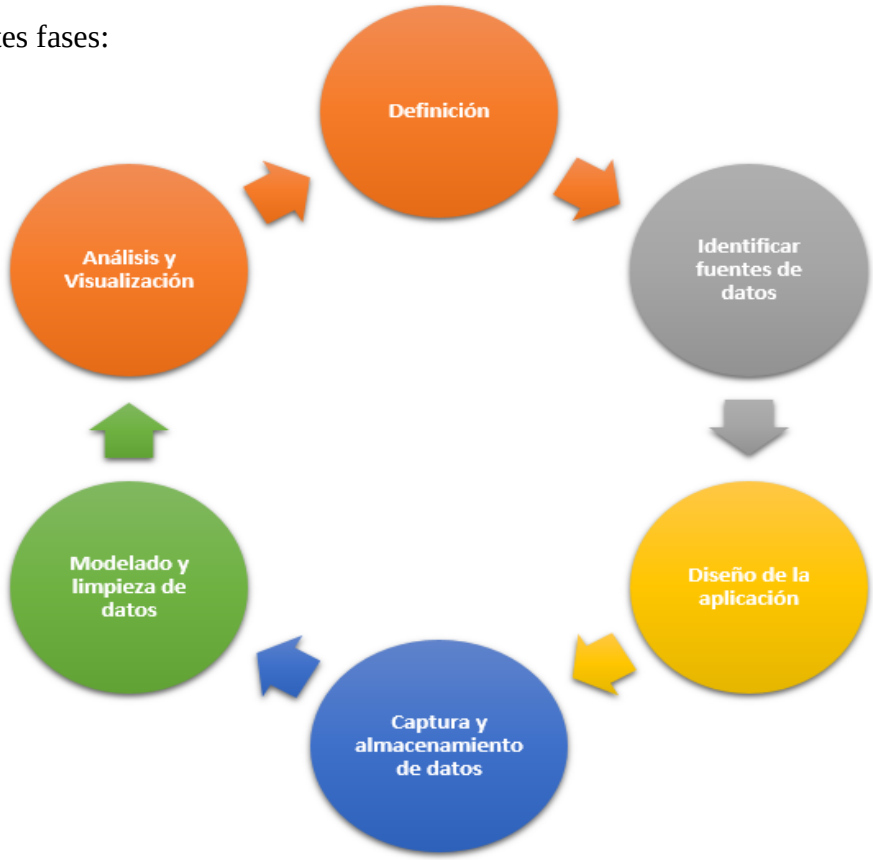


Figura 01: Metodología del Proyecto

Fuente: (Martínez , 2018)

FASE 1: DEFINICIÓN

En esta etapa, se definen los requisitos del negocio. El equipo de desarrollo del proyecto debe colaborar con las áreas de negocio para refinar y perfeccionar continuamente sus necesidades de información. Los objetivos del proyecto deben definirse lo más claramente posible para evitar confusiones, ya que se aclararán muchas dudas y se entenderán mejor los beneficios a medida que avancemos.

Debido al fuerte componente técnico, existe una tendencia a definir los resultados solo en términos de tecnología. Los equipos de desarrollo líderes se centraron en obtener datos y luego no saber qué hacer con ellos. Esta forma de pensar conduce a la dilución de la responsabilidad por el resultado final, que también debe pertenecer al negocio. En conclusión, los objetivos no pueden centrarse en la tecnología, deben ampliarse para cubrir el uso final de los datos y su traducción en valor comercial.

Por lo tanto, es necesario identificar a los usuarios finales y saber qué preguntas responder para tomar mejores decisiones. Además de las preguntas comerciales, también tuvimos que responder varias preguntas, como:

- ¿De dónde provinieron los datos?
- ¿Qué tan grandes son estos datos y qué tan rápido están creciendo?
- ¿Con qué frecuencia se producen y con qué frecuencia se necesitan?
- ¿Qué tipo de datos se necesitan?
- ¿Cómo garantizamos su fiabilidad y precisión?
- ¿Cómo se almacenarán?
- ¿Es necesario analizarlos en tiempo real?
- ¿Podemos combinar datos internos con otros datos externos para ayudar a encontrar relaciones valiosas?

Como nueva implementación, se recomienda comenzar con un tema específico que no cubra muchas áreas del conjunto de datos y que pueda evaluarse fácilmente para monitorear el proyecto y corregir errores. Por otro lado, cuando se trata de cubrir un área muy amplia y ambiciosa, el resultado puede ser más difícil de evaluar, y si quieres adaptarlo, puede ser más complicado.

FASE 2: IDENTIFICAR FUENTES DE DATOS

Una vez que se identifican las necesidades comerciales, se deben encontrar las fuentes de información necesarias para responder las preguntas comerciales.

En primer lugar, deben identificarse para que solo se analicen y utilicen aquellos datos relevantes para los objetivos previamente establecidos. En general, la mayor cantidad de fuentes de información en una organización son internas, es decir, se pueden encontrar en almacenes de datos, bases de datos relacionales, bases de datos NoSQL, CRM, ERP, archivos de texto plano, libros de Excel, etc.

Una de las ventajas de utilizar la tecnología Big Data es que permite enriquecer la información interna disponible en la organización con información de fuentes externas en la web, redes sociales, otras empresas, datos abiertos, etc. Sin embargo, no toda la información externa es útil o de calidad suficiente para lograr los objetivos de nuestro análisis y permitirnos obtener un conocimiento fiable. Si el conocimiento adquirido no es confiable, puede llevar a tomar decisiones equivocadas sobre los procesos de negocio que pretendemos mejorar, lo que lleva a pérdidas financieras y fallas en los proyectos. De hecho, es muy importante comprobar la calidad y la coherencia entre las fuentes de datos internas y externas utilizadas.

Por otro lado, es necesario determinar qué tipo de datos se necesitarán, por ejemplo, la mayoría de los datos estructurados se pueden encontrar en fuentes de datos internas, aunque también puede haber datos semiestructurados y no estructurados en correos electrónicos, documentos de texto. etc. Para fuentes externas, la mayoría de los datos serán semiestructurados y no estructurados.

FASE 3: DISEÑO DE LA SOLUCION

El análisis realizado en la etapa anterior permitirá elegir una arquitectura que se ajuste a las necesidades de la empresa. Los objetivos claros y los tipos identificados de fuentes de datos ayudarán a comprender qué se debe promover en la arquitectura. Aunque inicialmente se recomienda centrarse en problemas comerciales específicos para obtener mejores resultados, se debe esperar que la arquitectura crezca y sea capaz de soportar casos de uso futuros (Rivera,

2017). Por lo tanto, el marco debe ser al menos escalable, flexible y tolerante a fallas.

En la entrada de datos, no todas las herramientas son adecuadas para todas las fuentes de datos y, en algunos casos, es mejor combinar varias herramientas para cubrir todas las situaciones. El procesamiento debe evaluar si el sistema está transmitiendo o empaquetando. La opción ideal es aprovechar el procesamiento de flujo proporcionado por Big Data. También se recomienda utilizar herramientas de gestión, seguimiento y gestión de la arquitectura que faciliten y centralicen diversas tareas.

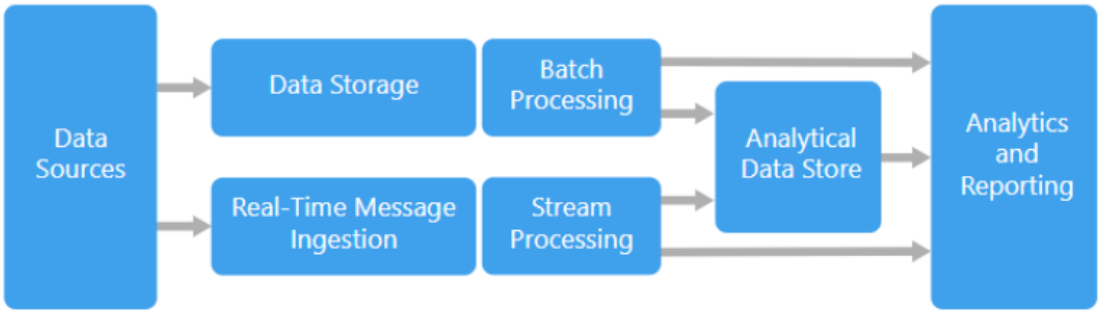


Figura N°02. Arquitectura para Big Data

Fuente: (Wasson, 2017)

FASE 4: CAPTURA Y ALMACENAMIENTO DE DATOS

Una vez que se diseña la implementación del clúster, lo siguiente a considerar es la carga de datos. En esta fase, los datos estructurados o no estructurados se utilizan sin acondicionamiento y se almacenan para su análisis. Los datos pueden almacenarse en su propia infraestructura o contratarse con un proveedor de servicios que proporcione una plataforma de big data basada en la nube.

Una arquitectura basada en procesamiento y almacenamiento distribuidos es una excelente solución para almacenar y procesar flujos continuos de datos, en

comparación con lo que las bases de datos relacionales pueden hacer muy poco. Hadoop puede almacenar y procesar datos de manera eficiente y puede procesar grandes cantidades de información rápidamente. Hadoop puede considerarse el mejor complemento a los datos estructurados almacenados en bases de datos relacionales o data warehouses, y utilizar estos últimos no significa abandonar las estrategias de data warehouse. También existen herramientas como Sqoop que te permiten importar datos desde una base de datos relacional en HDFS, Hive o HBase. Alternativamente, los archivos HDFS se pueden exportar a bases de datos relacionales.

Los datos semiestructurados y no estructurados cuentan con diferentes herramientas NoSQL que permiten almacenarlos y agregar valor con características propias que Hadoop no tiene. Ambas son soluciones de big data para almacenar big data que son complementarias y compatibles entre sí, así como con las bases de datos relacionales tradicionales. La integración entre los sistemas NoSQL y Hadoop es casi natural, cada base de datos NoSQL también tiene su propia interfaz.

FASE 5: MODELADO Y LIMPIEZA DE DATOS

En esta fase, los datos están listos para el análisis y luego se procesan en un formato estructurado que se puede consultar con herramientas analíticas. El almacén de datos analíticos utilizado para atender estas consultas probablemente será un almacén de datos relacional estilo Kimball, como el que se encuentra en la mayoría de las soluciones tradicionales de inteligencia empresarial (BI).

Los datos de baja calidad conducirán a un conocimiento de baja calidad. Por lo tanto, en esta etapa es fundamental el preprocesamiento de datos, cuyo objetivo principal es obtener conjuntos de datos finales que puedan considerarse confiables y útiles para la siguiente etapa de análisis de datos. El

preprocesamiento de big data es una tarea desafiante debido al tamaño del conjunto de datos. A medida que se recopilan grandes cantidades de datos, se necesitan mecanismos más sofisticados para analizarlos.

FASE 6: ANÁLISIS Y VISUALIZACIÓN

Una vez que toda la información se recopila en un repositorio común, es posible comenzar a analizar la gran cantidad de información utilizando técnicas avanzadas de análisis predictivo y minería de datos para encontrar patrones de comportamiento comunes para diferentes segmentos de clientes e identificar patrones que nos ayuden a predecir el futuro y alcanzar resultados. nuestros objetivos fijados Tendencia con la alternativa más adecuada.

En esta etapa entran en juego científicos de datos, analistas de datos, expertos en minería de datos, etc. Basándose en ideas, matemáticas, estadísticas y herramientas computacionales, pueden realizar análisis inteligentes de big data. Y a la medida de los objetivos de la organización, utilizando la tecnología para encontrar soluciones, hacer predicciones, brindar información a la que se pueda acceder en tiempo real a través de diversos canales, con una visualización simple e intuitiva de los resultados alcanzados, para que luego los usuarios del negocio puedan probar y utilizar.

En resumen, la fase de análisis le permite interactuar con datos que han sido preprocesados y almacenados en algún tipo de almacén de datos analíticos para obtener inteligencia comercial.

3. MATERIALES Y MÉTODOS

3.1. MATERIAL

3.1.1. Población

02 personas, Jefe del área comercial y el Gerente..

3.1.2. Muestra

Se seleccionará solo a los tomadores de decisiones.

3.1.3. Unidad de análisis

Los datos proporcionados por el Área de comercial de la empresa.

3.2. MÉTODO

3.2.1. Tipo de investigación

Aplicada.

3.2.2. Diseño de Investigación

Diseño Pre-experimental con pre-prueba y post-prueba

Diseño del modelo pre-experimental	**G -> O_1 -> X -> O_2**
G (Grupo a investigar)	Datos de la empresa
X (Tratamiento)	Solución de Big Data
O (Observación)	O_1 : Observación pre-test
	O_2: Observación post-test

Tabla N° 01. Diagrama de investigación

3.2.3. Variables de estudio

- Variable Independiente (VI): Big data
- Variable Dependiente (VD): Procesos de ventas de Inversiones Santa María utilizando el ecosistema de Apache Hadoop y MS Azure.

Variable	Dimensión	Indicador	Unidad de medida	Instrumento de Investigación
VI	Tiempo	Tiempo en obtener registros desde la solución de big data	Minutos	Hoja de captura de tiempos
	Grado de satisfacción de los usuarios	Grado de satisfacción de los sobre los reportes de la solución	% grado satisfacción	Hoja resumen de porcentajes de satisfacción
VD	Oportunidad	Tiempo para analizar información	Minutos	Hoja de captura de datos

Tabla N° 02: Operacionalización de las variables

3.2.4. Técnicas e instrumentos de recolección de datos

3.2.4.1.Técnicas

- ✓ Observación
- ✓ Análisis Documental

3.2.4.2.Instrumentos

- ✓ Cuestionarios
- ✓ Hojas de cálculo.

3.2.5. Técnicas de procesamiento y análisis de datos

3.2.5.1.Procesamiento de datos

Se analizará a través de tablas dinámicas y gráficos.

3.2.5.2.Análisis de datos

El análisis de datos se desarrollara en cuadros estadísticos y Pruebas de hipótesis nula y alternativa, así como las pruebas Z.

4. RESULTADOS: APLICACIÓN DE LA METODOLOGÍA

4.1. FASE 1: DEFINICIÓN

4.1.1. DESCRIPCIÓN DEL NEGOCIO

Inversiones Santa María fue creada en 2012 brindando servicios de venta al por mayor y menor de materiales de construcción, grifería y equipos y materiales de plomería.

Inversiones Santa María cuenta con una gerencia que controla los distintos procesos que se realizan, así como la toma de decisiones, un área de mantenimiento y su área principal de ventas a través de ferreterías, inventario y la utilidad de los productos vendidos y que también administra y entregas de productos con seguimiento móvil e inventario de estos productos.

a. Descripción de la organización

- **Razón Social:** Inversiones & Negociaciones Santa Maria S.A.C.
- **Ruc:** 20603756321
- **Ubicación:** Mza. I Lote. 4 Sector Natasha Alta - La Libertad – Trujillo.
- **Rubro Económico:** Venta de Materiales de Construcción y Ferretería.
- **Clientes:** Estado, Empresas Privadas.
- **Competidores:** Todas las Empresas ligadas en el Rubro situadas en la localidad de Trujillo.

b. Visión – Misión

- **Visión:**

  Brindar servicios de alta calidad para satisfacer la demanda del mercado.

- **Misión:**

  Brindar servicios de alta calidad, satisfacer necesidades de

nuestros clientes y fortalecer el estatus de la organización siendo líderes en la región norte.

c. Organigrama

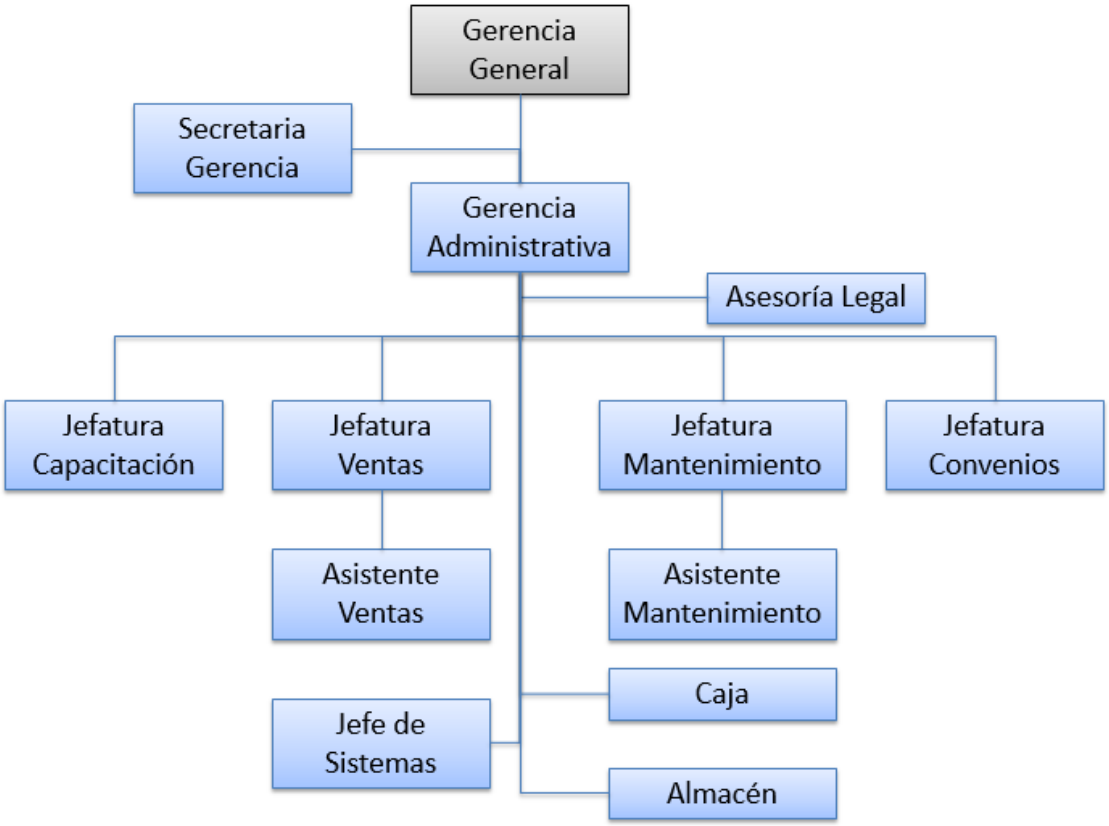


Figura N°03: Organigrama (Fuente: empresa)

4.1.2. OBJETIVOS DEL NEGOCIO

- Tener una mejor gestión de las ventas.
- Tener una mejor gestión de las compras.

4.1.3. DESCRIPCIÓN DE LOS PROCESO INVOLUCRADOS EN EL PROBLEMA

4.1.3.1.Descripción del proceso

El cliente va a la ferretería y pregunta al vendedor sobre el producto que necesita, se chequea en el sistema con su almacén el producto deseado, se envía al cajero, el cajero pregunta cuál es el pago método, efectivo o tarjeta, si es efectivo, se le da el recibo junto con su artículo, si es tarjeta, pide pagar a plazos y le entrega el recibo y el artículo.

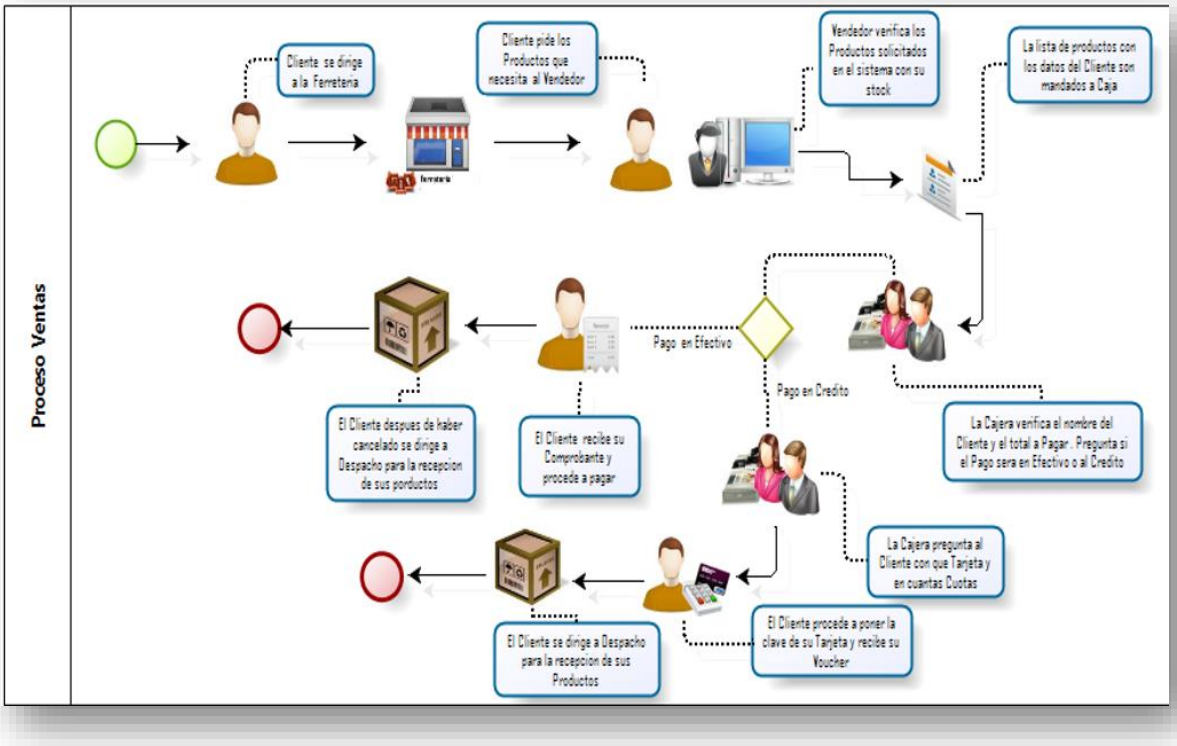


Figura N° 04. Proceso de ventas

Fuente: propia

4.1.3.2. Descripción del proceso

El comprador solicita al proveedor visitar la tienda de equipos de cómputo, el proveedor ofrece diferentes precios para el producto que el comprador solicita, luego de analizar los precios se llega a un acuerdo con el proveedor, el comprador elabora una guía de productos para ser recibidos y entregados a el proveedor, el propietario recibe el producto, el comprador verifica el producto recibido y firma la instrucción de entrega.

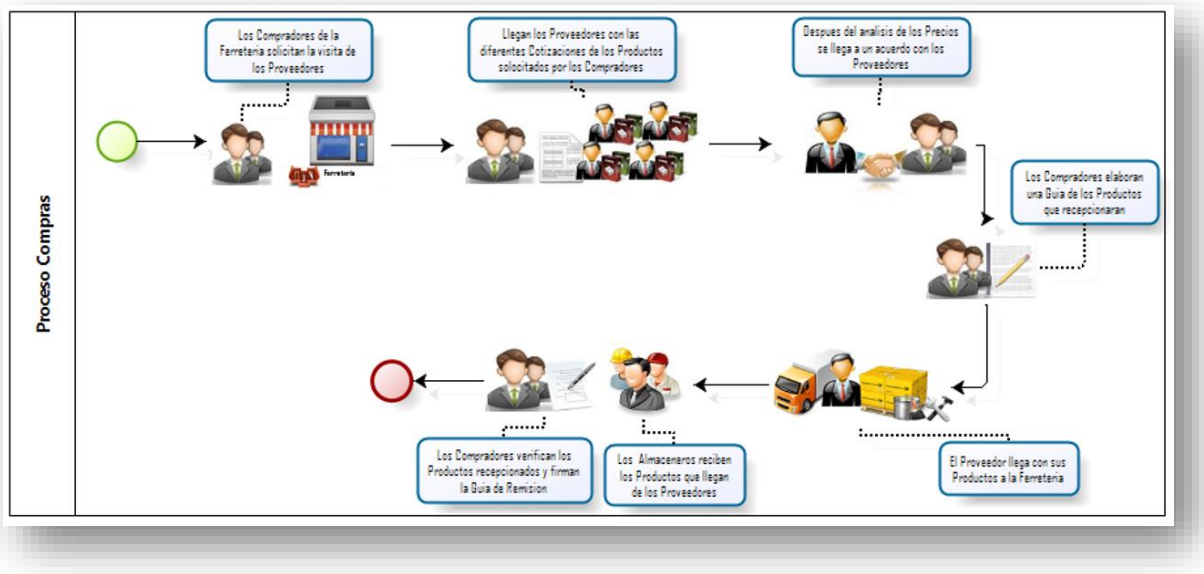


Figura N°05. Proceso de compras

Fuente: propia

4.1.4. GENERACIÓN DE REPORTES EN LA ACTUALIDAD

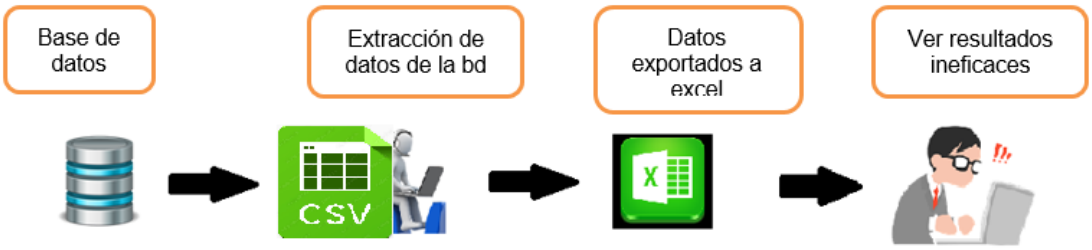


Figura N° 06. Proceso de generación de reportes

Fuente: propia

4.1.5. DEFINICIÓN DE LOS REQUISITOS DEL PROYECTO

A través de entrevistas y reuniones con los empleados de la empresa, así como reuniones que involucran varios procesos, se determinaron los siguientes requisitos para las ventas y compras.

NRO	REQUERIMIENTOS	INDICADOR	DIMENSIONES CANDIDATAS
1	Cuantos Productos han sido comprados y de que Proveedores en un Tiempo determinado	**Cant productos**	**Proveedor Tiempo**
2	Conocer los Precio total de pedidos por Proveedores en un tiempo determinado	**Monto productos**	**Material Producto Tiempo**
3	Resumen de cantidad de Compras	**Cant Compras**	**Sucursal Tiempo**
4	Resumen de los Proveedores frecuentes en un mes determinado	**Cant proveedores**	**Proveedor tiempo**
5	Conocer cuántos Pedidos de productos se hacen un mes determinado	**Cant Productos**	**Pedido Producto tiempo**
6	Conocer los egresos en un determinado semestre	**Monto Egresos**	**Sucursal tiempo**
7	Conocer los Productos más vendidos en un tiempo determinado	**Productos**	**Movimiento Productos tiempo**
8	Conocer los Productos Recibidos en un tiempo determinado	**Cant productos**	**Producto sucursal Tiempo**

9	Conocer la Categoría de producto y su demanda en un trimestre determinado	**Cant Productos**	**Producto tiempo**
10	Conocer las condiciones de los Productos en un mes determinado	**Cant Productos**	**Producto tiempo**
11	Conocer la cantidad de Productos que entraron en un tiempo determinado	**Cant productos**	**Sucursal Producto tiempo**
12	Resumen de los Productos en buen estado en un tiempo determinado	**Productos**	**Producto tiempo**
13	Conocer el record de Ventas de cada Empleado	**Monto Total ventas**	**Empleado Sucursal**
14	Conocer cantidad de productos vendidos por trimestre y año	**Cant productos**	**Tiempo**
15	Conocer la cantidad de productos compradas por clientes más frecuentes en un trimestre determinado	**Cant productos**	**Cliente productos tiempo**
16	Conocer los Productos comprados por Cliente en un mes determinado	**Cant Productos**	**Producto Cliente tiempo**
17	Conocer el Resumen de Ventas en un determinado mes	**Monto total ventas**	**Tiempo sucursal**
18	Reporte del promedio de ventas por tiempos determinados	**Promedio ventas**	**Productos Sucursal tiempo**

4.2. FASE 2: IDENTIFICAR FUENTES DE DATOS

4.2.1. Origen de datos:

Los datos que se van a ingesta en la solución del big data están en formatos CSV obtenidos desde la base de datos de la empresa.

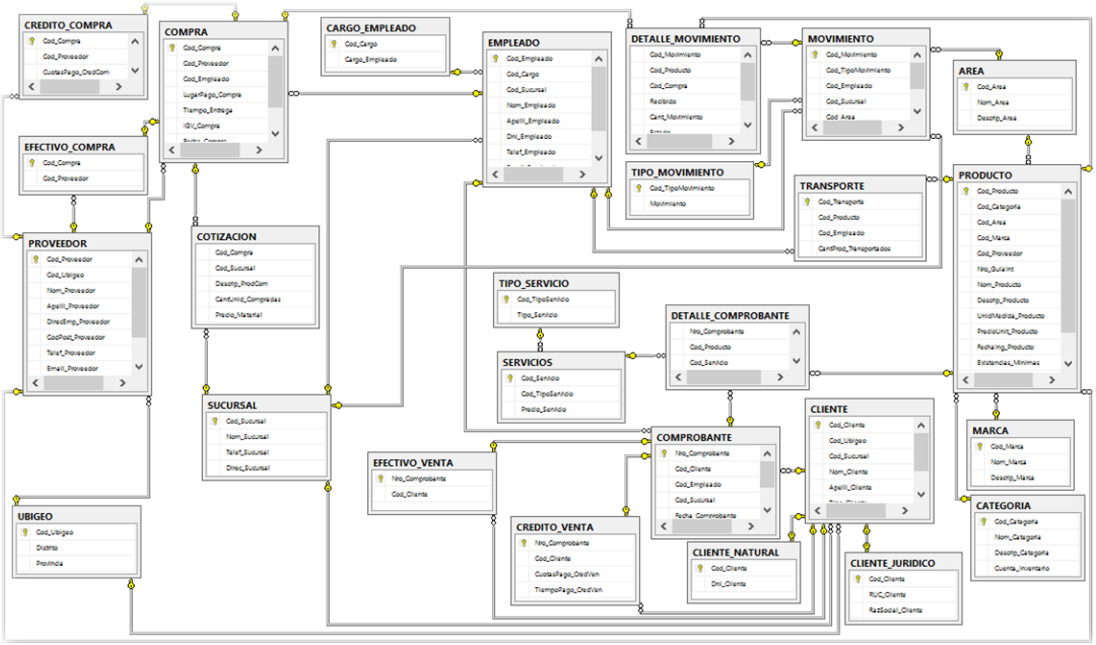
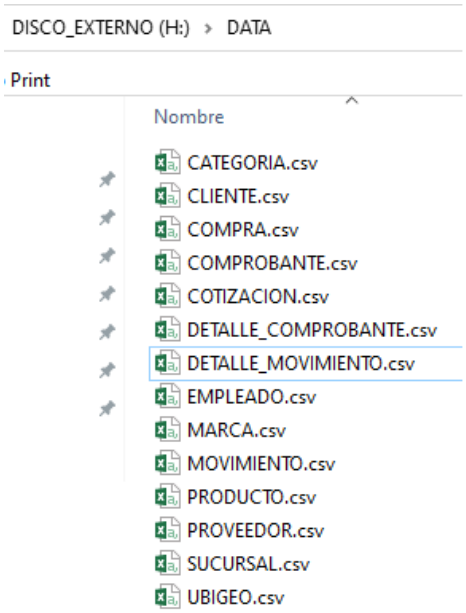


Figura N° 07. Base de datos de la empresa.

Archivos CSV a utilizar extraídos desde la base de datos



4.2.2. Estructura de los Archivos a utilizar

- ✓ **Archivo: CATEGORIA.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Categoria	Código de la Categoría del producto	A012
Nom_Categoria	Nombre de la categoría	Aridos,Cemento para todo tipo de suelos salitrosos
Descrip_Categoria	Descripción de la categoría	convencionales
Cuenta_Inventario	Número de conteo en el inventario	12

- ✓ **Archivo: CLIENTE.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Cliente	Código del cliente	1
Cod_Ubigeo	Código del Ubigeo	U003
Cod_Sucursal	Código de sucursal	SM01
Nom_Cliente	Nombre del cliente	Graciela Natural,
Apelli_Cliente	Apellido del cliente	Abanto Méndez
Tipo_Cliente	Tipo de cliente	Cliente Natural
NroCuenta_Cliente	Número de cuenta cliente	

- ✓ **Archivo: MARCA.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Marca	Código de la marca	1
Nom_Marca	Nombre de la marca	Vencedor
Descrip_Marca	Descripción de la marca	Pinturas

- ✓ **Archivo: COMPRA.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Compra	Código de compra	20-1163

Cod_Proveedor	Código de Proveedor	1
Cod_Empleado	Código del empleado	16
LugarPago_Compra	Lugar donde se pagó la compra	Ferretería SM
Tiempo_Entrega	Tiempo de entrega de la compra	1 Semana
IGV_Compra	IGV generado por la compra	19
Fecha_Compra	Fecha de la compra	2019-10-08
Monto_Compra	Monto de la compra	3125

- ✓ **Archivo: COMPROBANTE.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Nro_Comprobante	Número del comprobante	30-1100
Cod_Cliente	Código del cliente	1
Cod_Empleado	Código del empleado	1
Cod_Sucursal	Código de la Sucursal	SM01
Fecha_Comprobante	Fecha del comprobante	2019-10-16
IGV_Venta	IGV de la venta	19

- ✓ **Archivo: SUCURSAL.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Sucursal	Código de la sucursal	SM03
Nom_Sucursal	Nombre de la sucursal	SM Matizados
Telef_Sucursal	Teléfono de la sucursal	44286562
Direc_Sucursal	Dirección de la sucursal	Emporio Albarracín Cdra.03

✓ **Archivo: COTIZACION.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Compra	Código de compra	20-1163
Cod_Sucursal	Código de sucursal	SM01
Descrip_ProdCom	Descripción del producto comprado	Vencedor Látex Vencelatex Colores Variados
CantUnid_Compradas	Cantidad de unidades compradas	25
Precio_Material	Precio del material	32

✓ **Archivo: DETALLE_COMPROBANTE.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Nro_Comprobante	Numero de comprobante	30-1100
Cod_Producto	Código del producto	P0001
Cod_Servicio	Código del servicio	S001
CantUnid_Vendidas	Cantidad de unidades vendidas	2

✓ **Archivo: DETALLE_MOVIMIENTO.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Movimiento	Código del movimiento	MOV0001
Cod_Producto	Código del producto	P0001
Cod_Compra	Código de compra	20-1163
Recibido	Estado del producto	True
Cant_Movimiento	Cantidad de movimiento	25
Estado	Estado del producto	Buen Estado
Condición	Condición del producto	En Almacén

- ✓ **Archivo: EMPLEADO.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Empleado	Código del empleado	1
Cod_Cargo	Código del cargo	V001
Cod_Sucursal	Código de la sucursal	SM01
Nom_Empleado	Nombre del empleado	Lilia
Apelli_Empleado	Apellido del empleado	Abanto Alcántara
Dni_Empleado	DNI del empleado	54786765
Telef_Empleado	Teléfono del empleado	44285459
Email_Empleado	Email del empleado	labantoa@hotmail.com
FechaIng_Empleado	Fecha de ingreso del empleado	2019-04-13

- ✓ **Archivo: MOVIMIENTO.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Movimiento	Código del movimiento	MOV0001
Cod_TipoMovimiento	Código del tipo de movimiento	TMOV1
Cod_Empleado	Código del empleado	26
Cod_Sucursal	Código de la sucursal	SM01
Cod_Area	Código del área	I04
Fecha_Movimiento	Fecha de movimiento	2019-10-15

- ✓ **Archivo: UBIGEO.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Ubigeo	Código del ubigeo	U001
Distrito	Distrito	NATASHA
Provincia	Provincia	Trujillo

✓ **Archivo: PRODUCTO.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Producto	Código del producto	P0001
Cod_Categoria	Código de la categoría	P001
Cod_Area	Código del Área	I04
Cod_Marca	Código del Marca	1
Cod_Proveedor	Código del proveedor	1
Nro_GuiaInt	Numero de guía interna	45-7764
Nom_Producto	Nombre del producto	Látex Vencelatex
Descrip_Producto	Descripción del producto	Colores Variados
UnidMedida_Producto	Unidad de medida del producto	GL ,
PrecioUnit_Producto	Precio unitario del producto	38
FechaIng_Producto	Fecha de ingreso del producto	2019-10-15
Existencias_Minimas	Existencia mínima de producto	3
Suspendido_Producto	Producto suspendido	False
Stock_Producto	Stock de productos	25

✓ **Archivo: PROVEEDOR.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Cod_Proveedor	Código del proveedor	1
Cod_Ubigeo	Código de ubigeo	U003
Nom_Proveedor	Nombre del proveedor	ASEFE
Apelli_Proveedor	Apellido de proveedor	Gomez Miño
DirecEmp_Proveedor	Direecion proveedor	Mz. D Lt.15 URB.La Arboleda

CodPost_Proveedor	Codigo Post Proveedor	B159
Telef_Proveedor	Telefono del proveedor	44284186
Email_Proveedor	Email del proveedor	asefedistribuidores@hotmail.com
NroCuenta_Proveedor	Numero cuenta del proveedor	1264736343556784

4.3. FASE 3: DISEÑO DE LA SOLUCIÓN

Para el diseño de la arquitectura de la solución a emplear se ha considerado los siguientes servicios de Microsoft Azure:

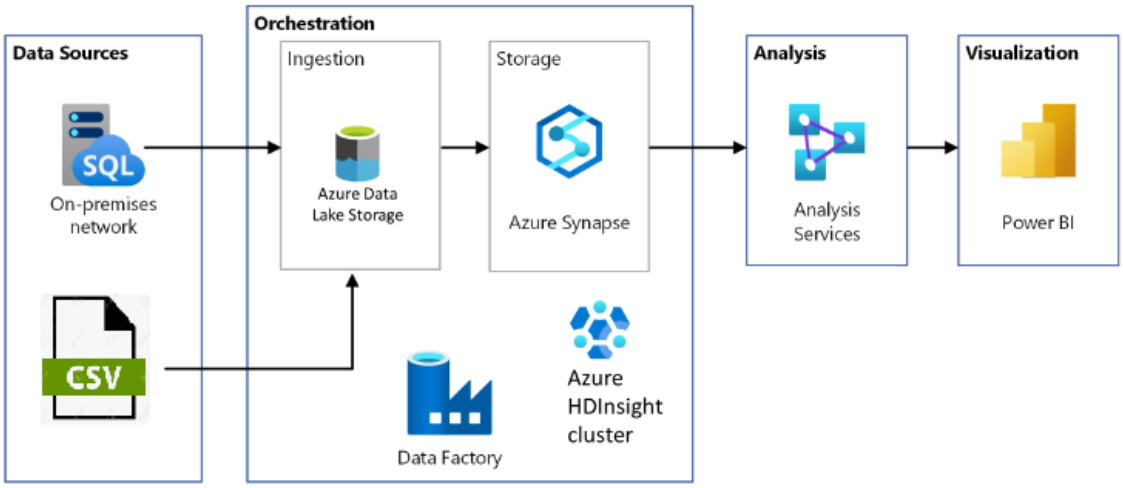
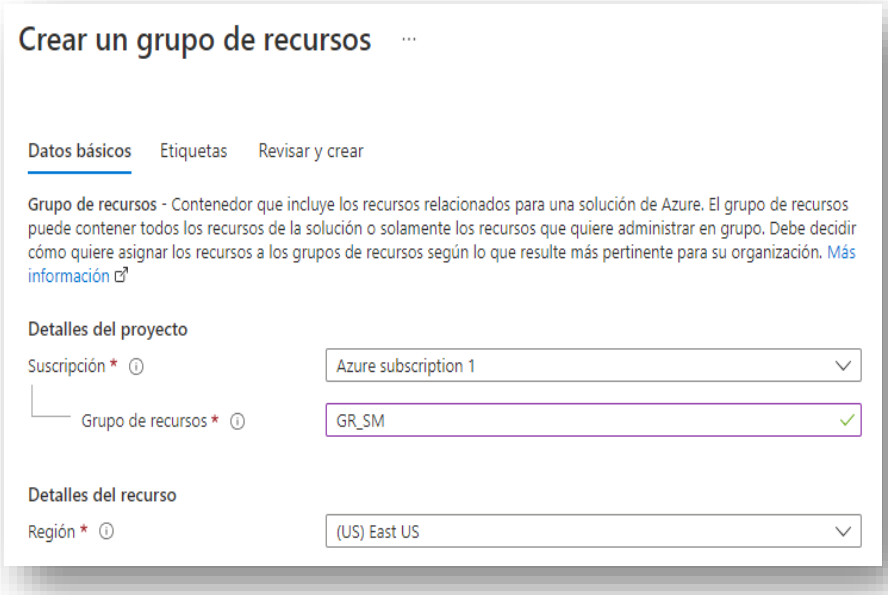


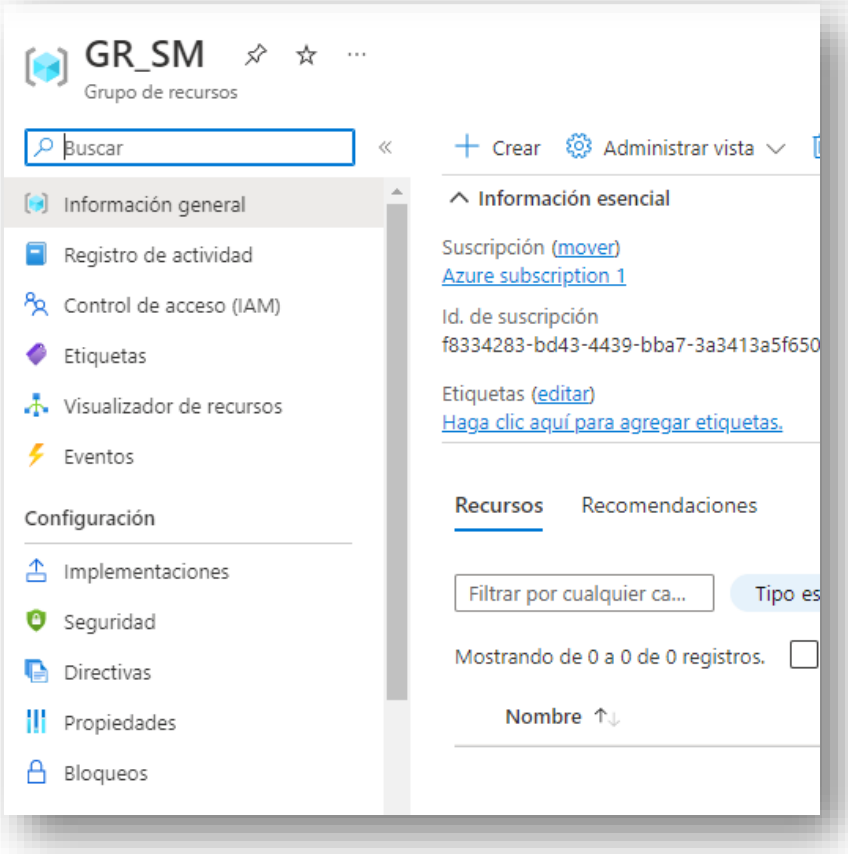
Figura N° 08: Diseño de la arquitectura de la solución

Fuente: Elaboración propia.

4.3.1. GRUPO DE RECURSOS EN AZURE

Para la creación de los servicios o recurso de Azure es necesario tener primero un grupo de recursos, por lo que el primer paso será crear y configurar un Grupo de Recursos para nuestra solución llamado “GR_SM”





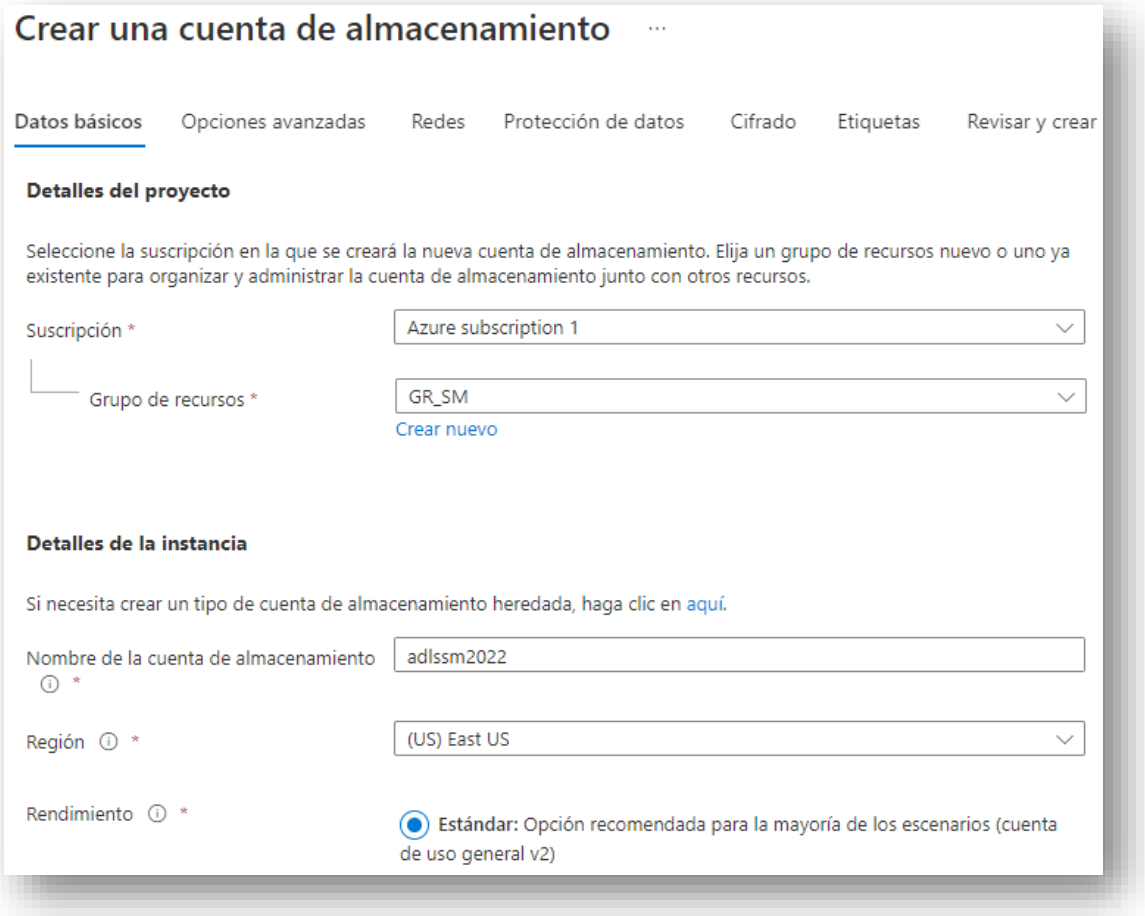
Finalmente, ya tenemos preparado el Grupo de recursos para los servicios a implementar.

4.3.2. AZURE DATA LAKE STORAGE

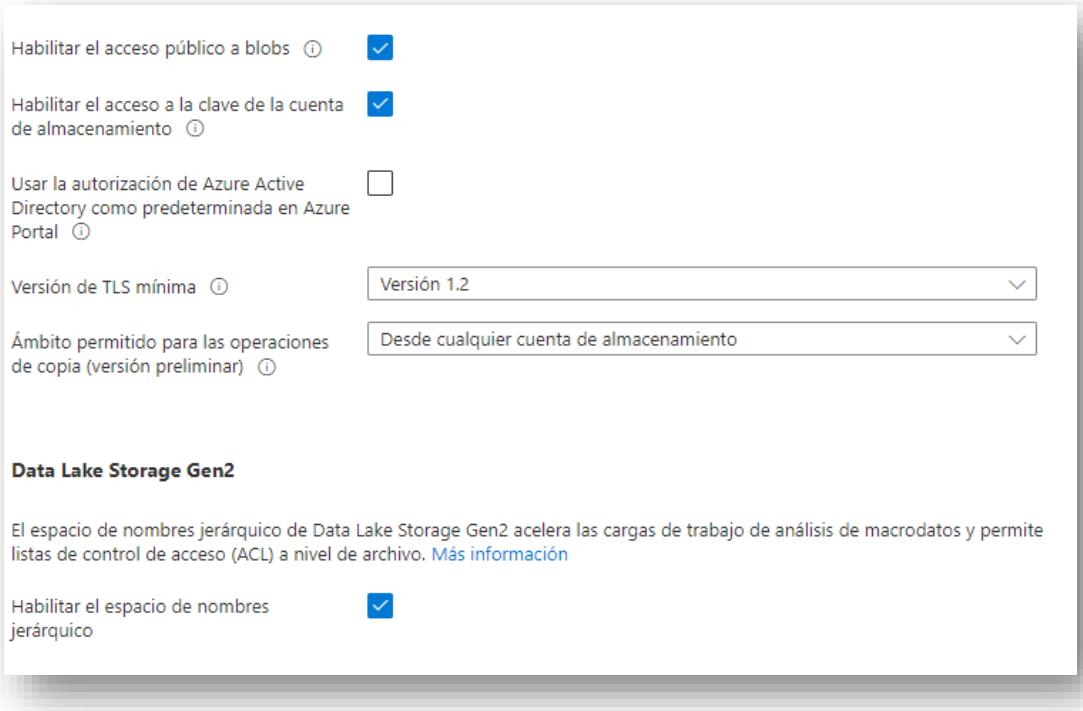
El recurso de data lake storage nos va a servir como un repositorio de almacenamiento para una gran cantidad de datos en bruto y que luego lo utilizaremos en la configuración de HDInsight.

La creación del Data Lake pasa por los siguientes pasos:

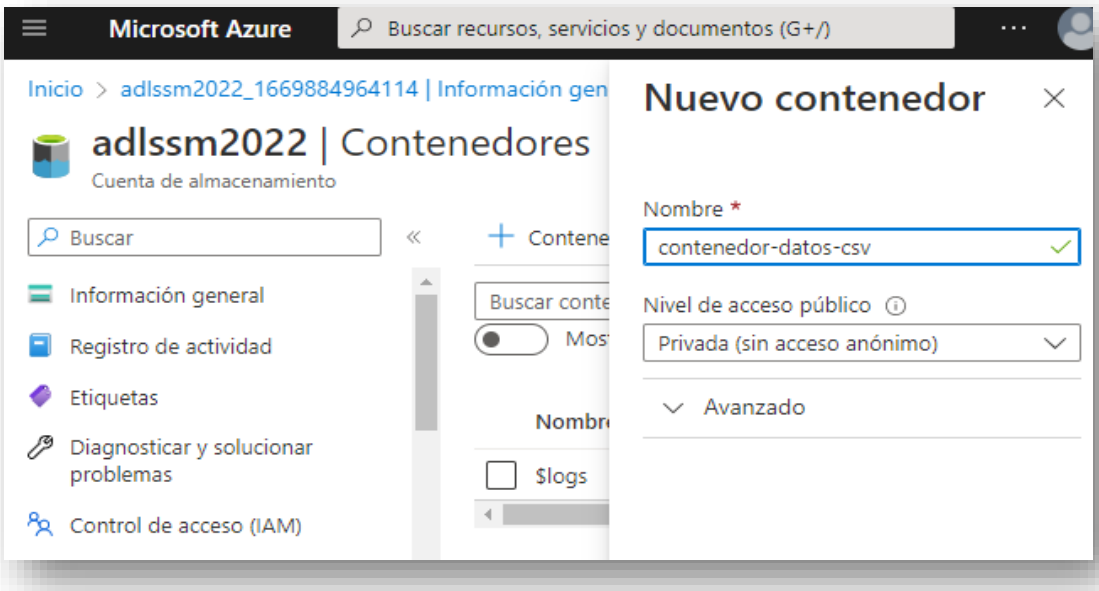
Primero creamos una “cuenta de almacenamiento”: adlssm2022



Luego para Habilitar el espacio de nombres jerárquico vamos a las opciones avanzadas y lo habilitamos, de esta manera se acelera las cargas de trabajo para el análisis de macrodatos y además permite utilizar listas de control de acceso (ACL) a nivel de archivo.

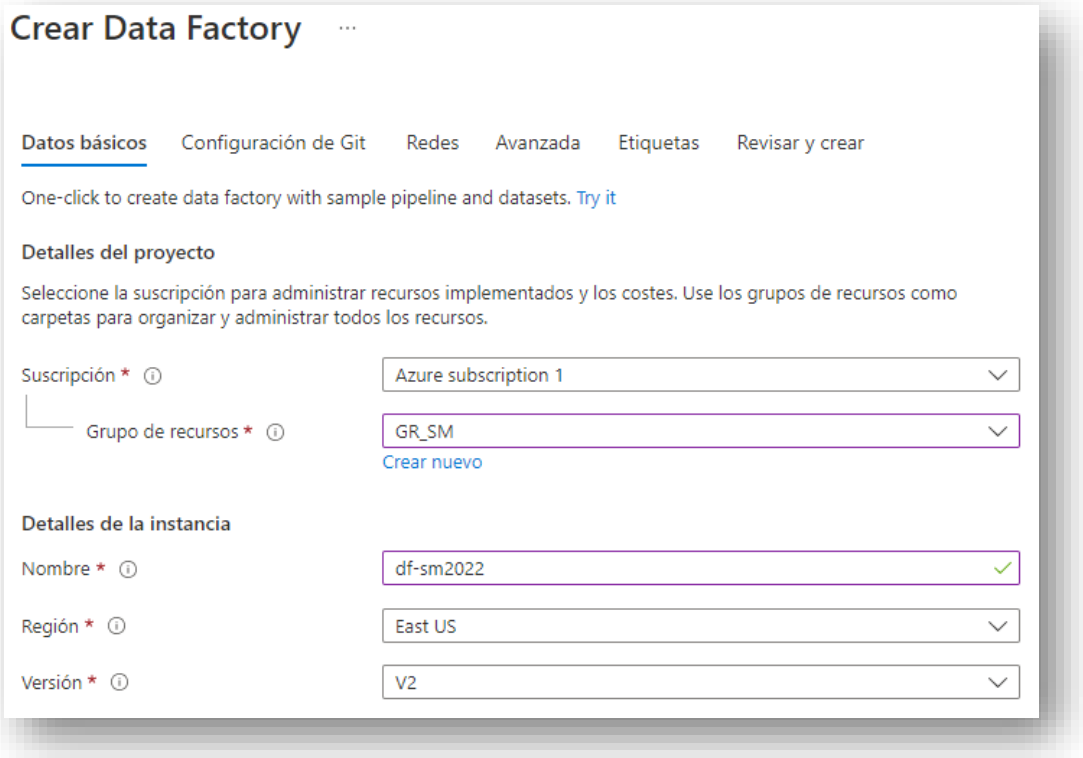


El Data Lake creado tendrá que tener contenedores don estarán los datos sin procesar y procesados.



4.3.3. AZURE DATA FACTORY

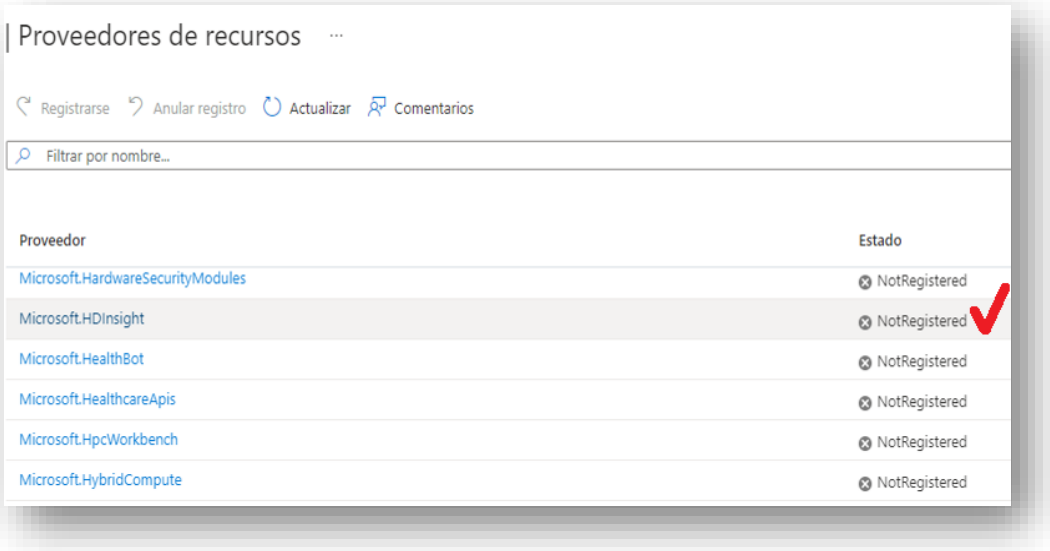
Azure Data Factory nos va permitir la integración y transformación de datos.



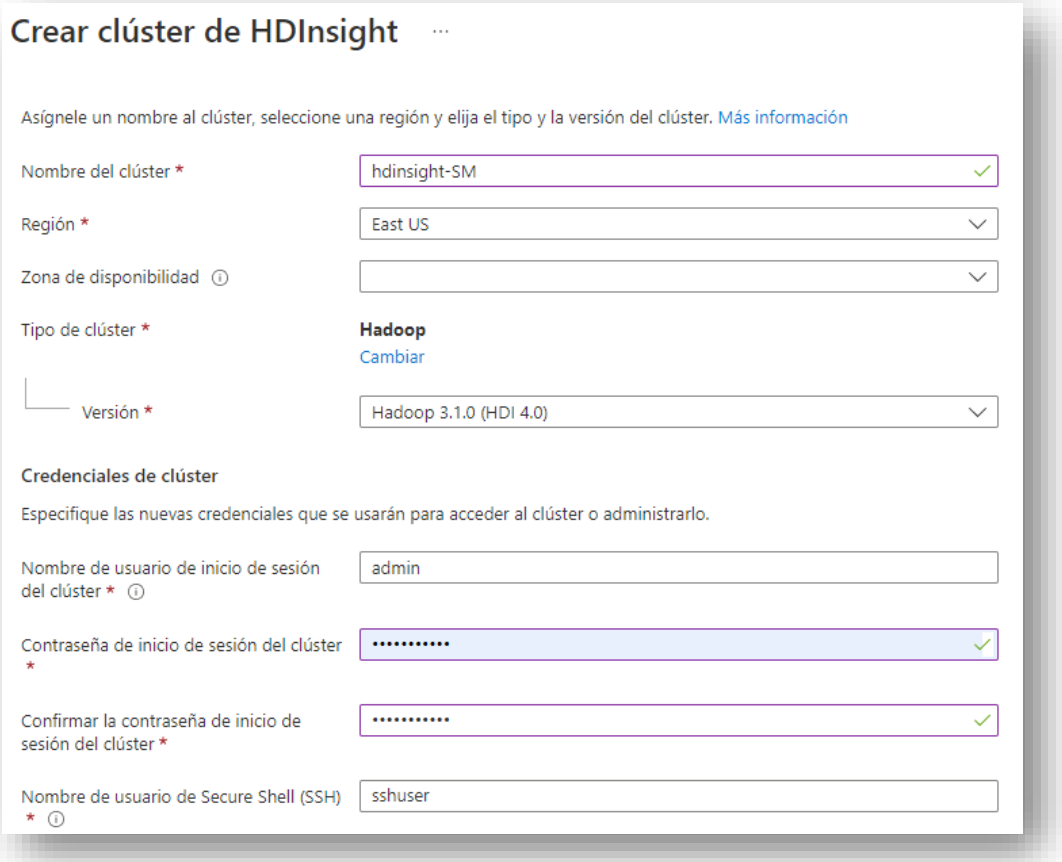
4.3.4. AZURE HDINSIGHT

Microsoft Azure integra dentro de sus servicios una distribución de nube con componentes de Apache Hadoop, llamado HDInsight. Este servicio permite que sea fácil, rápido y rentable procesar grandes cantidades de datos de manera personalizable.

Además, HDInsight permite crear clústeres para Hadoop, Spark, Interactive query (LLAP), Kafka y HBase en Azure. En nuestro caso para nuestra solución crearemos cluster en Hadoop, pero antes de poder utilizar el recurso debemos de registrar la suscripción al tipo de recurso para habilitar su creación y uso.

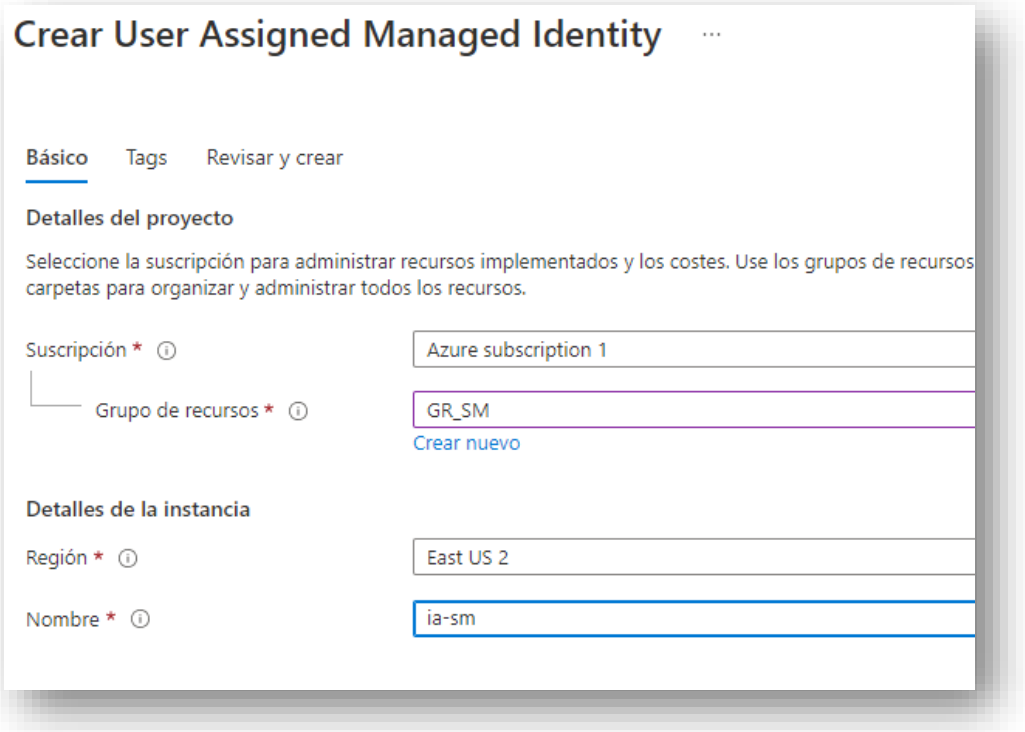


Luego de registrarlo recién se puede realizar la creación del cluster de Hadoop

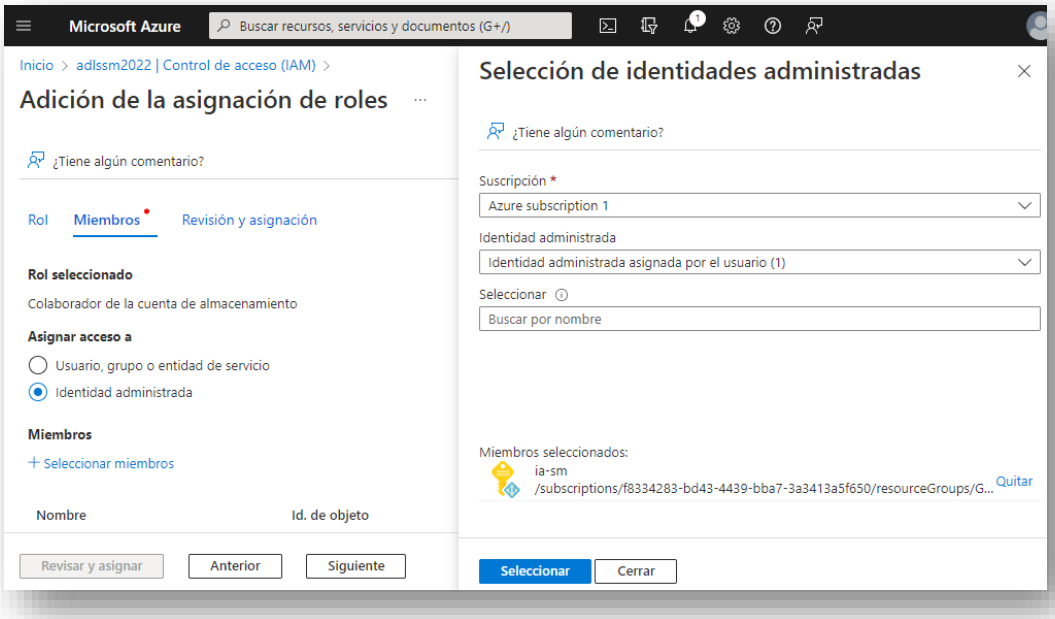


Para representar el clúster con acceso a la cuenta de almacenamiento de Azure Data Lake Gen2, se debe de seleccionar una identidad administrada asignada por el usuario. En caso de no tenerla se le debe de configurar.

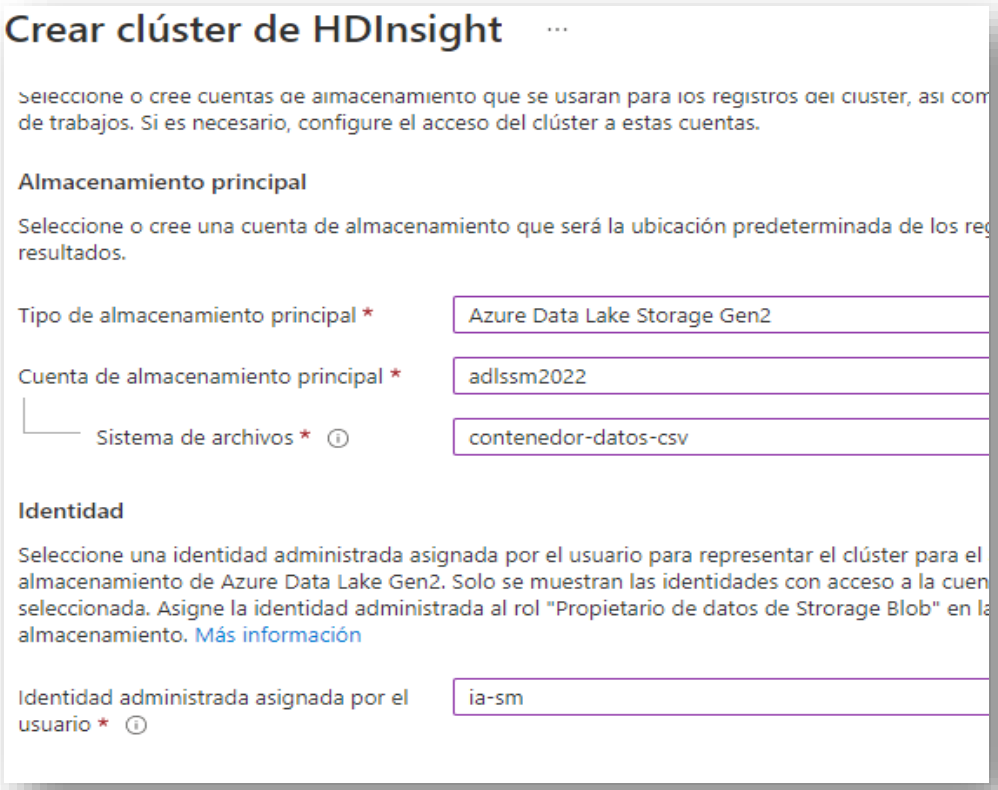
Para nuestro recurso asignamos la identidad administrada al rol "Propietario de datos de Strorage Blob" en la cuenta de almacenamiento.



Luego se debe de agregar en el Data Lake el acceso de control al usuario asignado para el administrador de identidades.



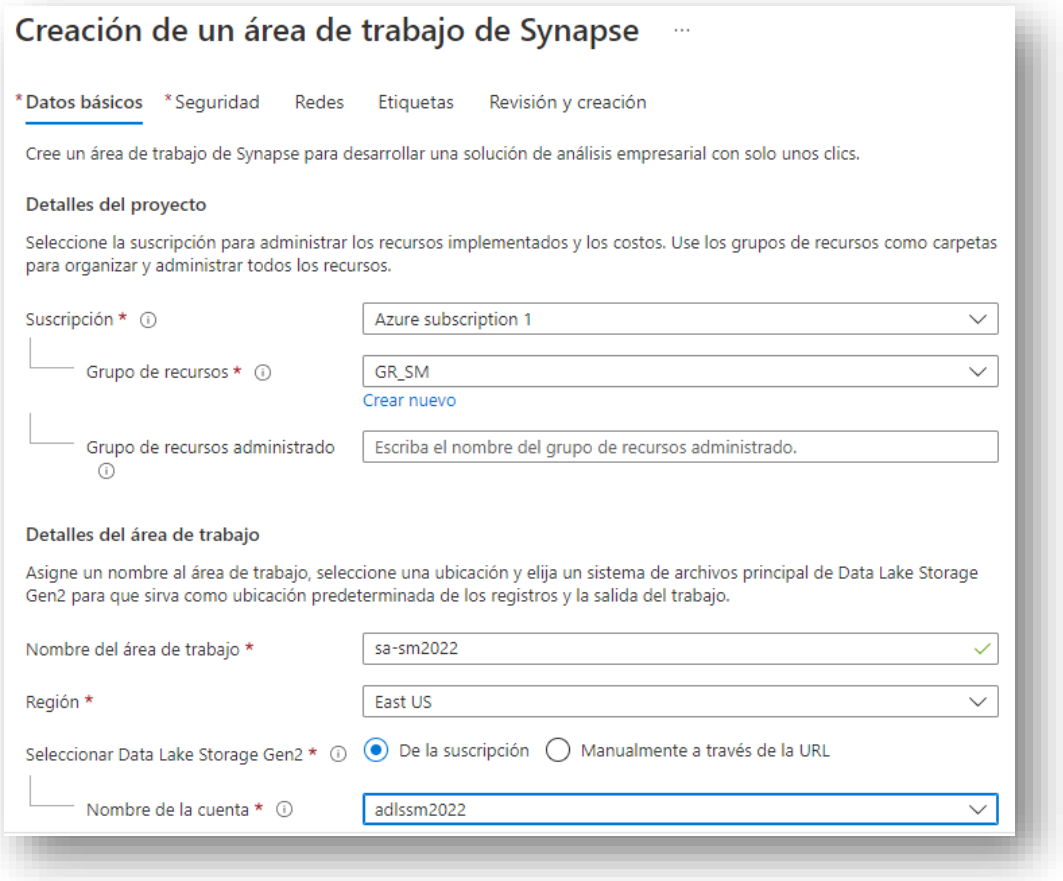
En la configuración del cluster de HDInsight se selecciona el Almacenamiento y la Identidad que lo va a administrar



4.3.5. AZURE SYNAPSE ANALYTICS

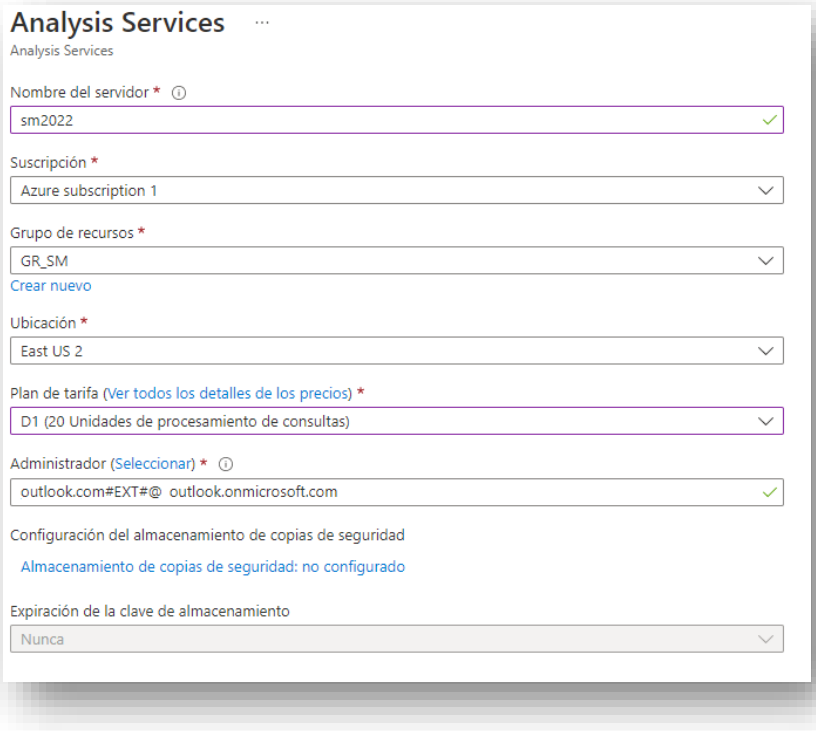
Azure Synapse Analytics e la evolución de SQL Data Warehouse, cuyo objetivo es unir las capacidades del datawarehousing empresarial con las capacidades de análisis de Big Data.

Crearemos el servicio para nuestra solución



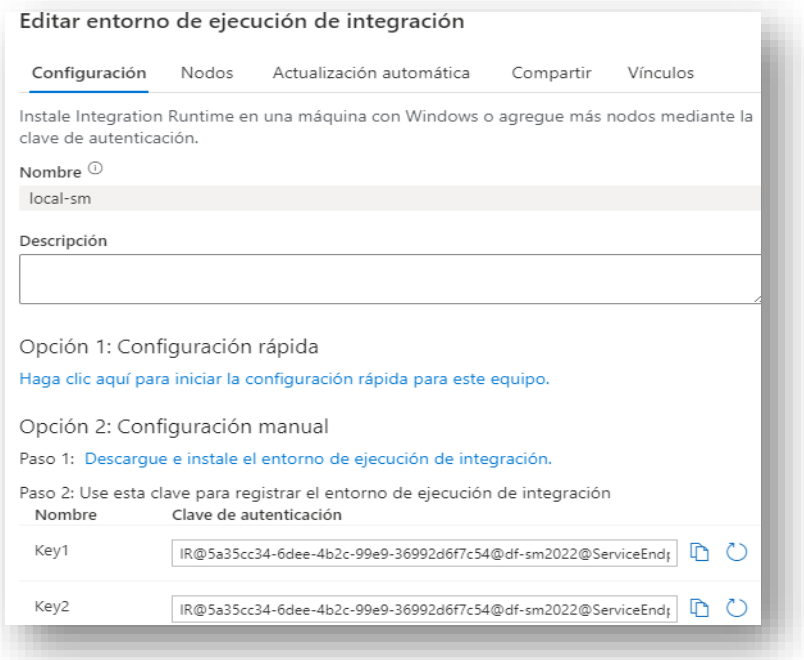
4.3.6. AZURE ANALYSIS SERVICES

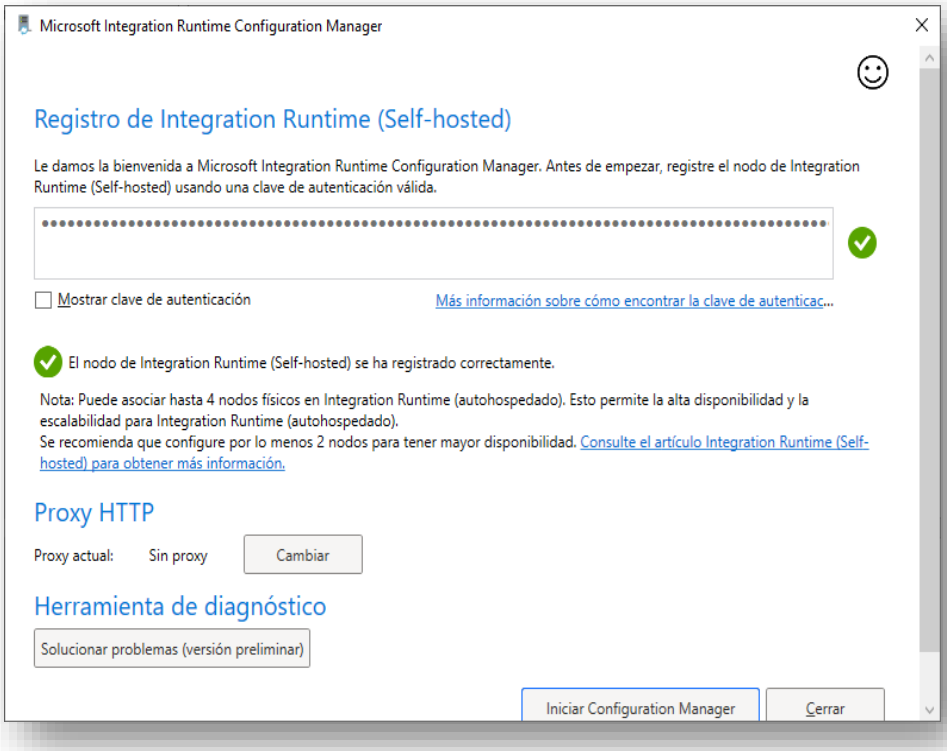
Azure Analysis Services es una plataforma como un servicio (PaaS) completamente administrada que proporciona modelos de datos en la nube de nivel empresarial.



4.4. FASE 4: CAPTURA Y ALMACENAMIENTO DE DATOS

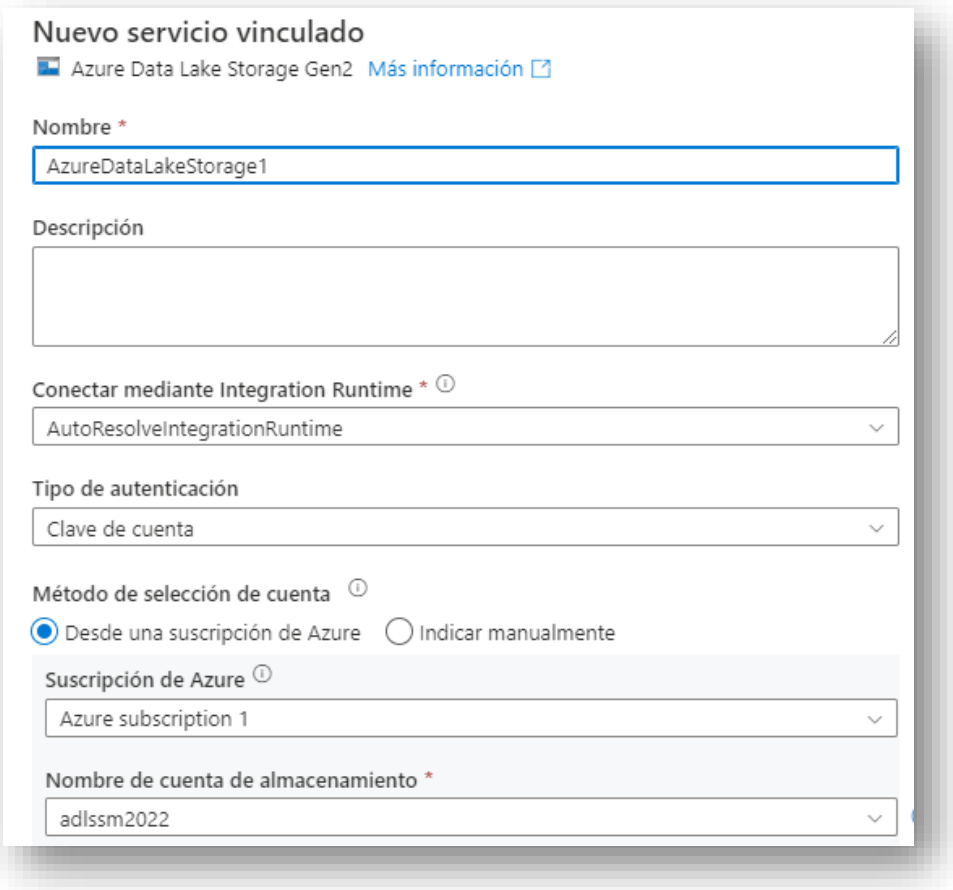
Para la ingesta de datos se va a utilizar el servicio de Data Factory, pero antes de debe de configurar un Registro de Integración Runtime y luego la conexiones con los servicios vinculados



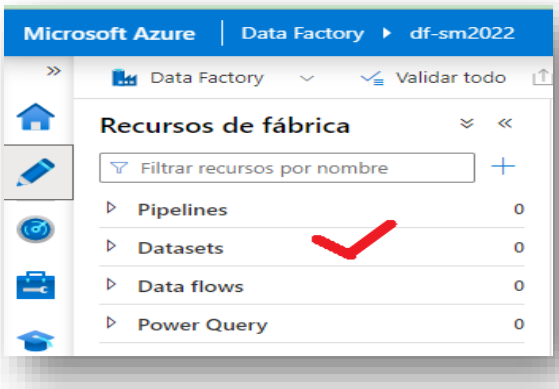


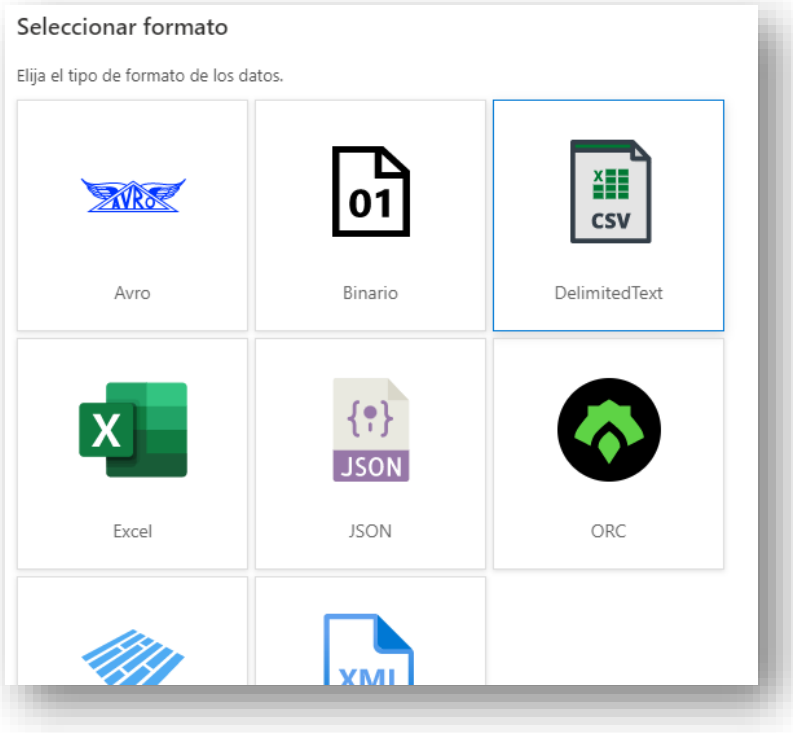
Luego se ha configurado un nuevo Linked Services que apunta a los archivos de orígenes de datos (Archivos csv que están en el host)

También se crea un Linked Service que enlace al Data Lake y al contenedor donde se alojaran los datos

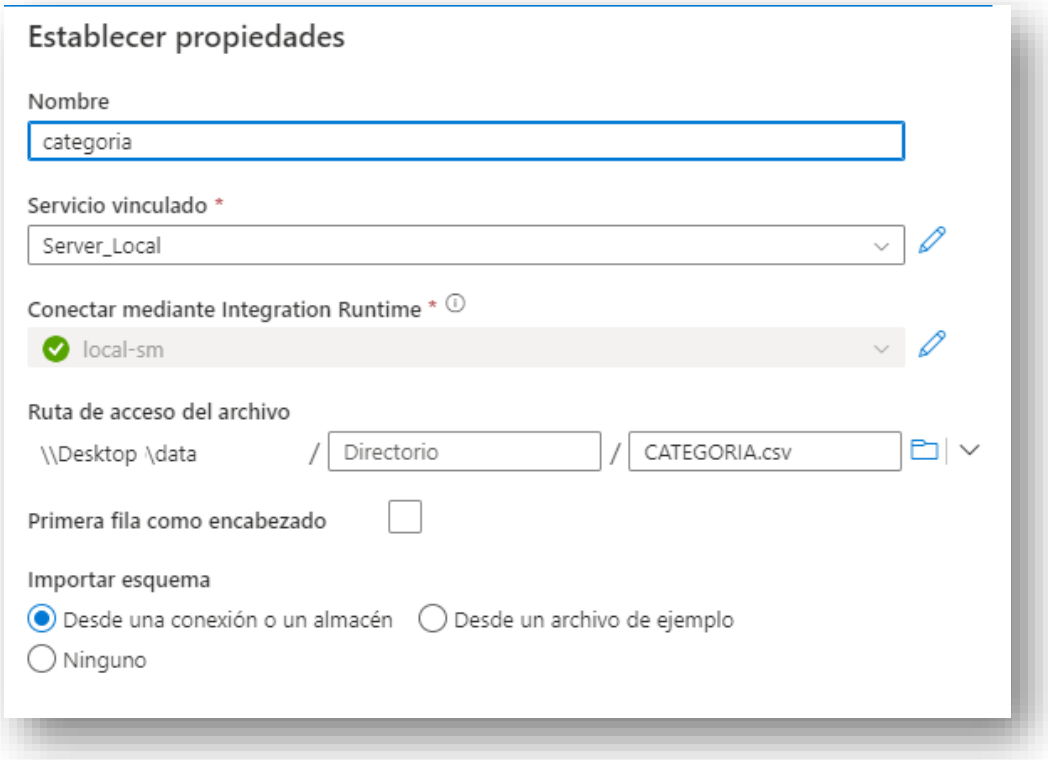


Para realizar la ingesta de los datos también es necesario crear las conexiones para la representación de los datasets, para luego crear una canalización par la ingesta de datos de manera directa.

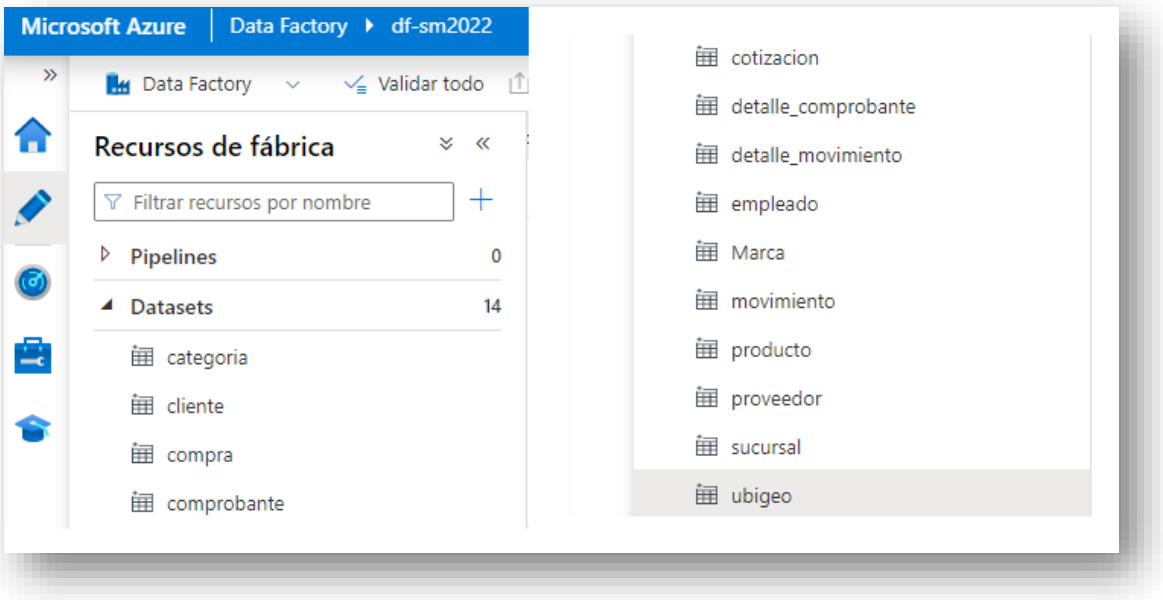




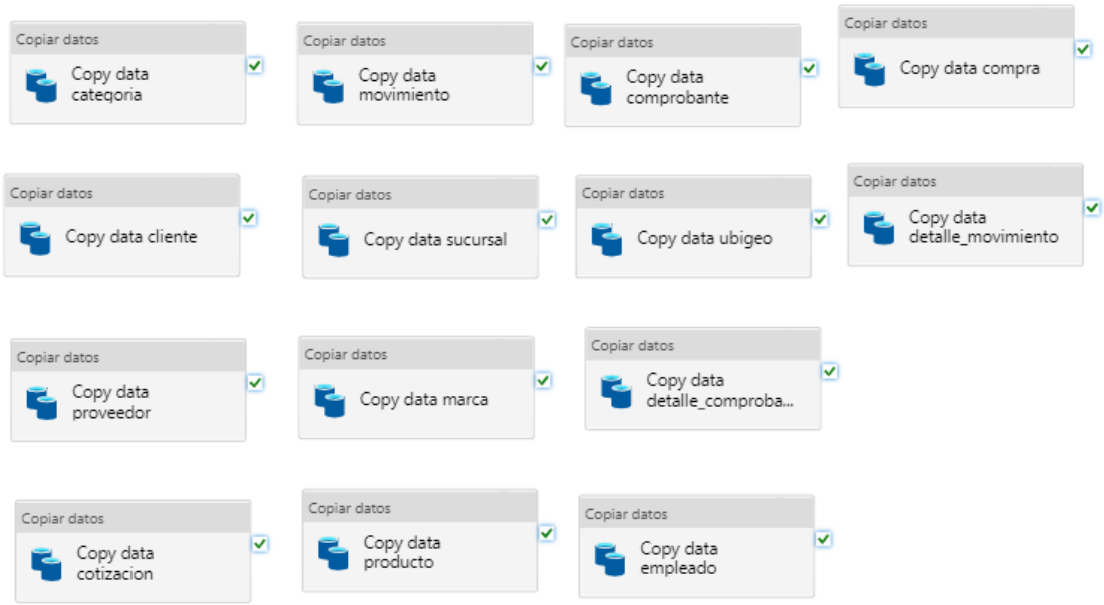
Se configura un Dataset para cada archivo csv:



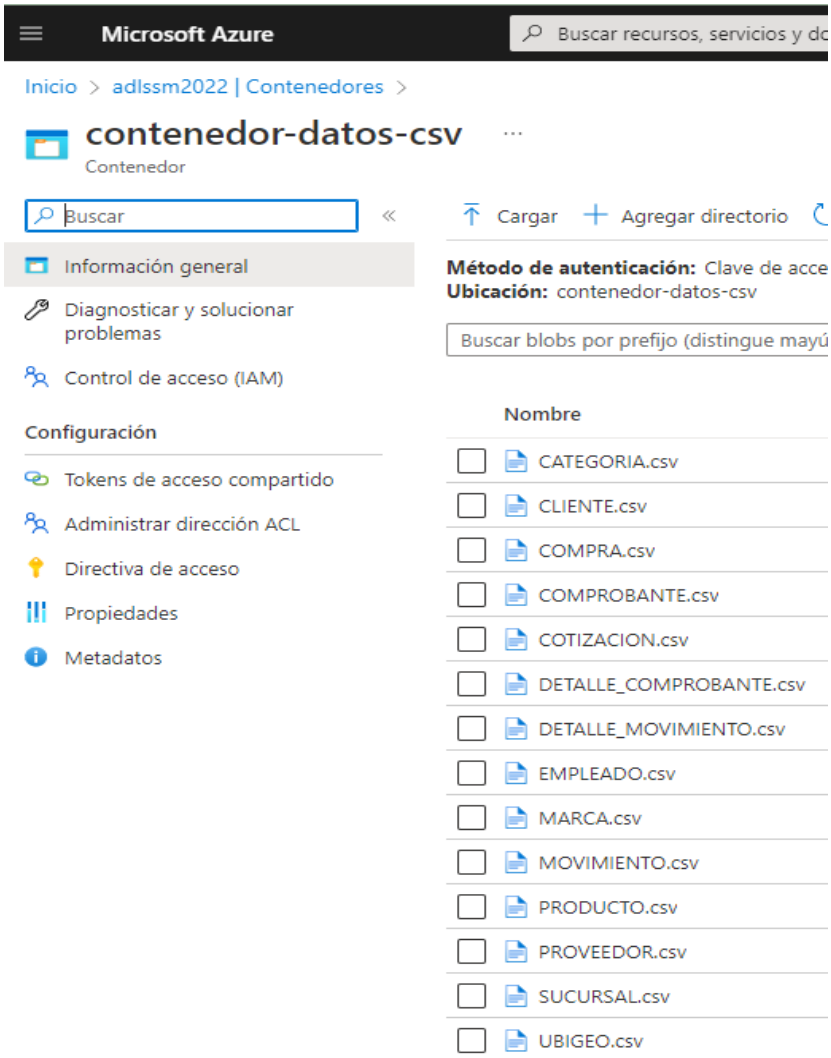
Lista de Conjuntos de datos a utilizar:



Se crea una canalización par la ingesta de datos de manera directa



Luego verificamos si se cargaron todos los datos, así como también se verifica que el contenedor haya recibido los archivos



4.5. FASE 5: MODELADO Y LIMPIEZA DE DATOS

4.5.1. Modelo Dimensional

Los criterios que se utilizarán para nombrar las dimensiones a usar:

- ✓ **"Dim":** Representa el inicio del nombre de la dimensión a utilizar.

4.5.2. Tipo de modelo lógico

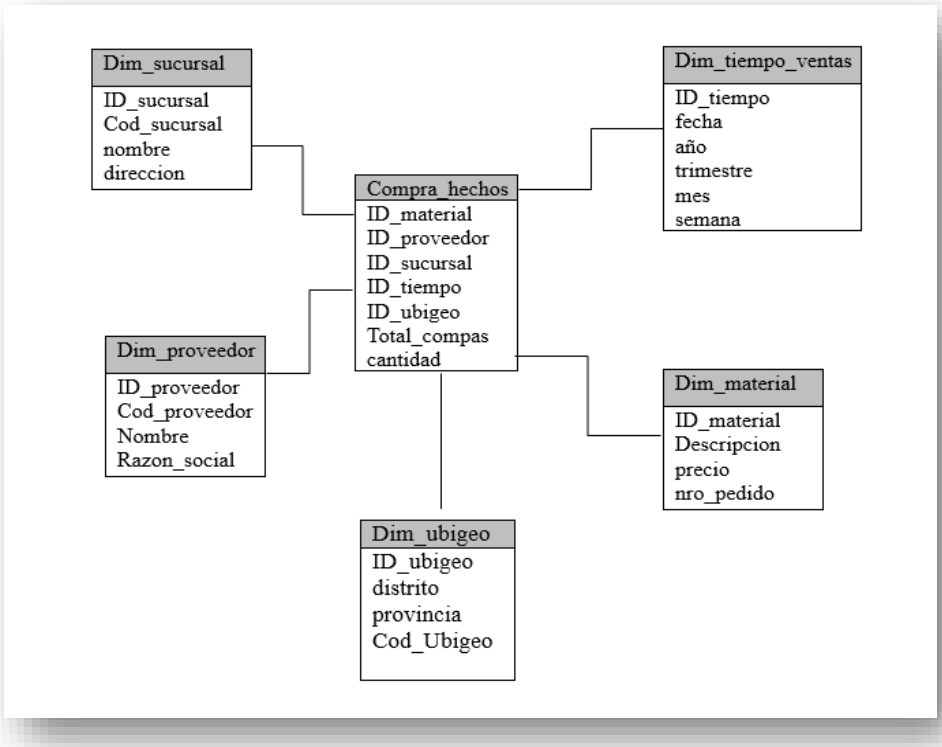


Figura N° 09. Modelo lógico de compras

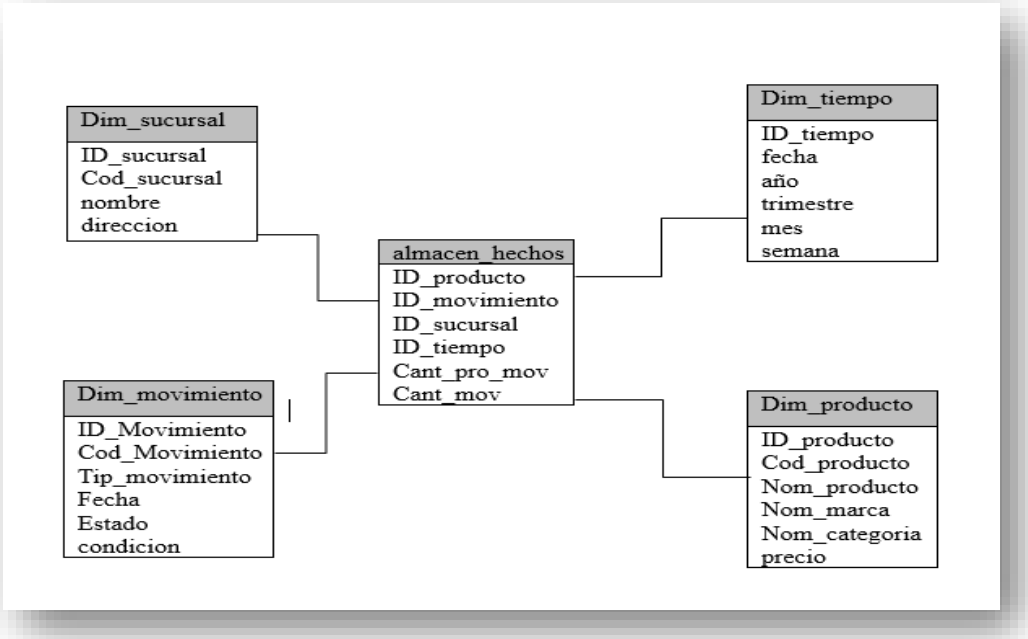


Figura N° 10. Modelo lógico de almacén

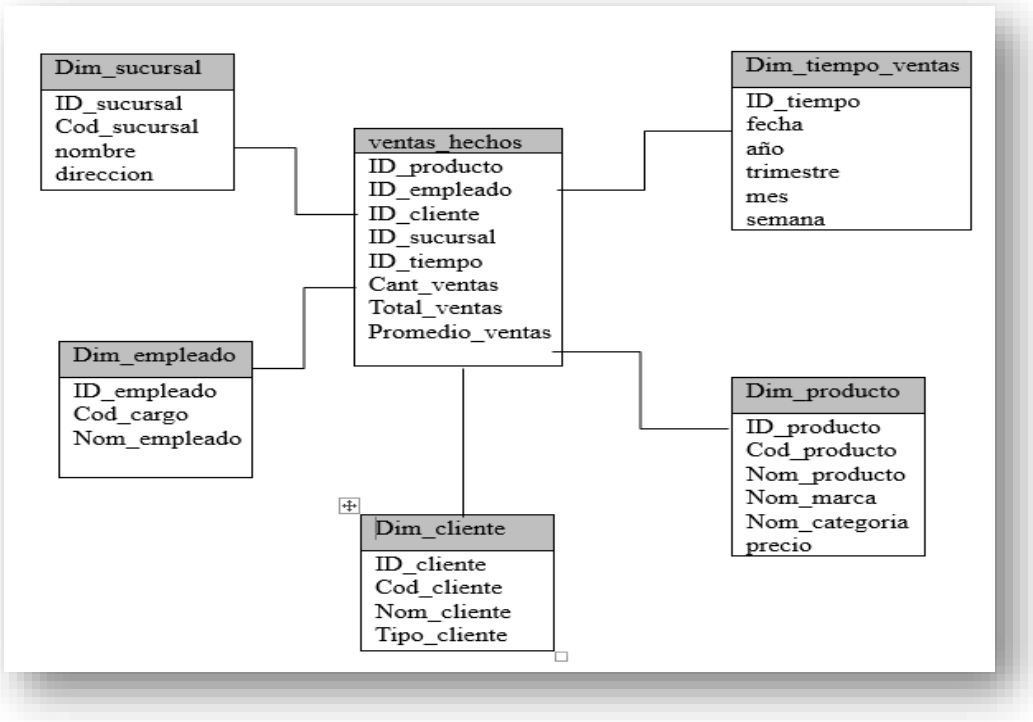
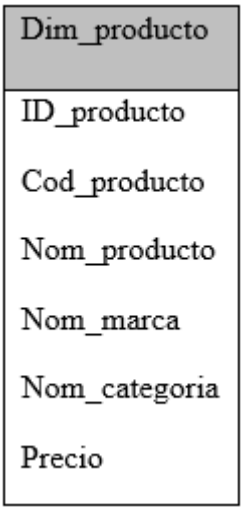


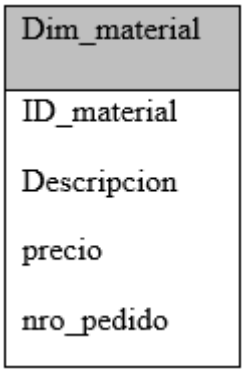
Figura N° 11. Modelo lógico de ventas

4.5.3. Tablas de dimensiones

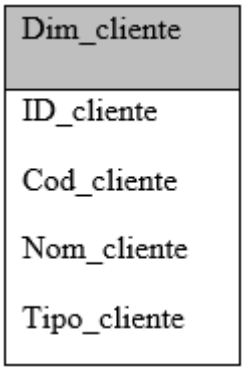
Dimensión producto :



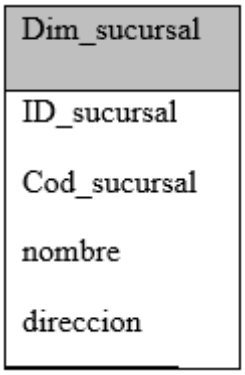
Dimensión Material : la dimension material se referirá a un producto en stock que se compró recientemente a un proveedor con los siguientes atributos.



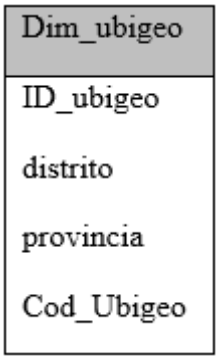
Dimensión cliente : La tabla de dimensión Clientes contendrá los clientes registrados para cada día que la ferretería Santa María tenga una venta y contendrá los siguientes atributos



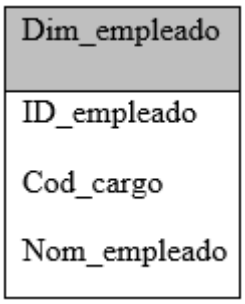
Dimensión sucursal:



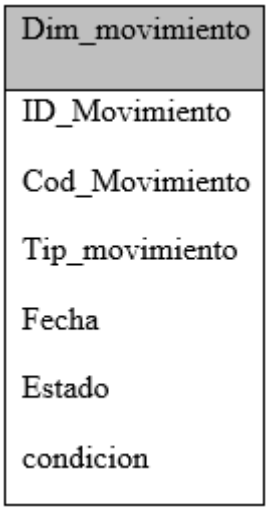
Dimensión ubigeo: Esta perspectiva contiene el área donde se encuentra la sucursal de la tienda y tendrá los siguientes atributos.



Dimensión Empleado: La tabla de dimensiones Empleados solo contendrá empleados que hayan completado ventas porque está configurada de la siguiente manera, tendrá los siguientes atributos:



Dimensión movimiento: Movimiento de productos.



Dimensión proveedor:

Dim_proveedor
ID_proveedor
Cod_proveedor
Nombre
Razon_social

Dimensión tiempo Esta dimensión utilizará 3 tablas con los mismos atributos, pero diferentes tiempos. En esta tabla, la fecha de compra venta y la fecha de almacenamiento se dividirán en diferentes tablas para que se vea más claramente el tiempo de producción de cada tabla. La mesa de procesamiento tendrá las siguientes propiedades

Estas propiedades se obtendrán por transformaciones derivadas de las respectivas fechas de las transacciones realizadas en cada proceso.

Dim_tiempo_compras
ID_tiempo
fecha
año
trimestre
mes
semana

Dim_tiempo_ventas
ID_tiempo
fecha
año
trimestre
mes
semana

Dim_tiempo_almacen
ID_tiempo
fecha
año
trimestre
mes
semana

4.5.4. Tablas de hechos

Tabla Ventas_ hechos: Contiene metricas para el proceso de Ventas y claves primarias. Metricas a utilizar:

- Cant _ventas
- Total _ventas
- Promedio de ventas

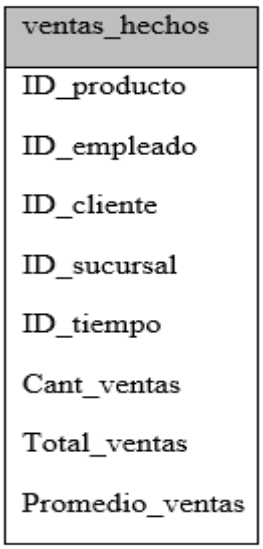


Tabla Compras_ hechos: Esta tabla tendrá 2 hechos:

- Total_compras
- Cantidad

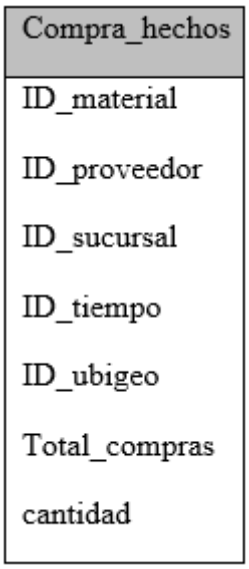
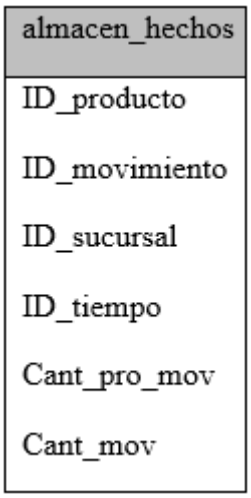


Tabla Almacen_ hechos:

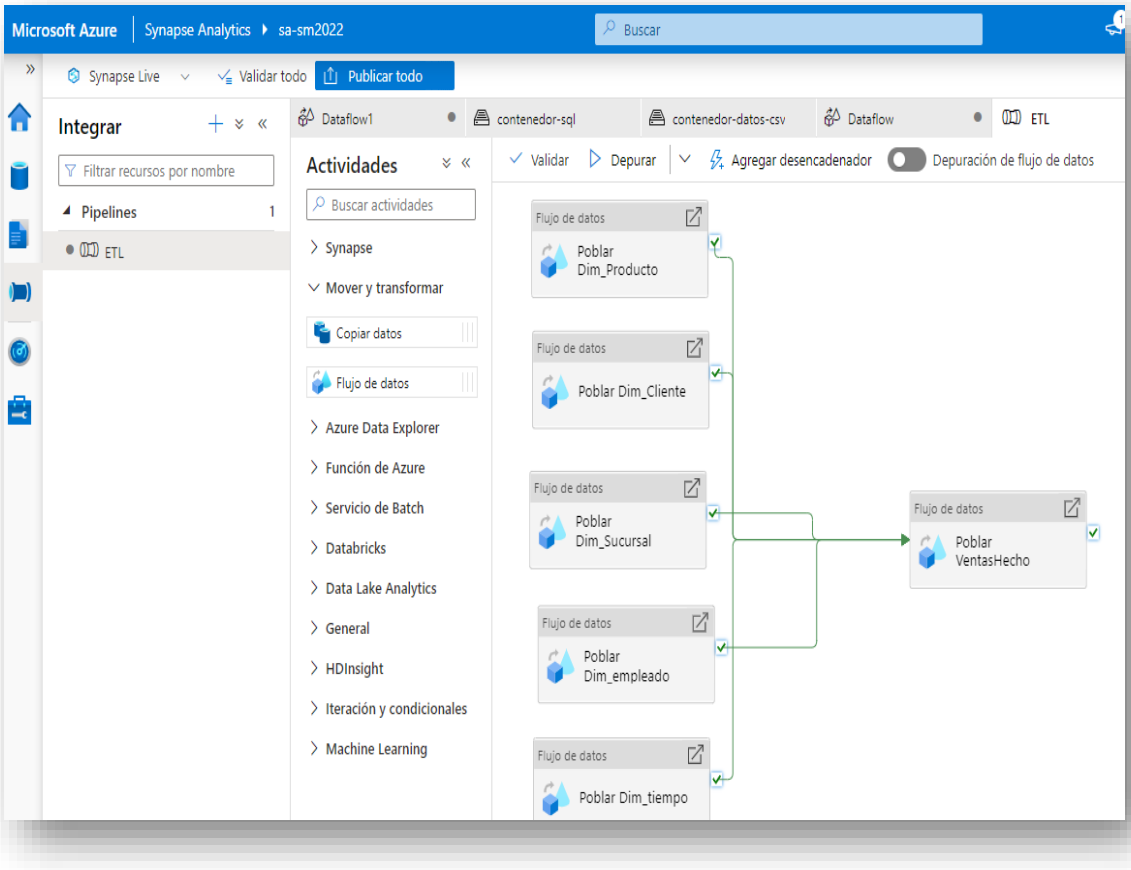
- Cant _ pro_mov
- Cant _ mov



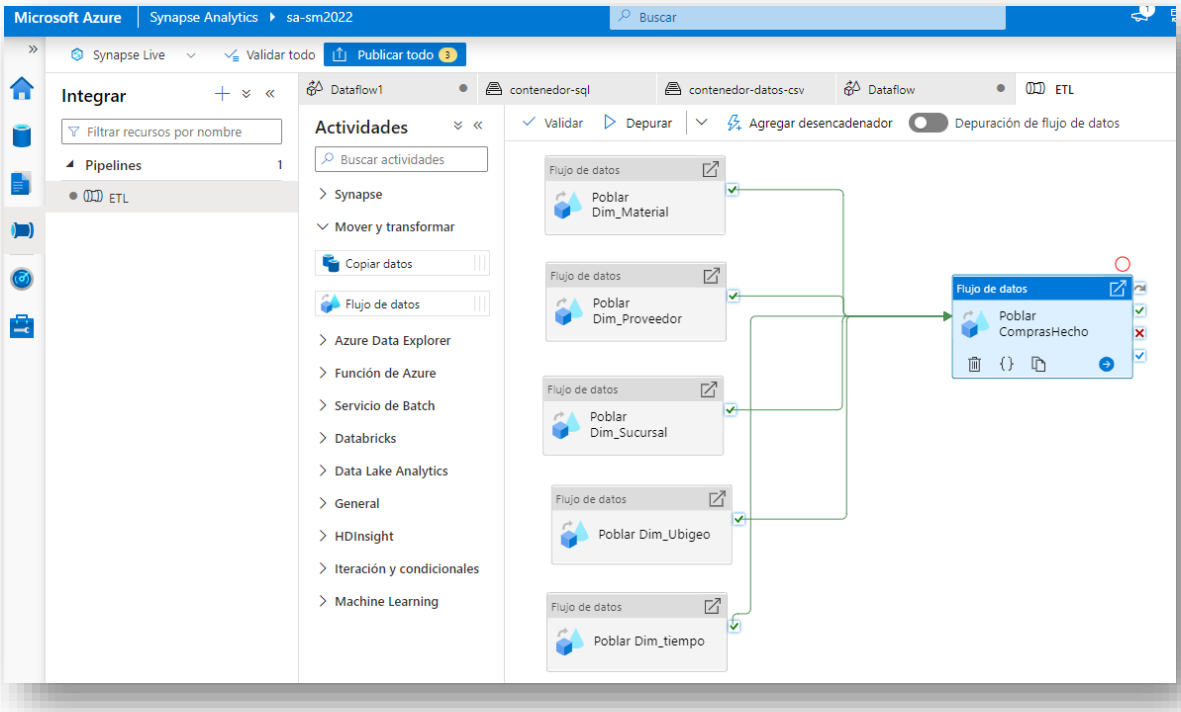
4.5.5. ETL (Extracción, Transformación y Limpieza de los Datos)

Para realizar el proceso ETL utilizaremos el servicio de Azure Synapse Analytics.

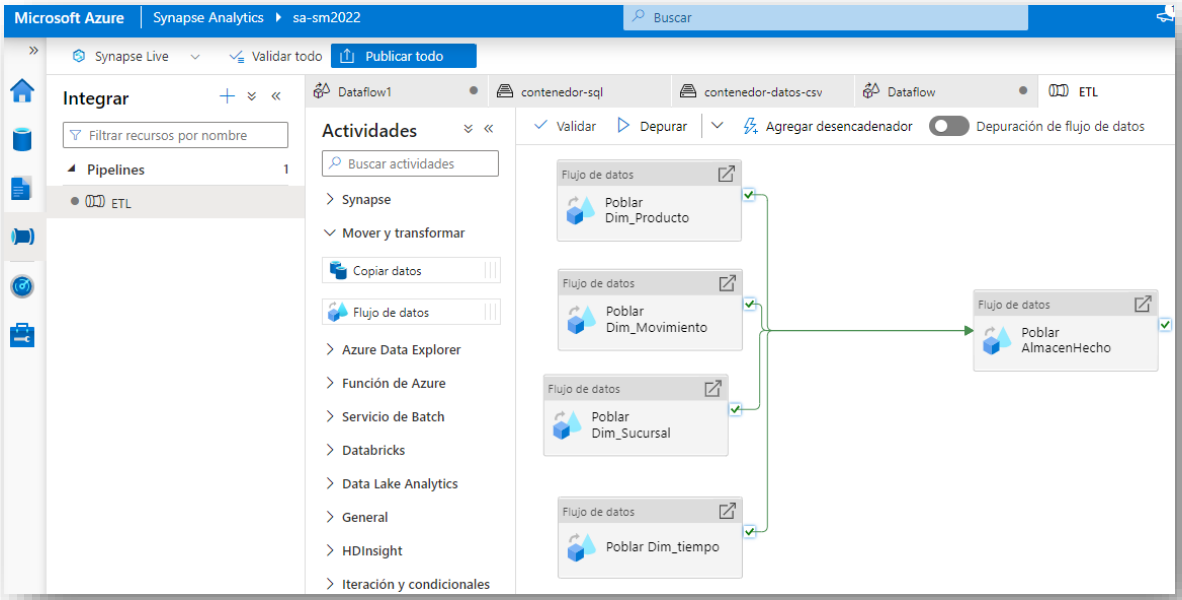
Poblamiento de Ventas_Hecho



Poblamiento de Compras_Hecho



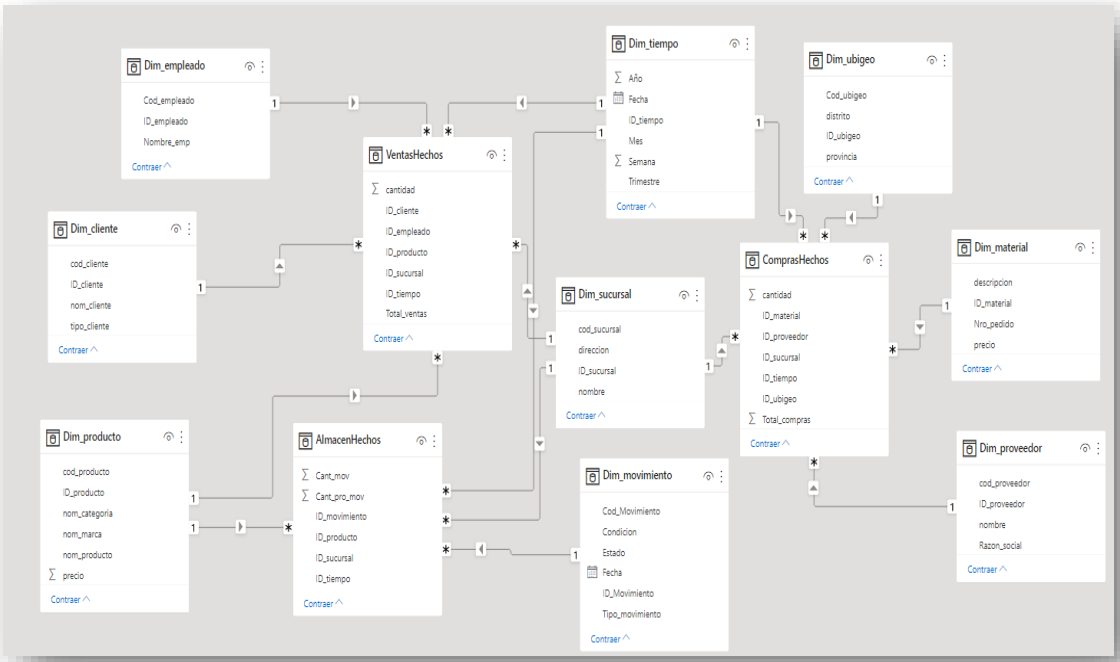
Poblamiento de Almacén_Hechos

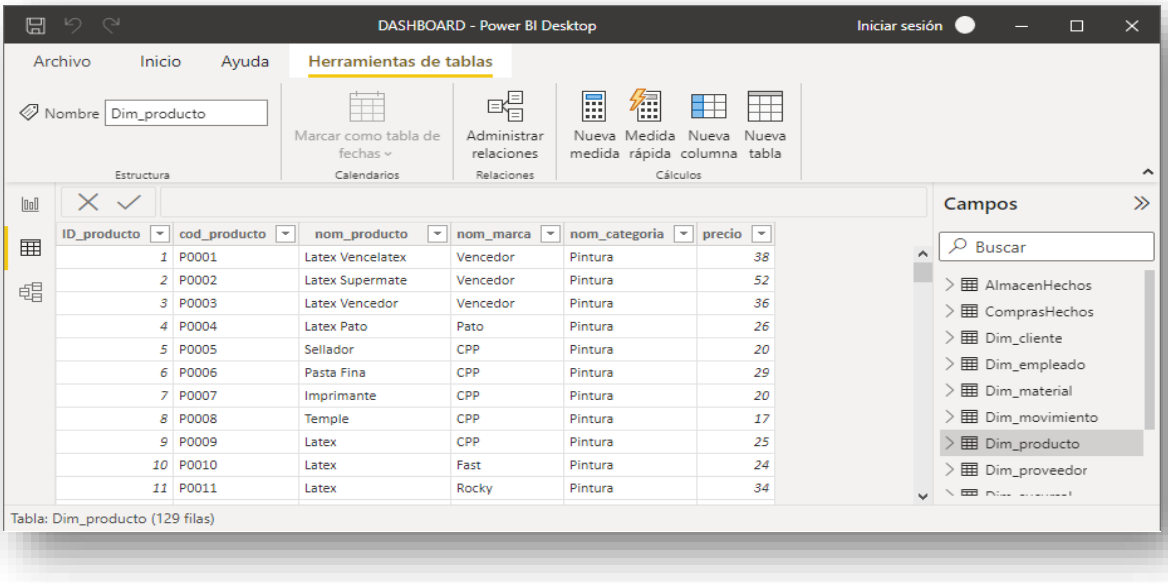


4.6. FASE 6: ANÁLISIS Y VISUALIZACIÓN

4.6.1. Modelo dimensional en Power BI

Se realizó la conexión con Power BI para crear el Modelo Dimensional siguiente:



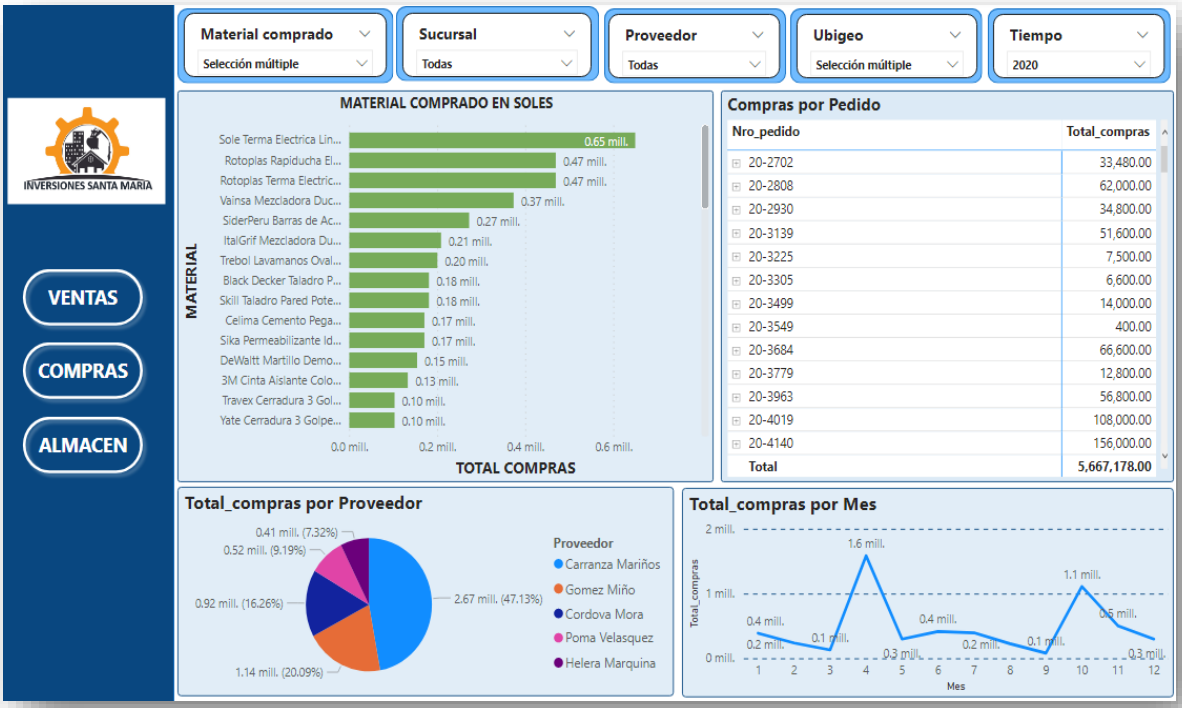


ID_producto	cod_producto	nom_producto	nom_marca	nom_categoria	precio
1	P0001	Latex Vencelatex	Vencedor	Pintura	38
2	P0002	Latex Supermate	Vencedor	Pintura	52
3	P0003	Latex Vencedor	Vencedor	Pintura	36
4	P0004	Latex Pato	Pato	Pintura	26
5	P0005	Sellador	CPP	Pintura	20
6	P0006	Pasta Fina	CPP	Pintura	29
7	P0007	Imprimante	CPP	Pintura	20
8	P0008	Temple	CPP	Pintura	17
9	P0009	Latex	CPP	Pintura	25
10	P0010	Latex	Fast	Pintura	24
11	P0011	Latex	Rocky	Pintura	34

4.6.2. Dashboard Ventas – Compras - Almacén

a. Perspectiva de Ventas

b. Perspectiva de Compra



c. Perspectiva de Almacén

5. DISCUSIÓN DE RESULTADOS

5.1.Formulación del Problema

¿Cómo brindar un mejor análisis de información al proceso de ventas en Inversiones Santa María?

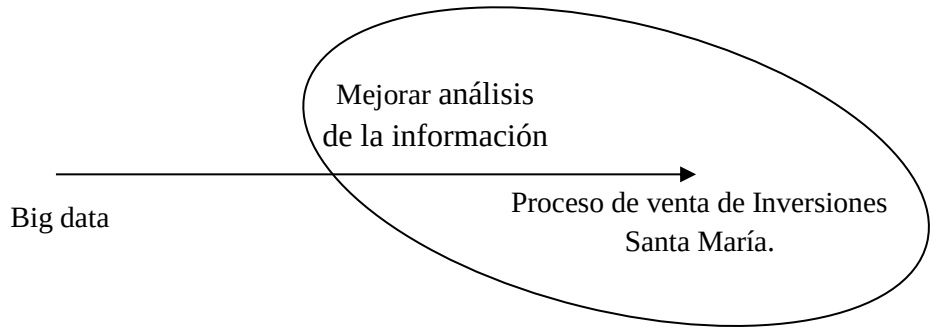
5.2.Hipótesis

H_0: "Una Solución de Big data no permite mejorar análisis de la información de los procesos de ventas de la empresa Inversiones Santa María utilizando el ecosistema de Apache Hadoop y MS Azure"

H_1: "Una Solución de Big data permite mejorar análisis de la información de los procesos de ventas de la empresa Inversiones Santa María utilizando el ecosistema de Apache Hadoop y MS Azure"

- ✓ Variable Independiente (VI): Big data
- ✓ Variable Dependiente (VD): Proceso de venta de Inversiones Santa María.

5.3.MANERA PRESENCIAL



5.4. DISEÑO PRE-EXPERIMENTAL , PRE-PRUEBA Y POST-PRUEBA

PRE-PRUEBA (O_1): Es la medición previa de X a G

POST-PRUEBA (O_2): Corresponde a la nueva medición de X a G

Se determinó usar el Diseño Pre Experimental Pre-Prueba y Post-Prueba, porque nuestra hipótesis se adecua a este diseño. Este diseño experimenta con un solo grupo de sujetos el cual es medido a través de un cuestionario antes y después de presentar el estímulo (Solución de Big Data). Este diseño se presenta de la siguiente manera:

G O_1 X O_2

Donde:

X: Tratamiento (Big Data)

O: Medición a sujetos

G: Grupo de sujetos

El espacio muestral utilizado para medir la métrica hipotética corresponde al total de personas que administrarán la solución, 2 de ellas: los jefes de los departamentos de compras y ventas y el administrador de la base de datos; antes (O1) y después de interactuar con la solución (O2), estas dos personas recibieron una encuesta de cuestionario. Al final del estudio se determina la diferencia entre O1 y O2 para determinar si los resultados obtenidos han aumentado.

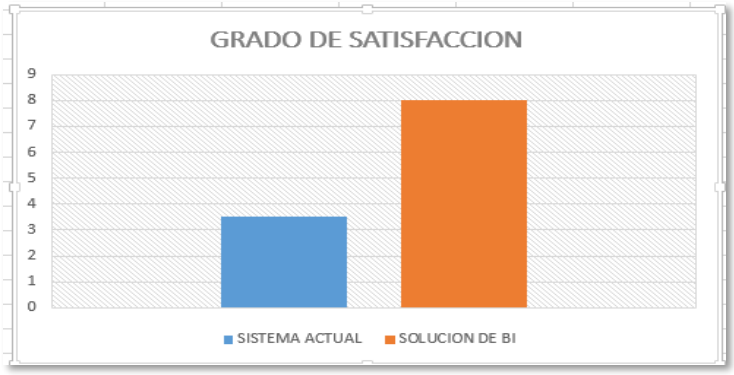
5.4.1. CÁLCULO DE LOS INDICADORES

Para calcular los indicadores en la solución Big Data (SBD) y el sistema existente (S.T), se realizó una encuesta de cuestionario y una evaluación del usuario después de la interacción con los resultados de la solución (dashboard). tabla a continuación. El rango de satisfacción mostrado utiliza el valor asignado por el usuario a las respuestas del cuestionario:

RANGO	GRADO DE SATISFACIÓN
0 – 2.5	Insatisfecho
2.6 – 5	Medianamente Satisfecho
5.1 – 7.5	Satisfecho
7.6 – 10	Muy Satisfecho

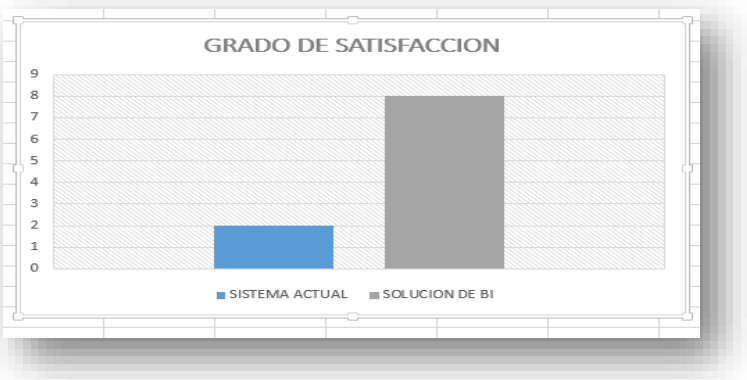
Tabla N° 03. Tabla de rangos de satisfacción

RQ N° 01



Resultado: Los valores obtenidos son: 3.5 (S.T) y 8 (SBD)

RQ N° 02



Resultados: Los valores que se obtuvieron son: 2 (S.T) y 8(SBD).

RQ N° 03

Resultado: Los valores obtenidos son: 2.5 (S.T) y 8.5 (SBD).

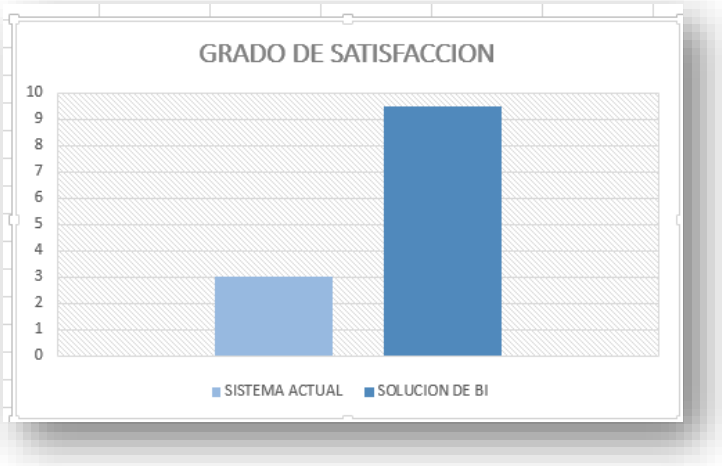
RQ N° 04

Resultados: los valores obtenidos fueron 0 (S.T) Y 9.5 (SBD).

RQ N° 05

Resultados: Los valores obtenidos fueron 2 (S.T) y 8 (SBD)**.**

RQ N° 06



Resultados: los valores obtenidos fueron 3 (S.T) y 9.5 (SBD) .

RQ N° 07

Resultados: los valores obtenidos son : 1.5 (S.T) 10 (SBD).

RQ N° 08

Resultados: Los valores obtenidos fueron 1 (S.T) y 8(SBDP).

RQ N° 09

Resultado: Se obtuvieron los siguientes valores 1.5 (S.T) y 9.5 (SBD).

RQ N° 10

Resultados: Los valores obtenidos son 2 (S.T) y 7.5 (SBD).

RQ N° 11

Resultados: Los valores obtenidos fueron 3 (S.T) y 6.5 (SBD).

RQ N° 12

Resultado: Los valores obtenidos son: 3.5 (S.T) y 8.5 (SBD).

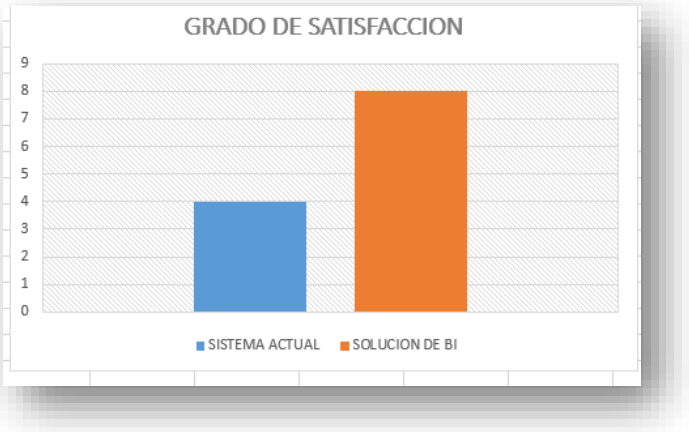
RQ N° 13

Resultado: Los valores obtenidos son 2 (S.T) 10 (SBD).

RQ N° 14

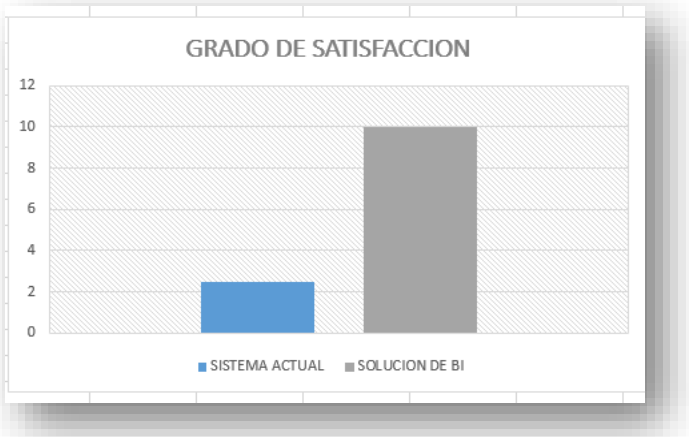
Resultado: Los valores obtenidos son: 3.5 (S.T) 9.5 (SBD).

RQ N° 15



Resultados: los valores obtenidos son: 4(S.T) y 8 (SBD).

RQ N° 16



Resultados: los valores obtenidos son: 2 (S.T) y 10 (SBD).

RQ N° 17

Resultado : los valores obtenidos fueron : 1 (S.T) y 10 (SBD) el S.T .

RQ N° 18

Resultados: Los valores obtenidos son: 0 (S.T) y 9.5 (SBD).

5.4.2. APLICACIÓN DEL RANGO DE SATISFACCIÓN A LOS INDICADORES DE LA HIPÓTESIS

Evaluación de los indicadores:

INDICADORES	S.T	SBD
Nro productos comprados x proveedor x tiempo	2	8
Precio total x proveedores x tiempo	3	8
Resumen de compras x sucursales x mes	0	8.5
Resumen de proveedores más frecuentes x mes	3	9.5
N° productos x mes	1	9.5
Egresos x sucursales x semestre	4	9.5
Productos más vendidos	1	9
N° productos recibidos x sucursales x tiempo	1	8.5
Categoría de producto vendidos x trimestre	2	9
Condiciones de productos x tiempo	3	6.5
Resumen de cantidad de productos que ingresaron x sucursales	2	7.5
Resumen de productos en buen estado x tiempo	3	7.5
Ventas totales x empleado x sucursales	2.5	9

Cantidad de productos vendidos x sucursal x tiempo	3.5	9.5
Cantidad total de productos comprados x cliente x tiempo	3	9
Productos comprados x cliente x mes	3	9
Resumen de ventas totales x mes x sucursales	2	9
Promedio de ventas x sucursales x tiempo	0	10
PROMEDIO	**2.09**	**8.90**

Tabla N° 04. Tabla de indicadores

5.4.3. ANÁLISIS ESTADÍSTICO PARA LA HIPÓTESIS

Diferencia de medias:

Descripción	**Medias**	**Varianzas**
Fórmula	$\mu_i = \frac{\sum X_i}{N}$	$\sigma^2 = \frac{\sum (X_i - \mu_i)^2}{N}$
Cálculo N=18	μ_1= 2.08 μ_2= 8.80	$\sigma^2{}_1$ = 1.25 $\sigma^2{}_2$ =1.03

Tabla N° 05 Tabla diferencia de las medias

Cálculo de la Prueba de Hipótesis:

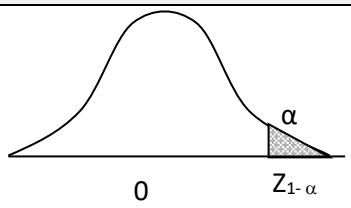
TIPO DE HIPÓTESIS	ESTADÍSTICA DE PRUEBA	REGIONES DE ACEPTACIÓN Y RECHAZO DE Ho	VALOR CRÍTICO
Hipótesis Nula Ho : $\mu_1 - \mu_2 = 0$ Nivel de signif α	$z_0 = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2{}_1 / n + \sigma^2{}_2 / n}}$		$\alpha = 0.05$ $Z_{1-\alpha} = 0.97$
Hipótesis Alterna $H_1 : \mu_1 > \mu_2$	**Zo = 17.60**	**Rechazar Ho si, Zo > $Z_{1-\alpha}$**	**17.60>0.97**

Tabla N° 06 Cálculo prueba de la hipótesis

CONCLUSIÓN: Ya que **Zo es mayor que $Z_{1-\alpha}$** se rechaza la hipótesis nula y se acepta H_1 , entonces se concluye que "Una Solución de Big data permite mejorar análisis de la información de los procesos de ventas de la empresa Inversiones Santa María utilizando el ecosistema de Apache Hadoop y MS Azure".

6. CONCLUSIONES

- Se obtuvo información del área crítica e importante en necesidad de información y análisis, pero el trabajo desarrollado no dejó de lado dos procesos que van de la mano con las ventas, como es el de compras y logístico (almacén).
- De acuerdo con las necesidades del personal relevante, se analizó el negocio y los procesos de ventas, compras, inventario y otros relacionados y se obtuvieron 18 requisitos para el proyecto. El análisis de estos requerimientos llevo a consolidar un modelo estrella constelación basado en 3 tablas de hechos y 9 dimensiones, y por la cantidad de datos que se tiene se propone una arquitectura basada en la nube de Azure.
- La arquitectura de Big Data propuesta contempla un clúster creado en HDInsight (Hadoop) en la nube de Microsoft Azure, permitiendo una mayor escalabilidad en servicios y procesamiento, pero en esta solución se configuró inicialmente los principales servicios para que de soporte a los requerimientos, entre los cuales fueron, un Data Lake con su contenedor, un Data Factory y un recurso de Synapse analytics.
- El proceso ETL se desarrolló usando Data Factory y Synapse Analytics realizando las conexiones entre los contenedores del Data Lake.
- Para la presentación de los datos se implementó un dashboard en Power BI, permitiendo visualizar la información de manera dinámica e acuerdo a los requerimientos planteados.

7. RECOMENDACIONES

- Se recomienda el uso de la metodología propuesta por Ángelo Martínez, por ser una metodología ágil y enfocada a los temas de soluciones de analítica de datos, teniendo 06 fases que describen claramente que actividades se tienen que hacer para estos tipos de proyectos.
- Para conocer y comprender la lógica del negocio es necesario mantener una estrecha relación con los usuarios relevantes, mantener una comunicación constante con el administrador, proponer un cronograma de reuniones con los empleados para involucrarlos y poder recolectar la información necesaria del área. donde el proyecto se llevará a cabo en la información de tiempo especificado.
- Se recomienda usar Azure HDInsight sobre las instancias locales de Hadoop, permitiendo manejar un bajo costo mediante la creación de clústeres a petición y pagando solo por lo que usa y también proporcionando flexibilidad al mantener el volumen de datos independiente del tamaño del clúster.
- Al realizar ETL en Synapse Analytics, debe considerar los conectores apropiados según los tipos de datos que esté usando y los contenedores o bases de datos en los que residen o se dirigirán a transformar.
- Para trabajos futuros se recomienda utilizar modelos predictivos a través de Machine learning, Inteligencia Artificial, que se encuentran en el ecosistema de Azure, los cuales permitan ampliar el análisis de los datos, conociendo las tendencias o pronosticando las ventas.

8. REFERENCIAS BIBLIOGRAFICAS

ETL. (05 de 08 de 2022).

Arimetrics. (05 de 04 de 2022). *Qué es Dashboard*. Obtenido de https://www.arimetrics.com/glosario-digital/dashboard

Azure HDInsight. (22 de 05 de 2019). *Azure HDInsight*. Obtenido de https://docs.microsoft.com/es-es/azure/hdinsight/

Bit. (19 de 05 de 2017). *Servicios de intelligence y analytics en microsoft azure*. Obtenido de https://www.bit.es/knowledge-center/servicios-de-intelligence-y-analytics-en-microsoft-azure-i/

Executrain. (03 de 10 de 2017). *Los 4 Tipos de Analítica de Datos que debes conocer*. Obtenido de https://www.executrain.com.mx/blog/big-data/item/los-4-tipos-de-analitica-de-datos-que-debes-conocer

Gartner. (02 de 2019). *Magic Quadrant for Analytics and Business Intelligence Platforms*. Obtenido de https://www.gartner.com/en/webinars/3900973/the-2019-analytics-and-bi-magic-quadrant-highlights

Iebschool. (15 de 02 de 2019). *Glosario Big Data*. Obtenido de https://www.iebschool.com/blog/glosario-big-data/

Ionos. (13 de 03 de 2019). *Apache Hadoop: sistema de archivos distribuido*. Obtenido de https://www.ionos.es/digitalguide/servidores/know-how/apache-hadoop-el-framework-para-big-data/

Martínez , Á. (2018). *BIG DATA APLICADO EN EL SECTOR BANCARIO TRADICIONAL PARA LOGRAR UNA MAYOR VENTAJA COMPETITIVA FRENTE A LAS FINTECH.* Argentina: Universidad de Palermo.

Microsoft. (08 de 05 de 2019). *¿Qué es Azure?* Obtenido de https://azure.microsoft.com/es-es/overview/what-is-azure/

Microsoft. (15 de 11 de 2020). *¿Qué es Azure?* Obtenido de https://azure.microsoft.com/es-es/overview/what-is-azure/

Microsoft. (05 de 04 de 2022). *¿Qué es Azure HDInsight?* Obtenido de https://learn.microsoft.com/es-es/azure/hdinsight/hdinsight-overview

Power BI. (17 de 05 de 2019). *Power BI*. Obtenido de https://powerbi.microsoft.com/es-es/

PowerData. (01 de 09 de 2019). *Integracion de datos*. Obtenido de http://blog.powerdata.es: http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/405060/Qu-significa-la-integraci-n-de-datos

Prometeusgs. (01 de 02 de 2019). *Sin análisis no hay información útil. La importancia del Data Analytics en tu negocio*. Obtenido de https://prometeusgs.com/analisis-de-datos-informacion-util/

RAYO, Á. M. (11 de 05 de 2017). *Servicios y productos Big Data en Microsoft Azure*. Obtenido de https://www.bit.es/knowledge-center/servicios-y-productos-big-data-en-microsoft-azure/

SAS. (15 de 03 de 2022). *Big Data*. Obtenido de https://www.sas.com/es_es/insights/big-data/what-is-big-data.html

Stratebi. (05 de 03 de 2019). *Power BI*. Obtenido de http://www.stratebi.com/power-bi

TableauPeru. (01 de 05 de 2022). *¿Qué es un dashboard?: Concepto y utilidad*. Obtenido de https://tableauperu.com/que-es-un-dashboard/

Talend. (29 de 01 de 2019). *¿En qué consiste la integración de datos?* Obtenido de https://es.talend.com/resources/what-is-data-integration/

Tecon. (25 de 04 de 2019). *¿Qué es Microsoft Azure? ¿Cómo funciona?* Obtenido de https://www.tecon.es/que-es-microsoft-azure-como-funciona/

Ticportal. (15 de 05 de 2019). *Microsoft Azure*. Obtenido de https://www.ticportal.es/temas/cloud-computing/microsoft-cloud/microsoft-azure

Wasson, M. (28 de 11 de 2017). *Big data architecture style*. Obtenido de https://learn.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data

Workana. (24 de 02 de 2020). *¿Qué es un Dashboard?* Obtenido de https://www.workana.com/i/glosario/que-es-un-dashboard/