LA INTELIGENCIA DE NEGOCIOS TOMANDO UN ENFOQUE CON SQOOP, FLUME Y HDFS EN HADOOP

BUSINESS INTELLIGENCE TAKING AN APPROACH WITH SQOOP, FLUME AND HDFS ON HADOOP

José Javier Mendoza Loor [1*]

[1] Universidad Técnica Luis Vargas Torres de Esmeraldas. Ecuador. ORCID: https://orcid.org/0000-0001-8623-872X. Correo: jose.mendoza.loor@utelvt.edu.ec

Washington Ramiro Bonilla Vimos [2]

[2] Escuela Superior Politécnica de Chimborazo .Carrera Tecnologías de la Información. Ecuador. ORCID: https://orcid.org/0000-0002-6834-7030. Correo: washington.bonilla@espoch.edu.ec

Segundo Isaías Quezada Valencia [3]

[3] Instituto Superior Tecnológico Sudamericano. Ecuador. ORCID: https://orcid.org/0009-0009-0074-0897. Correo: iquiss@gmail.com

María Belén Torres Bueno [4]

[4] Instituto Superior Tecnológico Quito (ITQ). Ecuador. ORCID: https://orcid.org/0009-0004-2462-611X. Correo: belen.torres@itq.edu.ec

*** Autor para correspondencia: jose.mendoza.loor@utelvt.edu.ec**

Resumen

En el ecosistema de Hadoop, Sqoop, Flume y Hdfs (*Hadoop Distributed File System*), han jugado roles importantes y complementarios en el manejo y procesamiento de grandes volúmenes de datos, la combinación de estas herramientas ha permitido a las organizaciones manejar grandes volúmenes de datos provenientes de diversas fuentes, tanto estructuradas como no estructuradas, de manera eficiente y escalable en el ecosistema Hadoop. La inteligencia de negocios (BI) con Hadoop ha ofrecido una forma poderosa y escalable de manejar, analizar y extraer valor de grandes volúmenes de datos, con su capacidad para procesar grandes cantidades de datos distribuidos en clusters de servidores, se han integrado con diversas herramientas de inteligencia de negocios para proporcionar datos empresariales valiosos. Como se sabe, Hadoop ha sido considerado como un *framework* de software de código abierto que ha permitido el procesamiento distribuido de volúmenes de

datos en clusters de ordenadores utilizando modelos de programación. Ha sido vista como una herramienta esencial en el ámbito del big data y ha sido diseñada para escalar desde servidores individuales hasta miles de máquinas, cada una ofreciendo computación y almacenamiento local. Estos entornos han analizado a los clientes y su comportamiento, han optimizado las cadenas de suministro analizando los datos en tiempo real mejorando la eficiencia operativa de las empresas, con su detección de fraude procesando grandes volúmenes de datos, además de haber segmentado clientes para campañas efectivas de marketing.

Palabras clave: análisis de datos; inteligencia de negocios

Abstract

In the Hadoop ecosystem, Sqoop, Flume and Hdfs (Hadoop Distributed File System), have played important and complementary roles in handling and processing large volumes of data, the combination of these tools has allowed organizations to handle large volumes of data from various sources, both structured and unstructured, in an efficient and scalable manner in the Hadoop ecosystem. Business intelligence (BI) with Hadoop has offered a powerful and scalable way to manage, analyze and extract value from large volumes of data, with its ability to process large amounts of data distributed across server clusters, they have been integrated with various business intelligence tools to provide valuable business data. As is known, Hadoop has been considered as an open source software framework that has allowed the distributed processing of data volumes in computer clusters using programming models. It has been seen as an essential tool in the field of big data and has been designed to scale from single servers to thousands of machines, each offering local computing and storage. These environments have analyzed customers and their behavior, optimized supply chains by analyzing data in real time, improving the operational efficiency of companies, with their fraud detection by processing large volumes of data, in addition to having segmented customers for effective marketing campaigns.

Keywords: *data analysis; business intelligence*

Fecha de recibido: 29/03/2024
Fecha de aceptado: 21/06/2024
Fecha de publicado: 01/07/2024

Introducción

La llegada del Big Data ha revolucionado el panorama del procesamiento de datos, brindando oportunidades sin precedentes para la generación de conocimiento y la toma de decisiones en diversos campos, sin embargo, con la gran cantidad de datos, la alta velocidad y la variedad, se debe garantizar la confiabilidad del sistema, el manejo se vuelve crítico, la tolerancia a fallos, la capacidad de un sistema de funcionar de manera óptima, se convierte en un requisito crítico en situaciones en las que hay un defecto o mal funcionamiento (Bastidas, 2022). La importancia de la tolerancia a fallos en el procesamiento de información en la inteligencia de

negocios, donde los conjuntos de datos a menudo superan la capacidad y la tolerancia a fallos de los sistemas informáticos tradicionales, desempeña un papel clave en el mantenimiento de la integridad de los datos (Wu, 2010).

Hadoop se ha convertido en una tecnología fundamental para manejar tareas de procesamiento de datos a gran escala, su arquitectura y la tecnología distribuida con sus mecanismos tolerantes a fallas la hacen ideal para procesar y analizar conjuntos de datos masivos al dividir las tareas en sub tareas más pequeñas y distribuirlas entre los nodos del clúster, Map Reduce permite el procesamiento paralelo, aumentando así la eficiencia y la escalabilidad, sin embargo no es inmune a fallas, los nodos o componentes pueden interrumpir las tareas de procesamiento en curso, lo que resulta en pérdida de datos y fallas del sistema (Wei Kuang, 2013).

La capacidad de gestionar y procesar grandes cantidades de datos es fundamental para las empresas que buscan obtener una ventaja competitiva. Hadoop es un marco de software de código abierto que se ha convertido en una de las tecnologías más fundamentales en este campo. Desarrollado por Apache Software Foundation, Hadoop revoluciona la forma en que se administran y analizan los datos al permitir el procesamiento distribuido de grandes conjuntos de datos en grupos de computadoras.

HDFS (*Hadoop Distributed File System*) actúa como el sistema de almacenamiento principal donde se guardan grandes volúmenes de datos de manera distribuida y segura. Sqoop se utiliza para la transferencia de datos entre sistemas de bases de datos relacionales y HDFS, facilitando la ingesta de datos estructurados en el entorno Hadoop. Flume es utilizado para la ingesta de datos no estructurados en tiempo real, como logs y eventos, hacia HDFS para su almacenamiento y posterior análisis. La combinación de estas herramientas permite a las organizaciones manejar grandes volúmenes de datos provenientes de diversas fuentes, tanto estructuradas como no estructuradas, de manera eficiente y escalable en el ecosistema Hadoop (Tomer, 2024).

A pesar de que el término Big Data se asocia principalmente con cantidades de datos exorbitantes, se debe dejar de lado esta percepción, pues Big Data no va dirigido solo a gran tamaño, sino que abarca tanto volumen como variedad de datos y velocidad de acceso y procesamiento. En la actualidad se ha pasado de la transacción a la interacción, con el propósito de obtener el mejor provecho de la información que se genera minuto a minuto, se ha dado cabida también a un nuevo concepto, Data Science o Ciencia de los Datos, que se usa de forma genérica para hacer referencia a la serie de técnicas necesarias para el tratamiento y manipulación de información masiva desde un enfoque estadístico e informático. Incluyendo también el surgimiento de un nuevo perfil profesional, el “Data Scientist” (Bastidas, 2022), las personas capacitadas en este perfil deben saber del negocio, de las herramientas computacionales y de análisis e interpretación estadística.

Ahora bien, pensando en la creación de soluciones que incluyan problemas enmarcados en este enfoque, se pueden encontrar cuatro fases donde se agrupan o clasifican las diferentes tecnologías de soporte, estas son: generación, adquisición, almacenamiento y análisis de datos (Maarten van Steen, 2006). Se define la primera fase, generación, como un proceso propio de diversas actividades de la sociedad, en estas se genera una cantidad inmensa de datos, que, según su naturaleza, puede estar almacenada y estructurada o puede corresponder a datos sin ninguna estructura, pero con características de gran valor (Vimala, 2019).

La fase de almacenamiento de Big Data ha generado la necesidad de generar estudios y propuestas de nuevas estrategias que permitan afrontar los tipos de datos que no se pueden gestionar con un sistema de gestión de bases de datos relacionales. Surgen entonces, tecnologías de almacenamiento de datos masivos como

almacenamiento con conexión directa y el almacenamiento en red, también diferentes motores NoSQL. Finalmente, la fase de análisis debe atender a la necesidad de extraer rápidamente información desde los datos masivos para poder generar valor en las organizaciones y facilitar procesos de toma de decisiones, se requiere de tecnologías que faciliten incluso el análisis en tiempo real.

Materiales y métodos

La inteligencia de negocios (BI) con Hadoop ofrece una forma poderosa y escalable de manejar, analizar y extraer valor de grandes volúmenes de datos. Hadoop, con su capacidad para procesar grandes cantidades de datos distribuidos en clusters de servidores, se integra con diversas herramientas de BI para proporcionar insights empresariales valiosos. A continuación se explica cómo Hadoop puede potenciar la inteligencia de negocios:

Componentes Clave de Hadoop en BI

HDFS (Hadoop Distributed File System):

- Almacenamiento Escalable: Proporciona una base para almacenar grandes volúmenes de datos en un entorno distribuido.
- Alta Disponibilidad: Los datos se replican en múltiples nodos, garantizando la disponibilidad y resiliencia frente a fallos de hardware.

MapReduce:

- Procesamiento Distribuido: Permite el procesamiento paralelo de grandes conjuntos de datos, lo que es crucial para análisis complejos y grandes volúmenes de datos.
- Modelado de Datos: Facilita la ejecución de tareas de análisis de datos, como agregaciones, filtrados y transformaciones.
- Hive: *Consultas SQL:* Proporciona una interfaz de consulta similar a SQL sobre datos almacenados en HDFS.
- Facilidad de Uso: Permite a los analistas de datos y usuarios de BI familiarizados con SQL realizar consultas complejas sin necesidad de programar en Java o Python.
- Pig:Lenguaje de Scripting: Ofrece un lenguaje de scripting (Pig Latin) para el procesamiento de datos a gran escala.
- Procesamiento de Datos: Utilizado para tareas de extracción, transformación y carga (ETL) de datos.
- HBase: Base de Datos NoSQL: Proporciona un almacén de datos NoSQL de baja latencia sobre HDFS.
- Acceso Rápido a Datos: Ideal para aplicaciones que requieren acceso rápido y en tiempo real a grandes conjuntos de datos.
- Impala y Presto: Consultas SQL en Tiempo Real: Ofrecen capacidades de consulta SQL de baja latencia sobre datos en HDFS y HBase.
- Análisis Interactivo: Permiten a los usuarios realizar análisis ad hoc y obtener resultados en tiempo casi real.

- Spark:Procesamiento en Memoria: Proporciona procesamiento de datos en memoria, lo que acelera significativamente las tareas de análisis.
- Componentes Múltiples: Incluye Spark SQL, MLlib (para machine learning), GraphX (para gráficos) y Streaming (para procesamiento en tiempo real).

Integración con Herramientas de BI

Herramientas de Visualización:

- Tableau, Power BI, Qlik: Se pueden integrar con Hadoop a través de conectores nativos y APIs para la visualización de datos.
- Dashboards Interactivos: Permiten crear dashboards interactivos y visualizaciones avanzadas sobre los datos procesados en Hadoop.
- ETL (Extract, Transform, Load):Talend, Informatica, Pentaho: Herramientas de ETL que se integran con Hadoop para la ingesta y transformación de datos.
- Flujos de Trabajo ETL: Facilitan la preparación de datos para su análisis en las plataformas de BI.

Análisis Avanzado y Machine Learning:

- Spark MLlib, TensorFlow, R: Integración con bibliotecas y plataformas de machine learning y análisis estadístico para desarrollar modelos predictivos y de análisis avanzados.
- Modelos Predictivos: Permiten a las organizaciones desarrollar modelos de machine learning sobre grandes conjuntos de datos almacenados en Hadoop.

Beneficios de Utilizar Hadoop para BI

- Escalabilidad: Permite a las organizaciones escalar horizontalmente añadiendo más nodos a su cluster de Hadoop, manejando así crecientes volúmenes de datos sin sacrificar el rendimiento.
- Costo-Eficiencia: Utiliza hardware común y open-source, reduciendo significativamente los costos en comparación con soluciones tradicionales de almacenamiento y procesamiento de datos.
- Flexibilidad: Capacidad para manejar datos estructurados, semi-estructurados y no estructurados, permitiendo una visión integral de los datos de la organización.
- Análisis en Tiempo Real: Con herramientas como Spark y Flume, se pueden realizar análisis en tiempo real y procesamiento de flujos de datos, proporcionando insights inmediatos.

Casos de uso en BI con Hadoop

- Análisis de Clientes: Integración de datos de diferentes fuentes (transacciones, interacciones en redes sociales, historiales de compra) para obtener una visión completa del comportamiento del cliente.
- Optimización de la Cadena de Suministro: Análisis de datos de sensores, logs de envío y datos de inventario para mejorar la eficiencia y reducir costos.
- Detección de Fraude: Análisis en tiempo real de transacciones y patrones de comportamiento para identificar y prevenir actividades fraudulentas.
- Marketing Personalizado: Uso de análisis predictivo para segmentar clientes y ofrecer campañas de marketing altamente personalizadas (Vimala, 2019).

En cuanto a técnicas de Big Data, cabe aclarar que existen diferentes clasificaciones y que muchas de estas técnicas se aplican tanto en soluciones Big Data como en otros enfoques (Bastidas, 2022). La clasificación de las técnicas de Big data incluyen: técnicas estadísticas, métodos de optimización, minería de datos, técnicas de machine learning (aprendizaje automático), técnicas de clasificación, clustering, técnicas de análisis y regresión. Clasificador brillante, rápido, simple y secuencial, capaz de aprendizaje on line en entornos exigentes.

Además, tiene implementaciones secuenciales y paralelas del algoritmo clásico de clasificación diseñado para modelar procesos del mundo real cuando el proceso de generación subyacente es desconocido. Diseñado para reducir el ruido en matrices grandes, haciendo con esto que sean más pequeñas y que sea más fácil trabajar con ellas. Enfoque de almacenamiento en clúster basado en modelo, que determina la propiedad con base en si los datos se ajustan al modelo subyacente. Es una familia de enfoques similares que usa un enfoque basado en gráficas para determinar la membresía a clúster. Utiliza una estrategia de hash para agrupar elementos similares, produciendo así clústeres de conocurrencia distribuida, SVD y mínimos cuadrados alternantes (Vimala, 2019).

Resultados y discusión

Hadoop proporciona una base robusta y escalable para la inteligencia de negocios, permitiendo a las organizaciones manejar y analizar grandes volúmenes de datos para tomar decisiones más informadas y estratégicas. Es importante también, tener en cuenta cómo en el área de la industria y los negocios se ha presentado una explosión en el número de datos, causada principalmente por el rápido desarrollo del internet, nuevos conceptos como el internet de las cosas y la computación en la nube.

Big data se ha constituido como un "tópico caliente" que atrae la atención no solo de la industria, sino también de la academia y del Gobierno. Los autores presentan desde diferentes perspectivas el significado y las oportunidades que nos brinda el ecosistema y dan una serie de condiciones necesarias para que un proyecto de esta magnitud sea exitoso. En primer lugar, se deben tener claros los requerimientos independientemente de si son técnicos, sociales o económicos. En segundo lugar, para trabajar de forma eficiente se requiere explorar y encontrar la estructura central o el kernel de los datos a ser procesados, ya que al tener esto se puede caracterizar el comportamiento y las propiedades subyacentes. Se debe adoptar un modelo de administración top-down, se puede considerar también un modelo bottom-up, sin embargo, solo serviría cuando se trata de problemas específicos, y luego tratar de unirlos para formar una solución completa es complejo.

Pentaho, Sqoop, Flume y MapReduce son componentes cruciales en el ecosistema Hadoop, cada uno desempeñando un papel específico en la gestión y el análisis de grandes volúmenes de datos.Se observa la instalación realizada en Windows 11, todo es por defecto descargado del link de la página oficial: https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform/pentaho-data-integration/pentaho-trial-download.html

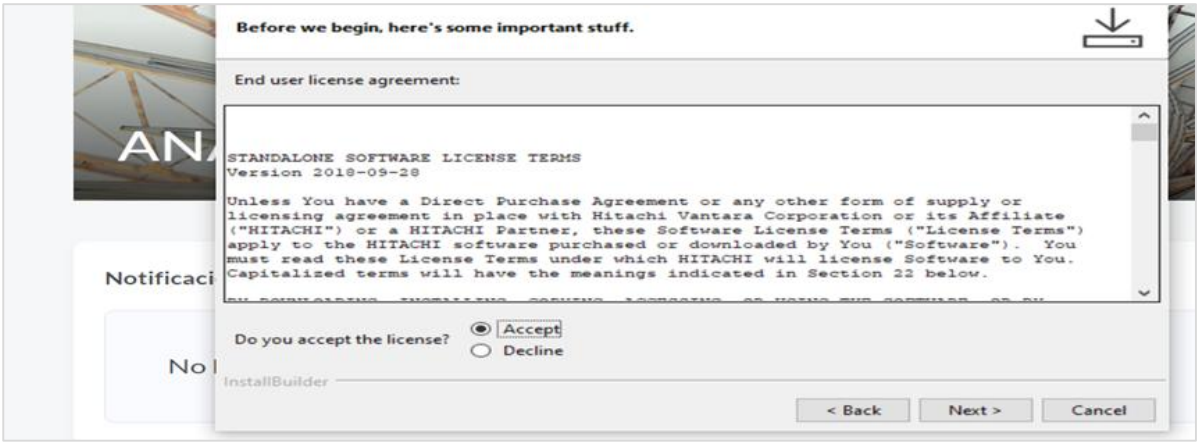

Figura 1 Instalación 01

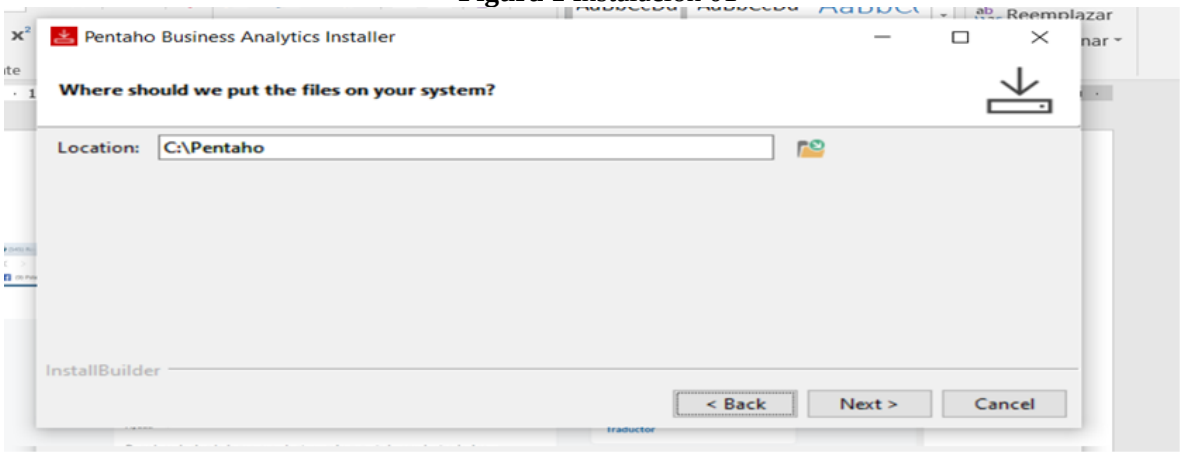

Figura 2. Instalación 02

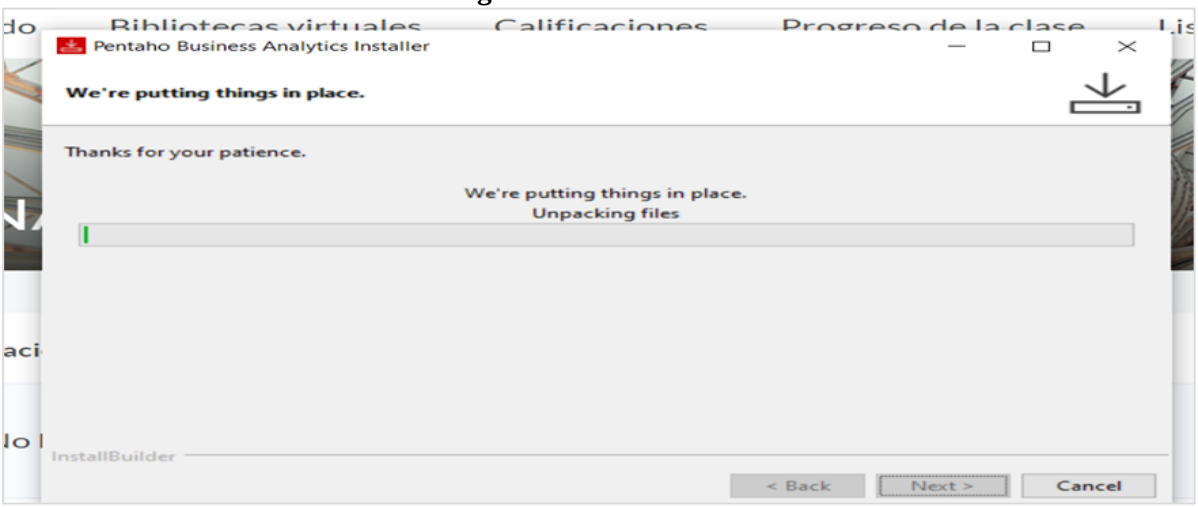

Figura 3 Instalación 03

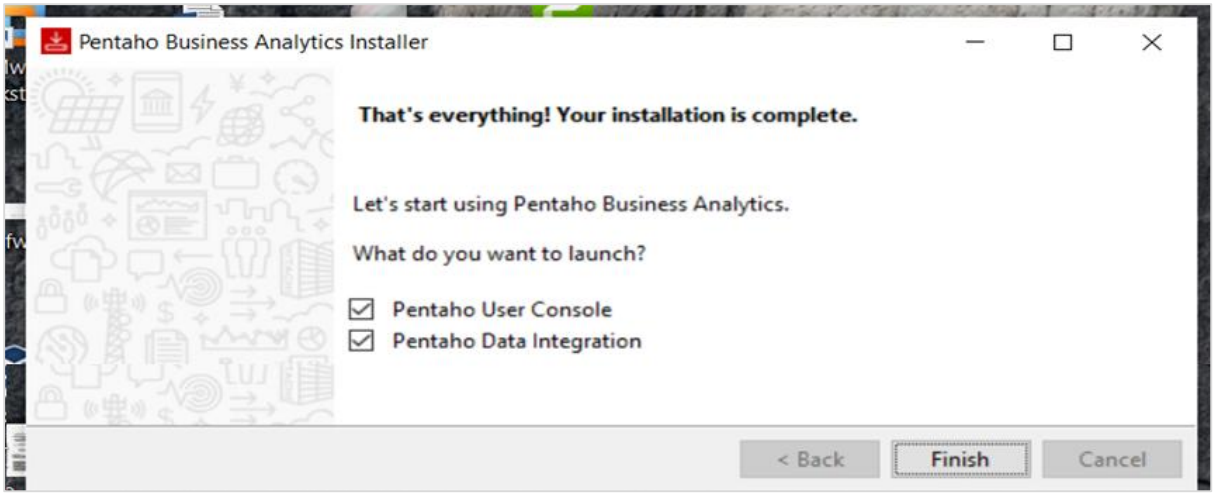

Figura 4 Instalación 04

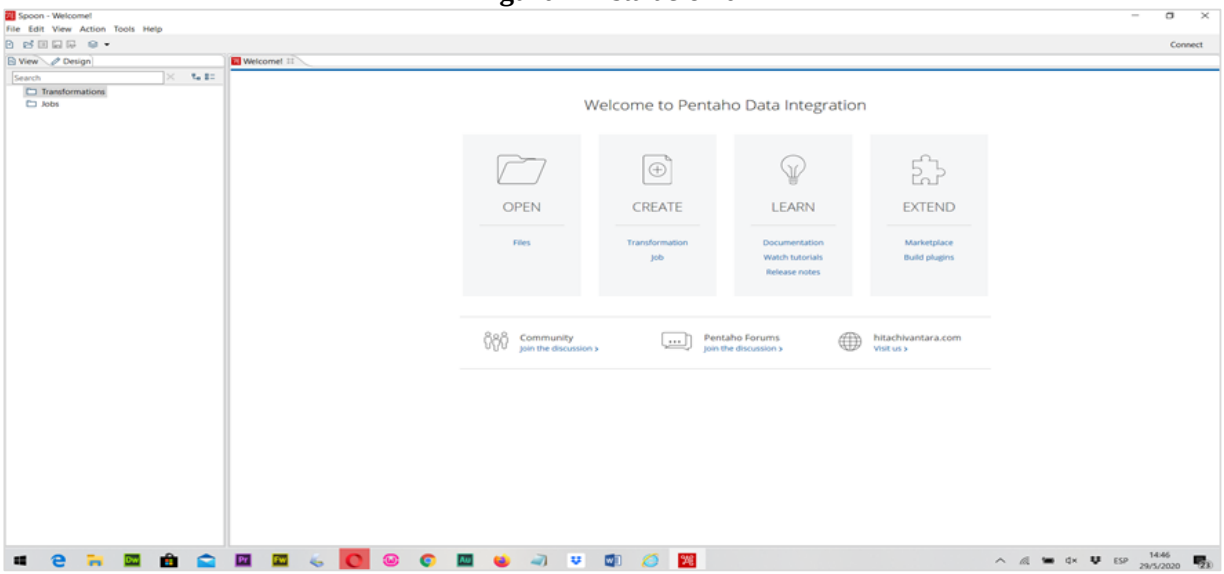

Figura 5 Instalación 05

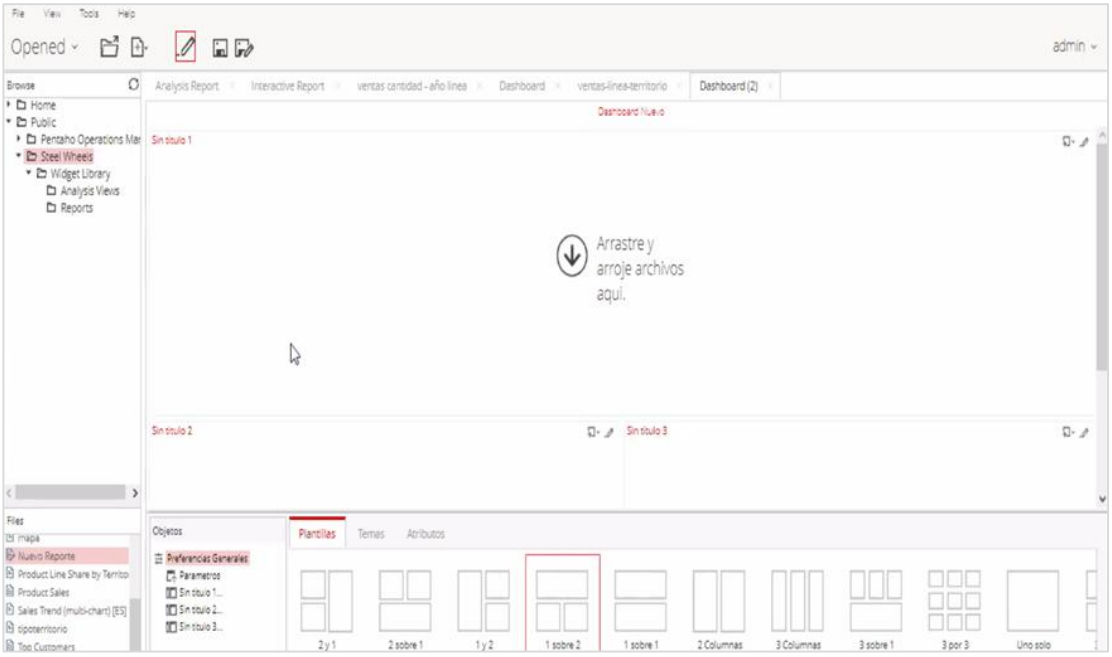

Figura 6 Pruebas 01

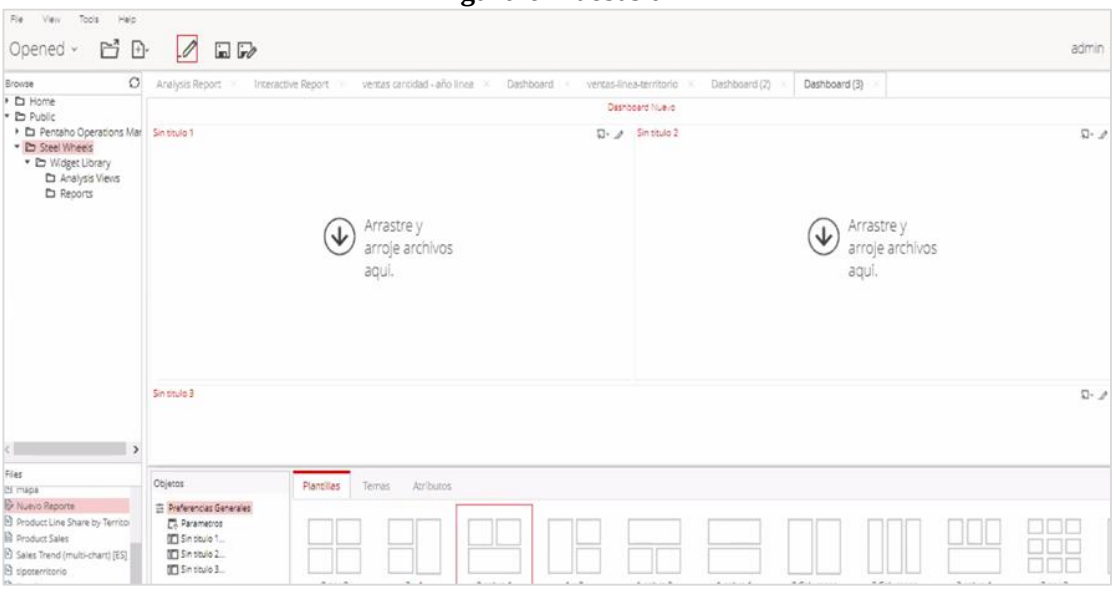

Figura 7 Pruebas02

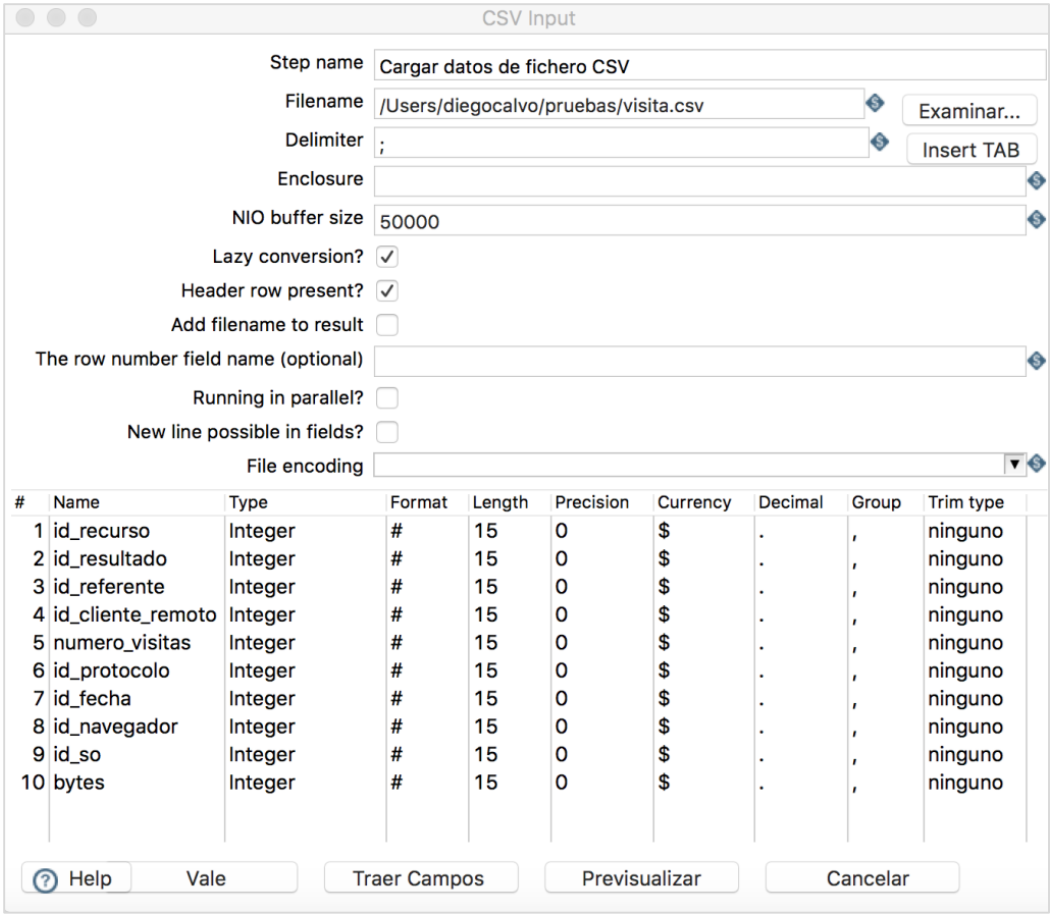

Figura 8. Insertar datos de un fichero CSV a una BBDD.

Debe tener en cuenta que no hace falta escribir los nombres, tipo, … y demás campos a cumplimentar ya que se autocompletan al dar al botón «Traer datos».Segundo, incluir el elemento «Table output» de la lista de la izquierda.Tercero, unir el elemento «CSV file input» con la «Table output», mediante el botón que parece poniéndonos encima del primero, se muestra imagen aclaratoria.

Discusión

Para la evaluación de las diferentes herramientas de procesamiento distribuido se seleccionaron las que actualmente ofrecen múltiples características del procesamiento de datos y se evalúan de acuerdo a los siguientes criterios como la reubicación de recursos que tratan de que las aplicaciones deben tener estrategias de reubicación de recursos que permitan la ejecución de las tareas evitando la pérdida de información y generando puntos de retorno.

El aislamiento de procesos que son ejecutados o pueden ser aislados para lograr un mejor control de la cantidad de procesadores asignados para la ejecución de una tarea, la cantidad de memoria RAM asignada, gestionar el acceso a la red, y administrar las características de almacenamiento (Bastidas, 2022). La escalabilidad con la adición de nuevos nodos a la arquitectura de computación distribuida es una de las

cualidades de un sistema distribuido, permitiendo el incremento de los recursos disponibles para el desarrollo de las tareas. La tolerancia a fallos en que las tareas deben ser reasignadas a nuevos nodos sin que estas se vean afectadas o interrumpidas, y el administrador de tareas que se envían a un sistema distribuido asignando las prioridades de ejecución y controlando las entradas y salidas de cada uno de los procesos, para que estos de la misma forma puedan ser ejecutados de forma paralela a lo largo de la arquitectura del sistema.

Conclusiones

La integración de Sqoop, Flume y HDFS dentro del ecosistema Hadoop ofrece un enfoque poderoso y eficaz para la inteligencia de negocios. Este enfoque permite a las organizaciones capturar, almacenar, procesar y analizar grandes volúmenes de datos de manera eficiente y escalable. Al proporcionar una infraestructura robusta y flexible, Hadoop junto con sus herramientas complementarias, habilita a las organizaciones para obtener ideas valiosas y tomar decisiones informadas que impulsen el crecimiento y la competitividad.

Referencias

Bastidas, D. J. (2022). Un ejemplo práctico de almacen de datos en sistemasdj. Polo del Conocimiento, 1712-1721.

Ding Zhang, Z.-Y. D.-P.-T.-H. (2024). A distributed data processing scheme based on Hadoop for synchrotron radiation experiments. Computer Programs, 1-11.

Maarten van Steen, M. H. (2006). Middleware 2006: ACM/IFIP/USENIX 7th International Middleware Conference, Melbourne, Australia, November 27 - December 1, 2006, Proceedings (Lecture Notes in Computer Science, 4290) 2006th Edición. Middleware 2006: ACM/IFIP/USENIX 7th International Middleware Conference (pág. 448). New York: Springer.

Mahmound, A. (2018). Developing Middleware in Java EE 8: Build robust middleware solutions using the latest technologies and trends . New York: Packt Publishing.

Qureshi, B. (2024). Optimizing Hadoop Scheduling inSingle-Board-Computer-Based Heterogeneous Clusters. Computation MDPI, 1-20.

Seibel, P. (2005). Practical Common Lisp. Berkeley: Apress.

Tomer, T. K. (2024). Innovative Advancements in Big Data Analytics: Navigating Future Trends and Direction with Hadoop Integration. School of Computer Science and Application, IIMT University Meerut U.P India, 173-181.

Vimala, K. A. (2019). A SURVEY ON BIG DATA PRIVACY USING HADOOP ARCHITECTURE. Karpagam University, 70-77.

Wei Kuang, L.-U. C. (2013). Towards a framework for large-scale multimedia data storage and processing on Hadoop platform. The Journal of Supercomputing , 20-30.

Wu, D. J. (2010). The Performance of MapReduce: An In-depth Study. School of ComputingNational University of Singapore, 4-17.