Exploración y visualización de datos de sensores de temperatura y humedad mediante tecnologías de Big Data y analítica.

John Henrry Rincón Gutiérrez

Asesor

Fernando Luis Carrascal Porras

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

Mayo de 2024

Dedicatoria

A mi querida esposa Johanna Morales Vega, quiero expresar mi profunda gratitud por su inquebrantable apoyo a lo largo de este proyecto aplicado. Su constante aliento y confianza en mí han sido la fuerza impulsora detrás de cada paso que he dado. Gracias a ti, he podido superar desafíos y limitaciones personales, creyendo en mi capacidad y visión. Este proyecto no solo es un logro personal, sino también un tributo a tu amor y apoyo incondicional. Tu presencia en mi vida es mi mayor inspiración y motivación. Te dedico este proyecto con todo mi cariño y agradecimiento.

Agradecimientos

Agradezco especialmente a Fernando Luis Carrascal Porras por su invaluable contribución como revisor crítico de este documento. Sus minuciosas revisiones y recomendaciones han enriquecido significativamente el contenido y la calidad del proyecto. Su dedicación y perspicacia han sido fundamentales para su mejora continua.

También expreso mi más sincero agradecimiento a Luis Ángel Anillo Arrieta por su destacado papel como asesor científico en materia. Sus conferencias inspiradoras y el liderazgo de círculos de interacción y participación académica y social (CIPAS) han sido de inestimable valor para el desarrollo y la consolidación de este proyecto aplicado. Su experiencia y sabiduría han sido una guía invaluable en mi camino hacia el éxito.

Además, quiero reconocer y agradecer a Dayana Alejandra Barrera Buitrago por su valioso papel como asesora científica, proporcionando orientaciones académicas fundamentales para el avance de este trabajo. Su compromiso y conocimientos han sido una luz orientadora en mi proceso de investigación y desarrollo. Agradezco sinceramente su apoyo continuo y dedicación a este proyecto.

Sin el aporte y la guía de estas personas, este proyecto no habría alcanzado su potencial máximo. Estoy profundamente agradecido por su compromiso y colaboración.

Resumen

El proyecto aplicado aborda una problemática crítica en la gestión de datos de sensores de temperatura y humedad en empresas colombianas. La complejidad inherente al manejo de grandes volúmenes de datos, combinada con la falta de experiencia, conduce a la pérdida de información valiosa y afecta la toma de decisiones estratégicas. La falta de exploración y visualización efectiva dificulta la identificación de patrones, tendencias y anomalías, impactando negativamente en la eficiencia operativa y la competitividad del negocio.

La solución propuesta se fundamenta en el uso de tecnologías de Big Data y analítica avanzada, empleando herramientas líderes como Apache Hadoop, Apache NiFi, Dremio, DBT (Data Build Tools) y Power BI. La metodología CRISP-DM servirá como guía durante todo el proceso, desde la comprensión del problema hasta la implementación de soluciones. Se adoptará un enfoque iterativo que permitirá realizar ajustes continuos en función de los resultados obtenidos y las necesidades emergentes.

El objetivo principal del proyecto es desarrollar una plataforma integral que facilite la extracción, procesamiento, almacenamiento y visualización eficiente de los datos de sensores de temperatura y humedad. Se espera que esta plataforma mejore significativamente la toma de decisiones en las empresas colombianas, incrementando su competitividad en el mercado y generando nuevas oportunidades de innovación y crecimiento empresarial.

Palabras clave: Gestión de datos, sensores, temperatura, humedad, Big Data, analítica, Apache Hadoop, Apache NiFi, Dremio, DBT, Power BI, metodología CRISP-DM, plataforma integral, extracción, procesamiento, almacenamiento, visualización.

Abstract

The applied project addresses a critical issue in the management of temperature and humidity sensor data in Colombian companies. The inherent complexity of handling large volumes of data, combined with the lack of experience, leads to the loss of valuable information and affects strategic decision making. The lack of effective exploration and visualization makes it difficult to identify patterns, trends and anomalies, negatively impacting operational efficiency and business competitiveness.

The proposed solution is based on the use of Big Data and advanced analytics technologies, using leading tools such as Apache Hadoop, Apache NiFi, Dremio, DBT (Data Build Tools) and Power BI. The CRISP-DM methodology will serve as a guide throughout the process, from problem understanding to solution implementation. An iterative approach will be adopted, allowing continuous adjustments based on the results obtained and emerging needs.

The main objective of the project is to develop a comprehensive platform that facilitates the efficient extraction, processing, storage and visualization of temperature and humidity sensor data. It is expected that this platform will significantly improve decision making in Colombian companies, increasing their competitiveness in the market and generating new opportunities for innovation and business growth.

Keywords: data management, sensors, temperature, humidity, Big Data, analytics, Apache Hadoop, Apache NiFi, Dremio, DBT, Power BI, CRISP-DM methodology, end-to-end platform, extraction, processing, storage, visualization.

Tabla de Contenido

Glosario 1

Introducción 4

Planteamiento del problema 5

Justificación 8

Objetivos 11

- Objetivo General 11
- Objetivos Específicos 11

Marco Conceptual 12

Marco Teórico 14

- Propuesta de Solución 22
  - Ciclo de Vida CRISP-DM 23
  - Aportes de la Solución 24
  - Aportes a Interrogantes Empresariales 26
  - Línea de Investigación Futura 27

Desarrollo Ciclo de Vida CRISP-DM 28

- Fase: Comprensión de Negocio 28
- Fase: Comprensión de los Datos 28
  - Fuente 29
  - Ingesta 32

Almacenamiento. 32

Procesamiento. 33

Consumo. 34

Fase: Preparación de Datos 35

Fase: Modelado 42

Fase: Despliegue 44

Conclusiones 51

Referencias 54

Lista de figuras

Figura 1 *Etapas de la arquitectura de datos* 29

Figura 2 *Funcionamiento de los sensores* 30

Figura 3 *Diagrama relacional de la aplicación web* 31

Figura 4 *Docker instalación herramientas* 36

Figura 5 *Inicio de sesión Apache NiFi* 36

Figura 6 *Resumen cluster Apache Hadoop* 37

Figura 7 *Página de inicio Dremio* 38

Figura 8 *Flujo de datos en Apache NiFi* 40

Figura 9 *Datos de sensores de temperatura y humedad en clúster Hadoop* 41

Figura 10 *Conjunto de datos consultado desde Dremio* 42

Figura 11 *Modelos dbt (Data Build Tools)* 43

Figura 12 *Reporte en Power BI sensores temperatura y humedad*.....45

Figura 13 *Power BI: sección de segmentación de datos* 46

Figura 14 *Power BI: sección de resumen* 46

Figura 15 *Power BI: contenido del informe* 47

Figura 16 *Modelo copo de nieve* 48

Glosario

Arquitectura de datos escalables: Diseño de sistemas que pueden crecer y adaptarse fácilmente a medida que aumentan los volúmenes de datos y las demandas de procesamiento.

Contenedores Docker: Entornos virtuales ligeros y portátiles que permiten empaquetar, distribuir y ejecutar aplicaciones de manera independiente de la infraestructura subyacente.

DataFrame basado en ML Pipelines y SparkR: Funcionalidades de Spark para realizar análisis de datos y machine learning utilizando pipelines y APIs específicas.

Dremio: Plataforma de datos que facilita la unificación, el análisis y la visualización de datos en diferentes fuentes y formatos.

ETL: Proceso de Extracción, Transformación y Carga, utilizado para mover datos desde múltiples fuentes a un repositorio centralizado para análisis.

Escalamiento vertical: Aumenta los recursos de hardware de un único servidor para mejorar el rendimiento del sistema. Se utiliza cuando se necesita un aumento inmediato en la capacidad de procesamiento y agregar servidores adicionales no es factible.

Gestión de datos: Proceso de organizar, controlar y administrar datos a lo largo de su ciclo de vida, desde su captura hasta su eliminación.

Hadoop es un software de código abierto utilizado para almacenar y procesar grandes cantidades de datos en computadoras agrupadas llamadas clústeres. Es útil cuando se tienen muchos datos que no caben en una sola computadora y se necesitan procesar de manera eficiente. Con Hadoop, los datos se dividen en partes más pequeñas y se distribuyen en múltiples

computadoras para un procesamiento más rápido. Es una herramienta importante en el campo del análisis de grandes volúmenes de datos.

Hive es una herramienta que facilita el análisis de grandes cantidades de datos almacenados en Hadoop utilizando un lenguaje de consulta familiar llamado SQL. Funciona como una especie de "base de datos virtual" sobre Hadoop, lo que permite a los usuarios consultar y analizar datos de manera similar a como lo harían en una base de datos tradicional, pero con el poder de procesamiento y almacenamiento distribuido de Hadoop. Esto hace que sea más fácil para los usuarios trabajar con grandes conjuntos de datos sin necesidad de aprender nuevas habilidades de programación complejas.

Machine Learning: Campo de la inteligencia artificial que utiliza algoritmos y modelos para capacitar a los sistemas informáticos a realizar tareas específicas sin instrucciones explícitas.

Modelo copo de nieve: Estructura de diseño de bases de datos analíticas utilizada en almacenamiento de datos. Consiste en una tabla central de hechos rodeada por múltiples tablas dimensionales, facilitando el análisis multidimensional y la generación de informes de manera eficiente. El diseño recibe su nombre de la apariencia visual, similar a una estrella, donde la tabla central de hechos está en el centro y rodeada por las tablas dimensionales.

NiFi: Plataforma de automatización de flujo de datos de código abierto utilizada para gestionar y transferir datos entre sistemas heterogéneos.

Parquet: Formato de archivo de almacenamiento columnar diseñado para eficiencia y rendimiento en procesamiento de Big Data.

Spark es un software que permite procesar datos de manera rápida y realizar análisis en tiempo real. Funciona principalmente en la memoria de la computadora, lo que significa que puede procesar datos mucho más rápido que los sistemas tradicionales que dependen de acceder a datos desde el disco duro. Esto hace que sea ideal para aplicaciones donde se necesita un procesamiento rápido de datos, como análisis en tiempo real, aprendizaje automático y procesamiento de datos en lotes.

DataFrames, DataSets, GraphFrames, Structured Streaming: Componentes de Apache Spark para procesamiento y análisis de datos estructurados y en streaming.

Introducción

El proyecto aplicado de exploración y visualización de datos de sensores de temperatura y humedad mediante tecnologías de Big Data y analítica surge de la necesidad de manejar grandes volúmenes de datos. Busca desarrollar soluciones prácticas para que empresas colombianas aprovechen estos datos, mejorando la toma de decisiones, anticipando problemas y optimizando procesos. Este proyecto es una respuesta estratégica a los desafíos de gestión y análisis de datos en un contexto empresarial orientado hacia el uso eficaz de la información.

El proyecto aplicado describe herramientas de licencia libre para diseñar soluciones accesibles a cualquier empresa, proporcionando una guía básica sobre cómo integrarlas para gestionar grandes volúmenes de datos. Esta iniciativa democratiza el acceso a soluciones tecnológicas avanzadas, permitiendo que empresas con recursos limitados aprovechen el potencial de sus datos. Ofreciendo orientación clara sobre la implementación de estas herramientas, se promueve la innovación y la competitividad empresarial, fomentando prácticas eficientes de gestión y análisis de datos en todos los sectores.

Planteamiento del problema

En la actualidad, las empresas colombianas que emplean sistemas de sensores de temperatura y humedad se encuentran ante desafíos críticos en la gestión de los datos generados por estos dispositivos. Esta problemática surge debido a la complejidad y la falta de experiencia para gestionar eficazmente los considerables volúmenes de datos producidos por los sensores. Domínguez et al., (2024). mencionan que los desafíos de almacenamiento, procesamiento y visualización de datos de internet de las cosas requieren soluciones especializadas para mantener la integridad y accesibilidad de los datos.

Es importante destacar que los sensores de temperatura y humedad, clasificados como dispositivos de internet de las cosas, son capaces de generar información en intervalos de tiempo sumamente cortos. Estos sensores producen grandes volúmenes de datos, llegando incluso a generar millones de datos en un periodo breve. La capacidad de generar datos de manera tan rápida y masiva plantea desafíos significativos en términos de almacenamiento, procesamiento y análisis. Las empresas deben adoptar soluciones avanzadas para manejar esta avalancha de información. Esto incluye el uso de bases de datos especializadas y herramientas de visualización robustas.

La dificultad para gestionar y aprovechar al máximo los datos generados por los sensores no solo implica un riesgo en términos de almacenamiento y recursos computacionales, no solo eso, también podría conducir a la perdida de información crucial y en la incapacidad para tomar decisiones oportunamente. Esto podría afectar adversamente la calidad de los productos y servicios que estas empresas ponen a disposición, lo que a su vez puede afectar su reputación y la satisfacción del cliente. En este contexto, Balla et al., (2017) destacan los desafíos específicos vinculados con la gestión de datos generados por dispositivos de internet de las cosas. Además,

sugieren que la implementación de una arquitectura de Big Data facilita la superación de estos desafíos.

La falta de una estrategia efectiva para la exploración y visualización de los datos generados por los sensores dificulta la identificación de patrones, tendencias y anomalías, lo que limita la capacidad de las empresas para anticipar problemas y tomar decisiones proactivas. Esto puede traducirse en costos innecesarios, tiempos de inactividad imprevistos y, en última instancia, la pérdida de ingresos.

James (2017) aborda directamente la ineficiencia de la visualización tradicional de datos en grandes volúmenes, proponiendo soluciones de visualización de datos. Este enfoque destaca la importancia de superar las limitaciones en la exploración y presentación de datos de sensores de temperatura y humedad, ofreciendo una perspectiva valiosa para mejorar la toma de decisiones en este contexto específico.

La gestión óptima de los datos provenientes de sensores plantea desafíos significativos debido a la complejidad, variedad y cantidad de información involucrada. Sin embargo, este desafío se ve agravado por el impacto negativo que tiene en la eficiencia operativa y la capacidad de optimización de procesos en las empresas. Según Zhang y Datcu (2015) señala que existe un desajuste entre la habilidad para recolectar información y la habilidad para gestionarla y analizarla, lo que resalta la urgente necesidad de desarrollar soluciones efectivas de exploración y visualización de datos utilizando tecnologías de Big Data y analítica. La analítica se presenta como el cerebro del Big Data y, por extensión, del Big Data de sensores, se trata de analizar extensas cantidades de datos de diversos tipos y de alta relevancia con el fin de encontrar relaciones no evidentes, patrones ocultos y cualquier otra información valiosa. (Joyanes, 2013)

La problemática identificada plantea un desafío significativo que debe abordarse para permitir a las empresas colombianas aprovechar al máximo los datos de sensores y, en última instancia, mejorar su competitividad en el mercado. El éxito en la gestión de estos datos no solo tiene el potencial de aumentar la eficiencia y la rentabilidad, sino que también puede abrir nuevas oportunidades de innovación y crecimiento empresarial en un mundo cada vez más impulsado por la toma de decisiones basada en datos.

Justificación

El proyecto de exploración y visualización de datos de sensores de temperatura y humedad mediante tecnologías de Big Data y analítica se presenta como una iniciativa estratégica y esencial en el contexto actual de las empresas colombianas. Esta justificación se basa en una serie de factores y consideraciones cruciales que respaldan la relevancia y la necesidad de llevar a cabo este proyecto:

Crecimiento exponencial de datos de sensores: En la era actual, la recopilación de datos de sensores, particularmente de temperatura y humedad, ha experimentado un crecimiento exponencial. Los sensores pueden generar información a tasas vertiginosas, lo que conduce a la acumulación rápida de grandes volúmenes de datos. Esta tendencia impulsa la necesidad de una gestión eficiente y efectiva de estos datos.

Impacto en la toma de decisiones empresariales: La información recopilada por sensores juega un papel fundamental en la toma instantánea de decisiones en diversos sectores industriales, incluyendo la manufactura, la logística, la agricultura, la salud y más. Una gestión inadecuada de estos datos puede dar lugar a decisiones deficientes o retrasadas, con consecuencias negativas para la eficiencia operativa y la rentabilidad. Ballesteros (2021) destaca que las tecnologías de Big Data permiten procesar información en tiempos reducidos para obtener soluciones aproximadas, especialmente en situaciones cruciales.

Competitividad en el mercado: Las empresas colombianas se enfrentan a una creciente competencia en los mercados globales. Aquellas que puedan aprovechar de manera efectiva la información generada por sensores de temperatura y humedad tienen la posibilidad de mejorar la eficiencia de sus operaciones, elevar el estándar de sus productos y servicios, y obtener una posición más sólida en el mercado.

Necesidad de analítica avanzada: La necesidad de Analítica Avanzada ha cobrado un papel fundamental en las organizaciones modernas, ya que la analítica de datos, especialmente la analítica avanzada y el uso de tecnologías de Big Data, se ha convertido en un activo estratégico esencial. Esta capacidad permite revelar patrones ocultos, identificar oportunidades de mejora y anticipar problemas antes de que se manifiesten. Es importante destacar que existen aplicaciones de licencia libre que facilitan el desarrollo de soluciones de Big Data, como lo señala Kumar (2021) sobre la aplicación de Hadoop.

Relevancia para la industria y la academia: Este proyecto tiene aplicaciones prácticas inmediatas para la industria y contribuirá al avance del conocimiento en la gestión de datos de sensores de temperatura y humedad y su visualización mediante tecnologías de Big Data y analítica. Esto enriquecerá la base de conocimientos académicos y contribuirá al desarrollo tecnológico del país. Se alinea con la temática explorada en el libro Aplicación de Big Data y Business Analytics de Kumari et al. (2021), centrado en el uso de herramientas analíticas avanzadas para abordar problemas en la sociedad, el medio ambiente y la industria. Nuestro proyecto también busca contribuir a este cuerpo de conocimiento y aplicarlo específicamente a la gestión de datos en el contexto colombiano.

El proyecto de exploración y visualización de datos de sensores de temperatura y humedad mediante tecnologías de Big Data y analítica responde a una necesidad crítica y actual en las empresas colombianas. Aborda la complejidad de manejar grandes volúmenes de datos generados por sensores de temperatura y humedad, lo que tiene un impacto directo en la toma de decisiones, la competitividad y la eficiencia operativa. Este proyecto no solo tiene relevancia

práctica sino también un valor significativo en términos de investigación y desarrollo tecnológico.

Este proyecto de exploración y visualización de datos de sensores de temperatura y humedad mediante tecnologías de Big Data y analítica tiene un claro aporte al campo de investigación de la automatización y herramientas lógicas dentro de la cadena de formación de Electrónica, telecomunicaciones y redes. A continuación, se explican las razones por las cuales este proyecto se clasifica como proyecto aplicado:

- El proyecto se enfoca en la aplicación práctica de tecnologías y métodos para resolver un problema concreto en el campo de la automatización: la gestión de datos de sensores de temperatura y humedad. En lugar de centrarse únicamente en la investigación teórica, busca desarrollar soluciones prácticas y útiles para abordar la acumulación de datos de sensores y su impacto en la toma de decisiones empresariales.
- El proyecto aborda un problema real y crítico que enfrentan las empresas colombianas en la gestión de grandes volúmenes de datos de sensores de temperatura y humedad. Esto demuestra su enfoque aplicado, ya que su objetivo principal es proporcionar soluciones tangibles que mejoren la eficiencia operativa y la competitividad de las organizaciones.
- El proyecto tiene un impacto directo en la industria al abordar la gestión de datos de sensores, lo que es esencial en una variedad de sectores, desde la manufactura hasta la agricultura. Su aplicación práctica puede generar mejoras significativas en la toma de decisiones y la eficiencia operativa de las empresas, lo que contribuirá a su competitividad en el mercado.

Este proyecto se clasifica como aplicado debido a su enfoque en la aplicación práctica de tecnologías y métodos para resolver problemas reales en la industria.

Objetivos

Objetivo General

Desarrollar una solución integral para la exploración y visualización eficiente de datos generados por sensores de temperatura y humedad mediante tecnologías de Big Data y analítica, con el propósito de mejorar la toma de decisiones en empresas colombianas y aumentar su competitividad en el mercado.

Objetivos Específicos

- Implementar un sistema de extracción y almacenamiento de datos de sensores de temperatura y humedad que sea escalable y eficiente en términos de recursos computacionales, garantizando la captura de información en múltiples ejecuciones diarias.
- Aplicar técnicas de preprocesamiento y limpieza de datos para garantizar la calidad y la integridad de los datos recopilados, reduciendo errores y ruido en la información.
- Crear visualizaciones interactivas y avanzadas que permitan a los usuarios explorar y comprender de manera efectiva los datos de sensores de temperatura y humedad, identificar patrones, tendencias y anomalías de manera intuitiva.

Marco Conceptual

Según Rajkumar et al., (2016), Big Data se caracteriza por ser agrupaciones extensas de datos rápidos o complejos que las herramientas de procesamiento de datos convencionales no pueden manejar de manera eficiente.

La comprensión y conceptualización del Big Data ha experimentado una evolución notable a lo largo del tiempo, reflejada en las diversas perspectivas presentadas por expertos destacados, Rajkumar et al., (2016) resaltan esta evolución al mencionar distintas formulaciones de las características fundamentales, conocidas como las "V" del Big Data. Estas perspectivas se detallan a continuación:

1. Volumen: Inicialmente, se enfatizó la importancia del tamaño masivo de los datos, destacando cómo las organizaciones deben gestionar cantidades enormes de información.
2. Velocidad: También se subrayó la rapidez con la que se generan y procesan los datos, resaltando la necesidad de sistemas capaces de manejar flujos de información en tiempo real.
3. Variedad: Se identificó la diversidad de formatos y tipos de datos, desde textos y números hasta imágenes y videos, como un desafío y una oportunidad para las tecnologías de procesamiento de datos.
4. Veracidad: Posteriormente, se añadió la dimensión de la confiabilidad y precisión de los datos, reconociendo que la calidad de la información es crucial para la toma de decisiones basada en datos.
5. Valor: Otra dimensión crítica es el valor que se puede extraer de los datos, subrayando la importancia de convertir la información en insights valiosos y acciones estratégicas.

6. Variabilidad: Finalmente, se introdujo la variabilidad, que reconoce la dinámica y los cambios constantes en los datos, así como la necesidad de gestionar datos en contextos variables.

La diversidad en las formulaciones de las "V" en el Big Data subraya la naturaleza multifacética y dinámica de este campo. Para nuestro proyecto aplicado, esta comprensión en evolución sirve como un marco conceptual clave. La consideración de las múltiples dimensiones de las "V" no solo enriquecerá la exploración y visualización de datos de sensores de temperatura y humedad, sino que también proporcionará una base robusta para la toma de decisiones informada en tiempo cortos, aportando así un valor significativo a nuestro proyecto.

Marco Teórico

En la actualidad, la generación de datos ha alcanzado proporciones extraordinarias, con millones de datos producidos en períodos breves. Las organizaciones deben comprender y analizar estos flujos para optimizar sus procesos y el manejo de datos, ahora visto como un recurso clave con potencial para impulsar ganancias. La gestión eficiente de esta avalancha de datos es crucial para el crecimiento económico de las empresas, que depende de la calidad de sus mecanismos de manipulación de datos.

Nitin (2021) menciona la creciente relevancia del Big Data en el ámbito empresarial, señalando tendencias y proyecciones que respaldan su trascendencia. Según una encuesta, el 97% de las empresas planea aumentar su inversión en tareas analíticas, reflejando el reconocimiento de la importancia estratégica del Big Data en la toma de decisiones empresariales.

El Big Data es esencial para el crecimiento e innovación en diversas áreas, optimizando procesos con grandes volúmenes de datos. Según Santos y Costa (2020), el Big Data ha crecido significativamente en sectores como ciudades inteligentes, manufactura, comercio minorista, finanzas, desarrollo de software, medio ambiente y medios digitales. Los autores proyectan que el almacenamiento tradicional, conocido como Data Warehouse, será sustituido por el Big Data Warehouse, enfrentando retos en el diseño e implementación de la capa lógica y física.

Webber y Zheng (2020) resaltan el papel fundamental del análisis de grandes conjuntos de datos en la administración de instituciones educativas. En el ámbito de la educación superior, donde los líderes enfrentan desafíos como el aumento de costos operativos y la competencia, la toma de decisiones basada en datos emerge como una herramienta estratégica. Esta observación subraya la universalidad y relevancia de la toma de decisiones basada en datos en diversos

contextos, no solo en la educación superior. La capacidad de abordar desafíos comunes, como el aumento de costos operativos y la intensificación de la competencia, es inherente a diversos sectores y mercados.

La toma de decisiones informada por datos es una herramienta estratégica fundamental para líderes administrativos y profesionales de investigación institucional en el ámbito educativo, así como para profesionales en otros campos. En un contexto empresarial en constante evolución y crecientemente complicado, la capacidad de analizar y utilizar en la actualidad, el manejo efectivo de los datos es crucial para sobrevivir y prosperar en el futuro.

Tanwar et al. (2020) abordan el papel fundamental del Big Data en el ámbito de la salud, delineando su impacto significativo en diversos aspectos cruciales de la atención médica. Enfocándose en los macrodatos, generados en parte por el Internet de las cosas y analizados con algoritmos especializados, la obra destaca su contribución invaluable en áreas de vital importancia. En particular, se dedica especial atención a la medicina preventiva, donde los macrodatos posibilitan prever y abordar problemas de salud antes de que se desarrollen por completo. La capacidad de recopilar y analizar datos a gran escala facilita la identificación de patrones y tendencias que pueden ser indicativos de riesgos para la salud, permitiendo intervenciones proactivas y estrategias preventivas más efectivas.

Además, los grandes conjuntos de datos mejoran el seguimiento de la salud en grupos específicos. Al aprovechar la recopilación continua de datos, especialmente aquellos provenientes de dispositivos conectados, se logra una comprensión más completa y detallada de la salud de poblaciones específicas. Esto no solo posibilita la personalización de los enfoques médicos, sino que también facilita la identificación de disparidades en la salud y la implementación de intervenciones más equitativas.

Estas capacidades de Big Data no se limitan al sector de la salud o la educación; trascienden estas fronteras y se convierten en una ventaja competitiva en cualquier mercado. Las organizaciones que pueden aprovechar el análisis de datos para tomar decisiones estratégicas están mejor posicionadas para adaptarse a cambios, identificar oportunidades y optimizar sus operaciones. En última instancia, la toma de decisiones basada en datos se erige como un catalizador clave para la eficacia y la eficiencia en un panorama empresarial en constante evolución.

Para enfrentar estos desafíos, las organizaciones han recurrido a soluciones tecnológicas avanzadas. Apache, uno de los pioneros en el manejo de grandes cantidades de datos, ha sido fundamental en esta transformación. Sus contribuciones al desarrollo de software de código abierto han revolucionado la gestión y el procesamiento de información a gran escala.

Las herramientas y proyectos de Apache como Hadoop, Spark, NiFi y Kafka permiten abordar los desafíos del Big Data, facilitando almacenamiento y procesamiento distribuido establecen bases para arquitecturas escalables. Santos y Costa (2020) señalan el papel crucial de Hadoop, especialmente cuando se integra con Spark para extracción, transformación y carga de datos. Además, Hive es esencial para un almacenamiento eficiente. La integración de estas tecnologías no solo potencia el procesamiento, sino que optimiza la gestión y almacenamiento de datos, fundamentales para un almacén de datos a gran escala.

Las herramientas de Apache son cruciales para la extracción de información. Wnęk y Boryło (2023) destacan el uso de Apache NiFi en redes 5G/6G para capturar datos a intervalos regulares. Apache NiFi permite la obtención de información en tiempo real, ofreciendo una solución configurable y escalable para manejar grandes volúmenes de datos de transceptores fotónicos. Esto optimiza el procesamiento y distribución eficientes, mejorando la gestión de

datos en tiempo real y facilitando la implementación de sistemas de monitoreo y control en redes ópticas de próxima generación.

En este contexto, Apache NiFi se posiciona como una herramienta ideal para la extracción de datos generados por sensores de temperatura y humedad. Su flexibilidad y capacidad de configuración permiten una integración efectiva, facilitando la recopilación y procesamiento de la información que generan los sensores.

Kumar (2021) propone una solución innovadora para el manejo eficiente de grandes volúmenes de datos a través de un proyecto de código abierto basado en Hadoop. Este enfoque se inspira en documentos publicados por Google, detallando los desafíos significativos que enfrentan los ingenieros al almacenar y procesar cantidades masivas de datos. La propuesta de Kumar no solo reconoce los obstáculos en el tratamiento de datos a gran escala, sino que también presenta una respuesta práctica y aplicable, destacando la relevancia de su enfoque en la gestión de información a gran escala.

Por otro lado, Ankam (2016) y Karim y Alla (2017) resaltan la importancia de integrar herramientas como Hadoop y Spark, proporcionando un enfoque detallado sobre los componentes clave de Spark y abordando temas contemporáneos en el ámbito del Big Data.

Business Wire (2020) informa sobre los reconocimientos que Dremio ha recibido en el campo de Big Data. Dremio es reconocida como una empresa líder en motores de lago de datos y ha sido elogiada por su contribución al ámbito de Big Data. Su plataforma ofrece mayor productividad y tiempos de obtención de conocimientos más rápidos para los equipos de inteligencia empresarial y ciencia de datos. Con consultas ultrarrápidas y una capa semántica de

autoservicio, Dremio facilita la integración y el análisis de datos procedentes de sensores de temperatura y humedad en entornos de almacenamiento de lago de datos.

Business Wire (2020) destaca los logros de Dremio en el ámbito del Big Data, resaltando su liderazgo en motores de lago de datos y su capacidad para mejorar la productividad y los tiempos de análisis. Los reconocimientos recibidos reflejan su habilidad para abordar desafíos en el procesamiento de datos en tiempo real, consolidando su relevancia en la industria.

Dremio es ideal para el proyecto aplicado enfocado en el tratamiento de datos generados por los sensores de temperatura y humedad, ofrece soluciones avanzadas para la gestión eficiente y el análisis de datos en tiempo real, siendo una contribución valiosa al campo del Big Data. Su capacidad para integrar y procesar datos de múltiples fuentes permite una mayor flexibilidad y precisión en la interpretación de los datos.

Los almacenes de datos modernos han simplificado este proceso al permitir a los analistas de datos realizar ingeniería de funciones mediante la transformación, validación y agregación de datos utilizando lenguajes de consulta estructurados como SQL. Una herramienta clave en este proyecto es Data Build Tool (DBT), una plataforma de código abierto diseñada para crear y orquestar canalizaciones SQL.

La integración de DBT en plataformas no solo simplifica la ingeniería de funciones, sino que también amplía la capacidad de realizar procesos de ETL (Extracción, Transformación y Carga) en datos generados por sensores de temperatura y humedad. Esto facilita que los analistas de datos manejen grandes cantidades de información de forma más eficiente y eficaz, lo que mejora su capacidad para analizar datos y tomar decisiones informadas en contextos de Big Data.

La visualización de datos desempeña un papel esencial en el análisis y la interpretación de la información en el entorno actual de abundancia de datos. Al presentar datos de manera gráfica e intuitiva, se facilita la identificación de patrones, tendencias y relaciones que podrían pasar desapercibidos en conjuntos de datos crudos. La capacidad de transformar datos complejos en representaciones visuales claras no solo simplifica la comprensión, sino que también permite una toma de decisiones más informada y eficiente.

Los beneficios de la visualización de datos son diversos y van más allá de la mera presentación estética. En primer lugar, mejora la comunicación de información, ya que las representaciones visuales son más accesibles y comprensibles para una audiencia diversa. Además, facilita la identificación rápida de patrones anómalos o tendencias emergentes, lo que puede ser crucial en entornos comerciales y científicos.

La capacidad de interactuar con visualizaciones de datos también es un aspecto valioso. Las herramientas interactivas permiten explorar detalles específicos y obtener información más detallada al hacer clic o interactuar con elementos específicos de la visualización. Este enfoque no solo optimiza cómo los usuarios interactúan, sino que también fortalece su capacidad para explorar datos de manera más efectiva.

Miller (2017) explora de manera detallada la esencialidad de abordar la visualización de grandes conjuntos de datos de manera efectiva. Destaca cómo el análisis convencional se ve directamente afectado por la magnitud del Big Data, exponiendo estrategias efectivas y eficientes para superar los desafíos visuales asociados, resalta la importancia crucial de la visualización en tiempo real y su aplicabilidad en una variedad de casos de uso.

También explora en detalle las herramientas de visualización de datos más utilizadas., como Splunk y Tableau. Asimismo, destaca la conexión entre el valor del Big Data visual y las herramientas de Business Intelligence, subrayando la sinergia entre la visualización de datos y la toma de decisiones estratégicas en entornos empresariales.

El análisis de grandes conjuntos de datos se ha convertido en un elemento clave para la toma de decisiones estratégicas y la mejora de procesos en múltiples campos. Su comprensión y manejo eficiente se vuelven imperativos para las organizaciones que buscan innovar y mantener su competitividad en un entorno empresarial en constante cambio. Como hemos visto, el análisis informado por datos no solo ofrece una ventaja competitiva, sino que también promueve la adaptabilidad y el crecimiento sostenible. Ahora, con una comprensión sólida de su relevancia, pasamos a explorar las metodologías clave que respaldan la aplicación práctica del Big Data.

Para aplicar eficazmente los conceptos teóricos del Big Data en proyectos reales, es vital adoptar enfoques metodológicos sólidos y adaptados a las necesidades específicas de cada proyecto. En esta sección, delinearemos las metodologías seleccionadas y justificaremos su elección en función de los objetivos establecidos. Desde la recolección y procesamiento de información hasta su evaluación y representación visual, cada paso en el proceso de gestión de datos requiere una estrategia cuidadosamente diseñada para garantizar resultados precisos y significativos. A continuación, se mencionan algunas de las metodologías utilizadas en proyectos que involucran grandes volúmenes de datos.

CRISP-DM, por sus siglas en ingles Cross Industry Standard Process for Data Mining: Esta metodología convencional ofrece un método organizado y sistemático para realizar proyectos de minería de datos. IBM (2021) indica que la metodología consta de seis fases principales: Entendimiento del negocio, Entendimiento de los datos, Preparación de los datos,

Modelado, Evaluación y Despliegue. Cada fase se subdivide en tareas específicas que guían al equipo a lo largo del proceso, desde la comprensión inicial del problema hasta la implementación de soluciones en un entorno operativo.

SEMMA, por sus siglas en ingles Sample, Explore, Modify, Model, Assess: Desarrollada por SAS Institute, SEMMA se compone de cinco etapas diseñadas para guiar proyectos de minería de datos de manera iterativa. Desde el muestreo y la exploración inicial de los datos hasta la evaluación de modelos y soluciones propuestas, esta metodología permite a los analistas explorar y comprender los datos de manera exhaustiva, construir modelos predictivos y evaluar su efectividad.

DMAMC, por sus siglas en español Definir, Medir, Analizar, Mejorar, Controlar: Aunque originalmente diseñada en el contexto de Six Sigma, la metodología DMAMC puede adaptarse para proyectos de datos. Consta de cinco etapas que incluyen articular el problema de manera precisa, recabar y examinar datos pertinentes, detectar oportunidades de mejora, aplicar soluciones efectivas y establecer medidas de seguimiento para asegurar la continuidad de los cambios implementados.

KDD, por sus siglas en ingles Knowledge Discovery in Databases: Este enfoque amplio abarca varias etapas, incluida la selección de datos, preprocesamiento, transformación, minería de datos e interpretación de resultados. Aunque menos prescriptivo que otras metodologías, el proceso KDD proporciona un marco general para la exploración y el descubrimiento de conocimientos en grandes conjuntos de datos.

Para el análisis de los datos de sensores de temperatura y humedad, la metodología que emerge como la más adecuada es CRISP-DM. Su estructura robusta y enfoque iterativo permiten

abordar sistemáticamente cada etapa del proceso de minería de datos, desde la comprensión inicial del problema hasta la evaluación de los resultados. Además, su flexibilidad permite adaptarse a las necesidades específicas de este proyecto, asegurando un enfoque personalizado y eficaz.

Propuesta de Solución

La exploración y visualización eficiente de datos provenientes de sensores de temperatura y humedad representan un desafío debido al alto volumen de datos que generan estos dispositivos. Para abordar esta problemática, se propone una solución integral fundamentada en tecnologías de Big Data y analítica, aprovechando herramientas líderes en el campo. Esta estrategia se estructura sobre una metodología sólida y tecnologías avanzadas, detalladas a continuación.

La metodología seleccionada para la ejecución del proyecto es CRISP-DM, un marco robusto y estructurado diseñado para guiar el proceso de Big Data desde la identificación inicial del problema hasta la implementación exitosa. Este enfoque metodológico, perfeccionado por un consorcio de expertos de la industria y la academia, se ha consolidado como el estándar preferido en el ámbito de Big Data. Su capacidad para adaptarse y proporcionar un enfoque sistemático en una diversidad de contextos y aplicaciones lo convierte en una elección especialmente pertinente para nuestro proyecto aplicado.

CRISP-DM se distingue por su estructura clara y paso a paso, facilitando la comprensión y ejecución del proceso de minería de datos. Proporciona un marco que abarca desde la comprensión inicial del problema hasta la fase de despliegue e iteración continua. Como guía flexible, permite ajustes y personalizaciones según las necesidades específicas del proyecto, asegurando una aplicación efectiva en diferentes entornos empresariales. Este enfoque iterativo

favorece la mejora continua y la adaptabilidad a los cambios en el entorno empresarial, proporcionando un camino claro y estructurado para la implementación exitosa del proyecto, respaldado por las mejores prácticas y la experiencia acumulada en el campo de la minería de datos.

Ciclo de Vida CRISP-DM

El ciclo de vida de CRISP-DM ofrece un marco sólido y estructurado para guiar el proceso de Big Data desde la concepción del proyecto hasta su implementación exitosa. Este enfoque metodológico, respaldado por una amplia experiencia en la industria y la academia, se compone de varias fases interrelacionadas, cada una vital para alcanzar los objetivos del proyecto aplicado:

Comprensión del negocio: En esta fase inicial, se trabaja en estrecha colaboración con los Stakeholders para identificar los objetivos comerciales y definir el problema en términos de negocio. Establecer una comprensión clara de los requisitos comerciales sienta las bases para las decisiones futuras y garantiza que el proyecto esté alineado con las necesidades del negocio.

Comprensión de los datos: Una vez establecido el contexto empresarial, se procede a adquirir y evaluar los datos relevantes. Aquí es donde entra en juego Apache NiFi, facilitando la captura eficiente de datos desde los sensores de temperatura y humedad. Esta herramienta permite la creación de flujos de trabajo que garantizan la transferencia efectiva de datos al sistema de Big Data, asegurando una ingesta de datos sin problemas.

Preparación de datos: Durante esta etapa crucial, se organiza y almacena los datos de manera adecuada utilizando sistemas como Hadoop. Además, se aplican técnicas de

preprocesamiento para abordar posibles irregularidades en los datos, asegurando que estén en un formato óptimo para su posterior análisis.

Modelado: En esta etapa, se emplean técnicas de modelado de datos para desarrollar modelos que resuelvan el problema identificado durante la fase de comprensión del negocio. Se exploran diferentes enfoques y se ajustan los modelos para mejorar su rendimiento. La transformación de datos se realiza con la ayuda de herramientas como Dremio y dbt (Data Build Tools), simplificando la consulta de datos y acelerando la creación de modelos para la generación de tablas materializadas.

Despliegue: Una vez desarrollados satisfactoriamente los modelos, se procede a su implementación en Power BI para simplificar la visualización y el análisis de los resultados. Se establece un plan de despliegue que incluye la integración con sistemas existentes y la gestión de cambios. La solución se monitorea de cerca en Power BI para garantizar su rendimiento continuo y su alineación con los objetivos del negocio. La capacidad de Power BI para crear paneles interactivos y reportes detallados permite a los stakeholders acceder fácilmente a la información relevante y tomar decisiones informadas basadas en los resultados del análisis de datos.

Al combinar CRISP-DM con estas tecnologías y herramientas, se asegura una implementación efectiva y precisa del proyecto aplicado, abordando de manera integral cada etapa del ciclo de vida de Big Data.

Aportes de la Solución

Esta estrategia integral proporciona una serie de beneficios significativos, específicamente diseñados para maximizar el valor de los datos de temperatura y humedad. La combinación de tecnologías como Hadoop, NiFi, Dremio, dbt (Data Build Tools) y Power BI se

traduce en una plataforma completa para la exploración y visualización efectiva de estos datos. Este enfoque responde a la necesidad imperiosa de abordar los desafíos actuales en la exploración de datos de sensores. De este modo, apoyando la toma de decisiones basadas en datos exactos y detallados.

El producto final, un informe personalizado, ofrece una presentación exhaustiva de los datos recopilados por los sensores de temperatura y humedad. Lo más destacado de este informe son los gráficos interactivos, una herramienta poderosa que no solo hace que la información sea más accesible, sino que también permite una exploración más profunda y significativa de los datos. Los beneficios clave son los siguientes:

- Identificación de patrones y tendencias: Los gráficos interactivos facilitan la rápida identificación de patrones y tendencias en los datos de temperatura y humedad. Esto permite a los usuarios detectar cambios significativos en el ambiente monitoreado y tomar medidas preventivas o correctivas según sea necesario.
- Análisis avanzado de series temporales: La capacidad de interactuar con los datos en forma de series temporales permite un análisis detallado y sofisticado. Los usuarios pueden examinar la evolución de las variables a lo largo del tiempo, identificar ciclos recurrentes o estacionales, y evaluar la eficacia de las intervenciones implementadas.
- Personalización según necesidades específicas: El informe personalizado se adapta a las necesidades particulares de cada usuario, permitiendo la inclusión de métricas específicas y la configuración de parámetros de visualización según sea necesario. Esto garantiza que la información presentada sea relevante y útil para cada contexto de aplicación.
- Facilidad de interpretación: Los gráficos interactivos ofrecen una representación visual intuitiva de los datos, lo que facilita su interpretación incluso para usuarios no técnicos.

Esto garantiza que todos los miembros del equipo puedan comprender fácilmente la información presentada y tomar decisiones informadas basadas en ella.

- Eficiencia en la toma de decisiones: Al proporcionar una visión clara y detallada de los datos de temperatura y humedad, el informe personalizado agiliza el proceso de toma de decisiones. Los usuarios pueden detectar rápidamente áreas que necesitan mejoras, prever posibles inconvenientes y tomar medidas correctivas a tiempo, lo que conduce a una gestión más eficiente y eficaz de los recursos.

El informe personalizado con gráficos interactivos no solo ofrece una presentación efectiva de los datos de temperatura y humedad, sino que también potencia el análisis y la toma de decisiones en diversos contextos, desde la gestión de la calidad del aire hasta la optimización de procesos industriales.

Aportes a Interrogantes Empresariales

Esta capacidad posibilita abordar interrogantes empresariales fundamentales, tales como:

¿Cuáles son las tendencias de temperatura y humedad a lo largo del tiempo?

El reporte puede mostrar gráficos de líneas o series temporales que ilustren la evolución de la temperatura y la humedad durante períodos específicos, días, semanas, meses o años.

¿Existen patrones estacionales en los datos de temperatura y humedad?

Mediante análisis de series temporales y visualizaciones estacionales, se pueden identificar patrones recurrentes a lo largo del año o en períodos específicos.

¿Cuáles son los valores máximos, mínimos y promedio de temperatura y humedad?

El reporte puede incluir métricas resumidas que muestren los valores máximos, mínimos y promedio de temperatura y humedad en diferentes períodos de tiempo.

¿Hay correlaciones entre la temperatura y la humedad?

Mediante gráficos de dispersión o análisis de correlación, se pueden identificar relaciones entre la temperatura y la humedad, lo que puede ser útil para comprender mejor el entorno.

¿Cómo varía la temperatura y la humedad en diferentes ubicaciones?

Los datos incluyen información de ubicación, se pueden crear mapas o gráficos geoespaciales que muestren la variación de la temperatura y la humedad en diferentes lugares.

Línea de Investigación Futura

Una dirección de investigación prometedora para el futuro implica en la aplicación y desarrollo de técnicas avanzadas de aprendizaje automático para predecir los datos de temperatura y humedad capturados por sensores. Al emplear algoritmos de Machine Learning en este contexto, no solo se busca comprender los datos existentes, sino también anticipar y prever sus fluctuaciones futuras con una alta precisión. Este enfoque tiene el potencial de mejorar significativamente la capacidad de tomar decisiones informadas y aplicadas al contexto de negocios. Al obtener análisis predictivos precisos sobre la temperatura y la humedad, las empresas pueden optimizar sus operaciones, gestionar de manera más eficiente sus recursos y anticiparse a las demandas del mercado. Además, este análisis avanzado permite identificar oportunidades comerciales emergentes y mitigar posibles riesgos, lo que resulta en una ventaja competitiva considerable en un entorno empresarial en constante cambio.

Desarrollo Ciclo de Vida CRISP-DM

Fase: Comprensión de Negocio

En esta fase, se aborda la comprensión de los objetivos comerciales y la problemática identificada. En nuestro caso, esto se resume en el objetivo general del proyecto aplicado: desarrollar una solución integral para la exploración y visualización eficiente de datos generados por sensores de temperatura y humedad, mediante tecnologías de Big Data y analítica, con el propósito de mejorar la toma de decisiones en empresas colombianas y aumentar su competitividad en el mercado.

Fase: Comprensión de los Datos

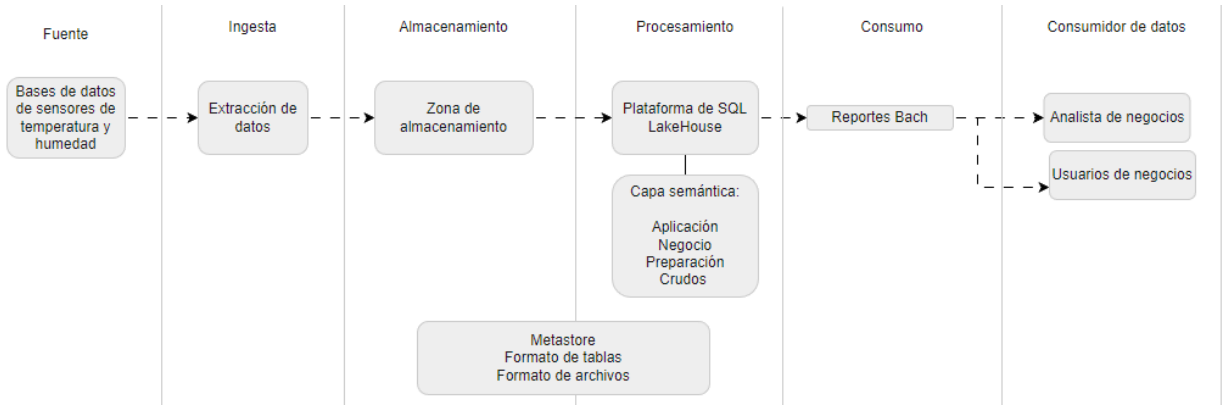
Aquí, es esencial obtener una visión global de los datos, lo cual se logra mediante una representación visual llamada arquitectura de datos. La arquitectura de datos es un esquema detallado que representa el trayecto de los datos desde su origen y recolección inicial, pasando por su almacenamiento, procesamiento, hasta su eventual consumo.

Este diseño detallado guía el flujo de los datos del proyecto aplicado, delineando cómo se recopilan, almacenan, manipulan y utilizan en distintas etapas y contextos, la Figura 1 brinda una

visión completa de la arquitectura de datos. Cada componente y su interacción están claramente delineados, lo que facilita la comprensión del flujo de los datos a lo largo de todo el proceso.

Figura 1

Etapas de la arquitectura de datos



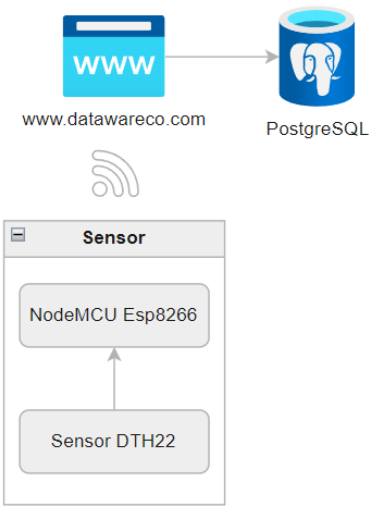
A continuación, se describen detalladamente cada una de las etapas de la arquitectura de datos:

Fuente. También conocida como "origen de datos", esta etapa se centra en identificar y describir todas las fuentes de información que nutrirán sistema. El principal origen de datos consiste en la información generada por los sensores de temperatura y humedad. La operatividad de estos sensores se ilustra a continuación.

La Figura 2 detalla el proceso de captura de datos de los sensores de temperatura y humedad. Estos dispositivos están compuestos por dos componentes principales: el sensor DTH22, encargado de medir temperatura y humedad, y NodeMCU Esp8266, que proporciona conectividad WiFi para enviar los datos a la plataforma web www.datawareco.com.

Figura 2

Funcionamiento de los sensores



Una vez adquiridos las variables ambientales por el sensor, los datos son transmitidos eficientemente al sitio web a través del NodeMCU Esp8266, garantizando una operación fluida y confiable. En www.datawareco.com, los datos son almacenados en una base de datos PostgreSQL, asegurando su organización y disponibilidad para análisis posterior.

Los datos se almacenan estructuradamente en la base de datos PostgreSQL de la aplicación web, organizados en varias tablas que se mencionan a continuación.

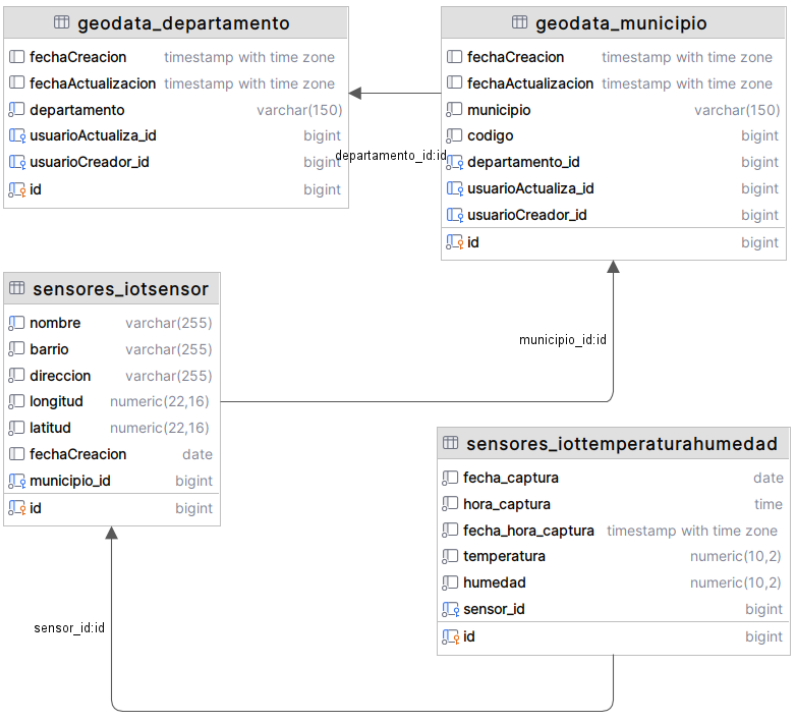
1. sensores_iotsensor: tiene los datos descriptivos de cada uno de los sensores.
2. sensores_iottemperaturahumedad: datos de temperatura y humedad.
3. geodata_departamento: datos geoespaciales de los departamentos de Colombia
4. geodata_municipio: datos geoespaciales de los municipios de Colombia

Estas tablas están interconectadas, lo que requiere realizar consultas adicionales para establecer las relaciones entre ellas. En la Figura 3 se muestra el diagrama relacional, que

permite identificar claramente las relaciones existentes entre las tablas de la base de datos en la aplicación web.

Figura 3

Diagrama relacional de la aplicación web



Las llaves primarias o foráneas en una base de datos se clasifican como datos cualitativos debido a su función principal de establecer relaciones simbólicas entre diferentes tablas. Aunque puedan estar representadas por valores numéricos, su significado no radica en la cantidad numérica en sí, sino en la conexión que establecen con la clave primaria de otra tabla. Las llaves foráneas actúan como enlaces cualitativos entre registros, permitiendo la integración y la referencia cruzada de información entre entidades relacionadas. Su naturaleza cualitativa radica en su capacidad para representar vínculos y asociaciones, más que en proporcionar valores cuantitativos independientes.

Las fechas, a pesar de estar representadas por valores numéricos, se clasifican como datos cualitativos debido a su interpretación basada en la calidad temporal más que en una cantidad numérica específica. Estas proporcionan información sobre el aspecto temporal de los eventos o registros, permitiendo la organización y referencia cronológica de los datos. Aunque se expresen en términos de días, meses y años, su importancia recae en la secuencia y la relación temporal entre diferentes eventos o situaciones. Las fechas actúan como marcadores cualitativos que facilitan la comprensión de la cronología de eventos, pero su esencia no reside en la magnitud numérica en sí misma.

Es importante entender el contexto de cómo se obtienen los datos. Para el proyecto aplicado, accederemos directamente a los datos almacenados en la base de datos de la aplicación web generados por los sensores de temperatura y humedad que están.

Ingesta. En esta fase, se lleva a cabo la tarea específica de extraer los datos generados por los sensores de temperatura y humedad, los cuales están almacenados en una base de datos PostgreSQL a través de la aplicación web. Esta operación se lleva a cabo utilizando Apache NiFi para facilitar el proceso de extracción y transferencia de datos. La utilización de Apache NiFi permite una gestión eficiente y automatizada de este flujo de datos, garantizando una transferencia segura y confiable. Esta integración entre la aplicación web y Apache NiFi ofrece una solución robusta y escalable para la recopilación de datos de temperatura y humedad, lo que contribuye significativamente a mejorar el análisis de estos datos en tiempo real.

Almacenamiento. El almacenamiento de datos se lleva a cabo a través de Apache Hadoop, un sistema distribuido ampliamente utilizado para gestionar grandes volúmenes de datos. Esta infraestructura desempeña dos funciones fundamentales: en primer lugar, almacena los datos consultados generados por los sensores de temperatura y humedad en un formato

parquet. En segundo lugar, dispone los datos para la fase procesamiento que se usara con la herramienta Dremio.

Además de almacenar los datos brutos, Hadoop también se encarga de almacenar los metadatos generados por Dremio. Estos metadatos incluyen información crucial sobre los datos almacenados, como los esquemas de las tablas, los formatos de los archivos y otros detalles importantes para su correcta interpretación y uso.

Ampliando esta idea, es importante destacar que el almacenamiento eficiente de datos es fundamental para garantizar un flujo de trabajo efectivo en cualquier sistema de gestión de datos. La capacidad de almacenar, gestionar y acceder a grandes volúmenes de datos de manera eficiente es un requisito indispensable en entornos modernos de análisis de datos, donde la toma de decisiones informada depende en gran medida de la disponibilidad y la calidad de los datos almacenados.

Procesamiento. En esta fase, se emplea la herramienta de Dremio, una plataforma de SQL lake house que facilita la interacción con una capa semántica. Esta capa permite clasificar los datos en diferentes categorías, como aplicaciones de negocio, datos preparados y datos en su estado crudo. Además de utilizar Dremio, se trabajará en colaboración con DBT (Data Build Tools). El objetivo de esta colaboración es crear tablas materializadas que permitan almacenar un histórico de los datos generados por los sensores de temperatura y humedad.

Esta estrategia integrada busca no solo gestionar eficazmente los datos generados por los sensores, sino también proporcionar un sistema robusto para su análisis y utilización. Al aprovechar las capacidades de Dremio para la gestión de datos con la capa semántica, junto con

la funcionalidad de DBT para la creación de tablas materializadas, se establece una base sólida para garantizar la disponibilidad, integridad y utilidad de los datos a lo largo del tiempo.

Además, la implementación de tablas materializadas permitirá no solo almacenar registros históricos de los datos, sino también agilizar el acceso y la consulta de información relevante. Esto contribuirá a mejorar la eficiencia en la toma de decisiones y el desarrollo de análisis predictivos basados en datos. La combinación de Dremio y DBT representa una estrategia integral para gestionar y aprovechar al máximo los datos generados por los sensores de temperatura y humedad, brindando una sólida base para la innovación y la mejora continua en el proyecto.

Consumo. La herramienta de analítica es Power BI. Esta plataforma desempeñará un papel fundamental al generar informes detallados y visualmente atractivos mediante consultas a los datos recopilados en las etapas anteriores. Power BI ofrecerá una visión completa y accesible de las ideas y tendencias clave extraídos del conjunto de datos, proporcionando así una base sólida para la toma de decisiones informadas. La habilidad del sistema para manejar grandes cantidades de información de manera eficiente, junto con su interfaz fácil de usar, permitirá a los usuarios finales interpretar los resultados con facilidad y obtener conclusiones valiosas de los datos generados por los sensores de temperatura y humedad.

Esta integración de herramientas y tecnologías garantiza un flujo de trabajo eficiente y robusto, desde la captura hasta el consumo de datos, mejorando así la calidad del análisis y la toma de decisiones en el proyecto aplicado.

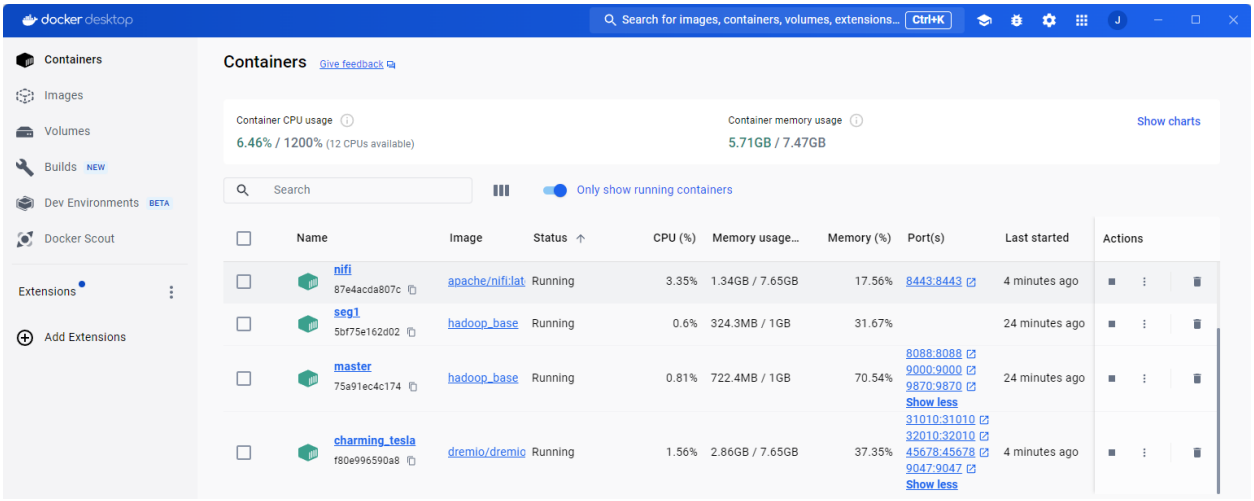
Fase: Preparación de Datos

En esta fase es fundamental establecer las bases técnicas necesarias. En este sentido, hemos optado por utilizar Docker para simplificar la instalación y configuración de las herramientas requeridas. Docker nos brinda la capacidad de crear contenedores que emulan entornos virtuales Linux, lo que nos proporciona un ambiente controlado y coherente para el despliegue de Hadoop, Dremio y NiFi. Estas herramientas se instalarán y configurarán dentro de sus respectivos contenedores, garantizando su disponibilidad inmediata para su uso en el proyecto.

En la Figura 4 se muestran los contenedores configurados específicamente para este proyecto aplicado. El contenedor "nifi" alberga la instalación de Apache NiFi, mientras que los contenedores "master" y "seg1" conforman un clúster de Apache Hadoop, estableciendo un nodo maestro y un nodo esclavo, respectivamente. Por último, en el contenedor "charming_tesla" se ha implementado y configurado Dremio. Cada contenedor está configurado con puertos de comunicación particulares, lo que permite el acceso a la interfaz web de cada herramienta para su interacción y gestión.

Figura 4

Docker instalación herramientas

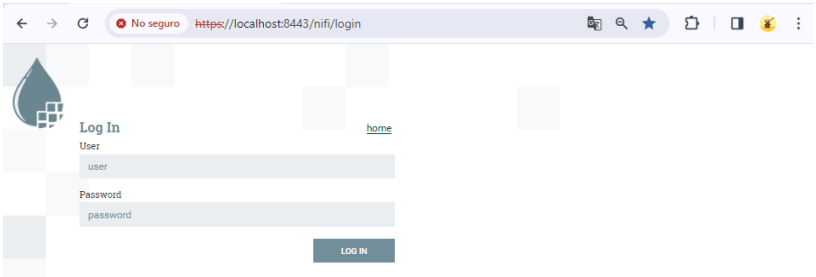


Esta configuración garantiza la disponibilidad de todas las herramientas necesarias en un entorno controlado y listo para el desarrollo del proyecto. Facilita la gestión y el análisis de los datos generados por los sensores de temperatura y humedad, permitiendo un flujo de trabajo eficiente y ordenado. Para verificar el funcionamiento de los servicios, se accede a la interfaz de cada herramienta y se realiza una comprobación de su operatividad.

Se puede observar la interfaz web de Apache NiFi, una herramienta esencial en este proceso, en la Figura 5. Apache NiFi se encarga de extraer datos desde la fuente de origen, capturando eficientemente los datos generados por los sensores de temperatura y humedad.

Figura 5

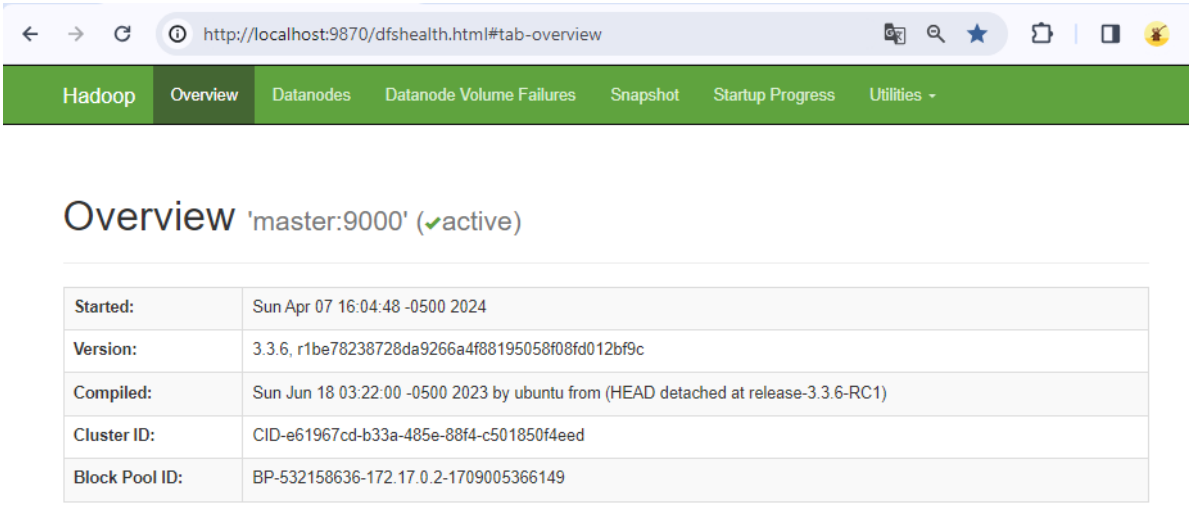
Inicio de sesión Apache NiFi



Se observa el acceso al clúster de Hadoop, el cual está compuesto por un nodo maestro y un nodo esclavo en la Figura 6. El nodo maestro cumple la función de coordinar las operaciones, mientras que los nodos esclavos se encargan del almacenamiento de datos. Esta configuración ofrece la posibilidad de realizar un escalamiento horizontal, lo que facilita la incorporación de nuevos nodos esclavos para mejorar tanto el rendimiento como el procesamiento de datos. Es importante destacar que Hadoop juega un papel fundamental en el almacenamiento de los datos una vez que han sido extraídos por Apache NiFi.

Figura 6

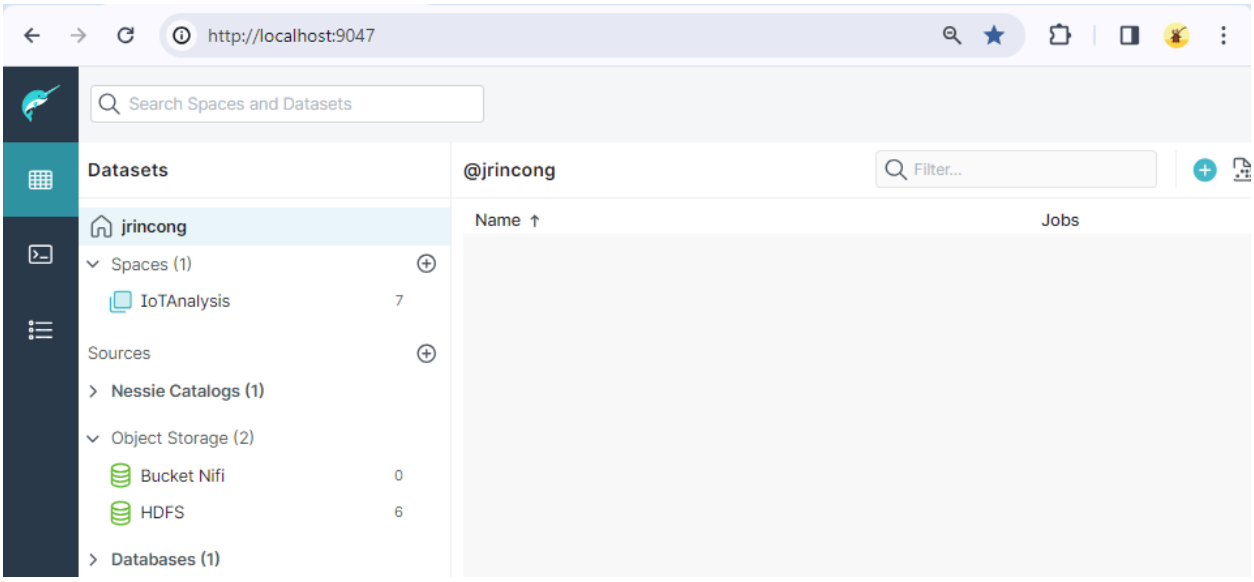
Resumen cluster Apache Hadoop



En la Figura 7, se aprecia el acceso a la interfaz web de Dremio, la cual desempeña un papel fundamental al acceder a los datos almacenados en el clúster de Hadoop. Dremio actúa como un punto central para la gestión de los datos provenientes de los sensores de temperatura y humedad, ofreciendo una interfaz SQL que permite realizar consultas declarativas sobre esta información. Esta funcionalidad simplifica la integración y el análisis eficiente de los datos recopilados, garantizando un acceso unificado y optimizado que atiende a todas las necesidades del proyecto de manera efectiva.

Figura 7

Página de inicio Dremio



Se procederá a instalar las herramientas DBT (Data Build Tool) y Power BI en la computadora local. Estas herramientas se integrarán con las aplicaciones alojadas en los contenedores de Docker mencionados anteriormente, lo que permitirá una interacción fluida y eficiente entre los distintos componentes del sistema.

Comprendiendo la base tecnológica, la extracción de los datos generados por los sensores de temperatura y humedad en Apache NiFi, plataforma diseñada específicamente para gestionar flujos de datos de manera eficiente. En este proceso, se crea un flujo de datos que establecerá una conexión con la base de datos de PostgreSQL de la aplicación web y consultará los datos generados por los sensores de temperatura y humedad. Posteriormente, estos datos se almacenarán en el clúster de Hadoop, todo dentro del mismo flujo de datos.

Este flujo de datos se programa para ejecutarse a intervalos específicos, ya sea por segundos, minutos u horas, según las necesidades del contexto. En nuestro caso, se configurará

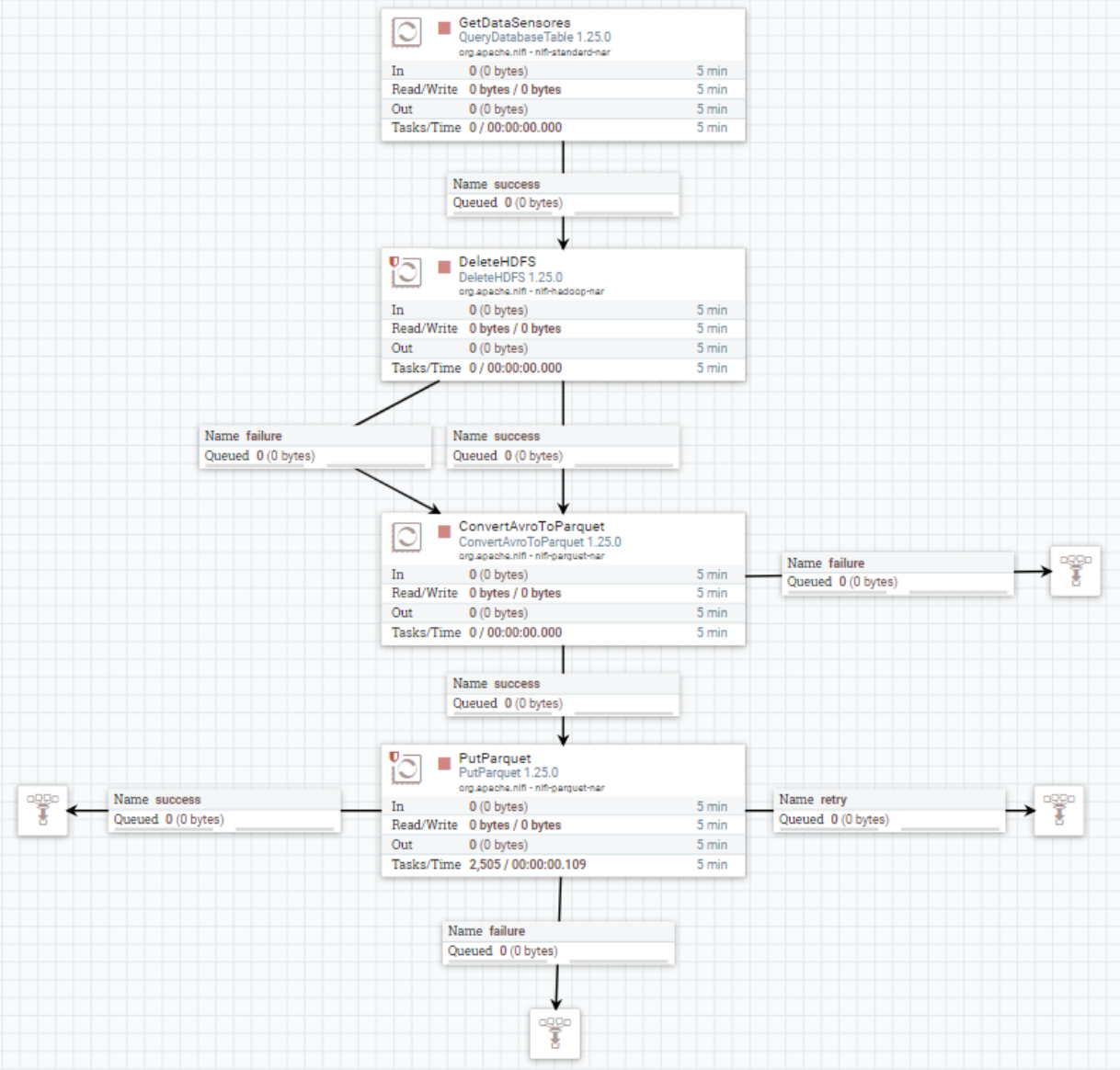
para que se ejecute cada 10 minutos, asegurando así una actualización constante de la información recopilada.

Este enfoque permite una extracción eficiente de los datos de los sensores, garantizando su disponibilidad y actualización periódica en el clúster. Además, proporciona una forma automatizada y escalable de gestionar la información generada por los sensores, facilitando su posterior análisis y utilización en diversas aplicaciones y procesos.

En la Figura 8 se presenta el flujo de datos dentro de la aplicación de Apache NiFi. Este flujo comienza con la obtención de los datos desde PostgreSQL generados por los sensores de temperatura y humedad. Posteriormente, se lleva a cabo una etapa de eliminación de archivos no deseados que puedan estar presentes en la zona de almacenamiento de Apache Hadoop. Luego, los registros obtenidos de los sensores son transformados al formato Parquet, un formato columnar eficiente para el almacenamiento de datos. Una vez convertidos, estos datos son almacenados en el clúster en una zona de almacenamiento designada.

Figura 8

Flujo de datos en Apache NiFi

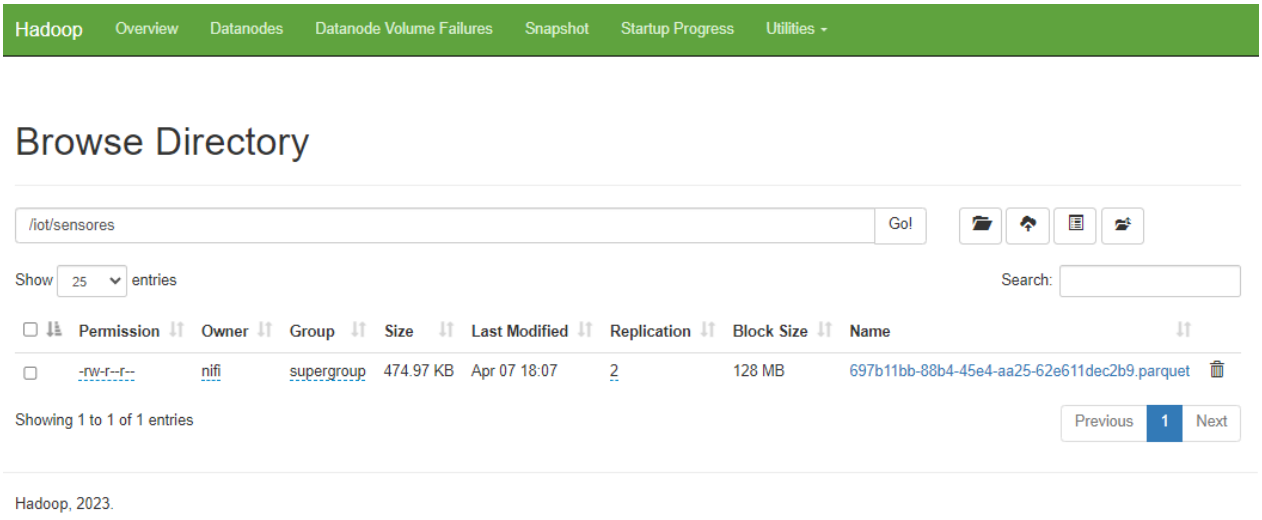


Este proceso ilustra la importancia de gestionar de manera eficiente los datos generados por los sensores, desde su adquisición hasta su almacenamiento final en el clúster. El uso de Apache NiFi facilita este flujo de trabajo al proporcionar herramientas para la adquisición, transformación y almacenamiento de datos de manera automatizada y escalable.

En la Figura 9 se presentan los datos recopilados por los sensores de temperatura y humedad, los cuales están almacenados en una zona específica dentro del clúster de Apache Hadoop en formato parquet. Se puede observar que esta información se encuentra ubicada en la carpeta designada para dispositivos iot/Sensores. Es importante destacar que, cada vez que Apache NiFi genera una consulta de los datos de estos sensores, asigna un identificador único para el nombre del archivo.

Figura 9

Datos de sensores de temperatura y humedad en clúster Hadoop



Los datos disponibles en esta zona de almacenamiento están listos para ser procesados en la siguiente fase, que implica el procesamiento de la información utilizando Dremio. Este paso es crucial para transformar los datos en ideas claves, útiles y significativas que puedan ser utilizados para tomar decisiones informadas y optimizar los procesos relacionados con la gestión de la temperatura y humedad.

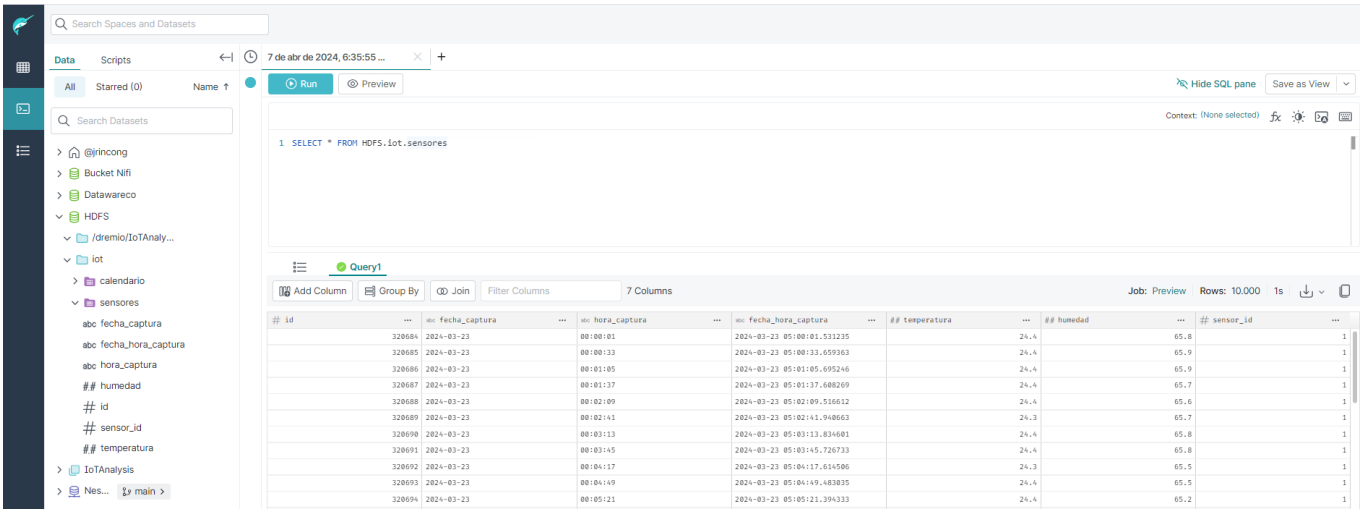
Fase: Modelado

En esta fase, nos centramos en transformar los datos mediante modelos que pueden ser utilizados para obtener insights significativos. Utilizamos herramientas avanzadas como Dremio y DBT (Data Build Tools) para organizar y estructurar los datos de manera eficiente. Nuestro objetivo es crear un entorno que facilite la consulta, visualización y análisis de los datos, asegurando que estén listos para apoyar la toma de decisiones informadas.

En la Figura 10 se presenta el conjunto de datos generado a partir de la extracción de datos en Apache NiFi y almacenado en el clúster de Apache Hadoop. Este conjunto de datos se consulta desde la aplicación Dremio. En la parte izquierda de la imagen, se observa que el conjunto de datos proviene de la fuente HDF, que hace referencia al clúster de Apache Hadoop, y se ha depositado en la carpeta "iot/sensores" en formato Parquet. En el centro de la ilustración, se muestra una vista previa de los datos.

Figura 10

Conjunto de datos consultado desde Dremio

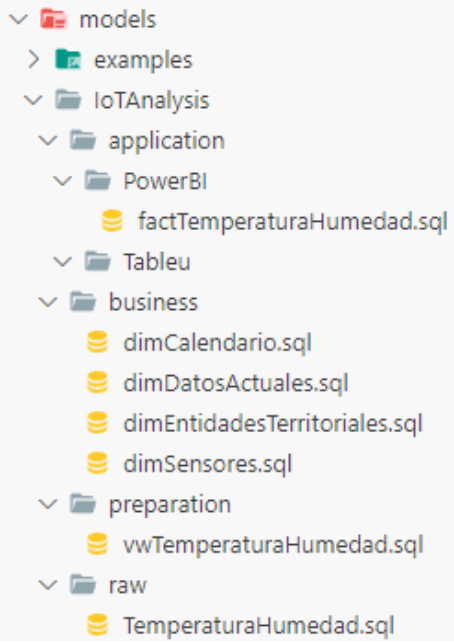


Teniendo en cuenta este proceso, se utiliza la capa semántica de Dremio, donde la clasificación de los datos se lleva a cabo con la herramienta DBT (Data Build Tools). En la

Figura 11, se puede observar que la capa semántica de Dremio se ha generado desde Data Build Tools denominada "IoTAnalysis" y está organizada en cuatro categorías: aplicaciones, negocio, preparación y datos crudos. Esta división facilita la identificación del acceso a la información dentro del modelo semántico, permitiendo a los usuarios entender claramente en qué parte del proceso se encuentran los datos que están consultando.

Figura 11

Modelos dbt (Data Build Tools)



DBT (Data Build Tools) ofrece una facilidad notable para la creación de modelos, permitiendo organizarlos en carpetas que siguen la misma estructura que la capa semántica en Dremio. Esto resulta particularmente útil al asociar los modelos con consultas SQL. Durante la ejecución de la transformación y carga de datos, se generan tablas diseñadas específicamente por consultas SQL orientadas a tablas de hechos, tablas dimensionales, vistas y datos crudos.

La capacidad de organizar el modelo en carpetas según la estructura semántica facilita la gestión y comprensión de los datos en el entorno de DBT. Además, al asociar los modelos con

consultas SQL, se promueve una mayor coherencia y eficiencia en el proceso de análisis de datos. Durante la ejecución de la transformación y carga, la generación de tablas enfocadas en diferentes tipos de datos permite una mayor flexibilidad y adaptabilidad en la manipulación y consulta de los datos, lo que contribuye a un proceso más eficiente y efectivo en el manejo de grandes volúmenes de información.

Fase: Despliegue

En esta fase final, se presenta al usuario un informe elaborado con Power BI, que incluye visualizaciones interactivas y avanzadas. Este informe permite a los usuarios explorar y comprender de manera efectiva los datos generados por los sensores de temperatura y humedad. Las visualizaciones permiten detectar patrones, tendencias y anomalías de forma intuitiva, ofreciendo una herramienta poderosa para tomar decisiones bien fundamentadas.

A continuación, se presenta el informe generado utilizando la herramienta analítica Power BI. Este informe ofrece una visión detallada y visualmente atractiva de los datos procesados, facilitando la interpretación y el análisis de la información recopilada.

En la Figura 12, se muestra el informe generado en Power BI, el cual consulta los datos del sistema de Big Data a través de la herramienta Dremio. En este informe, se presentan métricas importantes, tales como el valor mínimo, máximo y promedio de la temperatura y humedad. Además de estas métricas fundamentales, el informe también incluye secciones específicas adicionales, las cuales se detallan a continuación.

Figura 12

Reporte en Power BI sensores temperatura y humedad

En la Figura 13 se observa que el informe dispone de una sección de segmentación de datos que permite aplicar filtros. Es importante destacar que, al actualizar estos filtros, las gráficas dentro del informe se actualizan automáticamente con la nueva información, lo que

mejora significativamente la experiencia del usuario. Esto proporciona una navegación fluida y una visualización de datos más dinámica y actualizada.

Figura 13

Power BI: sección de segmentación de datos

En la Figura 14 se observa a la implementación una sección de resumen que se actualizará dinámicamente en función de los filtros aplicados. En esta sección, los usuarios pueden observar los detalles de los sensores, incluyendo su ubicación geográfica, así como un resumen de la cantidad de registros, la temperatura y la humedad. Este resumen se enfoca en métricas como el valor máximo, promedio y mínimo de temperatura y humedad.

Figura 14

Power BI: sección de resumen

En la Figura 15 se observa la sección final del informe, donde se pueden observar gráficos de tendencia que muestran la evolución de la temperatura, la humedad y la cantidad de registros a lo largo del tiempo, estos gráficos van enfocados a la necesidad del proyecto.

Por último, y no menos importante, se encuentra el modelo de copo de nieve con el cual está construido el informe, el cual se puede observar en la siguiente ilustración.

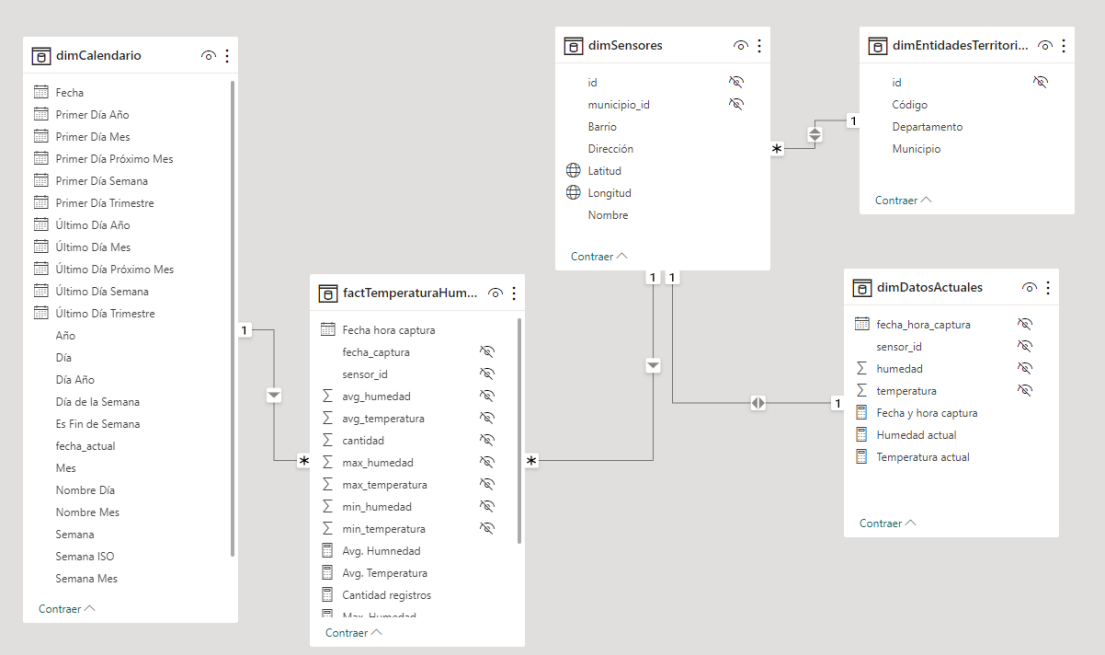
Figura 15

Power BI: contenido del informe

El modelo de copo de nieve es una estructura de base de datos diseñada para normalizar eficientemente los datos al dividir las tablas dimensionales en subtablas más pequeñas, reduciendo así la redundancia y optimizando el rendimiento de las consultas. Esta metodología ofrece varios beneficios, como una mayor eficiencia en el almacenamiento al minimizar el espacio requerido, una facilitación en el mantenimiento permitiendo actualizaciones más específicas y localizadas, y una optimización de las consultas mediante la reducción del número de uniones necesarias para recuperar la información, se puede observar el modelo de copo de nieve implementado en la Figura 16.

Figura 16

Modelo copo de nieve



Además, al reducir la redundancia de datos, el modelo de copo de nieve promueve la integridad y consistencia de los datos, resultando en una mayor calidad y fiabilidad de la información almacenada. El modelo de copo de nieve proporciona importantes ventajas en cuanto a eficiencia, facilidad de mantenimiento y calidad de datos en entornos de bases de datos dimensionales.

La usabilidad del reporte en Power BI se centra en la rapidez con la que se puede obtener información para tomar decisiones basadas en datos. Es fundamental que el diseño y la interactividad del reporte permitan una navegación fluida y una comprensión clara de los datos presentados. Esto facilita a los usuarios la identificación de tendencias, patrones y anomalías, permitiéndoles tomar decisiones informadas y rápidas para impulsar acciones estratégicas y operativas en la organización.

En relación con el informe mencionado anteriormente, se destaca la presencia de una sección de filtros que permite cambiar dinámicamente la visualización de datos y enfocarlos desde diferentes perspectivas. Esta capacidad facilita el análisis del conjunto de datos generados por los sensores de temperatura y humedad, ofreciendo una mayor flexibilidad y precisión en la exploración y comprensión de la información.

Las gráficas de tendencia para la temperatura y la humedad, complementadas con la visualización de métricas como máximo, promedio y mínimo, ofrecen beneficios significativos. Estas representaciones visuales permiten una comprensión más profunda y rápida del comportamiento de estos parámetros ambientales clave en diversos contextos. Facilitan la identificación de patrones y tendencias a lo largo del tiempo, ayudando a detectar ciclos recurrentes, cambios estacionales y otras variaciones importantes.

Además, la inclusión de métricas como máximo, promedio y mínimo en las gráficas brinda una visión más completa del rango de valores observados. Esto permite una evaluación más precisa de las condiciones ambientales y ayuda a identificar extremos, como picos de temperatura o períodos de alta humedad, que podrían requerir intervención inmediata.

Una ventaja significativa de estas gráficas es su capacidad para detectar anomalías o eventos inusuales. Al comparar los datos actuales con los valores históricos o esperados, los usuarios pueden identificar desviaciones significativas que podrían indicar problemas en el sistema relacionados con los niveles de temperatura y humedad fuera de lo común. Esta habilidad para detectar tempranamente tales desviaciones es crucial para prevenir daños o pérdidas, especialmente en entornos sensibles como laboratorios, almacenes de alimentos o instalaciones industriales.

Las gráficas de tendencia para la temperatura y la humedad, enriquecidas con métricas clave, ofrecen una herramienta poderosa para comprender, analizar y gestionar eficazmente las condiciones ambientales. Al proporcionar una representación visual clara y detallada de los datos, estas gráficas facilitan la toma de decisiones informadas y la implementación de medidas correctivas, contribuyendo así a la optimización de procesos.

Conclusiones

La implementación de un sistema completo de extracción y almacenamiento de datos de sensores requiere una cuidadosa selección de herramientas y una metodología estructurada para garantizar la eficiencia y escalabilidad del sistema. Es esencial contar con una representación visual clara de la arquitectura de datos para comprender cómo fluyen los datos y asegurar un diseño eficiente y escalable. La utilización de Apache NiFi para la extracción y transferencia de datos garantiza una gestión eficiente y automatizada del flujo de datos, contribuyendo a la escalabilidad y confiabilidad del sistema. Además, el uso de Apache Hadoop para el almacenamiento de datos permite gestionar grandes volúmenes de información de manera eficiente.

La combinación de Dremio y DBT facilita el procesamiento y preparación de datos, permitiendo la creación de tablas materializadas para análisis posteriores. Power BI desempeña un papel crucial al generar informes detallados y visualmente atractivos, proporcionando una visión completa de los datos y facilitando la toma de decisiones informadas. La integración de tecnologías como Docker, Apache NiFi, Apache Hadoop, Dremio, DBT y Power BI proporciona una solución completa que garantiza la captura y utilización efectiva de datos, promoviendo la eficiencia y escalabilidad del sistema.

El proyecto aplicado se enfocó en seleccionar herramientas que se complementaran entre sí y se adaptaran a las necesidades específicas de la gestión de datos. Por ejemplo, la adopción de Apache NiFi para la extracción y transferencia de datos fue fundamental para asegurar una gestión eficiente del flujo de datos. Asimismo, se utilizó Dremio como plataforma para la manipulación de los datos y Hadoop para el almacenamiento de estos. Este proyecto no solo ha contribuido teóricamente al campo de la gestión de datos, sino que también ha proporcionado

una hoja de ruta clara para la implementación de sistemas de extracción y almacenamiento de datos. Esto permite a las organizaciones evitar errores comunes y maximizar el retorno de inversión en tecnología.

Estas tecnologías se combinaron para abordar el desafío de manejar grandes volúmenes de datos generados por sensores de temperatura y humedad. Gracias a esta integración, se obtuvieron mejoras tangibles, incluyendo una mayor eficiencia en el procesamiento de datos y la generación de informes más precisos. Una dirección de investigación prometedora para el futuro implica en la aplicación y desarrollo de técnicas avanzadas de aprendizaje automático para predecir los datos de temperatura y humedad capturados por sensores. Al emplear algoritmos de Machine Learning en este contexto, no solo se busca comprender los datos existentes, sino también anticipar y prever sus fluctuaciones futuras con una alta precisión.

Gracias a esta integración, se han logrado mejoras tangibles, como una mayor eficiencia en el procesamiento de datos y la generación de informes más precisos, evidenciados claramente en el reporte de Power BI. Esta mejora sustancial ha simplificado la toma de decisiones informadas y optimizado la gestión de la información generada por los sensores de temperatura y humedad. Además, una dirección de investigación prometedora implica el desarrollo y aplicación de técnicas avanzadas de aprendizaje automático para predecir los datos capturados por los sensores, anticipando sus fluctuaciones futuras con alta precisión. Este enfoque tiene el potencial de mejorar significativamente la capacidad de tomar decisiones informadas y aplicadas al contexto de gestión de recursos ambientales y optimización de procesos industriales.

El proyecto enfrentó el desafío crucial de seleccionar herramientas que no solo cumplieran con las necesidades específicas del proceso, sino que también se integraran de manera efectiva dentro del ecosistema de herramientas libres. Esta tarea implicó una cuidadosa

evaluación de la funcionalidad y compatibilidad de cada herramienta, asegurando que trabajaran de manera conjunta para garantizar la eficiencia y escalabilidad del sistema de extracción y almacenamiento de datos de sensores de temperatura y humedad. Este proceso de selección y configuración adecuada fue fundamental para optimizar el flujo de datos y maximizar la utilidad de tecnologías como Apache NiFi, Apache Hadoop, Dremio, DBT y Power BI. Superar este desafío no solo permitió al proyecto cumplir con sus objetivos técnicos, sino también establecer un modelo replicable para futuras implementaciones en gestión de datos y análisis.

Referencias

A.J. Tallón-Ballesteros. (2021). Modern Management Based on Big Data II and Machine Learning and Intelligent Systems III : Proceedings of MMBD 2021 and MLIS 2021. IOS Press.

Balla, L., Reddy, C. R. K., & Sujith, A. V. L. N. (2017). BigData Analytical Challenges with IOT. International Journal of Distributed & Cloud Computing, 5(1), 27–31.

Business Wire (2020). Dremio Recognized for Product Innovation and Industry Leadership in Big Data and Data Analytics with Recent Awards.

Domínguez Bolaño, T., Barral, V., Escudero, C. J., & García-Naya, J. A. (2024). An IoT system for a smart campus: Challenges and solutions illustrated over several real-world use cases. Internet of Things, 25, 101099. https://doi.org/10.1016/j.iot.2024.101099

IBM (2021). Conceptos básicos de ayuda de CRISP-DM. Conceptos Básicos de Ayuda de Crisp-DM. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

James D. Miller. (2017). Big Data Visualization : Learn Effective Tools and Techniques to Separate Big Data Into Manageable and Logical Components for Efficient Data Visualization. Packt Publishing.

Joyanes, L. (2013). Big data. Análisis de grandes volúmenes de datos en organizaciones.

Karen L. Webber, & Henry Y. Zheng. (2020). Big Data on Campus : Data Analytics and Decision Making in Higher Education. Johns Hopkins University Press.

Maribel Yasmina Santos, & Carlos Costa. (2020). Big Data : Concepts, Warehousing, and Analytics. River Publishers.

Md. Rezaul Karim, & Sridhar Alla. (2017). Scala and Spark for Big Data Analytics : Harness the Power of Scala to Program Spark and Analyze Tonnes of Data in the Blink of an Eye! Packt Publishing.

Nitin Kumar. (2021). Big Data Using Hadoop and Hive. Mercury Learning and Information.

Poonam Tanwar, Vishal Jain, Chuan-Ming Liu, & Vishal Goyal. (2020). Big Data Analytics and Intelligence : A Perspective for Health Care. Emerald Publishing Limited.

Rajkumar Buyya, Rodrigo N. Calheiros, & Amir Vahid Dastjerdi. (2016). Big Data : Principles and Paradigms. Morgan Kaufmann.

Sneha Kumari, K.K. Tripathy, & Vidya Kumbhar. (2021). Application of Big Data and Business Analytics. Emerald Publishing Limited.

Venkat Ankam. (2016). Big Data Analytics. Packt Publishing.

Wnęk, K., & Boryło, P. (2023). A Data Processing and Distribution System Based on Apache Nifi. Photonics, 10(2). https://doi-org.bibliotecavirtual.unad.edu.co/10.3390/photonics10020210

Zhang, L., Du, Q. y Datcu, M. (2015). Management and analytics of remote sensing big data. Journal of Applied Remote Sensing, 9(1)