**Trabajo Final del Módulo Parte 1: “Introducción y Conceptos Data Science”**

GRUPO 4:

|  |  |  |
| --- | --- | --- |
| Bárbaro Ayala | Antonio Gallardo | Amadeo Delgado |
| Juan Carlos Cárdenas | Pedro Llull | Boris Victoria |

# **IMPLEMENTACIÓN DE ARQUITECTURA DE DATOS Y BIG DATA EN EL SECTOR ENERGÉTICO**

**I. Introducción**

El presente informe tiene como objetivo presentar una propuesta de implementación de arquitectura de datos y Big Data para EnergíaSostenible S.A., una empresa líder en el sector energético. En este informe, se abordará la problemática actual, los casos de estudio, las áreas funcionales involucradas, las tecnologías necesarias para cada área, así como los pros y contras de la implementación.

**II. Presentación de la empresa**

EnergíaSostenible S.A. es una empresa dedicada a la generación, distribución y comercialización de energía eléctrica. Con una amplia experiencia en el sector, la empresa se ha consolidado como líder en el mercado, ofreciendo servicios energéticos confiables y sostenibles a sus clientes.

**III. Problemática**

A pesar del éxito y la trayectoria de EnergíaSostenible S.A., la empresa se enfrenta a diversos desafíos en su gestión de datos y toma de decisiones. Algunos de los problemas identificados incluyen:

Gestión ineficiente de datos: La empresa se encuentra ante la necesidad de recopilar, almacenar y procesar grandes volúmenes de datos relacionados con la demanda, el consumo, la generación y otros factores relevantes. Sin embargo, la gestión de estos datos se realiza de manera fragmentada y poco eficiente, lo que dificulta la obtención de información valiosa para la toma de decisiones.

Falta de análisis avanzado: EnergíaSostenible S.A. carece de herramientas y tecnologías que le permitan realizar análisis avanzados en tiempo real. Esto limita su capacidad para identificar patrones, predecir demandas futuras, optimizar la generación y distribución de energía, y tomar decisiones fundamentadas.

**IV. Casos de estudio**

**Caso de estudio 1:** Gestión de la demanda y generación de energía En esta área funcional, la implementación de la arquitectura de datos y Big Data afecta de la siguiente manera:

Fase de Extracción, Transformación y Carga (ETL): Se utilizará una herramienta de ETL como Apache Nifi para extraer datos de diversas fuentes, transformarlos según las necesidades y cargarlos en el almacenamiento estructurado (MySQL) y no estructurado (MongoDB). Esto permite consolidar y organizar los datos para su posterior análisis.

Almacenamiento en la nube: AWS como infraestructura cloud para los elementos de la arquitectura.

Almacenamiento estructurado: MySQL se utilizará para almacenar y administrar datos relacionados con la demanda y generación de energía. Permite un acceso rápido y eficiente a los datos estructurados, lo que facilita su análisis y consulta.

Almacenamiento no estructurado: MongoDB será utilizado para almacenar datos no estructurados como registros de sensores y comentarios de clientes. Proporciona flexibilidad y escalabilidad para manejar datos no estructurados y semiestructurados, lo que facilita su procesamiento posterior.

Procesamiento distribuido: Apache Hadoop, junto con Apache Spark, se utilizará para procesar grandes volúmenes de datos de manera eficiente y realizar análisis avanzados en paralelo. Permite realizar tareas de procesamiento intensivas en recursos y computacionalmente costosas.

Herramientas adicionales: Para el procesamiento de datos en tiempo real, se incorporará Apache Kafka para la ingesta y procesamiento de streams de datos, lo que permite una respuesta más ágil a los eventos y situaciones cambiantes.

Visualización: Tableau se utilizará para crear paneles de control interactivos y visualizaciones intuitivas de los datos de demanda y generación de energía. Esto proporciona una visualización clara y accesible de los resultados del análisis, facilitando la toma de decisiones informadas.

**Caso de estudio 2:** Mantenimiento y gestión de activos En esta área funcional, la implementación de la arquitectura de datos y Big Data afecta de la siguiente manera:

Fase de Extracción, Transformación y Carga (ETL): Se utilizará Apache Nifi para extraer datos de sensores y otros dispositivos, transformarlos según los requisitos y cargarlos en el sistema de almacenamiento.

Almacenamiento estructurado: MySQL se utilizará para almacenar datos estructurados relacionados con los activos de la empresa, como información de mantenimiento, histórico de fallas y programación de mantenimiento. Facilita la gestión y consulta eficiente de estos datos.

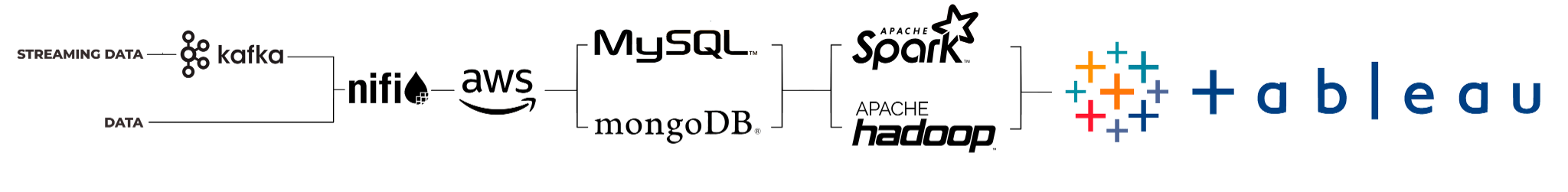
Almacenamiento no estructurado: MongoDB será utilizado para almacenar datos no estructurados como registros de sensores y documentación técnica. Permite la gestión flexible de datos no estructurados y su posterior procesamiento.

Procesamiento distribuido: Apache Hadoop y Apache Spark se utilizarán para procesar grandes volúmenes de datos de sensores y realizar análisis en tiempo real. Esto permite b

Herramientas adicionales: Para la ingesta y procesamiento de datos en tiempo real, se incorporará Apache Kafka para gestionar y distribuir los streams de datos generados por los sensores y dispositivos.

Visualización: Tableau se integrarán con la arquitectura de datos para presentar los resultados del análisis de manera visual y accesible, facilitando la supervisión y gestión centralizada de los activos.

**Figura 1: Arquitectura:**



**V. Áreas funcionales involucradas**

La implementación de la arquitectura de datos y Big Data impactará varias áreas funcionales dentro de EnergíaSostenible S.A., incluyendo:

**Área de IT:** Implementación de herramientas de ETL (Extract, Transform, Load) como Apache Nifi para la extracción, transformación y carga de datos desde diversas fuentes hacia el Data Lake y el Data Warehouse.

Configuración y administración del Data Lake utilizando tecnologías como Apache Hadoop y Apache Spark para el almacenamiento y procesamiento distribuido de grandes volúmenes de datos en diferentes formatos.

Implementación de Apache Kafka para la ingesta y procesamiento de streams de datos en tiempo real provenientes de diversas fuentes.

Configuración y administración del Data Warehouse utilizando tecnologías como MySQL para el almacenamiento de datos estructurados y la generación de informes y análisis.

**Área de Operaciones:** Utilización de datos almacenados en el Data Lake y el Data Warehouse para el análisis y la optimización de la demanda y generación de energía.

Implementación de herramientas de procesamiento distribuido como Apache Spark para el análisis en tiempo real de grandes volúmenes de datos y la detección de patrones de demanda.

Uso de Apache Kafka para el consumo de datos en tiempo real y la toma de decisiones basada en información actualizada.

Utilización de herramientas de visualización de datos como Tableau para la creación de paneles de control interactivos y la presentación de resultados y métricas relevantes.

**Área de Mantenimiento:** Utilización de datos almacenados en el Data Lake y el Data Warehouse para el monitoreo y mantenimiento de activos.

Implementación de herramientas de análisis de datos como Apache Spark para el procesamiento y análisis de grandes volúmenes de datos relacionados con el rendimiento de los activos.

Uso de Apache Kafka para la ingestión y procesamiento de datos en tiempo real provenientes de sensores y sistemas de monitoreo.

Utilización de bases de datos NoSQL como MongoDB para el almacenamiento de datos no estructurados relacionados con el mantenimiento de activos.

**Área de Facturación:** Utilización de datos almacenados en el Data Warehouse para el cálculo y generación de facturas precisas.

Implementación de herramientas de análisis y reporte como MySQL y Tableau para el procesamiento de datos de facturación y la generación de informes y métricas.

Uso de ETL (Extract, Transform, Load) para la extracción, transformación y carga de datos desde el Data Lake y otras fuentes hacia el Data Warehouse.

**Área de Postventa:** Utilización de datos almacenados en el Data Warehouse y Data Marts para brindar servicios de atención al cliente personalizados y recomendaciones de eficiencia energética.

Implementación de herramientas de análisis de datos como Apache Spark y MySQL para identificar patrones de consumo y generar recomendaciones adaptadas a las necesidades de los clientes.

Uso de ETL para la extracción, transformación y carga de datos desde diferentes fuentes hacia los Data Marts, donde se almacenan los datos segmentados y específicos para cada área funcional.

**VI. Pros y contras**

**Pros de la implementación de la arquitectura de datos y Big Data:**

Mayor capacidad para gestionar y procesar grandes volúmenes de datos en tiempo real. Análisis avanzado y detección de patrones que permiten la toma de decisiones fundamentadas. Mayor eficiencia operativa y optimización de la generación y distribución de energía. Mejora de la experiencia del cliente y usuarios a través de servicios personalizados.

La incorporación de las herramientas Lambda y Kappa en el caso de estudio 1 puede ofrecer una mayor flexibilidad y capacidad para procesar y analizar streams de datos en tiempo real, permitiendo una respuesta más ágil a los eventos y situaciones cambiantes.

La inclusión de Apache Kafka en el caso de estudio 2 facilitará la gestión y distribución de los streams de datos generados por los sensores y dispositivos, mejorando así la ingesta y procesamiento de datos en tiempo real.

**Contras de la implementación de la arquitectura de datos y Big Data:**

Requiere una inversión inicial significativa en tecnología y capacitación del personal. Posible resistencia al cambio por parte de los empleados y la necesidad de una cultura de datos y análisis. Desafíos en la privacidad y seguridad de los datos, lo que requiere medidas adicionales de protección.

**VII. Conclusiones**

La implementación de la arquitectura de datos y Big Data en EnergíaSostenible S.A. tiene el potencial de transformar la forma en que la empresa gestiona y utiliza sus datos. Mediante el uso de tecnologías como MySQL, MongoDB, Apache Hadoop, Apache Spark, Apache Kafka, Tableau , la empresa podrá superar los desafíos actuales en la gestión de datos, mejorar la toma de decisiones y optimizar sus operaciones.

La incorporación de las herramientas Lambda y Kappa en el caso de estudio 1, así como Apache Kafka en el caso de estudio 2, permitirá una mejor gestión y procesamiento de los datos en tiempo real, ofreciendo una mayor agilidad y capacidad de respuesta a los eventos y necesidades cambiantes.

Sin embargo, es importante tener en cuenta los desafíos y considerar los pros y contras antes de embarcarse en esta implementación. Se recomienda llevar a cabo un análisis detallado de los requerimientos, evaluar la viabilidad económica y técnica, y contar con un plan sólido de capacitación y gestión del cambio.

**En resumen,** la implementación de la arquitectura de datos y Big Data, junto con las herramientas Lambda, Kappa, Apache Kafka y Apache Stark, brinda una oportunidad única para EnergíaSostenible S.A. de destacarse en el sector energético, mejorar su eficiencia operativa y ofrecer servicios energéticos más personalizados y sostenibles a sus clientes.