Taller Big-Data

Docente: Jesus Ariel

Estudiante: Paula Andrea Terrios

Corporacion Universitaria del Huila (CORHUILA)

Ciencia de Datos

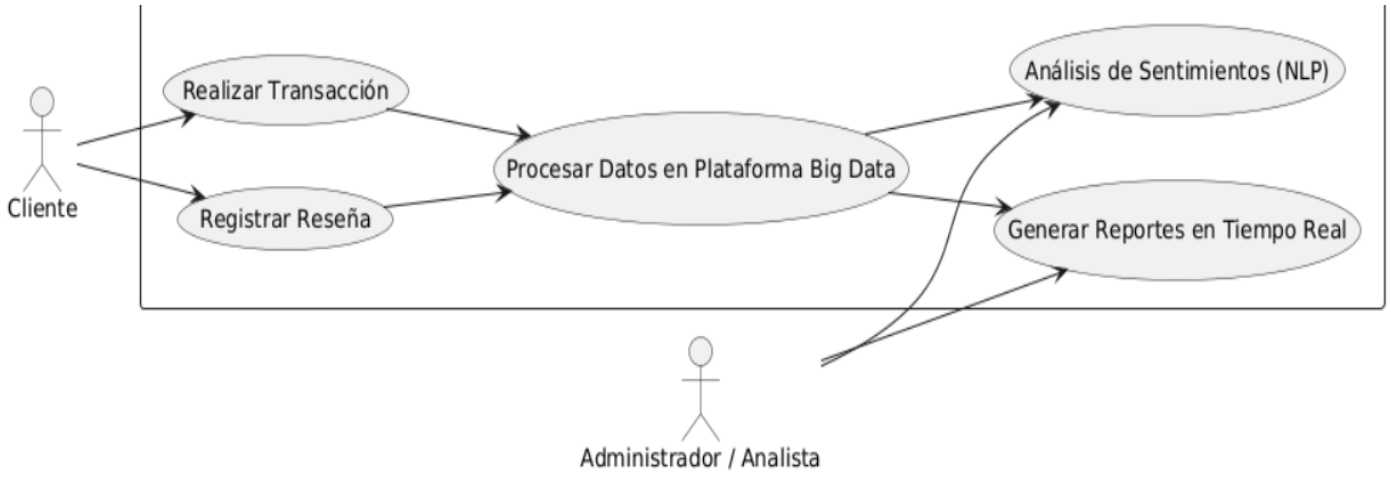
Neiva – Huila

2025

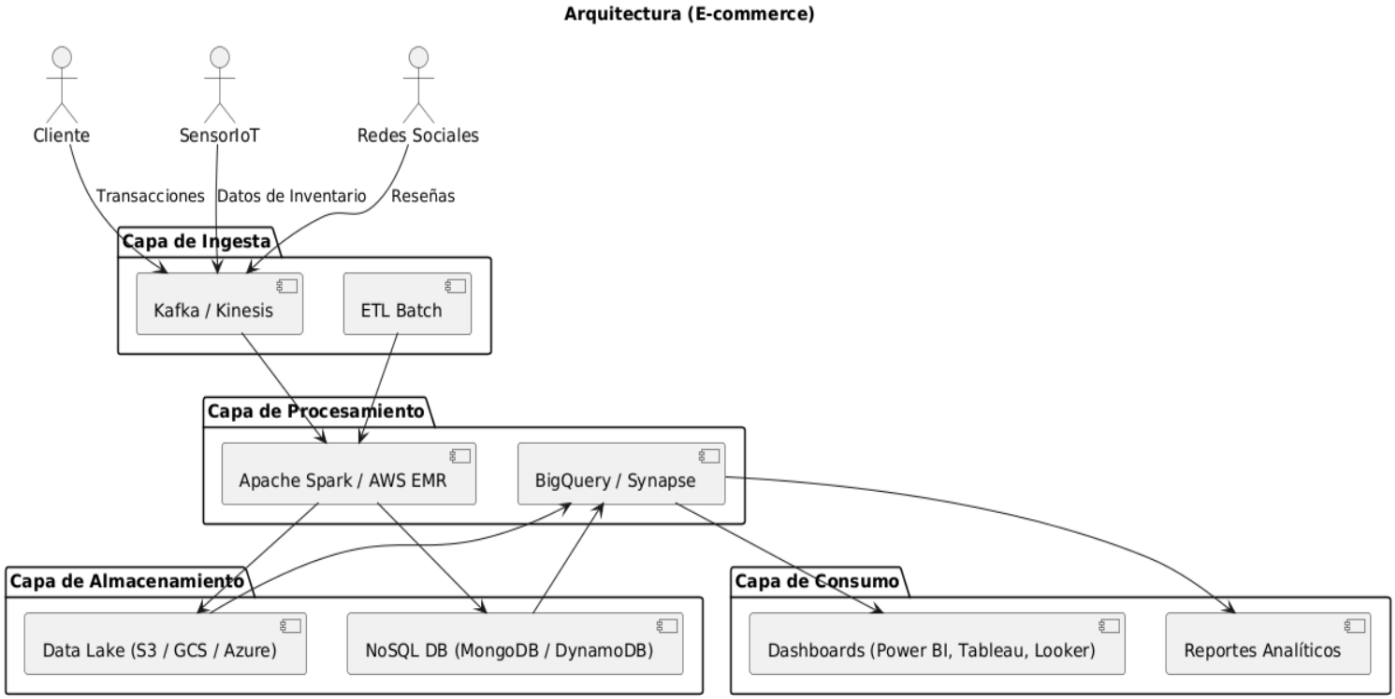
Cuadro comparativo

	A	B	C	D
1	**ASPECTOS**	**HADOOP**	**SPARK**	**Arquitecturas en la Nube (ej. AWS, GCP, Azure)**
2	**Procesamiento**	Batch con MapReduce (procesamiento por lotes).	Procesamiento en memoria; soporta batch y streaming en tiempo real.	Batch y streaming gestionado con servicios integrados (ej. AWS Kinesis, BigQuery).
3	**Escalabilidad**	Escalabilidad horizontal, pero requiere configuración manual de nodos.	Escalabilidad horizontal con cluster manager (YARN, Mesos, Kubernetes).	Escalabilidad automática según demanda (auto-scaling).
4	**Facilidad de uso**	Alta complejidad en programación y administración (MapReduce requiere más código).	APIs más amigables (Python, Scala, R, Java); integración con librerías como MLlib y SparkSQL.	Interfaces gráficas simplificadas y servicios administrados listos para usar.
5	**Costos**	Inversión alta en infraestructura propia y mantenimiento de clúster.	Costo de mantener clúster dedicado; más eficiente que Hadoop por menor tiempo de ejecución.	Pago por uso (modelo pay-as-you-go); sin inversión en hardware.
6	**Casos de uso**	Procesamiento batch de grandes volúmenes de datos históricos.	Analítica interactiva, aprendizaje automático, análisis en tiempo real.	Analítica empresarial en tiempo real, integración con IoT, BI en la nube.
7	**Requerimientos de hardware**	Clúster distribuido con múltiples nodos, almacenamiento local/ HDFS.	Clúster con nodos de memoria alta; depende de YARN/Mesos.	No requiere hardware propio; depende del proveedor cloud.
8	**Complejidad de gestión**	Alta, se necesitan administradores especializados en Hadoop.	Media, con mayor abstracción pero requiere conocimientos de clúster.	Baja, gestión automática por el proveedor.
9	**Métricas de rendimiento**	Ejecución más lenta en batch (comparado con Spark).	Procesa hasta 100x más rápido que Hadoop en memoria y 10x más en disco.	Variable, depende del servicio; tiempos optimizados con escalado automático.

Caso de Uso



Diagrama


Arquitectura (E-commerce)
Cliente
SensorIoT
Redes Sociales
Transacciones
Datos de Inventario
Reseñas
Capa de Ingesta
Kafka / Kinesis
ETL Batch
Capa de Procesamiento
Apache Spark / AWS EMR
BigQuery / Synapse
Capa de Almacenamiento
Data Lake (S3 / GCS / Azure)
NoSQL DB (MongoDB / DynamoDB)
Capa de Consumo
Dashboards (Power BI, Tableau, Looker)
Reportes Analíticos

Conclucion

La arquitectura propuesta presenta un flujo organizado de datos, que va desde la ingesta en tiempo real y por lotes, hasta el almacenamiento, procesamiento y consumo de información mediante dashboards y reportes. Esto permite responder a las necesidades de un e-commerce moderno, donde la velocidad y la precisión en el análisis de datos son claves para mejorar la experiencia del cliente, con unas pautas:

Escalabilidad: El uso de un Data Lake junto con bases NoSQL facilita la integración de grandes volúmenes de datos heterogéneos (transacciones, IoT, redes sociales).

Flexibilidad: Se combinan procesamientos batch y streaming, lo que permite tanto análisis históricos como respuestas en tiempo real.

Orientación al negocio: La capa de consumo traduce la complejidad técnica en información útil para la toma de decisiones estratégicas (ej. gestión de inventarios, personalización de ofertas).

Complejidad técnica: Requiere un equipo especializado en Big Data, lo que implica altos costos iniciales de implementación y mantenimiento.

Seguridad y cumplimiento: El manejo de datos sensibles de clientes obliga a cumplir normativas (ej. GDPR, PCI-DSS), lo que añade carga administrativa y tecnológica.

Dependencia tecnológica: La arquitectura depende de múltiples servicios (Kafka, Spark, BigQuery, etc.), lo que puede generar riesgos de integración y dependencia de proveedores cloud.

Si se optimiza la gobernanza de datos y se implementan políticas de seguridad desde el diseño, esta arquitectura puede evolucionar hacia un ecosistema de inteligencia artificial, incorporando modelos predictivos y sistemas de recomendación en tiempo real para anticipar la demanda y personalizar aún más la experiencia de compra.

En conclusión, la arquitectura propuesta no solo satisface las necesidades actuales del e-commerce, sino que también abre la puerta a una evolución futura hacia analítica avanzada e inteligencia de negocio en tiempo real, aunque exige una estrategia sólida en talento, seguridad y gestión de costos.

Bibliografia

https://www.oracle.com/latam/big-data/what-is-big-data/

https://www.obsbusiness.school/blog/big-data-y-sus-principales-aplicaciones-beneficios-y-ejemplos

https://www.teamcore.net/es/blog/5-empresas-que-usan-big-data-y-han-conseguido-los-mejores-resultados/

https://www.planttext.com/?utm_source=chatgpt.com