DATAHACK

Proyecto Apache Spark - Jorge de Dios

EJERCICIO 11

Diseño de arquitectura de datos para manejar el sistema de inspecciones de sanidad.

1. Dominio de Datos relevantes:

- Datos de inspecciones: Información detallada sobre cada inspección realizada, incluyendo resultados, fechas, ubicaciones, etc.
- Restaurantes inspeccionados: Información de las empresas con datos de ubicatión y riesgos, almacenadas en formato Delta Lake.
- Datos de Inspectores: Registro y seguimiento de los inspectores, incluyendo métricas de rendimiento.

2. Propietarios del Dominio

- Datos de Inspecciones: Responsabilidad del equipo de salud pública o departamento de inspecciones.
- Empresas Registradas: Mantenido por el departamento de registros comerciales o equivalente.
- Datos de Inspectores: Administrado por recursos humanos o un departamento de gestión de personal.

3. Infraestructura Autónoma

Cada dominio gestionará su infraestructura de la siguiente manera:

- Ingestión de Datos: Utilización de Apache Kafka para ingestión de datos en tiempo real desde fuentes externas como API REST y otros sistemas.
- Almacenamiento: Utilización de Delta Lake en Databricks para el almacenamiento robusto y transaccional de datos.
- Procesamiento: Apache Spark para el procesamiento distribuido de grandes volúmenes de datos y cálculo de métricas.
- Exposición de Datos: APIs REST para exposición controlada de datos a ciudadanos y sistemas externos.
- Dashboard y Reportes: Herramientas de visualización como Tableau o Power BI para métricas y reportes internos.

4. Interoperabilidad

- Estándares y Protocolos: Uso de JSON y Avro para el intercambio de datos entre sistemas internos y con fuentes externas. Integración mediante Kafka Connect para conectividad con sistemas externos.
- Comunicación entre Dominios: Eventos y mensajes en Kafka para la comunicación entre dominios y la sincronización de datos.

5. Gobernanza

- Calidad de Datos: Validación y limpieza de datos en etapas de ingestión y almacenamiento utilizando Apache Spark.
- Seguridad: Acceso basado en roles de Databricks y políticas de seguridad para proteger datos sensibles.
- Cumplimiento: Cumplimiento con regulaciones de privacidad como GDPR y HIPAA.

6. Tecnologías Utilizadas

- Apache Kafka: Ingestión de datos en tiempo real.
- Apache Spark: Procesamiento distribuido y cálculo de métricas.
- Delta Lake: Almacenamiento transaccional y manejo de versiones de datos.
- Databricks: Plataforma unificada para ejecutar Apache Spark y Delta Lake.
- APIs REST: Para exposición de datos a ciudadanos y sistemas externos.
- Tableau / Power BI: Herramientas de visualización para dashboards y reportes.

7. Evolución y Escalabilidad

- Escalamiento Horizontal: Uso de clústeres de Databricks escalables para manejar incrementos en el volumen de datos y la carga de trabajo.
- Monitoreo y Optimización: Implementación de monitoreo continuo y ajuste de recursos según necesidades para mantener el rendimiento óptimo.
- Actualización Continua: Iteración y mejora continua de la arquitectura en respuesta a cambios en requisitos y tecnologías emergentes.

Diagrama de la arquitectura:

- Justificación del Diseño:

  Este diseño fue elegido por su capacidad para manejar grandes volúmenes de datos en tiempo real, asegurando la integridad y disponibilidad de la información crítica para la salud pública.

  La utilización de tecnologías como Apache Kafka y Delta Lake garantiza la robustez y escalabilidad necesaria para cumplir con los requisitos de latencia y precisión en las métricas de inspección.

  La separación de dominios y la asignación clara de responsabilidades facilita la gobernanza y el cumplimiento normativo, mientras que la interoperabilidad asegura una integración fluida entre diferentes sistemas y la exposición controlada de datos a usuarios y aplicaciones externas.

  Este diseño está diseñado para ser flexible y escalable, permitiendo futuras expansiones y adaptaciones según las necesidades cambiantes del sistema de inspecciones de sanidad y las tecnologías emergentes.