TD7: Ingeniería de Datos

Trabajo Práctico Grupal

Segunda entrega: arquitectura y flujos de datos

Esta entrega está enfocada en ejercitar los conceptos revisados en la segunda parte de la materia. El trabajo consiste en tomar el modelo implementado en la primera entrega y enmarcarlo en el contexto de una organización que tiene esa base de datos, por lo cual será necesario diseñar una arquitectura, flujos de datos y procesos de validación de los mismos.

La fecha de entrega del trabajo práctico será el **Viernes 27 de Junio**.

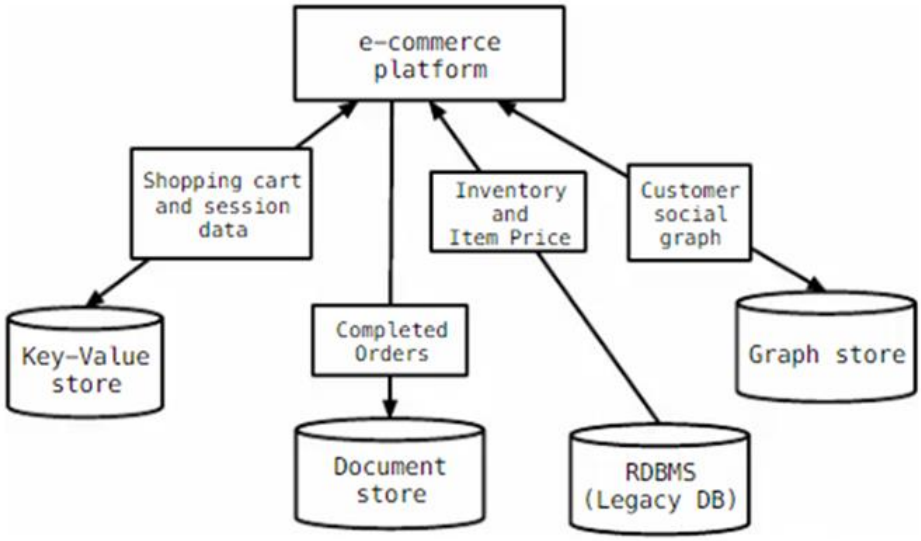

Figure 13.3 *Example implementation of polyglot persistence*

Nuevos requerimientos exceden las prestaciones de vuestro RDBMS: ahora es necesario pensar un sistema que ofrezca persistencia políglota. A partir del DER diseñado, de la base implementada en PostgreSQL y de los datos cargados para la PoC, deberán generar datos que sean relevantes al negocio de manera de ampliar la prueba de concepto a las siguientes aplicaciones y sistemas de procesamiento.

Se espera que el entregable principal sea un **informe** que detalle cómo es la resolución de cada punto, incluyendo la descripción de los flujos implementados, el código y los casos de uso relevados. Además del informe, se esperan los archivos de soporte.

TD7: Ingeniería de Datos

Trabajo Práctico Grupal

1. **Arquitectura**. Pensar el modelo definido en la parte 1 en el contexto de una organización real. Contemplar:
    a. ¿Cuál es el origen de los datos? ¿Qué cadencia tienen, con qué formato arriban, qué volumen se espera?
    b. ¿Cómo es el linaje de los datos desde su origen hasta que las distintas aplicaciones los consumen? ¿Cómo son los procesos intermedios y qué patrones siguen?
    c. ¿Cuáles son los usos que tienen los datos en las distintas etapas de procesamiento?
    d. Imaginen un uso de este sistema. ¿Cuáles son los distintos roles que lo usan y qué tipos de permisos tienen sobre los datos?
2. **Map Reduce con Spark**
   Armar al menos tres consultas que se resuelvan con sus correspondientes procesamientos MapReduce, explicitando en cada caso la Fase de Map, la Fase de Reduce y la Fase de Recolección de datos e Impresión. Se debe entregar la notebook correspondiente en Google Colaboratory. Atención: las consultas y los datos utilizados deben ser referidos al DER con el que ya cuentan.
3. **Spark SQL**
   Utilizando SQL realizar al menos dos consultas en data sets mediante Spark. Mostrar los resultados como tablas. Entregar la notebook correspondiente.
4. **Great Expectations: definir al menos 5 reglas que resulten de interés para validar los datos de entrada.**
   Sólo se piden las reglas a ejecutar, no es necesario integrarlas en el flujo.
5. **Redis**
   1. Usar el modelo KV de Redis para guardar información específica de vuestro modelo. Mostrar el manejo de datos de este modelo seteando y consultando valores.
   2. Crear una lista de tareas con datos pertinentes del dominio. Mostrar operaciones de gestión de la lista: agregar tareas, recuperarlas, eliminarlas, etc.
   3. Elegir datos del dominio a los que sea adecuado aplicarles un tiempo de expiración. Crear claves con TTL con estos datos e imprimir mensajes pertinentes por pantalla. Entregar la notebook correspondiente.

6- MongoDB

Aquí deben elegir una de las siguientes dos opciones:
-Crear una instancia de MongoDB en la nube y conectarse.
o bien
-Usar mongoDB en docker (en máquinas locales).

Con la instancia funcionando, crear colecciones de documentos pertenecientes al dominio específico y realizar consultas sobre los documentos.
Entregar la documentación necesaria según la opción elegida.

Además del informe, se esperan los archivos de soporte.
También los datos presentados en el primer tp; enunciar si se consideran datos de otras fuentes adicionales.
Se puede utilizar **Faker** para generar datos sintéticos.