ICC0012 - Almacenamiento y Procesamiento Masivo de Datos
Fecha de entrega: Domingo 20 de Agosto, 18:00hrs

Esta tarea es individual, y consiste en un ejercicio práctico de almacenamiento en sistemas SQL y NoSQL, más investigación de arquitectura de sistemas para grandes compañías.

Parte uno

- Usando Twitter API, descargue un conjunto de tweets. Como mínimo, rescaten un par de gigas de datos. Tip: Comiencen tan pronto como sea posible.
- Almacene los datos en una base de datos relacional: Mysql, PostgreSQL o MariaDB y en una base de datos NoSQL (MongoDB, CouchDB). Tip: recolecten datos en un sistema de almacenamiento NoSQL y luego migren esos datos a una base de datos relacional.
- Estudie y analice los datos que capture. Para esto, calcule algunas métricas básicas del conjunto de datos y deje un registro del tiempo que toma a su equipo calcular estas estadísticas.
    - Métricas :
        - Tweets por hora: cuantos tweets por cada hora de referencia puede encontrar ("timestamp_ms")
        - Algunos tweets tienen la variable ubicación. Intente calcular alguna estadística geográfica. (i.e. Cond1<=Variable<=Cond2)
        - Alguna otra métrica interesante para Ud.
- Compare el tiempo necesario para calcular las métricas entre cada sistema (SQL vs NoSQL)
- Investigue que es un índice en Mongo y en la base de datos tradicional de su elección. Ajuste ambos sistemas con un índice adecuado para su métrica. Y vuelva a calcular las estadísticas.
    - Compare los resultados con la parte anterior.
- Monitoree el comportamiento de su computador y vea si detecta uso intensivo de escritura/lectura de disco o memoria RAM en todas las situaciones.
- Escriba sus resultados y conclusiones en un reporte:
    - Qué dificultades tuvo que sortear para realizar esta parte de la tarea.
    - Después de cada cambio realizado y métrica calculada, cuáles eran sus expectativas, y compárelas con los resultados obtenidos.
    - Explique en sus palabras cuál es el rol de un Índice en la base de datos.

Parte dos

Investigue cómo funciona la infraestructura de datos de alguna empresa tecnológica conocida y trate de explicar cómo la compañía aplica los **principios (Tolerancia a fallas, baja latencia, escalabilidad, etc.) y la arquitectura lambda** que revisamos en clases. Describa que partes de la infraestructura corresponden a cada capa de la arquitectura vista en clases. Recuerde, no todas las compañías seguirán los mismos principios.

En sus palabras, **explique las tecnologías más importantes presentes en la compañía elegida** (no más de un párrafo por cada tecnología)

En su reporte, indique las fuentes que utilizó para realizar este análisis.

Empresas de ejemplo: Spotify, Ebay, Facebook, Pinterest, Amazon, Netflix, Twitter, etc.

Entrega:

- Un repositorio en Github publicando todos los códigos que haya escrito para trabajar en esta tarea.
- Un reporte explicando el proceso de trabajo y análisis realizado de ambas partes, los problemas que enfrentó, y la respuestas y conclusiones a las preguntas realizadas.
- Ud. Debe indicar las fuentes de información que haya utilizado para completar esta tarea.

Evaluación:

La nota final de la tarea se evaluará con un 60% para la parte uno y en un 40% para la parte 2. En su reporte, indique las fuentes y referencias utilizadas para su investigación. Se evaluará el proceso, el trabajo realizado, y sus conclusiones. **No se evaluará el volumen de datos procesados.**

Artículos de referencia:

https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f169c#.jcxk9eypa

https://docs.mongodb.com/manual/reference/method/db.stats/

Anexo:
Para Obtener sus claves de acceso en Twitter, deben acceder a https://apps.twitter.com/ y crear una nueva aplicación en Twitter.

Necesitan cuatro claves de acceso: 2 para identificar la aplicación y 2 para identificar al usuario que utiliza la aplicación.