Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación
Primer Semestre de 2023

IIC 2440 – Procesamiento de Datos Masivos
Examen

Instrucciones.

El examen dura hasta las 18:00 hrs. No se permiten preguntas, salvo en las siguientes ventanas, y en voz alta:

- De 15:40 a 16:00
- De 17:00 a 17:20

No se permite consulta a ningún tipo de apunte o robot digital.

Resuelve cada uno de los problemas en una hoja separada. Marcala con tu appelido y tu nombre. Las preguntas cortas generalmente tienen una mayor probabilidad de estar correctas.

La nota del examen se calcula como la suma de los puntos, mas un punto base.

Problema 1 (1 pto) Hace un tiempo tuviste que ayudar a una de tus amigas que había abierto una tienda de venta de artículos musicales en línea. En aquella ocasión su negocio había despegado y ya era costoso hacer análisis de datos en la base de datos que mantenía la operación del producto, por lo que le sugeriste tener un Data Warehouse, cosa que ha estado considerando. El esquema era el siguiente:

```
Usuarios(uid PRIMARY KEY, nombre, direccion, edad)
Productos(pid PRIMARY KEY, nombre, tipo, precio)
Compras(cid PRIMARY KEY, uid, fecha_compra, fecha entrega)
Compras_Productos(pid, cid, cantidad)
```

Y el tipo de consultas es el siguiente:

1. La vista que le muestra a un usuario todas las compras que ha realizado.
2. La vista que le permite a tu amiga seleccionar un producto y ver todas las compras en las que se adquirió ese producto.
3. La vista que le permite a tu amiga contar el número de instrumentos comprado por tipo.
4. La vista que le muestra a un usuario todos los productos adquiridos en una compra, junto con el total de la compra.

Parte 1.1 (0.5 pts). Como ha pasado un tiempo, tu amiga olvidó algunos de los conceptos que le entregaste, así que ella acude a ti con las siguientes dudas, que son las que tienes que responder en esta pregunta:

1. ¿Por qué es mala idea hacer análisis de datos en una base de datos transaccional (por ejemplo, PSQL) que sustenta la operación del producto?
2. ¿Por qué las bases de datos como BigQuery funcionan bien para hacer Data Warehousing?

Parte 1.2 (0.5 pts). Además, tu amiga escuchó que hace un tiempo estaba muy de moda utilizar Apache Spark para hacer análisis de datos y que ha considerado utilizarlo en su stack de tecnologías. En base a eso, responde la siguiente pregunta:

¿Cuál es la diferencia entre hacer análisis de datos con una base de datos como BigQuery, versus hacer el análisis en Apache Spark? Entrega los principales pros y contras de cada caso. Además, señala en qué se parece BigQuery con Apache Spark

Problema 2 (2 ptos) En clases estudiamos los fundamentos de Apache Spark, principalmente la estructura RDD y las funciones asociadas a sus objetos. Un punto importante de estos algoritmos es que varias instrucciones de alto nivel, como una consulta SQL, se pueden traducir a algoritmos distribuidos que usan funciones como `map`, `reduceByKey`, `join`, entre otras.

En esta pregunta queremos que expliques algunos conceptos relacionados a cómo se traducen consultas SQL a instrucciones para un entorno distribuido con varios *workers*, como lo hace por ejemplo Spark SQL o BigQuery.

1. **(0.4 pts)** Primero, explica la diferencia de cómo funciona una base de datos transaccional como PSQL vs. cómo funciona una base de datos en un entorno distribuido como BigQuery. Queremos que te enfoques comparando las técnicas en las que se basa cada paradigma para responder una consulta lo más rápido posible.
2. **(0.3 pts)** Ahora explica cómo lo hace un sistema como BigQuery para evaluar una consulta de *join*. Puedes apoyarte en las instrucciones asociadas a los RDD.
3. **(0.3 pts)** Explica cómo lo hace un sistema como BigQuery para responder una consulta con agregación. Entrega un ejemplo.

4. **(0.4 pts)** Explica cómo lo hace un sistema como BigQuery para responder consultas con Window Functions. ¿Crees que responder estas consultas en un entorno distribuido es eficiente?. Entrega al menos un ejemplo para fundamentar tu respuesta.
5. **(0.3 pts)** ¿Cómo crees que un sistema como BigQuery responde una consulta más compleja, en la que se mezclan joins, filtros y operadores de analítica como Window functions?
6. **(0.3 pts)** ¿Qué tipo de consultas crees que no se pueden responder de forma eficiente en un sistema como BigQuery?

Problema 3 (1 pto.) En el curso vimos una serie de herramientas que estaban basadas en argumentos probabilísticos, como sampleos o funciones de hash. En esos casos, uno normalmente consigue acelerar los cómputos y/o procesos, pero el costo a pagar es una pequeña probabilidad de error en alguna de las operaciones.

Alguien, por supuesto, podría decir que el introducir probabilidad de error es un costo demasiado alto, pues en el manejo de datos siempre se necesitan todas las respuestas, y un calculo que sea completamente determinístico.

En esta pregunta, elabora el siguiente contra-argumento: Muchas veces esa probabilidad de error igual da paso a algoritmos cuya respuesta es completamente determinísitca.

Problema 4 (1 pto.) ¿Por qué usar una base de datos de grafos siendo que tenemos networkx a nuestra disposición? Da dos argumentos a favor de las bases de datos.

Problema 5 (1 pto.) Ya es tiempo de despedirnos de tu amiga y su tienda de Música. Pero antes te pide un último consejo.

Finalmente tu amiga implemento un warehouse en la nube. Uso Google cloud y Big Query. El problema es que al poco tiempo sus consultas comenzaron a crecer en complejidad. Hoy tiene consultas de 1000 líneas que nadie puede entender, y procesos que son muy difíciles de modificar y de revisar posibles bugs o errores.

¿Qué le recomendarías a tu amiga?