TICS575 - Técnicas para Big Data

Actividad 04 - Apache Spark

En esta actividad vamos a realizar dos aplicaciones con Apache Spark. La primera es leer un archivo de logs que puede contener distintos errores de una aplicación, y hay que generar un reporte de estos errores. La segunda, es usar Apache Spark y GraphX para crear un procedimiento con Pregel que compute el PageRank de un grafo. Para correr los programas puedes hacerlo en la Spark Shell o utilizando Databricks.

Importante: lee atentamente el formato esperado de la entrega al final de este enunciado.

1. Pregunta I: Procesar un archivo de *logs* [2 pts]

Junto con este enunciado vas a encontrar un archivo llamado `logs.txt` que contiene los siguientes tipo de mensajes:

- Error de la base de datos.
- Error de la aplicación web.
- Operación correcta en la base de datos.
- Operación correcta en la aplicación web.

Además los mensajes de error de la aplicación web indican de qué tipo es. Para esta parte de la tarea tienes que hacer un programa en Spark que imprima:

- Desplegar los errores de la base de datos.
- El número de errores por tipo de la aplicación web.
- El número de operaciones correctas.

2. Pregunta II: GraphX y Pregel [4 pts]

En esta parte de la tarea vas a crear un programa en GraphX, basado en Pregel, que compute el PageRank de un grafo. Tienes que considerar el archivo `PageRankPregel.scala` publicado junto a este enunciado, donde hay un grafo de ejemplo. Además, te recomendamos apoyarte en las *slides* de la clase.

Para la primera parte de esta tarea deberás explicar a alto nivel cómo utilizar Pregel para computar PageRank, para luego entregar un código en la segunda parte que compute el PageRank de un grafo con la función Pregel implementada en GraphX.

2.1. Parte I: Pregel y PageRank [2 pts]

Para esta parte, tienes que responder las siguientes preguntas, relacionadas a cómo se computa PageRank utilizando Pregel.

- Dado el grafo definido en el archivo `PageRankPregel.scala`, ¿cómo crees que deberías preprocesar el grafo para entregarlo como *input* a la función `pregel` de GraphX? ¿Van a votar los nodos por detenerse en algún momento?
- En cada iteración de Pregel, ¿cuál es el mensaje que manda cada nodo?
- Después de enviar los mensajes, ¿cuál es la función para agrupar los mensajes por nodo en este caso?
- Después de que cada nodo reciba los mensajes, ¿cuál es la operación que va a ejecutar cada nodo?

2.2. Parte II: PageRank en GraphX [2 pts]

Ahora, tienes que tomar el archivo de base `PageRankPregel.scala` que vas a encontrar junto a este enunciado. Luego, debes hacer un programa que compute el PageRank del grafo con la función `pregel` de GraphX. **Importante: no puedes usar la función PageRank implementada en GraphX, se debe hacer con Pregel**.

Para realizar esta tarea, te recomendamos ver las *slides* de clase, donde encontrarás cómo:

- Pre-procesar el grafo.
- Definir las funciones `vprog`, `sendMsg` y `mergeMsg`.
- Cómo se llama a la función `pregel en este caso`.

Detalles Académicos. La tarea se entrega el día 29 de junio hasta las 20:00 horas. Pueden realizar la tarea en grupos de hasta dos personas. La entrega es un archivo en formato **.pdf** a través de un buzón en Webcursos que debe contener:

- Un pantallazo del *output* del programa de la Pregunta I.
- La respuesta a la Parte I de la Pregunta II.
- Un pantallazo del *output* del programa de la Parte II de la Pregunta II.

Además, **debes subir el código de la Pregunta I y de la Parte II de la Pregunta II**.