Procesamiento de Grandes Volúmenes de Datos - MII 775
1er Semestre 2012
Tarea #1

Wenceslao Palma <wenceslao.palma@ucv.cl>

El procesamiento de logs se ha convertido en una operación importante en el contexto de análisis de grandes volúmenes de datos. Como parte de dicho análisis a menudo se realiza un equi-join entre un log y uno o más tablas de referencia las cuales contienen información relativa a usuarios, direcciones, etc. Por ejemplo, Facebook almacena $6TB$ de logs por día y posee $4TB$ de tablas de referencia.
Considere un equi-join $L \bowtie_{L.k=R.k} R$ entre un archivo de log L y una tabla de referencia R, donde $|L| \gg |R|$. En la presente tarea se debe codificar una estrategia clásica de join llamada **Partitioned Sort-Merge Join** usando Hadoop. Si bien las estrategias clásicas de join han sido muy estudiadas durante los últimos 30 años su implementación eficiente en Hadoop no es tan evidente. Para la implementación consideren como guía el pseudocódigo de **Partitioned Sort-Merge Join**, en su versión estándar y mejorada, publicada en el artículo *A Comparison of Join Algorithms for Log Processing in MapReduce.* S. Blanas, J. Patel, V. Ercegovac, J. Rao, E. Shekita and Y. Tan. SIGMOD 2010.

RESTRICCIONES

- La tarea debe ser codificada usando Hadoop.
- Deben codificar la versión estándar y mejorada de **Partitioned Sort-Merge Join**.
- Tanto L como R deben ser archivos de texto generados considerando las condiciones propuestas en la sección 4.1 del artículo mencionado anteriormente.
- Anexe al código fuente un informe (formato pdf) de no más de tres páginas que detalle las decisiones de diseño tomadas en la implementación. Incorpore un gráfico que muestre a grandes rasgos las tareas que realizan Mappers y Reducers así como también las salidas de cada fase. Tomen como ejemplo la figura de la página 11 del documento **Hadoop: Programming** publicado en la página de la asignatura.
- Sólo se consideran las tareas que cumplan con las especificaciones para los datos de entrada y salida.
- La revisión de la tarea incluye una interrogación.
- Solo se recibirán tareas fuera de plazo dentro de las 24 horas siguientes a la fecha de entrega. Nota máxima es un 5.0

FECHA DE ENTREGA: Martes 15 de Junio. Enviar por email código fuente e informe hasta las 24h00 a <wenceslao.palma@ucv.cl>.