Tarea 1

September 1, 2017

1 Instrucciones

- La presente tarea debe ser entregada mediante el Blackboard el día **Viernes 8 de Setiembre a máximo las 17.59 horas.**
- La tarea puede ser realizada en grupos de 4 personas.
- Deberá presentar un documento WORD de 2 páginas (como máximo) conteniendo los resultados de su investigación.
- Deberá preparar un PPT con los resultados de su investigación así como la respuesta a la pregunta que se le hace en la parte de implementación.
- Deberá mostrar los resultados de su procesamiento mapReduce (o realizar el procesamiento en tiempo real). Para esto puede utilizar sus propios computadores o coordinar con Giancarlo para procesarlo en el datacenter.

2 Investigación

Puntaje: 8 puntos

Se le pide leer los siguientes papers:

- http://vldb.org/pvldb/vol5/p1436_alexanderhall_vldb2012.pdf
- http://www.vldb.org/pvldb/vldb2010/papers/R29.pdf
- https://static.googleusercontent.com/media/research.google.com/es//archive/mapreduce-osdi04.pdf

Se le pide realizar un cuadro comparativo donde se comparen las distintas tecnologías de base de datos mencionadas en los papers. En el cuadro comparativo debe de hacer un resumen de la tecnología y ver sus similitudes, diferencias, beneficios y problemas. Al final debe elegir alguna de ellas para el contexto de poder realizar dashboards para una cadena de venta por departamento.

3 Implementación

Puntaje: 12 puntos

En MongoDB, implementar una función mapReduce que me permita contar las palabras de los mails de los empleados de la empresa ENRON.

El dataset lo podrá descargar de la siguiente URL:

- https://www.kaggle.com/wcukierski/enron-email-dataset

Para la carga de la data en MongoDB puede utilizar cualquier lenguaje de programación (Java, Python, R, etc.).

Debe presentar la colección de salida que debe de tener el apellido de cada empleado así como la cantidad de palabras de todos sus mails así como responder la siguiente pregunta:

- £Qué empleado de ENRON escribió más palabras?