Universidad de los Andes Facultad de Ingeniería	Ingeniería de Sistemas y Computación **Maestría** ISIS4517 – Análisis de Big Data – Sección 01 http://sistemas.uniandes.edu.co/~isis4517 Semestre: 2014-10 Horario: Martes – 18:00 a 20:50	Accredited ABET Engineering Accreditation Commission

Laboratorio 1 - Hadoop

Objetivos

- Desarrollar las habilidades básicas de uso de un *cluster* Hadoop
- Desarrollar las habilidades básicas en procesamiento de información utilizando mapReduce
- Realizar una práctica de procesamiento básico paralelo sobre un *dataset* real

Prerrequisitos

- Disponibilidad de recursos en un cluster configurado con Hadoop
- Conocimiento básico de Java
- Conocimiento básico de Unix
- Conocimiento del esquema de desarrollo de una aplicación mapReduce

Herramientas

- Eclipse
- Hadoop 1.1.2.21
- *Dataset* de pruebas

Enunciado

1) Acceso a la infraestructura del cluster

a) Ingrese al cluster previsto para el curso, con las credenciales asignadas a su grupo. La configuración de las máquinas del cluster es igual a la encontrada en la máquina virtual inicial y la asignada en el servidor de admonsis.

b) Ubique en la máquina los recursos asignados para el laboratorio: Instalación de Hadoop y datos de trabajo.

c) Se trabaja en esta práctica con el dataset de noticias de Reuters-21578, disponible en forma gratuita en internet. Contiene noticias de 1987 publicadas por dicha cadena internacional. Encuentre los datos en la carpeta `hdfs /datos/reuters`. Revise rápidamente el contenido de alguno de los archivos que se encuentran en el *dataset*, con el fin de comprender su estructura y contenido.

2) Ejecución de un proceso mapReduce

a) Adjunto a este enunciado encuentra un proyecto Java que permite hacer un proceso mapReduce básico. Importe dicho proyecto en su ambiente de desarrollo y tome el archivo `.jar` correspondiente para hacer las primeras pruebas de un proceso mapReduce sobre Hadoop.

b) Ejecute el proceso en Hadoop.

c) (5%) Explique de forma concisa y asertiva cómo es la respuesta y cómo es el comportamiento del proceso en términos del sistema de archivos. Documente con fotos de pantalla el proceso de ejecución y el resultado..

3) Desarrollo de procesos mapReduce

Para cada uno de los trabajos mapReduce propuestos:

- ✓ Desarrolle y ejecute el trabajo.
- ✓ Documente su trabajo adjuntando el código fuente desarrollado (archivos .java), una rápida descripción del resultado obtenido y una foto de pantalla de la respuesta encontrada.
- ✓ Adjunte los archivos respuesta obtenidos.

a) (15%) Cuente cuántas noticias hay en el *dataset*

b) (40%) Cuente cuántas veces aparece cada palabra en la primera línea del título de las noticias para todos los títulos de noticias.

c) (40%) Indique el título de cada noticia y cuántas palabras aparecen el cuerpo de la misma y encuentre el título de la noticia con más palabras en el cuerpo de la misma (probablemente el `TextInputFormat` no es el indicado…)

d) BONO (20%) Sólo se tienen en cuenta el bono si **TODOS** los retos anteriores están adecuadamente desarrollados.

En UNA pasada sobre el *dataset*, genere un archivo con los títulos y etiquetas de las noticias relacionadas con una etiqueta dada (ver asignación de etiquetas por grupo) y otro archivo con el título y las 10 palabras más frecuentes en el cuerpo de las noticias del rango de fechas indicado para su grupo (ver asignación por grupo).

Metodología

- Este laboratorio es una actividad presencial. La no asistencia se ve reflejada en una nota de 0.0/5.0
- El laboratorio se desarrolla en los grupos de estudiantes que están previstos para el desarrollo de trabajos prácticos.
- Se espera una participación equitativa de los integrantes del grupo

Entrega de laboratorio

Fecha y hora límite: Martes 25 de febrero, 8:50 pm.

Archivo de entrega: **<Lab1_NN_login1_login2_login3>.zip.** NN es el número del grupo y los *logins* corresponden a los integrantes presentes y participantes en el desarrollo.

Contenido del archivo:

- Informe de resultados, en formato `.pdf`, que sigue la plantilla de informes técnicos prevista para el curso. Contiene las explicaciones y documentación solicitada para cada uno de los puntos del laboratorio.

  Nombre del archivo: **<Lab1_NN_login1_login2_login3>.pdf**

- Archivos adjuntos de respuestas, adecuadamente identificados con respecto al punto del enunciado.
- Archivos `.java` desarrollados.

El no seguimiento del formato de entrega del taller tiene una penalización de **0.5/5.0** en la nota final.

Asignación de retos por grupo

Grupo	People	Fechas
1	Fidel Castro	1-01-1987 a 15-04-1987
2	thatcher	16-02-1987 a 31-05-1987
3	stoltenberg	1-04-1987 a 15-07-1987
4	balladur	16-05-1987 a 31-08-1987
5	nakasone	1-07-1987 a 31-10-1987
6	sumita	1-09-1987 a 15-12-1987
7	james-baker	16-02-1987 a 30-06-1987
8	douglas	1-03-1987 a 15-07-1987
9	conable	15-06-1987 a 30-09-1987
10	dauster	1-08-1987 a 31-12-1987
11	mancera-aguayo	1-06-1987 a 30-09-1987