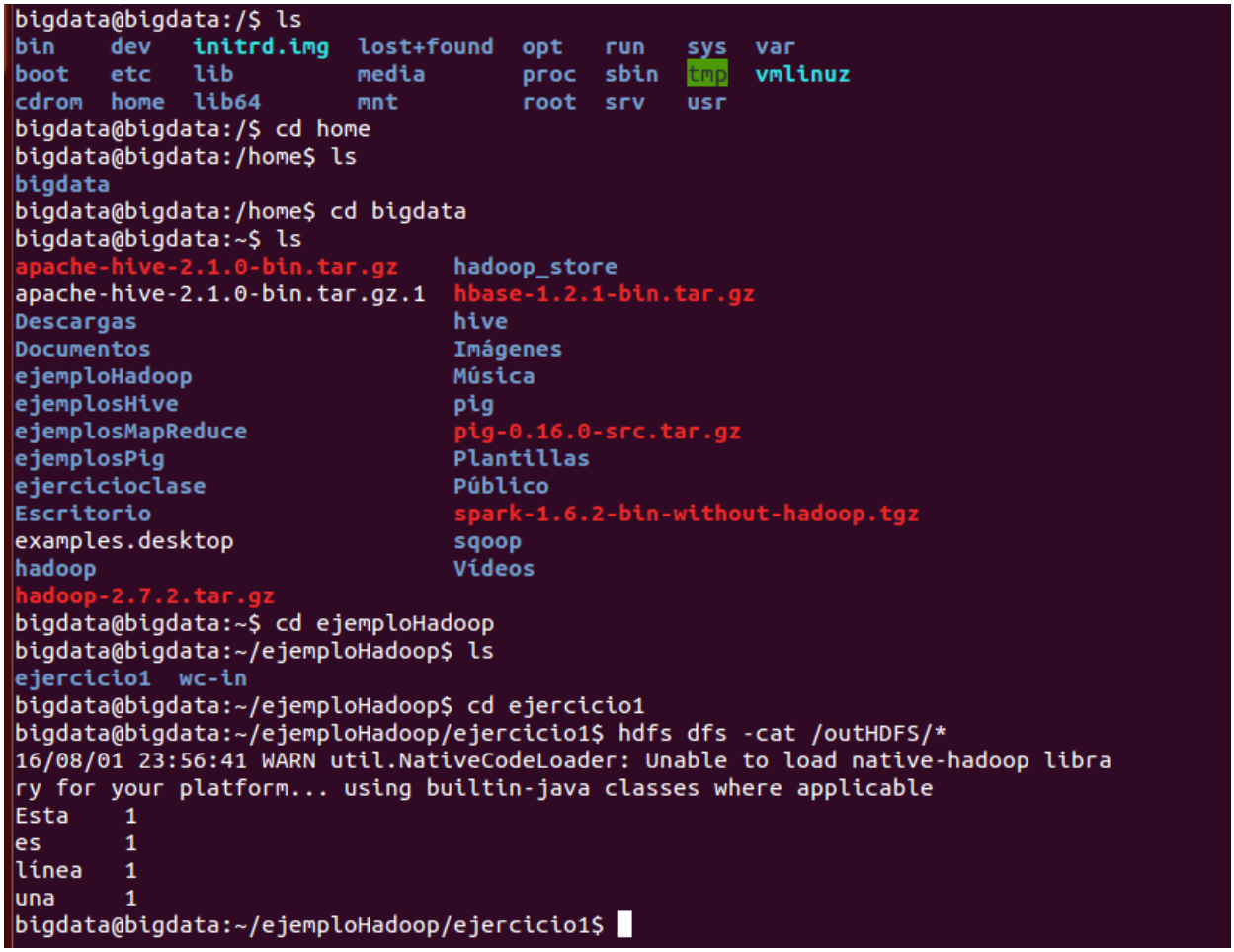
EJERCICIOS DE HADOOP

1. Devolver el resultado de ejecutar el siguiente comando
$HADOOP_HOME/bin/hdfs dfs -cat /outHDFS/*

Tengo creada carpeta bigdata/ejemploHadoop/ejercicio1 y ejecuto el comando en esa carpeta.
bigdata@bigdata:~/ejemploHadoop/ejercicio1$ hdfs dfs -cat /outHDFS/*



→Las respuestas marcadas en negrita

2. ¿Cuál de los siguientes funcionalidades pertenecen a un NameNode de HDFS?

a. Transferir bloques de datos de los datanodes a los clientes

b. Mantener el árbol del sistema de archivos y los metadatos de todos los ficheros y directorios

c. Controlar los procesos de Map Reduce

d. Almacenar bloques de datos

e. Ninguna de las opciones es correcta

Recursos:

https://wiki.apache.org/hadoop/NameNode
https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

3. ¿Cuál de los siguientes funcionalidades pertenece a un DataNode de HDFS?

a. Mantener el árbol del sistema de archivos y los metadatos de todos los ficheros y directorios.

b. Controlar la ejecución de una tarea de mapeo o de reduce individual.

c. Gestionar el sistema de espacios de nombres de los archivos.

d. Almacenar y recuperar bloques cuando los clientes o el NameNode lo solicita.

e. Ninguna de las opciones es correcta.

Recursos:

https://wiki.apache.org/hadoop/DataNode
https://hadoop.apache.org/docs/r1.2.1/hdfs_design.HTML

4. ¿Cuál de las siguientes frases es cierta con respecto a YARN?

a. Implementa un gestor de ficheros para todos los frameworks de Hadoop.

b. Permitir acceso a los datos de HDFS a programas que no estén desarrollados en Hadoop.

c. Permitir a múltiples Namenodes con sus propios namespaces, compartir el pool de Datanodes.

d. Usar el JournalNode para decidir el NameNode activo.

e. Ninguna de las anteriores es correcta.

"YARN is a software rewrite that decouples MapReduce's resource management and scheduling capabilities from the data processing component, enabling Hadoop to support more varied processing approaches and a broader array of applications.
...
YARN combines a central resource manager that reconciles the way applications use Hadoop system resources with node manager agents that monitor the processing operations of individual cluster nodes"

Recursos:

http://searchdatamanagement.techtarget.com/definition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator
http://www.happyminds.es/apache-hadoop-introduccion-a-yarn/#sthash.OheEiN0j.dpbs

5. HDFS está diseñado para

a. Ficheros grandes, acceso continuo a los datos y hardware de grandes prestaciones.

b. Ficheros pequeños, acceso continuo a los datos y hardware básico.

c. Ficheros grandes, baja latencia de acceso y hardware básico.

d. Ficheros grandes, acceso continuo a los datos y hardware básico.

e. Ninguna de las anteriores es correcta.

Recursos:

http://www.happyminds.es/apache-hadoop-introduccion-a-hdfs/#sthash.yQ6KWFkV.dpbs

6. ¿Qué es común a Pig y Hive?

a. Permiten múltiples y aleatorias escrituras y lecturas.

b. Traducen lenguajes de alto nivel a trabajos de Map Reduce.

c. Todas operan con estructura de datos JSON.

d. Todas son lenguajes de flujos de datos.

e. Ninguna de las anteriores es correcta.

Recursos:

http://stackoverflow.com/questions/3356259/difference-between-pig-and-hive-why-have-both
http://blogs.solidq.com/es/big-data/hive-sqoop-y-pig/

http://formacionhadoop.com/aulavirtual/pluginfile.php/32/course/summary/Desarrollador%20Apache%20Hadoop%20Cap%C3%ADtulo%203%20-%20Hadoop%20Conceptos%20B%C3%A1sicos.pdf

7. ¿Qué es Flume?

a. Un sistema de archivos distribuido.

b. Una plataforma de ejecución de tareas de MapReduce.

c. Un lenguaje de programación que traduce queries de alto nivel en tareas de map reduce.

d. Un servicio para mover grandes cantidades de datos en un cluster una vez los datos se han generado.

e. Ninguna de las anteriores.

Recursos:

http://hortonworks.com/apache/flume/

8. ¿En cuál de los siguientes escenarios usarías Hadoop?

a. Analizar los signos vitales de un bebé en tiempo real.

b. Obtener las tendencias de acciones bursátiles cada minuto.

c. Procesar un sensor meteorológico para predecir la trayectoria de un huracán.

d. Procesar billones de mensajes de email para ejecutar análisis de texto.

e. Ninguna de las anteriores.

9. ¿Cuál de las siguientes frases es cierta?

a. Hadoop es una nueva tecnología diseñada para reemplazar las bases de datos relacionales.

b. Hadoop incluye componentes open source y closed source.

c. Hadoop se puede usar para bigdata, DSS y OLTP.

d. Todas las anteriores son correctas.

e. Ninguna de las anteriores es correcta.

Hadoop no pretende sustituir las bases de datos relacionales

OLTP(Online trasnaction processing) – Hadoop no suele usarse para procesamiento online

"Cloudera's enterprise Hadoop distribution, CDH3, is open source, while MapR's enterprise Hadoop distribution, M5, is closed"

Recursos:

http://www.allinterview.com/showanswers/8480/difference-between-dss-oltp.html

http://siliconangle.com/blog/2011/08/26/rejecting-%E2%80%9Cclosed-source%E2%80%9D-label-mapr-contrasts-its-hadoop-approach-to-cloudera/

Otros recursos

http://www.dummies.com/how-to/content/hadoop-distributed-file-system-shell-commands.html
https://dzone.com/articles/top-10-hadoop-shell-commands
http://itwiz.pl/hadoop-czyli-przetwarzanie-rozproszone-open-source/