Apache Hadoop

Week 6 Tasks

MCQ

Q 1 - In a Hadoop cluster, what is true for a HDFS block that is no longer available due to disk corruption or machine failure?

A - It is lost for ever
B - It can be replicated form its alternative locations to other live machines.
C - The namenode allows new client request to keep trying to read it.
D - The Mapreduce job process runs ignoring the block and the data stored in it.

Q 2 - Which one of the following is not a main component of HBase?

A - Region Server.
B - Nagios.
C - ZooKeeper.
D - Master Server.

Q 3 - When a client communicates with the HDFS file system, it needs to communicate with

A - only the namenode
B - only the data node
C - both the namenode and datanode
D - None of these

Q 4 - Which of the following is not a Hadoop operation mode?

A - Pseudo distributed mode
B - Globally distributed mode
C - Stand alone mode
D - Fully-Distributed mode

Q 5 - The information mapping data blocks with their corresponding files is stored in

A - Data node
B - Job Tracker
C - Task Tracker
D – Namenode

Q 6 - HDFS stands for

A - Highly distributed file system.
B - Hadoop directed file system
C - Highly distributed file shell
D - Hadoop distributed file system.

Q 7 - The source of HDFS architecture in Hadoop originated as

A - Google distributed filesystem
B - Yahoo distributed filesystem
C - Facebook distributed filesystem
D - Azure distributed filesystem

Q 8 - The current limiting factor to the size of a hadoop cluster is

A - Excess heat generated in data center
B - Upper limit of the network bandwidth
C - Upper limit of the RAM in namenode
D - 4000 data nodes

Q 9 - The namenode loses its only copy of fsimage file. We can recover this from

A - Datanodes
B - Secondary namenode
C - Checkpoint node
D – Never

Q 10 - Which of the following technologies is a document store database?

A - HBase
B - Hive
C - Cassandra
D – CouchDB

Problem Statement:

Split the Data into Columns

Dataset:
https://docs.google.com/spreadsheets/d/1rkJL6jIQLDSHB3ypeZ_bNTUmmJDgfvGn/edit#gid=422718792

Assignment: A CSV data file including movie names and their release years is supplied to you. Your assignment as follows:

1) You must divide the information into two columns: “ReleasedYear” and “MovieTitle.
2) Identify the duplicate entries.
3) Find out how many times the word “red” appears in a title.

Thank You!