coreservlets.com – Hadoop Course

HDFS Overview

This exercise provides a chance to answer questions about HDFS to reinforce concepts presented in HDFS Overview lecture.

Approx. Time: 15 minutes

Answer

1. What are the three daemons that manage HDFS? What is the purpose of each?
2. What is the block replication in HDFS? What is the default replication factor?
3. What can you do to affect Namenode's memory usage?
4. What are two common HDFS access patterns?

Solution

1. The Daemons are

   (1) Namenode: manages the filesystem's file blocks; runs on one machine to several machines

   (2) Datanode: stores and retrieves data blocks; reports to Namenode; Runs on many machines

   (3) Secondary Namenode: Performs housekeeping work so Namenode doesn't have to; requires similar hardware as Namenode machine; not used for high-availability, not a backup to Namenode

2. Files are split into blocks and replicated across machines at load time. Replication is used to assure fault-tolerance and provide access. Default replication is 3.

3. Changing HDFS's default block size will affect the number of blocks that a Namenode can handle and significantly reduce Namenode's memory usage.

4. Common HDFS access patterns are

   (1) Direct: Communicate with HDFS directly through native client such as Java, C++

   (2) Proxy Server: Access HDFS through a Proxy Server, middle man, such as REST, Thrift, and Avro Servers