CS455 – Lab Session 10

NAMAN SHAH
03/31/2017

Quiz-7

1. Consider a client that is generating data for a file to be stored in HDFS. The file is being stored with a default replication level of 3. The client creates a pipeline with 3 replicas of a block, and writes to ONLY the first replica within this pipeline. 
   [True/False]

2. HDFS supports UNIX session semantics – changes to a file are immediately visible. 
   [True/False]

Agenda

• Quiz-7 review
• Spark vs. Hadoop (Speed)
• Course Logistics
• Spark execution DAG

Quiz-7

3. Placing all replicas on single machine offers speed but no redundancy. 
   [True/False]

4. According to default replication strategy in HDFS, the 2nd replica is placed on different rack from the first. 
   [True/False]
Quiz-7

5. 'distcp' uses reducers to copy data in parallel. [True/False]

6. The namenode may redirect client requests to a corrupted replica even if a datanode has previously reported that replica as corrupt. [True/False]

7. Disabling checksum is possible in HDFS. [True/False]

8. Bzip2 is a splittable compression format. [True/False]

9. Hadoop uses built-in serialization mechanisms provided by Java. [True/False]

10. HDFS stores 1GB of a compressed gzip file as 16 blocks. Hadoop can process this data using 16 concurrent maps one for each of the 16 blocks. [True/False]

Spark vs. Hadoop (Speed)

Hadoop

Spark
Spark execution DAG

- **Job**
  - Reads data from input (DFS)
  - Executes code chunks
  - Writes data to output/driver
  - Identified by action performed on RDD

- **Stages**
  - Jobs are divided into stages
  - Stage consists of multiple operations, most of which are map*
  - Stage boundary defined by operations that require shuffle
  - All operations in single stage are performed in single task.

Spark execution DAG

- **Task**
  - Each stage has multiple tasks
  - One task is created per stage per RDD partition and executed by Executor
  - Initially, the tasks are in a job, i.e., a stage * N by N partition in these stages.

Course Logistics

- HW4 - Team submissions (Deliverable 2)
  - Team submissions ASAP
- HW4 – Project proposal (Deliverable 1)
  - Using checkin
  - Folder set aside for this is TP-D1
Next Session

- Quiz - B review
- one or two vanilla spark examples

QUESTIONS?