

CS 686: Special Topics in Big Data

HDFS

Lecture 6

Today's Agenda

- Some notes on HDFS
- Group & Class Discussion: HDFS
- (Small) lab: P1 design documents

9/6/17

CS 686: Big Data

2

Today's Agenda

- Some notes on HDFS**
- Group & Class Discussion: HDFS
- (Small) lab: P1 design documents

9/6/17

CS 686: Big Data

3

Namespace & Image

- There seems to be a complicated relationship between these two concepts, but it's fairly simple
 - All about the active (in memory) state of the system and the persisted (on disk) state
- Sometimes the Namespace is used to refer to an "instance" of HDFS
 - Newer versions support federation

9/6/17

CS 686: Big Data

4

Name/Backup/Checkpoint Node

- These nodes share a lot of the same functionality, so distinguishing between them can be difficult!
- The NameNode can also be configured to take on more or less of the overall set of responsibilities
- Newer versions of HDFS support high availability (HA) with two NameNodes
 - In HA mode, the JournalNode takes care of logging the edit stream

9/6/17

CS 686: Big Data

5

Node Functionality

- Checkpoint Node – responsible for managing the on-disk representation of the file system
 - fsimage + edits
 - Sends updated checkpoints back to NameNode
- Backup Node – same as the Checkpoint node but also maintains in-memory state of the file system
 - You can't have a Checkpoint + Backup Node
 - You can configure NameNode to only maintain in-memory state and make Backup responsible for disk

9/6/17

CS 686: Big Data

6

Deploying HDFS

- The main thing to remember: you can run HDFS with just a NameNode
 - Recovery times will be long (edit log not merged)
 - NN goes down, your system is useless
- I've seen installations that actually put the BackupNode on the same machine as the NameNode
 - Don't do this! ☹
- You **can** run the CheckpointNode on the same machine as the NameNode if you must

9/6/17

CS 686: Big Data

7

Today's Agenda

- Some notes on HDFS
- **Group & Class Discussion: HDFS**
- (Small) lab: P1 design documents

9/6/17

CS 686: Big Data

8

Groups

- Group 1 – Sections II-A, II-B
- Group 2 – Sections II-C, II-D
- Group 3 – Sections II-E, II-F, II-G
- Group 4 – Section III-A
- Group 5 – Sections III-B, III-C

9/6/17

CS 686: Big Data

9

Today's Agenda

- Some notes on HDFS
- Group & Class Discussion: HDFS
- **(Small) lab: P1 design documents**

9/6/17

CS 686: Big Data

10
