facebook

The High Availability story for HDFS so far

Dhruba Borthakur
dhruba@apache.org

Presented at ApacheCon at Oakland, California
November 5, 2009

How infrequently does the NameNode (NN) stop?

- **Hadoop Software Bugs**
  - Two directories in fs.name.dir, but when a write to first directory failed, the NN ignored the second one (once)
  - Upgrade from 0.17 to 0.18 caused data corruption (once)
- **Configuration errors**
  - Fsimage partition ran out of space (once)
  - Network Load Anomalies (about 10 times)
- Maintenance:
  - Deploy new patches (once every month)

What does the SecondaryNameNode do?

- **Periodically merges Transaction logs**
- **Requires the same amount of memory as NN**
- **Why is it separate from NN?**
  - Avoids fine-grain locking of NN data structures
  - Avoids implementing copy-on-write for NN data structures
- **Renamed as CheckpointNode (CN) in 0.21 release.**

Shortcomings of the SecondaryNameNode?

- **Does not have a copy of the latest transaction log**
- **Periodic and is not continuous**
  - Configured to run every hour
- **If the NN dies, the SecondaryNameNode does not take over the responsibilities of the NN**

BackupNode (BN)

- NN streams transaction log to BackupNode
- BackupNode applies log to in-memory and disk image
- BN always commit to disk before success to NN
- If BN restarts, it has to catch up with NN
- Available in HDFS 0.21 release

Limitations of BackupNode (BN)

- **Maximum of one BackupNode per NN**
  - Support only two-machine failure
- **NN does not forward block reports to BackupNode**
- **Time to restart from 2 GB image, 20M files + 40 M blocks**
  - 3 – 5 minutes to read the image from disk
  - 30 min to process block reports
  - BN will still take 30 minutes to failover!

Overlapping Clusters for HA

- “Always available for write” model
- Two logical clusters each with their own NN
- Each physical machine runs two instances of DataNode
- Two DataNode instances share the same physical storage device
- Application has logic to failover writes from one HDFS cluster to another
- More details at http://hadoopblog.blogspot.com/2009/06/hdfs-scribe-integration.html

HDFS+Zookeeper

- HDFS can store transaction logs in Zookeeper/Bookeeper
  - http://issues.apache.org/jira/browse/HDFS-234
- Transaction log need not be stored in NFS filer
- A new NN will still have to process block reports
  - Not good for HA yet, because NN failover will take 30 minutes