Hadoop Tools

- Abhay Dandekar

Agenda

1. What is Hadoop streaming?
2. How it works?
3. Options for Hadoop streaming
4. Example

5. What is distcp?
6. How it works?
7. Options for distcp
8. Example

Hadoop streaming...

1. Utility with hadoop
2. Create and run MR jobs with any executable as mapper / reducer
3. Provide input from HDFS and output into HDFS
4. Example :
   hadoop jar hadoop-streaming-3.3.1.jar
   -input /user/abhay/README.txt
   -output /user/abhay/output_streaming_5
   -mapper /bin/cat
   -reducer /usr/bin/wc
5. Link:
   https://hadoop.apache.org/docs/current/hadoop-streaming/HadoopStreaming.html

Command line options

-input	Mapper’s Input location on HDFS
-output	Reducer’s output location on HDFS
-mapper	Mapper command
-reducer	Reducer command
-files	Files to be packaged with Mapper and Reducer. These files will be available inside the Mapper/Reducer container
-inputformat	Input format (default TextInputFormat)
-outputformat	Output format (default TextInputFormat)
-verbose	

Command line - 2

-partitioner <JAVA_CLASS>	Java class to determine the partition information
-combiner	Combiner for Map output
-cmdenv name=value	Environment variable for streaming commands
-lazyOutput	Generates output in a lazy fashion
-numReduceTasks	Number of reducers
-mapdebug	Script to call on Map failure (on debugging only)
-reducedebug	Script to call on Reduce failure (on debugging only)
-inputreader	Specified a input reader class (in older version versions)

Distcp

1. Tool for large inter/intra cluster copying.
2. Uses map reduce in background to copy across data
3. Example:

   hadoop distcp
   hdfs://nn1:9000/foo/bar
   hdfs://nn2:9000/bar/foo
4. Link: https://hadoop.apache.org/docs/current/hadoop-distcp/DistCp.html

Distcp - Diagram DFD