Hadoop Revision Notes

Version 1.0.0

Author: HadoopExam Learning Resource

Print Screen Hadoop Exam Simulator

Hadoop Revision Notes

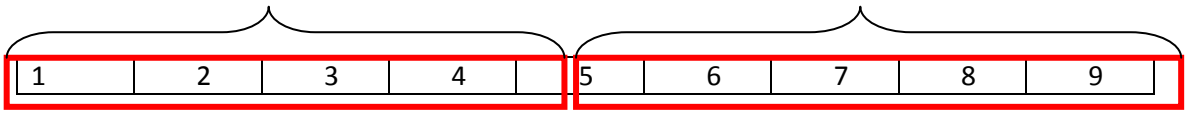
Core Concepts

- The responsibility of the TaskTracker is to Instantiating and monitoring of individual map and reduce task.
- JobTracker is responsible to send individual task to each Node.
- Hadoop Daemon does not share the JVM.
- Technically all the Hadoop Node can run, all the daemons on single node, however it is not advised.
- Daemons (JobTracker, NameNode and Secondary NameNode) are considered as Master Node
- DataNode and TaskTracker are slave node

Input/Output

- Reducer input must have the same type as the Map output.
- Context object is used for emitting key-value pairs.

- By default hash partition algorithm is used to decide in which partition key-value should be sent and is based on the keys hash values.
- We can run MapReduce job without setting Mapper and Reducer and it will produce line-offset (Which is not a line number) as key and line as a value in output directory.
- The default input format for the MapReduce job is **TextInputFormat**, which produce **LongWritable** as key and **Text** as value type in a Tab separated form.
- Each partition is processed by a reduce task, therefore number of partion is equal to number of reducer.
- Example for calculating the partion
    - **(key.hashCode() & Integer.Max_Value)%numberOfReducer**



HDFS

- HDFS actively monitors any failed datanode(s) and upon failure detection immediately schedules re-replication of blocks (if needed).
- With 3 replication factor and multiple racks HDFS will write one copy of the block on the same node and other two copy on different rack in different machine.

MapReduce

- An Instance of Configuration class represent collection of configuration properties.
- Configuration properties can be defined in terms of system properties but they can not be used until System Properties are redefined using Configuration properties.

Hive

- Hive gives SQL like interface for data stored in HDFS
- Hive provides programtic access like JDBC/ODBC for data stored in HDFS

Data Join

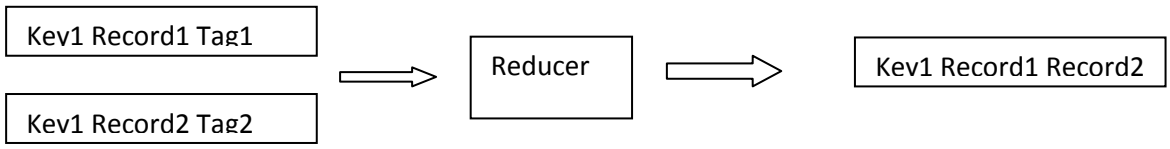
- Map-join map side join is done in map phase and done in memory. The map-side join technique allows splitting map file between different data nodes. The data will be loaded into memory. This allows very fast performance for join.

HBase

- It is a column family database
- Can store massive data e.g. TeraBytes
- It gives high throughput

Algorithm and Features

- Counters are useful channel for gathering statistics about the Job, for quality control or for application level statistics.
- TaskCounter : Updated as task progresses.
- JobCounter : Updated as Job progresses.



Note : PappuPass.com is www.HadoopExam.com

Contact Us:

Email : admin@hadoopexam.com

hadoopexam@gmail.com

Phone: 022-42669636

Mobile : +91-8879712614

HadoopExam Learning Resources

B-902 Shah Trade Center

Malad-E Mumbai 400097