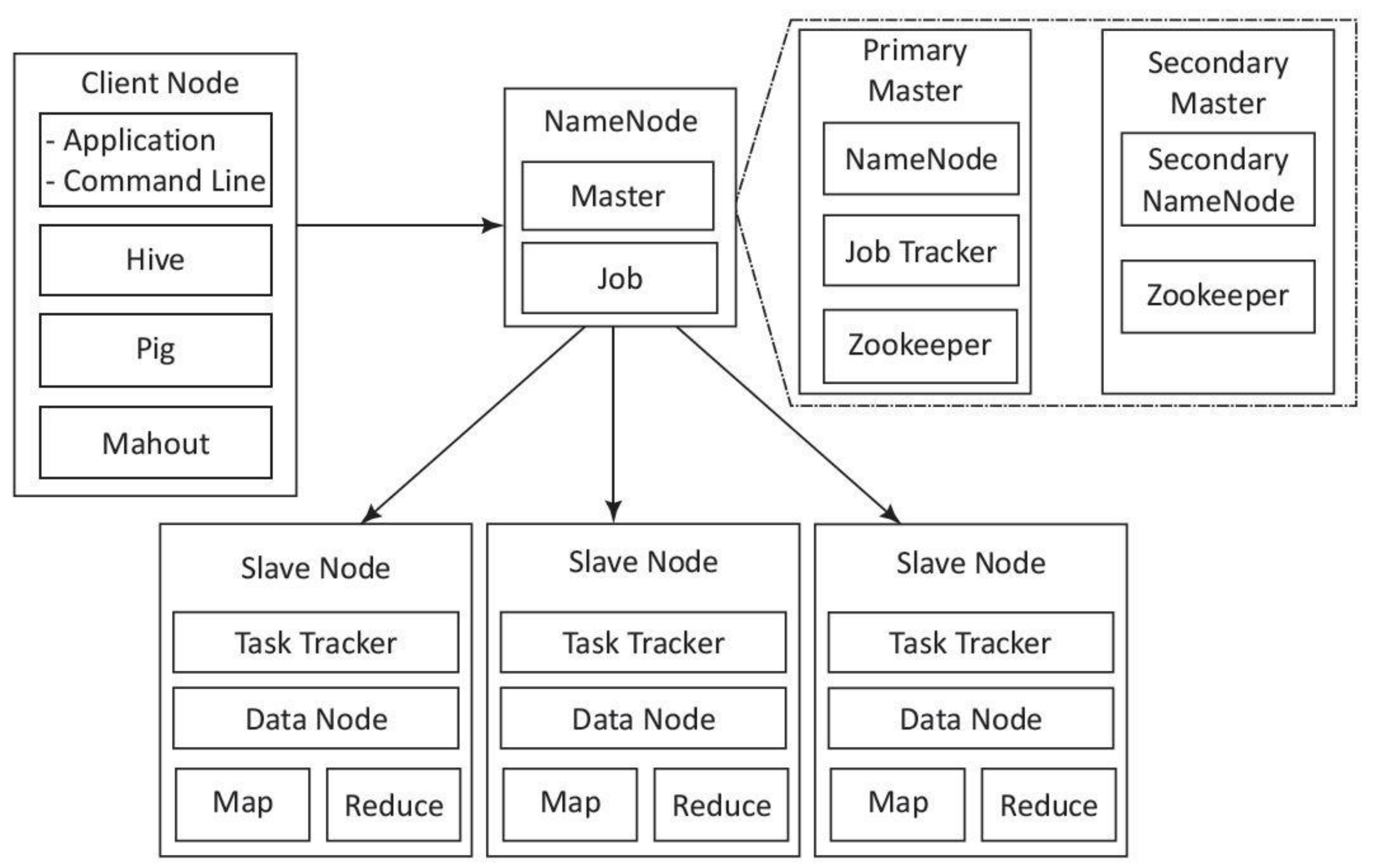
Lesson 3

MapReduce Framework and Programming Model

MapReduce— an Integral Part of Hadoop Physical Organization



Mapper

- Means software for doing the assigned task after organizing the data blocks imported using the keys
- A key is specified in a command line of Mapper
- The command maps the key to the data

Reducer

- Means software for reducing the mapped data by using the aggregation, query or user-specified function
- Reducer provides a concise cohesive response for the application
-

Aggregation Function

- Means the function that groups the values of multiple rows together to result a single value which provides more significant meaning
- For example, function such as count, sum, maximum, minimum, deviation and standard deviation

Querying function

- Means a function that finds the desired values
- For example, function for finding a best student of a class who has shown the best performance in examination

MapReduce For Application Tasks

- The tasks send the MapReduce for processing,
- Reliably processes the huge amounts of data, in parallel, on large clusters of servers
- The cluster size does not limit as such to process in parallel.

Parallel Programming

- The parallel programs of MapReduce useful for performing large scale data analysis using multiple machines in the cluster

Features of MapReduce Framework

1. Provides automatic parallelization and distribution of computation based on several processors

2. Processes data stored on distributed clusters of DataNodes and racks

4. Provides scalability for usages of large number of servers

Features of MapReduce Framework

5. Provides MapReduce batch-oriented programming model in Hadoop version 1

6. Provides additional processing modes in Hadoop 2 YARN-based system and enables required parallel processing of 3V characteristics data

Features of MapReduce Framework

For example enables required parallel processing for queries, graph databases, streaming data, messages, real-time OLAP and ad hoc analytics with Big Data 3V characteristics in Hadoop 2.

Hadoop MapReduce Framework

- The processing tasks are submitted to the Hadoop
- The Hadoop framework in turns manages the task of issuing jobs, job completion, and copying data around the cluster between the DataNodes with the help of JobTracker

MapReduce Daemon Feature

- Daemon refers to a highly dedicated program that runs in the background in a system
- The user does not control or interact with that

MapReduce

- Runs as per assigned Job by JobTracker, which keeps track of the job submitted for execution and runs TaskTracker for tracking the tasks
- MapReduce programming enables job scheduling and task execution

Job Tracker

- A client node submits a request of an application to the JobTracker
- A JobTracker is a Hadoop daemon (background program)

Steps at MapReduce

(i) estimate the need of resources for processing that Job request

(ii) analyze the states of the slave nodes

(iii) place the mapping tasks in queue

(iv) monitor the progress of task, and on the failure, restart the task on slots of time available

Mapper and Reducer Roles

- Deploys map tasks on the slots
- Map tasks assign to those nodes where the data for the application is stored
- The Reducer output transfers to the client node after the data serialization using AVRO.

Summary

We learnt

- Mapper does the assigned task after organizing the data blocks imported using the keys
- Reducer reducing the mapped data by using the aggregation, query or user-specified function

… Summary

We learnt:

- Enables required parallel processing at MapReduce in Hadoop 2
- Place the mapping tasks in queue
- Monitor the progress of task
- Reducer output transfers to the client node after the data serialization using AVRO

End of Lesson 3 on

MapReduce Framework and Programming Model