Agenda

To understand the major components of Hadoop 1.0 architecture

- Name node
- Secondary name node
- Job Tracker
- Task tracker
- Data nodes

Key Terms

greatlearning
Learning for Life

Some key terms used while discussing Hadoop

- Commodity hardware: PCs which can be used to make a cluster
- Cluster/grid: Interconnection of systems in a network
- Node: A single instance of a computer
- Distributed System: A system composed of multiple autonomous computers that communicate through a computer network
- ASF: Apache Software Foundation
- HA: High Availability
- Hot standby : Uninterrupted failover

Master-Slave Architecture

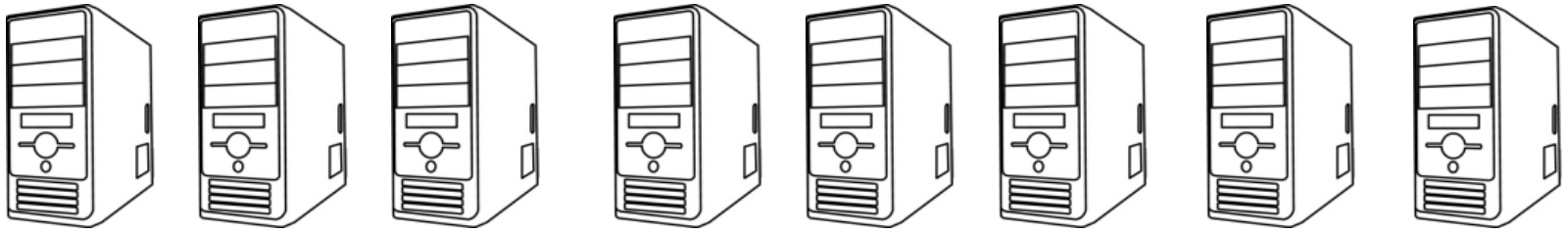
Machines in MASTER MODE

Name node

Secondary name node

JOB tracker

Data nodes in slave mode



HADOOP Deployment Modes

HADOOP supports 3 configuration modes when its is implemented on commodity hardware:

**Standalone mode : ** All services run locally on single machine on a single JVM (seldom used)

Pseudo distributed mode : All services run on the same machine but on a different JVM (development and testing purpose)

Fully distributed mode: Each service runs on a seperate hardware (a dedicated server). Used in production setup.

Note: Service here reffers to namenode, secondary name node, job tracker and data node.

Functionality of Each Component of HADOOP

Master nodes

- **Name node** : Central file system manager
- **Secondary name node :** Data backup of name node (not hot standby)
- **Job tracker** : Centralized job scheduler

Slave nodes and deamons/software services

- **Data node** : Machine where files gets stored and processed
- **Task tracker :** A software service which monitors the state of job tracker
- **Note :** Every slave node keeps sending a heart beat signal to the name node once in every 3 seconds to state that its alive. What happens when a data node goes down would be discussed in the subsequent slides.

What is a job in HADOOP eco system?

- A job usually is some **task** submitted by the user to the Hadoop cluster.

- The job is in the form of a **program or collection of programs** (a JAR file) which needs to be executed.

- A job would have the following attributes to it

  1)The actual program/s

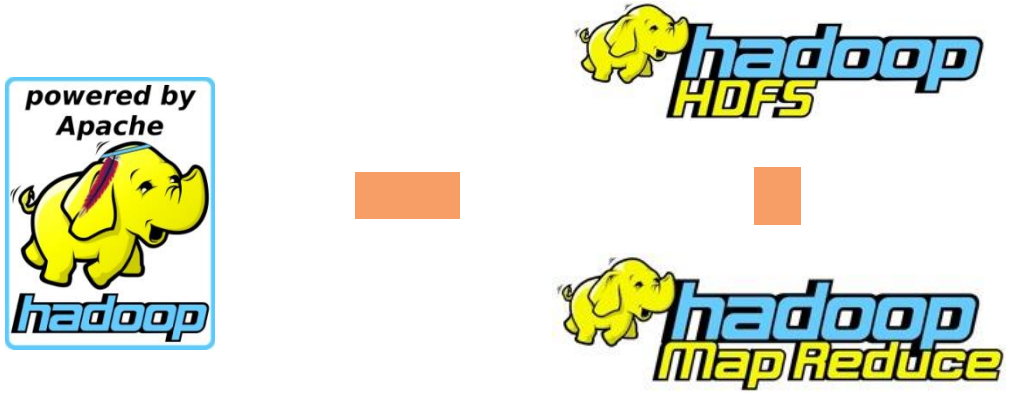
  2)Input data to the program (a file or a collection of files in a directory)

  3)The output directory where the results of execution is collected in a file/s

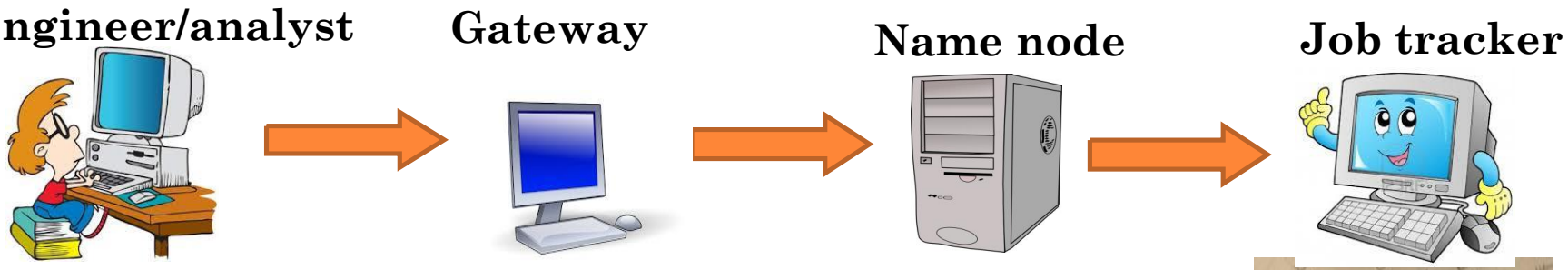
Apache HADOOP Core Features

Core features of Apache Hadoop are:

- HDFS (Hadoop Distributed File System) – data storage
- MapReduce Framework – compute in distributed environment
  - A Java framework responsible for processing jobs in distributed mode
  - User-defined map phase, which is a parallel, share-nothing processing of input
  - User-defined reduce phase aggregates of the output of the map phase



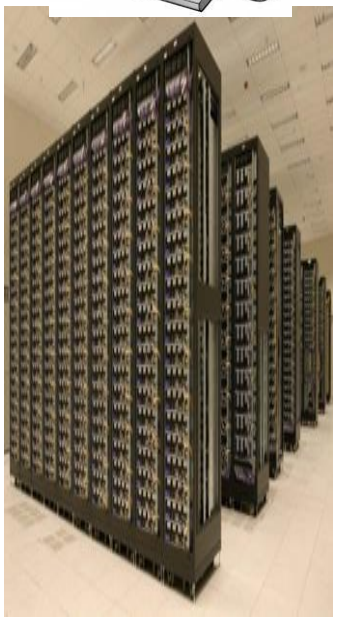
Submitting and executing a job in a hadoop cluster



The engineer/analyst's machine is not a part of Hadoop cluster. Usually Hadoop would be installed in pseudo-distributed mode on his/her machine. The job (program/s) would be submitted to the gateway machine

The gateway machine would have the necessary configuration to communicate to the name node and job tracker.

Job gets submitted to the name node and eventually job tracker is responsible for scheduling the execution of the job on the data nodes in the cluster.



Data nodes

Summary

An overview of the core components of HADOOP 1.0

Basic understanding of the components namenode, secondary name node, job tracker, data node and task tracker

What is a job in a HADOOP eco system ?

How to submit a job in a HADOOP cluster ?