SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)
Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

Department of MCA

Topic: Hadoop Setup & Installation

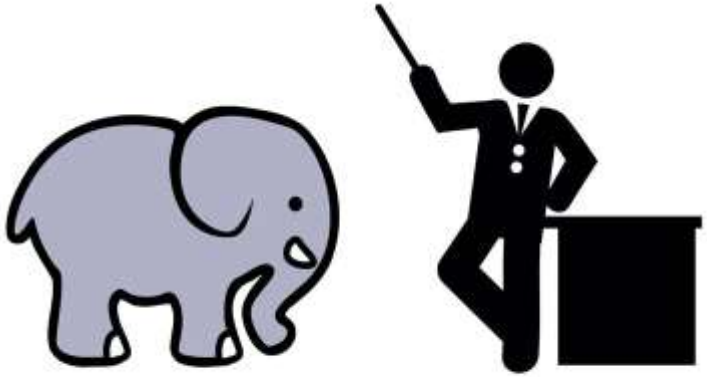
Course: 16CAT702 - Big Data Analytics

UNIT II : Hadoop

III Semester / II MCA

Session Objectives

- Create, setup and configure Hadoop cluster
- Manage environment setting and properties



Hadoop Installation

- Installation can be done in two ways
- One is
  - Install java version 6 or above
  - Create a separate user for Hadoop
  - Download and unpack Apache Hadoop distribution in a sensible location
  - Test the installation

Hadoop Installation

- Another one is
  - linux RPMs or Debian packages
  - use an automated installation method like Red Hat Linux's Kickstart / Debian's Fully Automatic Installation

SSH Configuration

- Hadoop control scripts rely on SSH to perform cluster-wide operations
- SSH needs to be set up to allow password-less login for the hadoop user from machines in the cluster
- Simple way is to generate a public/private key pair, and it will be shared across the cluster using NFS
- Use ssh-agent to avoid the need to enter a password for each connection
- Private key is in the file ~/.ssh/id_rsa, and the public key is stored in a file with the same name with .pub appended, ~/.ssh/id_rsa.pub

Hardware Configuration

- Files for controlling the configuration of a Hadoop installation
- Stored in conf directory

Hadoop Installation

- Hadoop does not have a single, global location for configuration information
- Node in the cluster has its own set of configuration files
- Hadoop provides facility for synchronizing configuration using rsync
- Hadoop is designed to have a single set of configuration files that are used for all master and worker machines

Control Scripts

- scripts for running commands, and starting and stopping daemons across the whole cluster
- To tell Hadoop which machines are in the cluster, there are two file
- Masters file is actually a misleading name, in that it determines which machine or machines should run a secondary namenode
- Slaves file lists the machines that the datanodes and tasktrackers should run on
- No differences between datanode and tasktracker, but identifies by using running script

Control Scripts

start-dfs.sh script, which starts all the HDFS daemons in the cluster, runs the namenode on the machine

1. Starts a namenode on the local machine (the machine that the script is run on)
2. Starts a datanode on each machine listed in the slaves file
3. Starts a secondary namenode on each machine listed in the masters file

Control Scripts

- start-mapred.sh, which starts all the MapReduce daemons in the clusterStarts a namenode on the local machine (the machine that the script is run on)
  1. Starts a jobtracker on the local machine
  2. Starts a tasktracker on each machine listed in the slaves file
- stop-dfs.sh and stop-mapred.sh scripts to stop the daemons started by the corresponding start script

Hadoop Configuration

Filename	Format	Description
hadoop-env.sh	Bash script	Environment variables that are used in the scripts to run Hadoop.
core-site.xml	Hadoop configuration XML	Configuration settings for Hadoop Core, such as I/O settings that are common to HDFS and MapReduce.
hdfs-site.xml	Hadoop configuration XML	Configuration settings for HDFS daemons: the namenode, the secondary namenode, and the datanodes.
mapred-site.xml	Hadoop configuration XML	Configuration settings for MapReduce daemons: the jobtracker, and the tasktrackers.
masters	Plain text	A list of machines (one per line) that each run a secondary namenode.
slaves	Plain text	A list of machines (one per line) that each run a datanode and a tasktracker.
hadoop-metrics.properties	Java Properties	Properties for controlling how metrics are published in Hadoop (see "Metrics" on page 306).
log4j.properties	Java Properties	Properties for system logfiles, the namenode audit log, and the task log for the tasktracker child process ("Hadoop User Logs" on page 156).

Hadoop Configuration

Table 9-3. Important HDFS daemon properties

Property name	Type	Default value	Description
`fs.default.name`	URI	`file:///`	The default filesystem. The URI defines the hostname and port that the namenode's RPC server runs on. The default port is 8020. This property should be set in *core-site.xml*.
`dfs.name.dir`	comma-separated directory names	`${hadoop.tmp.dir}/dfs/name`	The list of directories where the namenode stores its persistent metadata. The namenode stores a copy of the metadata in each directory in the list.
`dfs.data.dir`	comma-separated directory names	`${hadoop.tmp.dir}/dfs/data`	A list of directories where the datanode stores blocks. Each block is stored in only one of these directories.
`fs.checkpoint.dir`	comma-separated directory names	`${hadoop.tmp.dir}/dfs/namesecondary`	A list of directories where the secondary namenode stores checkpoints. It stores a copy of the checkpoint in each directory in the list.

Hadoop Configuration

Table 9-4. Important MapReduce daemon properties

Property name	Type	Default value	Description
`mapred.job.tracker`	hostname and port	`local`	The hostname and port that the jobtracker's RPC server runs on. If set to the default value of `local`, then the jobtracker is run in-process on demand when you run a MapReduce job (you don't need to start the jobtracker in this case, and in fact you will get an error if you try to start it in this mode).
`mapred.local.dir`	comma-separated directory names	`$ {hadoop.tmp.dir} /mapred/local`	A list of directories where the MapReduce stores intermediate data for jobs. The data is cleared out when the job ends.
`mapred.system.dir`	URI	`$ {hadoop.tmp.dir} /mapred/system`	The directory relative to `fs.default.name` where shared files are stored, during a job run.
`mapred.task tracker.map.tasks. maximum`	`int`	`2`	The number of map tasks that may be run on a tasktracker at any one time.
`mapred.task tracker.reduce.tasks. maximum`	`int`	`2`	The number of reduce tasks that may be run on a tasktracker at any one time.
`mapred.child.java.opts`	`String`	`-Xmx200m`	The JVM options used to launch the tasktracker child process that runs map and reduce tasks. This property can be set on a per-job basis, which can be useful for setting JVM properties for debugging, for example.

Environment Setting

- Discuss about how to set the variables in hadoop-env.sh
- HADOOP_HEAPSIZE property in hadoop-env.sh used to allocate memory (by default, 1000MB)
- apred.tasktracker.map. tasks.maximum property is used to set maximum number of map tasks that will be run on a tasktracker at a time
- mapred.tasktracker.reduce.tasks.maximum property is used to set maximum number of reduce tasks that will be run on a tasktracker at a time
- So, by default, 2,800 MB of memory for a worker machine

Environment Setting

JVM	Default memory used (in MB)
Datanode	1,000
Tasktracker	1,000
Tasktracker child map task	2 × 200
Tasktracker child reduce task	2 × 200

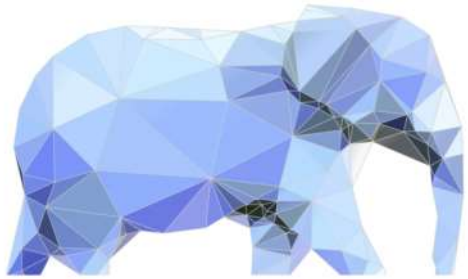
- Number of tasks that can be run simultaneously on a tasktracker is governed by the number of processors available on the machine
- JAVA_HOME property set location of the Java implementation in hadoop-env.sh

Environment Setting

- To run HDFS, we need to designate one machine as a namenode
- fs.default.name is an HDFS filesystem URI, whose host is the namenode's hostname/IP address, and port is the port that the namenode will listen
- Hadoop does not have a single, global location for configuration information

Other Hadoop Properties

- Cluster membership – update dfs.hosts, mapred.hosts and dfs.hosts.exclude
- Buffer size – by default, 4 KB, but for performance benefits, it may be increased to 64/128 KB by setting io.file.buffer.size property in core-site.xml
- HDFS block size -64 MB by default, but many clusters use 128 MB or even 256 MB by setting dfs.block.size property in hdfs-site.xml



Other Hadoop Properties

- Reserved storage space - set dfs.datanode.du.reserved to the amount, in bytes, of space to reserve
- Trash – moves deleted files to Trash directory. Minimum period can set by fs.trash.interval configuration property in core-site.xml

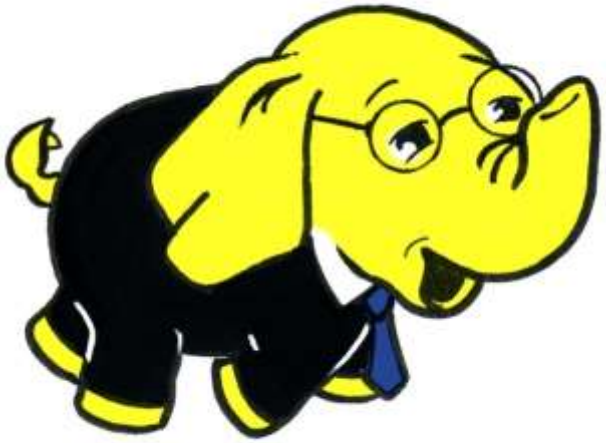
References

Book

- Tom White, “ Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 4th Edition, 2012

Web Resources

- https://www.edureka.co/blog/install-hadoop-single-node-hadoop-cluster
- https://www.tutorialspoint.com/hadoop/hadoop_enviornment_setup.htm
- https://phoenixnap.com/kb/install-hadoop-ubuntu



Thank you