HADOOP AND MAPREDUCE CHEAT SHEET

Hadoop & MapReduce Basics

Hadoop

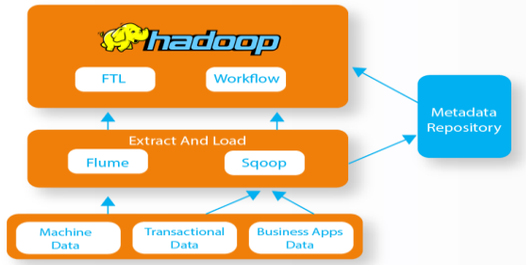
Hadoop is a framework basically designed to handle a large volume of data both structured and unstructured

HDFS

Hadoop Distributed File System is a framework designed to manage huge volumes of data in a simple and pragmatic way. It contains a vast amount of servers and each stores a part of file system

In order to secure Hadoop, configure Hadoop with the following aspects

- Authentication:
  - Define users
  - Enable Kerberos in Hadoop
  - Set-up Knox gateway to control access and authentication to the HDFS cluster
- Authorization:
  - Define groups
  - Define HDFS permissions
  - Define HDFS ACL's
- Audit:
  Enable process execution audit trail
- Data protection:
  Enable wire encryption with Hadoop



Hadoop HDFS List File Commands	Tasks
hdfs dfs –ls /	Lists all the files and directories given for the hdfs destination path
hdfs dfs –ls –d /hadoop	This command lists all the details of the Hadoop files
hdfs dfs –ls –R /hadoop	Recursively lists all the files in the Hadoop directory and al sub directories in Hadoop directory
hdfs dfs –ls hadoop/ dat*	This command lists all the files in the Hadoop directory starting with 'dat'

Hdfs basic commands	Tasks
hdfs dfs -put logs.csv /data/	This command is used to upload the files from local file system to HDFS
hdfs dfs -cat /data/logs.csv	This command is used to read the content from the file
hdfs dfs -chmod 744 /data/logs.csv	This command is used to change the permission of the files
hdfs dfs -chmod –R 744 /data/logs.csv	This command is used to change the permission of the files recursively
hdfs dfs -setrep -w 5 /data/logs.csv	This command is used to set the replication factor to 5
hdfs dfs -du -h /data/logs.csv	This command is used to check the size of the file
hdfs dfs -mv logs.csv logs/	This command is used to move the files to a newly created subdirectory
hdfs dfs -rm -r logs	This command is used to remove the directories from Hdfs
stop-all.sh	This command is used to stop the cluster
start-all.sh	This command is used to start the cluster
Hadoop version	This command is used to check the version of Hadoop
hdfs fsck/	This command is used to check the health of the files
Hdfs dfsadmin –safemode leave	This command is used to turn off the safemode of namenode
Hdfs namenode -format	This command is used to format the NameNode
hadoop [--config confdir]archive -archiveName NAME -p	This command is used to create a Hadoop archieve
hadoop fs [generic options] -touchz <path> ...	This is used to create an empty files in a hdfs directory
hdfs dfs [generic options] -getmerge [-nl] <src> <localdst>	This is used to concatenate all files in a directory into one file
hdfs dfs -chown -R admin:hadoop /new-dir	This is used to change the owner of the group

Commands	Tasks
yarn	This command shows the yarn help
yarn [--config confdir]	This command is used to define configuration file
yarn [--loglevel loglevel]	This can be used to define the log level, which can be fatal, error, warn, info, debug or trace
yarn classpath	This is used to show the Hadoop classpath
yarn application	This is used to show and kill the Hadoop applications
yarn applicationattempt	This shows the application attempt
yarn container	This command shows the container information
yarn node	This shows the node information
yarn queue	This shows the queue information

MapReduce

MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of systems referred as clusters. Basically, it is a processing technique and program model for distributed computing based on Java

Mahout

Apache **Mahout** is an open source algebraic framework used for data mining which works along with the distributed environments with simple programming languages

Components of MapReduce

PayLoad: The applications implement Map and Reduce functions and form the core of the job

MRUnit: Unit test framework for **MapReduce**

Mapper: Mapper maps the input key/value pairs to the set of intermediate key/value pairs

NameNode: Node that manages the **HDFS** is known as **namednode**

DataNode: Node where the data is presented before processing takes place

MasterNode: Node where the **jobtrackers** runs and accept the job request from the clients

SlaveNode: Node where the Map and Reduce program runs

JobTracker: Schedules jobs and tracks the assigned jobs to the task tracker

TaskTracker: Tracks the task and updates the status to the **job tracker**

Job: A program which is an execution of a Mapper and Reducer across a dataset

Task: An execution of Mapper and Reducer on a piece of data

Task Attempt: A particular instance of an attempt to execute a task on a **SlaveNode**

Commands used to interact with MapReduce	
Commands	**Tasks**
hadoop job -submit <job-file>	used to submit the Jobs created
hadoop job -status <job-id>	shows map & reduce completion status and all job counters
hadoop job -counter <job-id> <group-name><countername>	prints the counter value
hadoop job -kill <job-id>	This command kills the job
hadoop job -events <job-id> <fromevent-#> <#-of-events>	shows the event details received by the job tracker for given range
hadoop job -history [all] <jobOutputDir>	Prints the job details, and killed and failed tip details
hadoop job -list[all]	This command is used to display all the jobs
hadoop job -kill-task <task-id>	This command is used to kill the tasks
hadoop job -fail-task <task-id>	This command is used to fail the task
hadoop job -set-priority <job-id> <priority>	Changes and sets the priority of the job
HADOOP_HOME/bin/hadoop job -kill <JOB-ID>	This command kills the job created
HADOOP_HOME/bin/hadoop job -history <DIR-NAME>	This is used to show the history of the jobs

Important commands used in MapReduce	
Usage: mapred [Generic commands] <parameters>	
Parameters	**Tasks**
-input directory/file-name	Shows Inputs the location for mapper
-output directory-name	Shows output location for the mapper
-mapper executable or script or JavaClassName	Used for Mapper executable
-reducer executable or script or JavaClassName	Used for reducer executable
-file file-name	Makes the mapper, reducer, combiner executable available locally on the computing nodes
-numReduceTasks	This is used to specify number of reducers
-mapdebug	Script to call when the map task fails
-reducedebug	Script to call when the reduce task fails