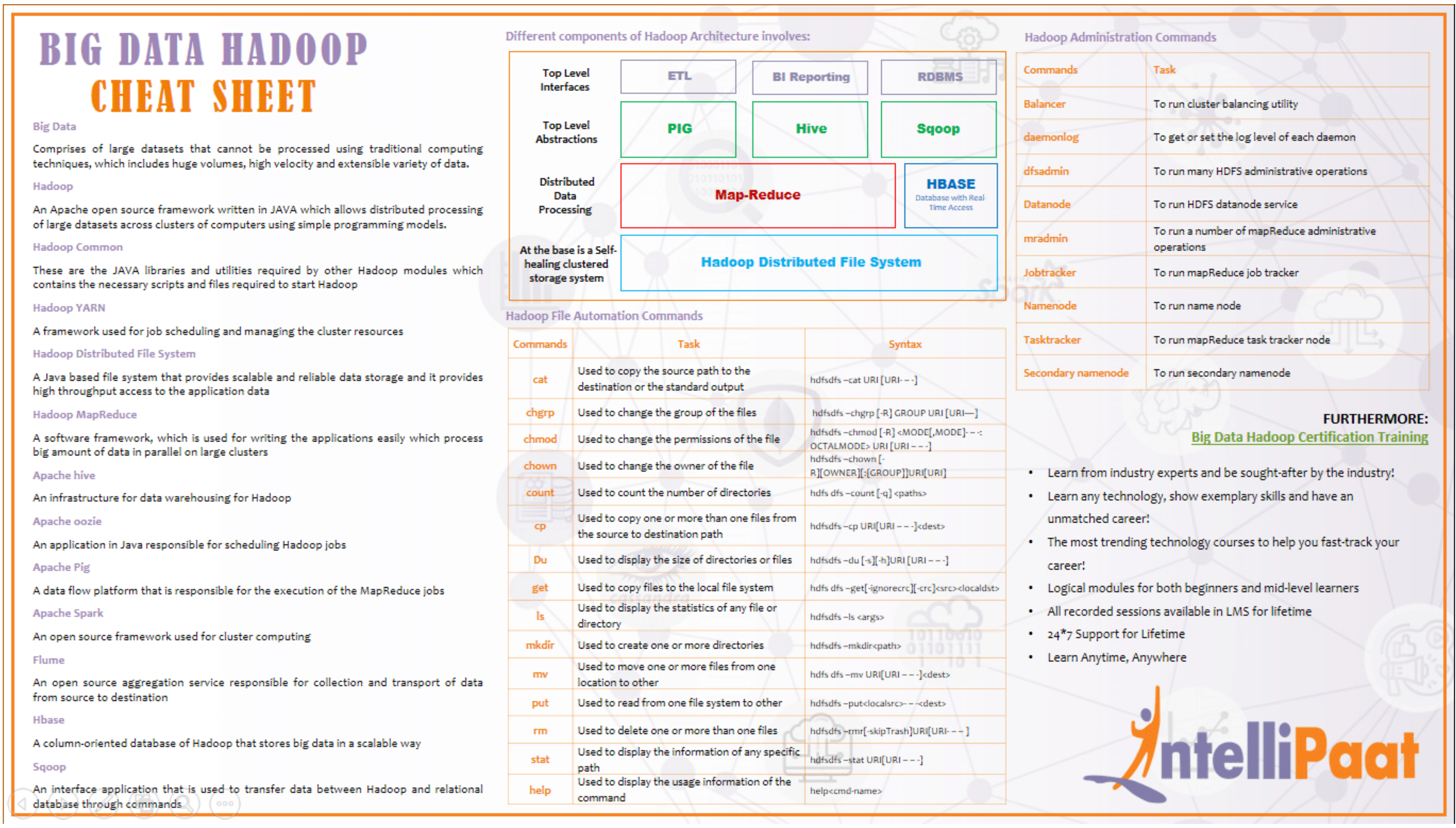
Big Data Hadoop Cheat Sheet:

In the last decade mankind has seen a pervasive amount of growth in data. Then we started looking for ways to put these data in use. Analyzing and Learning from these data has opened many doors of opportunities. That is how Big Data became a buzzword in IT industry. Then we are introduced to different technologies and platforms to learn from these enormous amounts of data collected from all kinds of sources. Now comes the question, ***"How do we process Big Data?"***. Apache Hadoop has filled up the gap, also it has become one of the hottest open source software.

Big Data cheat sheet will guide you through the basics of the Hadoop and important commands which will be helpful for new learners as well as for those who wants to take a quick look at the important topics of Big Data Hadoop.



BIG DATA HADOOP
CHEAT SHEET

Big Data

Comprises of large datasets that cannot be processed using traditional computing techniques, which includes huge volumes, high velocity and extensible variety of data.

Hadoop

An Apache open source framework written in JAVA which allows distributed processing of large datasets across clusters of computers using simple programming models.

Hadoop Common

These are the JAVA libraries and utilities required by other Hadoop modules which contains the necessary scripts and files required to start Hadoop

Hadoop YARN

A framework used for job scheduling and managing the cluster resources

Hadoop Distributed File System

A Java based file system that provides scalable and reliable data storage and it provides high throughput access to the application data

Hadoop MapReduce

A software framework, which is used for writing the applications easily which process big amount of data in parallel on large clusters

Apache hive

An infrastructure for data warehousing for Hadoop

Apache oozie

An application in Java responsible for scheduling Hadoop jobs

Apache Pig

A data flow platform that is responsible for the execution of the MapReduce jobs

Apache Spark

An open source framework used for cluster computing

Flume

An open source aggregation service responsible for collection and transport of data from source to destination

Hbase

A column-oriented database of Hadoop that stores big data in a scalable way

Sqoop

An interface application that is used to transfer data between Hadoop and relational database through commands

Different components of Hadoop Architecture involves:

Top Level Interfaces	ETL	BI Reporting	RDBMS
Top Level Abstractions	PIG	Hive	Sqoop
Distributed Data Processing	Map-Reduce		HBASE Database with Real Time Access
At the base is a Self-healing clustered storage system	Hadoop Distributed File System		

Hadoop File Automation Commands

Commands	Task	Syntax
cat	Used to copy the source path to the destination or the standard output	hdfsdfs –cat URI [URI- – -]
chgrp	Used to change the group of the files	hdfsdfs –chgrp [-R] GROUP URI [URI—]
chmod	Used to change the permissions of the file	hdfsdfs –chmod [-R] <MODE[,MODE]- – -: OCTALMODE> URI [URI – – -]
chown	Used to change the owner of the file	hdfsdfs –chown [-R][OWNER][:{GROUP]]URI[URI]
count	Used to count the number of directories	hdfs dfs –count [-q] <paths>
cp	Used to copy one or more than one files from the source to destination path	hdfsdfs –cp URI[URI – – -]<dest>
Du	Used to display the size of directories or files	hdfsdfs –du [-s][-h]URI [URI – – -]
get	Used to copy files to the local file system	hdfs dfs –get[-ignorecrc][-crc]<src><localdst>
ls	Used to display the statistics of any file or directory	hdfsdfs –ls <args>
mkdir	Used to create one or more directories	hdfsdfs –mkdir<path>
mv	Used to move one or more files from one location to other	hdfs dfs –mv URI[URI – – -]<dest>
put	Used to read from one file system to other	hdfsdfs –put<localsrc>- – -<dest>
rm	Used to delete one or more than one files	hdfsdfs –rmr[-skipTrash]URI[URI- – –]
stat	Used to display the information of any specific path	hdfsdfs –stat URI[URI – – -]
help	Used to display the usage information of the command	help<cmd-name>

Hadoop Administration Commands

Commands	Task
Balancer	To run cluster balancing utility
daemonlog	To get or set the log level of each daemon
dfsadmin	To run many HDFS administrative operations
Datanode	To run HDFS datanode service
mradmin	To run a number of mapReduce administrative operations
Jobtracker	To run mapReduce job tracker
Namenode	To run name node
Tasktracker	To run mapReduce task tracker node
Secondary namenode	To run secondary namenode

FURTHERMORE:
Big Data Hadoop Certification Training

- Learn from industry experts and be sought-after by the industry!
- Learn any technology, show exemplary skills and have an unmatched career!
- The most trending technology courses to help you fast-track your career!
- Logical modules for both beginners and mid-level learners
- All recorded sessions available in LMS for lifetime
- 24*7 Support for Lifetime
- Learn Anytime, Anywhere

IntelliPaat

Big Data: Big data comprises of large datasets that cannot be processed using traditional computing techniques, which includes huge volumes, high velocity and extensible variety of data.

Hadoop: Hadoop is an Apache open source framework written in JAVA which allows distributed processing of large datasets across clusters of computers using simple programming models.

Hadoop Common: These are the JAVA libraries and utilities required by other Hadoop modules which contains the necessary scripts and files required to start Hadoop

Hadoop YARN: Yarn is a framework used for job scheduling and managing the cluster resources

Hadoop Distributed File System: HDFS is a Java based file system that provides scalable and reliable data storage and it provides high throughput access to the application data

Hadoop MapReduce: It is a software framework, which is used for writing the applications easily which process big amount of data in parallel on large clusters

Apache hive: It is an infrastructure for data warehousing for Hadoop

Apache oozie: It is an application in Java responsible for scheduling Hadoop jobs

Apache Pig: It is a data flow platform that is responsible for the execution of the MapReduce jobs

Apache Spark: It is an open source framework used for cluster computing

Flume: Flume is an open source aggression service responsible for collection and transport of data from source to destination

Hbase: Apache Hbase is a column-oriented database of Hadoop that stores big data in a scalable way

Sqoop: Scoop is an interface application that is used to transfer data between Hadoop and relational database through commands

Hadoop Ecosystem:

Hadoop Ecosystem represents various components of the Apache software. Typically, it can be divided into the following categories.

- Top Level Interface
- Top Level Abstraction
- Distributed Data Processing

- Self Healing Clustered Storage System

Hadoop file automation commands:

Cat: Cat command is used to copy the source path to the destination or the standard output

```
hdfsdfs –cat URI [URI- – -]
```

chgrp: This command is used to change the group of the files.

```
hdfsdfs –chgrp [-R] GROUP URI [URI—]
```

chmod: This command is used to change the permissions of the file.

```
hdfsdfs –chmod [-R] <MODE[,MODE]- – -: OCTALMODE> URI [URI – – -]
```

chown: This command is used to change the owner of the file

```
hdfsdfs –chown [-R][OWNER][:{GROUP]]URI[URI]
```

cp: This command can be used to copy one or more than one files from the source to destination path

```
hdfsdfs –count [-q] <paths>
```

Du: It is used to display the size of directories or files

```
hdfsdfs –cpURI[URI – – -]<dest>
```

get: This command can be used to copy files to the local file system

```
hdfsdfs –get[-ignorecrc][-crc]<src><localdst>
```

ls: It is used to display the statistics of any file or directory

```
hdfsdfs –ls <args>
```

mkdir: This command is used to create one or more directories

```
hdfsdfs –mkdir<path>
```

mv: It is used to move one or more files from one location to other

```
hdfsdfs –mv URI[URI – – -]<dest>
```

put: This command is used to read from one file system to other

```
hdfsdfs –put<localsrc>- – -<dest>
```

rm: This command is used to delete one or more than one files

```
hdfsdfs –rmr[-skipTrash]URI[URI- – – ]
```

stat: It is used to display the information of any specific path

```
hdfsdfs –stat URI[URI – – -]
```

help: It is used to display the usage information of the command

```
help<cmd-name>
```

Hadoop Administration commands:

The commands which can be used only by the Hadoop Administrators are mentioned below with the operations performed by them.

Balancer: To run cluster balancing utility
Daemonlog: To get or set the log level of each daemon
Dfsadmin: To run many HDFS administrative operations
Datanode: To run HDFS datanode service
mradmin: To run a number of mapReduce administrative operations
Jobtracker: To run mapReduce job tracker
Namenode: To run name node
Tasktracker: To run mapReduce task tracker node
Secondary namenode: To run secondary namenode