Session #3

Hadoop In Action

Enough taking about Big data and Hadoop and let's see how Hadoop works in action.
We will locate a real dataset, ingest it to our cluster, connect it to a database, apply some queries and data transformations on it , save our result and show it via BI tool.

Hadoop:

- Hadoop quick definition.
- Why Hadoop?
- Hadoop ecosystem.
- Tools to be used.

Practical part:

- What's the current setup?
- Ambari look.
- Current installed systems.
- Use case high-level description.
- Steps to develop the use case?

Use case:

- Locating the data.
- Ingest the data into the HDFS
- See how the files got created in HDFS
- Feed other data from DB.
- Data querying via Hive and MapReduce
- Hive table creation.
- Running transudation job via Pig.
- Check the Hive metastore.
- Connect BI to Hadoop.
- Sqoop basic commands
- End to End look solution.

presented by
Mahmoud Yassin

When

Thursday 30-03-2017
08:00 PM –10:00 PM

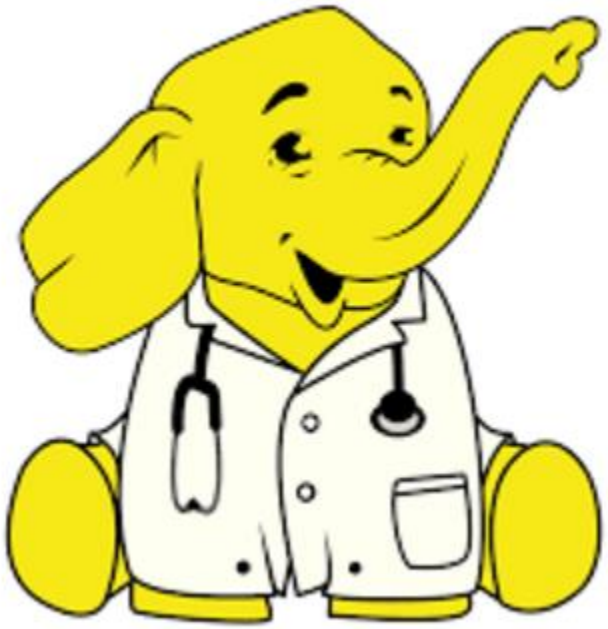
Where

Online via
WebEx

@Techie_bits

slack Big Data-KSA

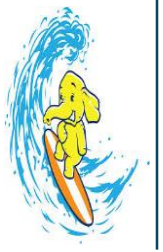
#Big_Data



Hadoop Hands On session

By Mahmoud Yassin

Agenda:



Hadoop:

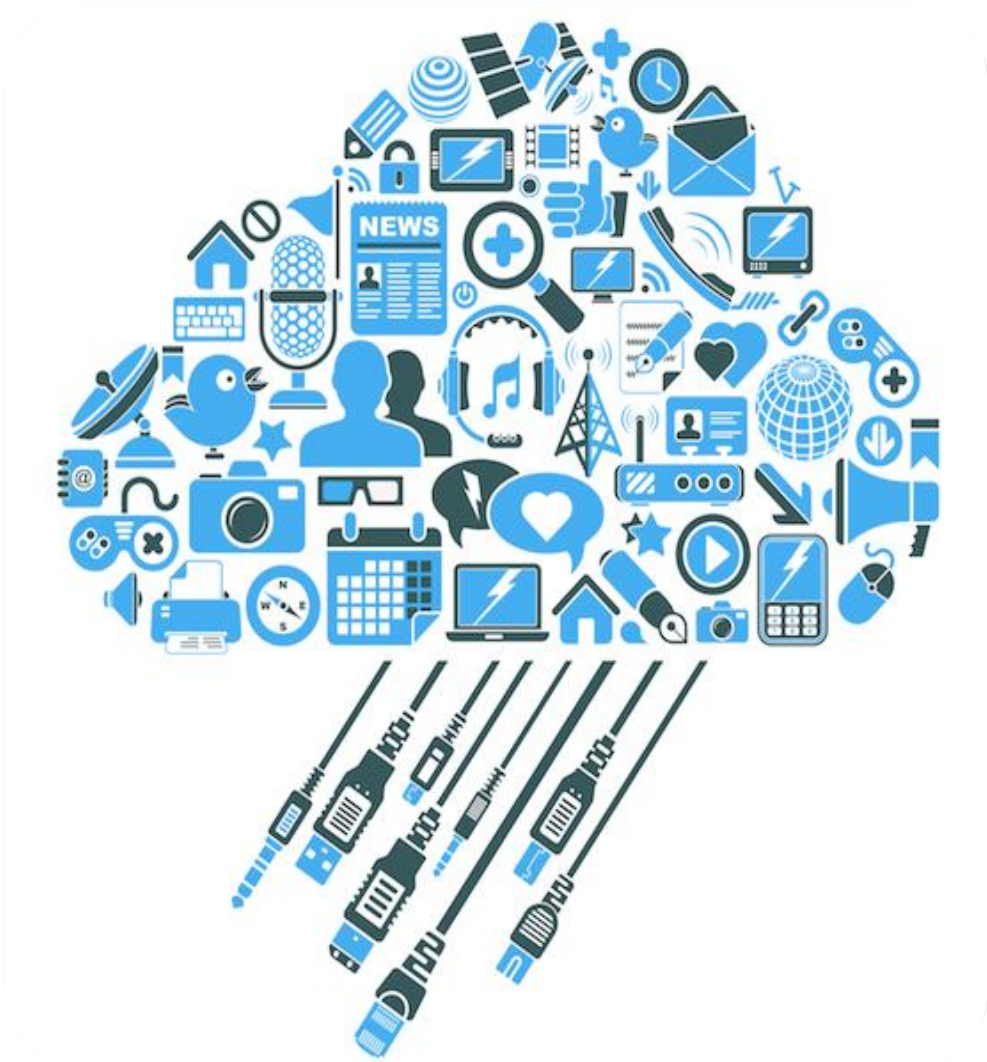
- Hadoop quick definition.
- Why Hadoop?
- Hadoop ecosystem.
- Tools to be used.

Practical part:

- What's the current setup?
- Ambari look.
- Current installed systems.
- Use case high-level description.
- Steps to develop the use case?

Use case:

- Locating the data.
- Ingest the data into the HDFS
- See how the files got created in HDFS
- Feed other data from DB.
- Data querying via Hive and MapReduce
- Hive table creation.
- Running transudation job via Pig.
- Check the Hive metastore.
- Connect BI to Hadoop.

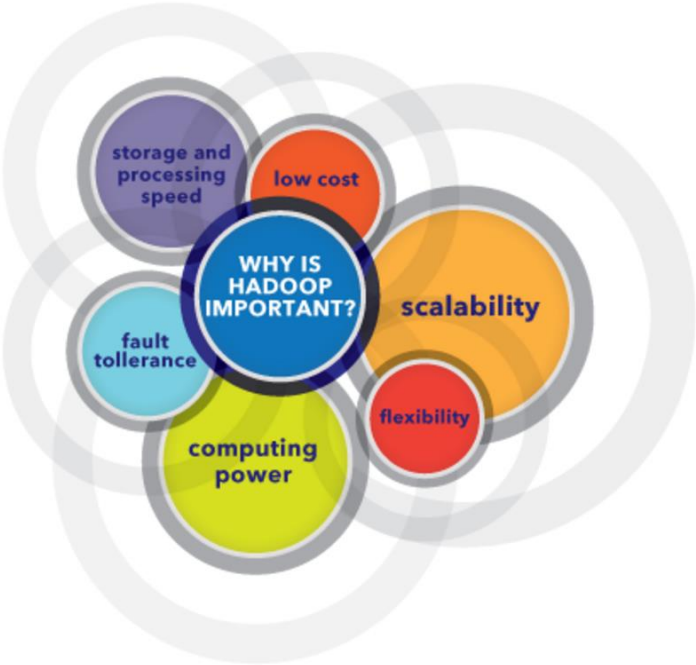


What is Hadoop

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

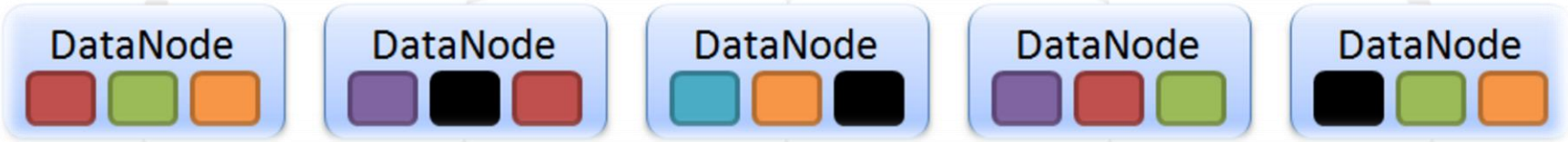
Why Hadoop is important ?



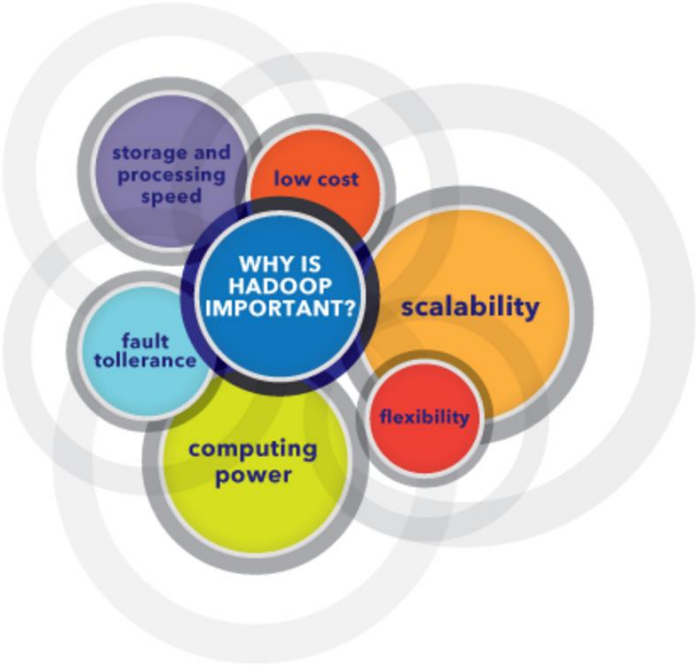
Ability to store and process huge amounts of any kind of data, quickly.
With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.

Computing power. Hadoop's distributed computing model processes big data fast. The more computing nodes you use, the more processing power you have.

Fault tolerance. Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.



Why Hadoop is important ?



Flexibility. Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.

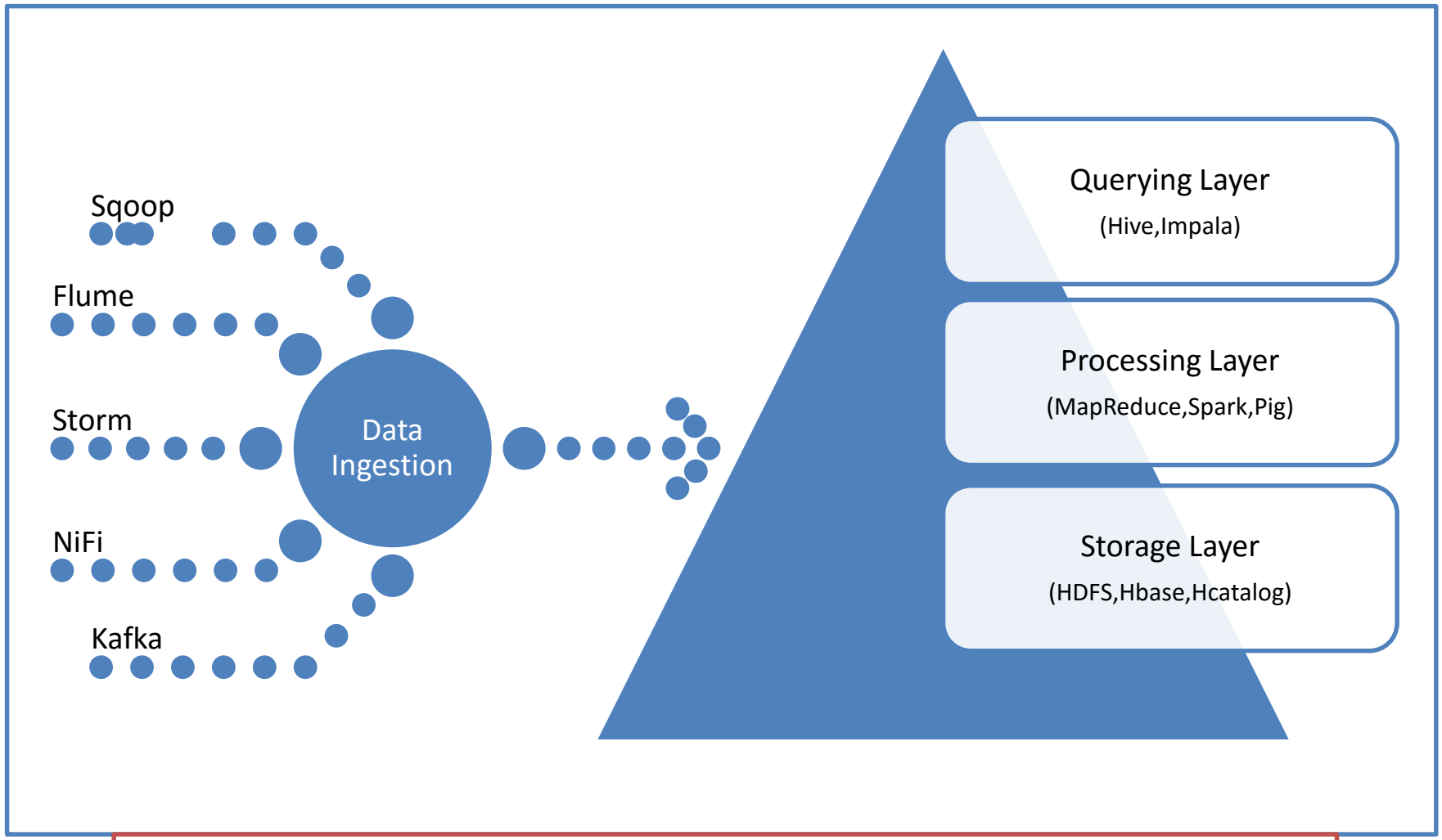
Low cost. The open-source framework is free and uses commodity hardware to store large quantities of data.

Scalability. You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

Scalability

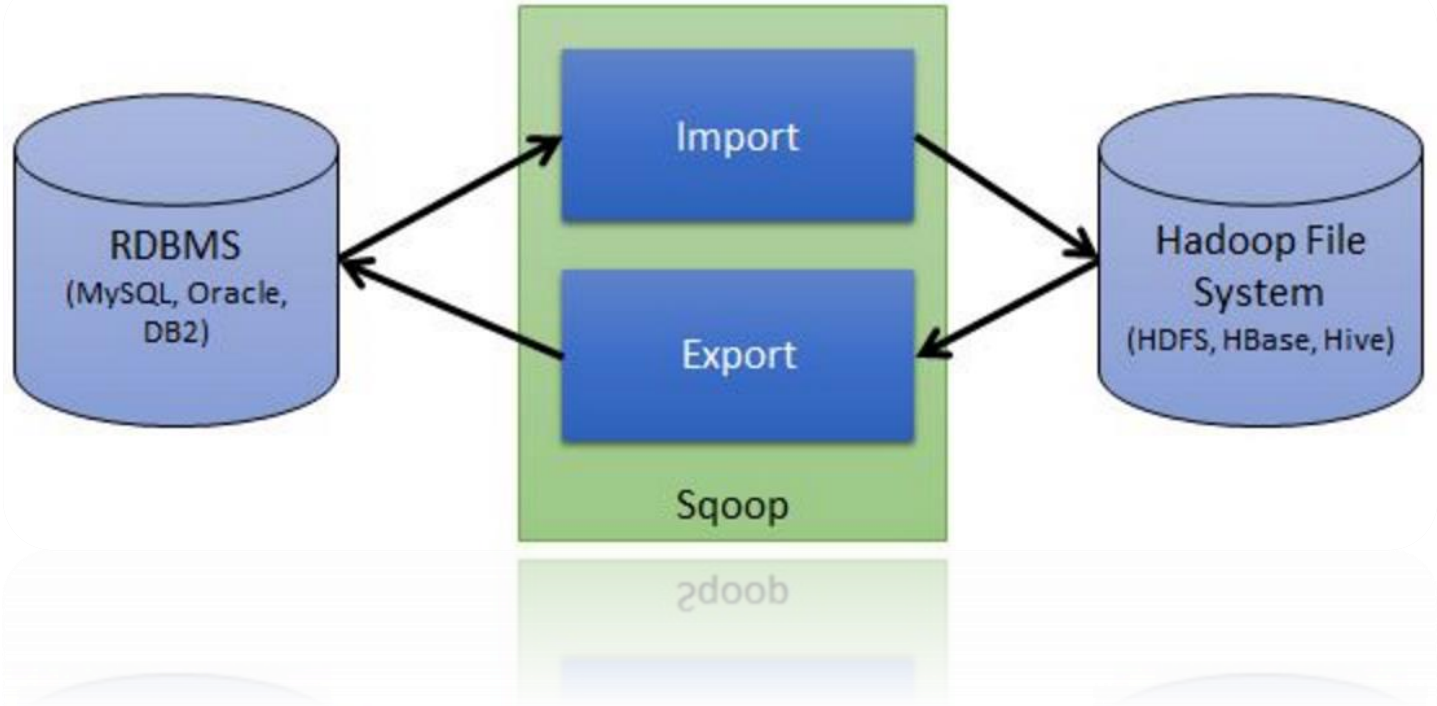
- Horizontal scaling means that you scale by adding more machines into your pool of resources
- Vertical scaling means that you scale by adding more power (CPU, RAM) to an existing machine

Hadoop ecosystem

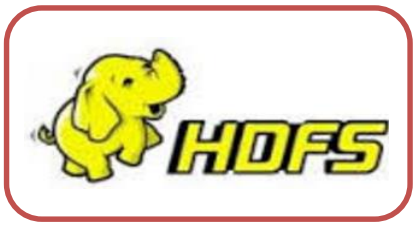


Hadoop | Data Ingestion

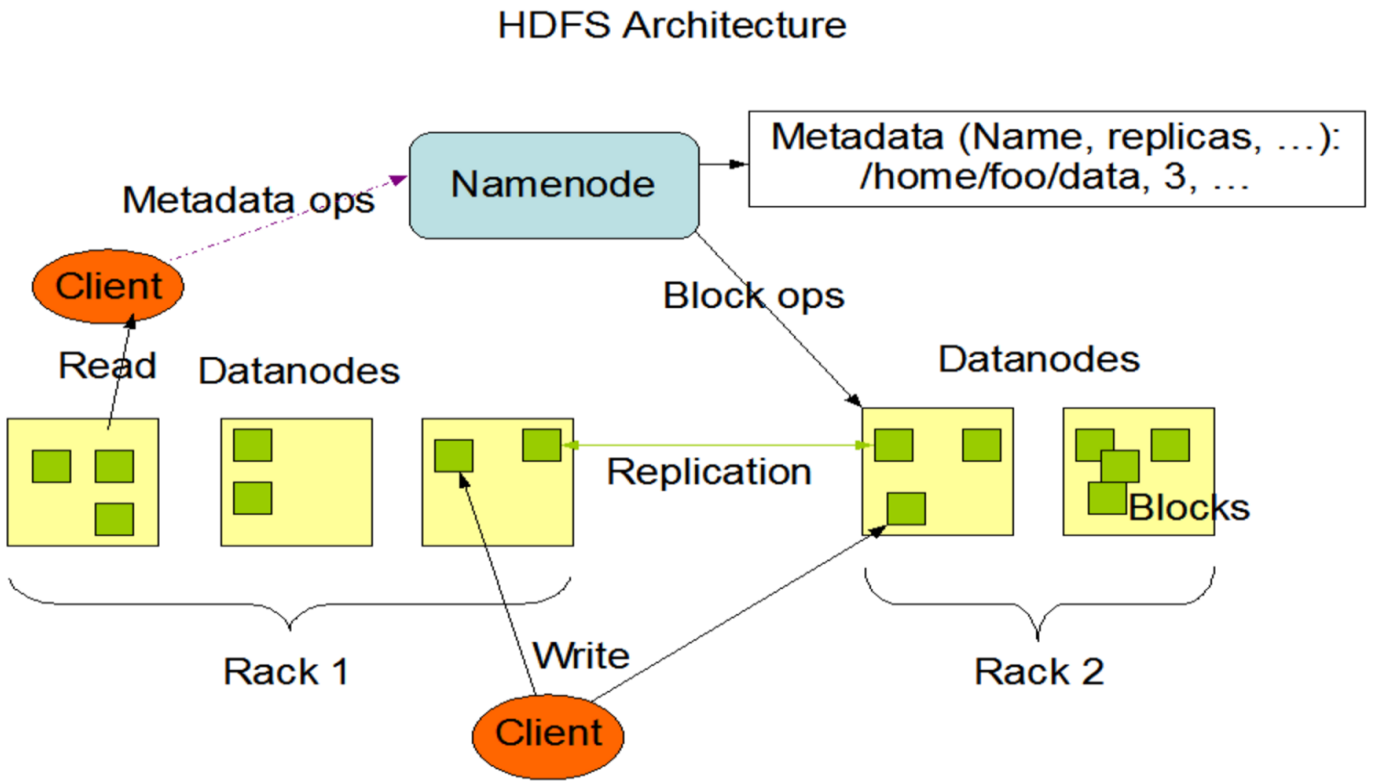
Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.



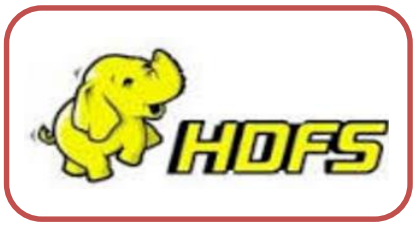
Hadoop | Data Storage Layer

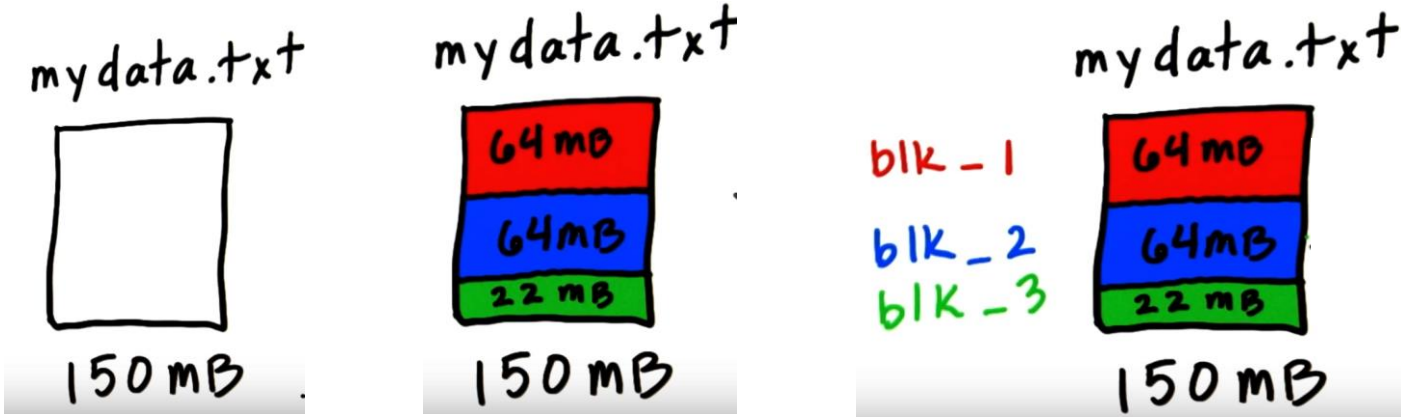


Hadoop Distributed File System (HDFS) offers a way to store large files across multiple machines. Hadoop and HDFS was derived from Google File System (GFS) paper.



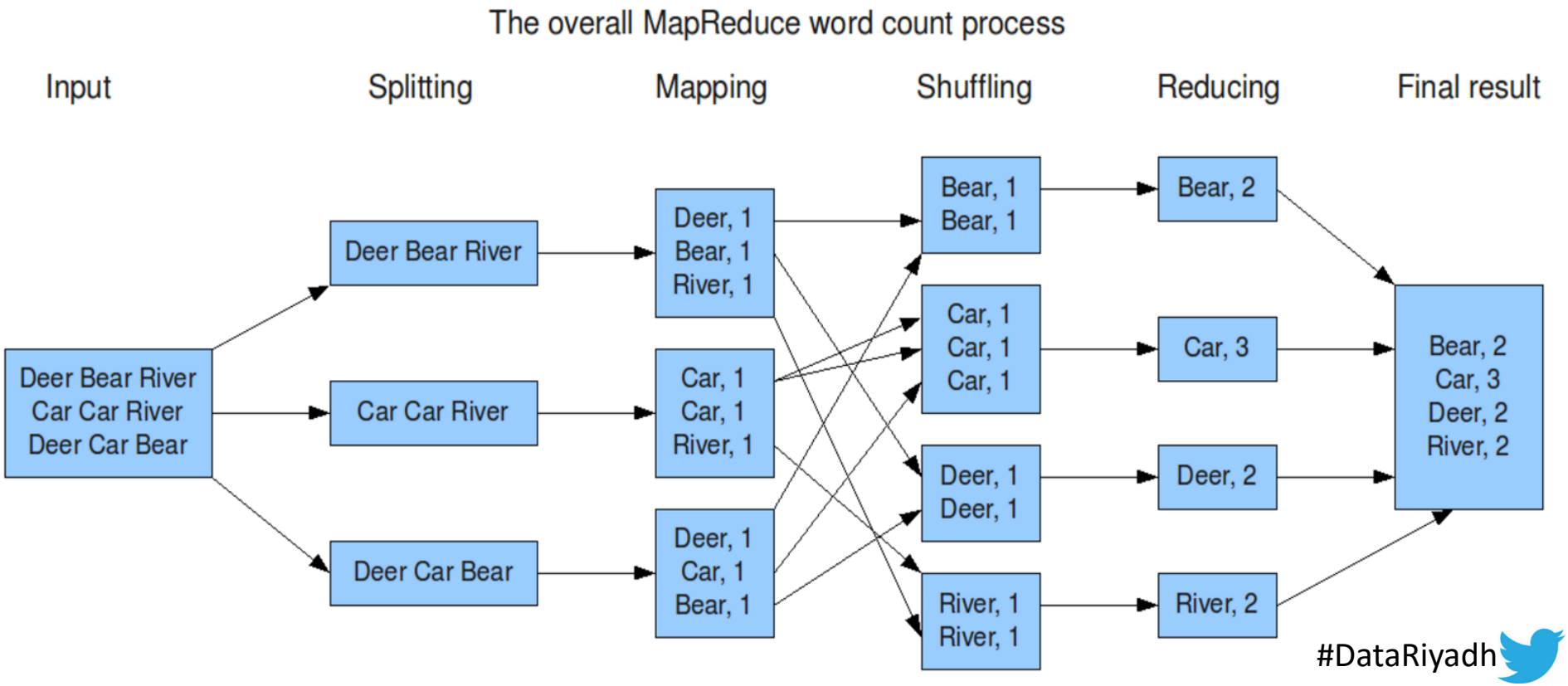
Hadoop | Data Storage Layer





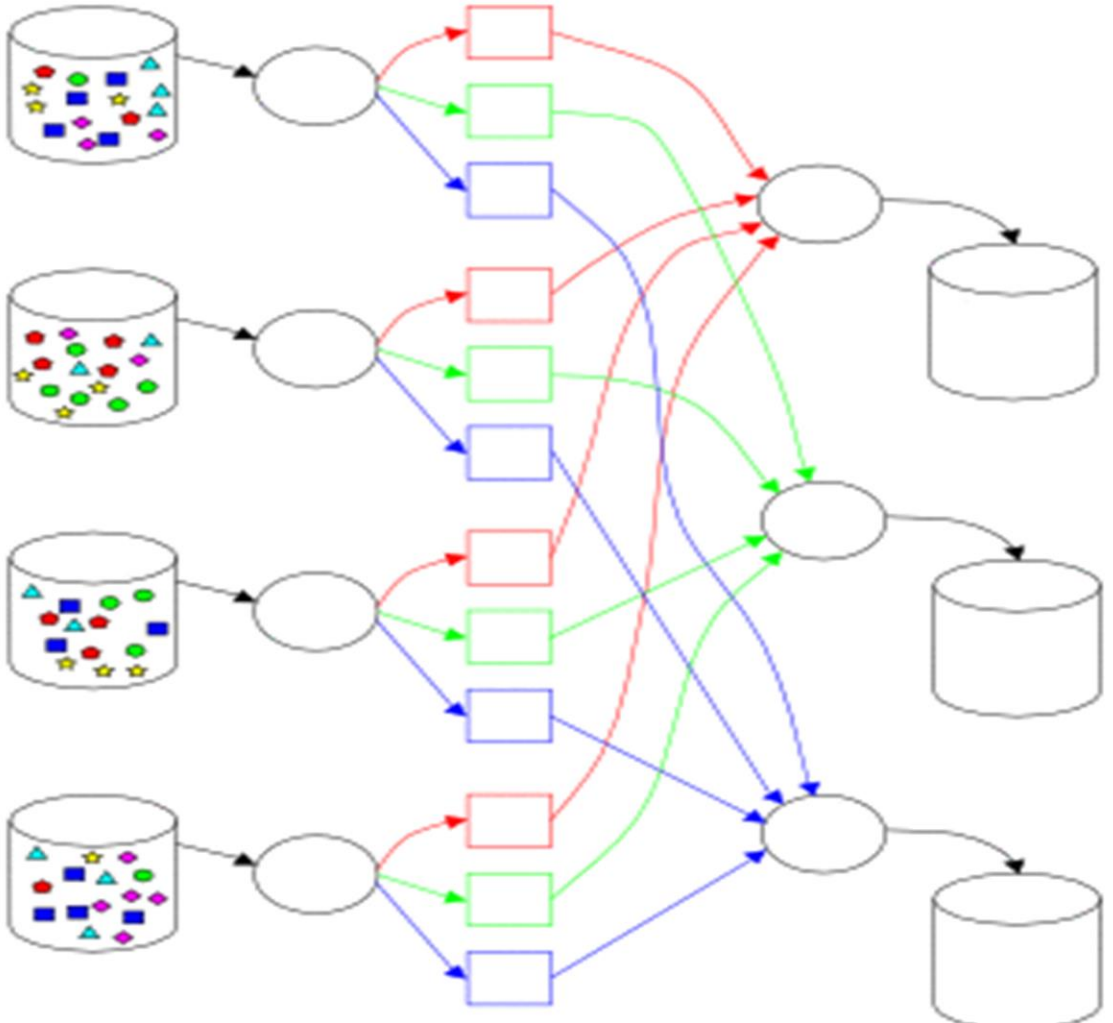
Hadoop | Data Processing Layer

MapReduce is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster with a parallel, distributed algorithm.



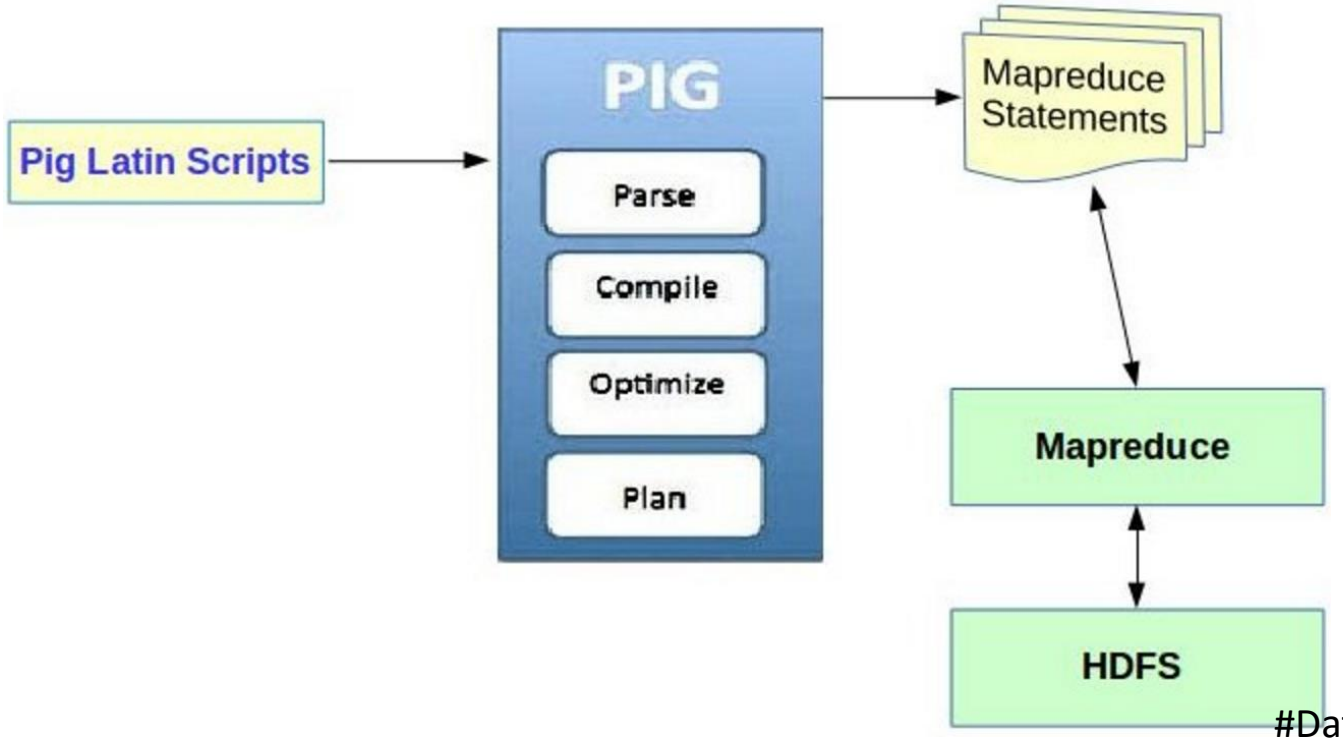
#DataRiyadh

Hadoop | Data Processing Layer



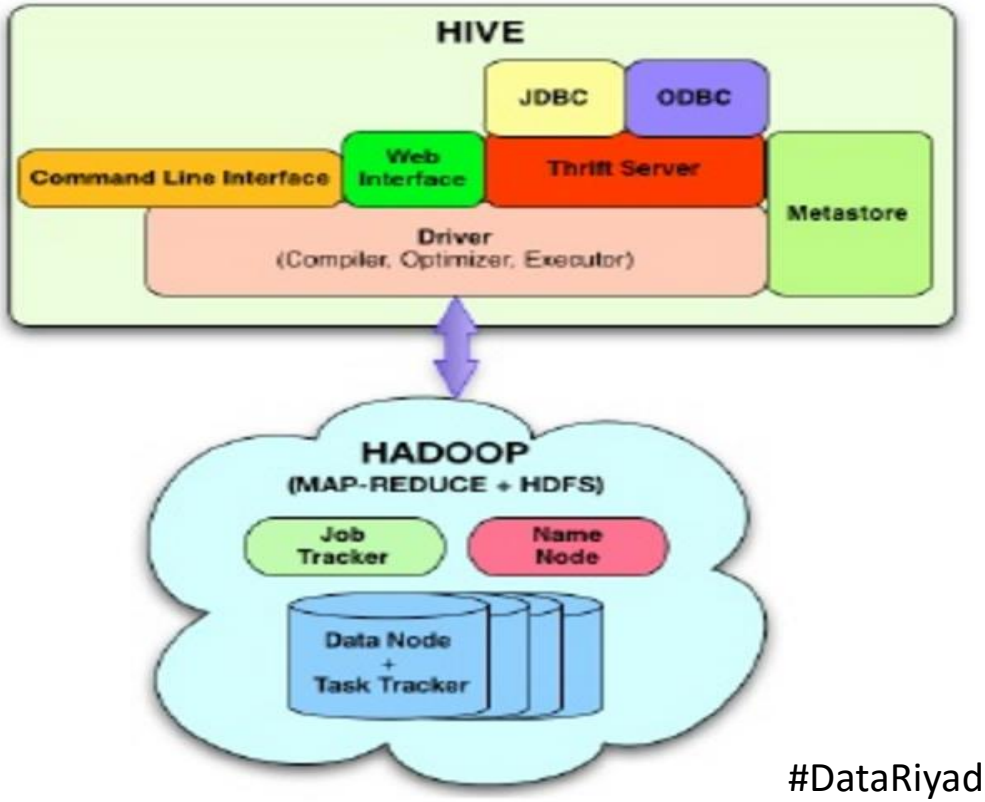
Hadoop | Data Processing Layer

A scripting SQL based language and execution environment for creating complex MapReduce transformations. Functions are written in Pig Latin (the language) and translated into executable MapReduce jobs. Pig also allows the user to create extended functions (UDFs) using Java.



Hadoop | Data Querying Layer

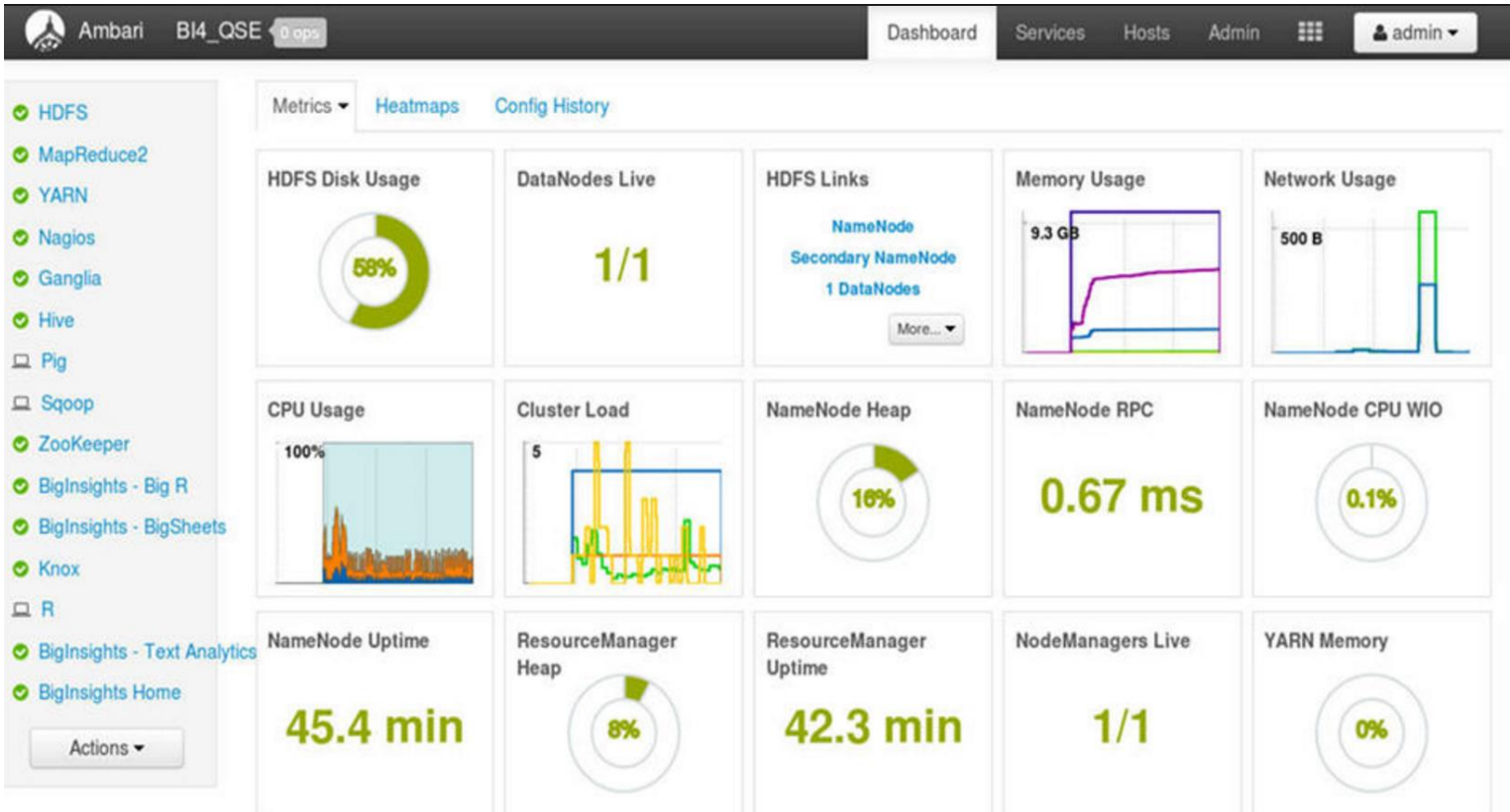
A distributed data warehouse built on top of HDFS to manage and organize large amounts of data. Hive provides a query language based on SQL semantic (HiveQL) which is translated by the runtime engine to MapReduce jobs for querying the data.



Hadoop | Management Layer

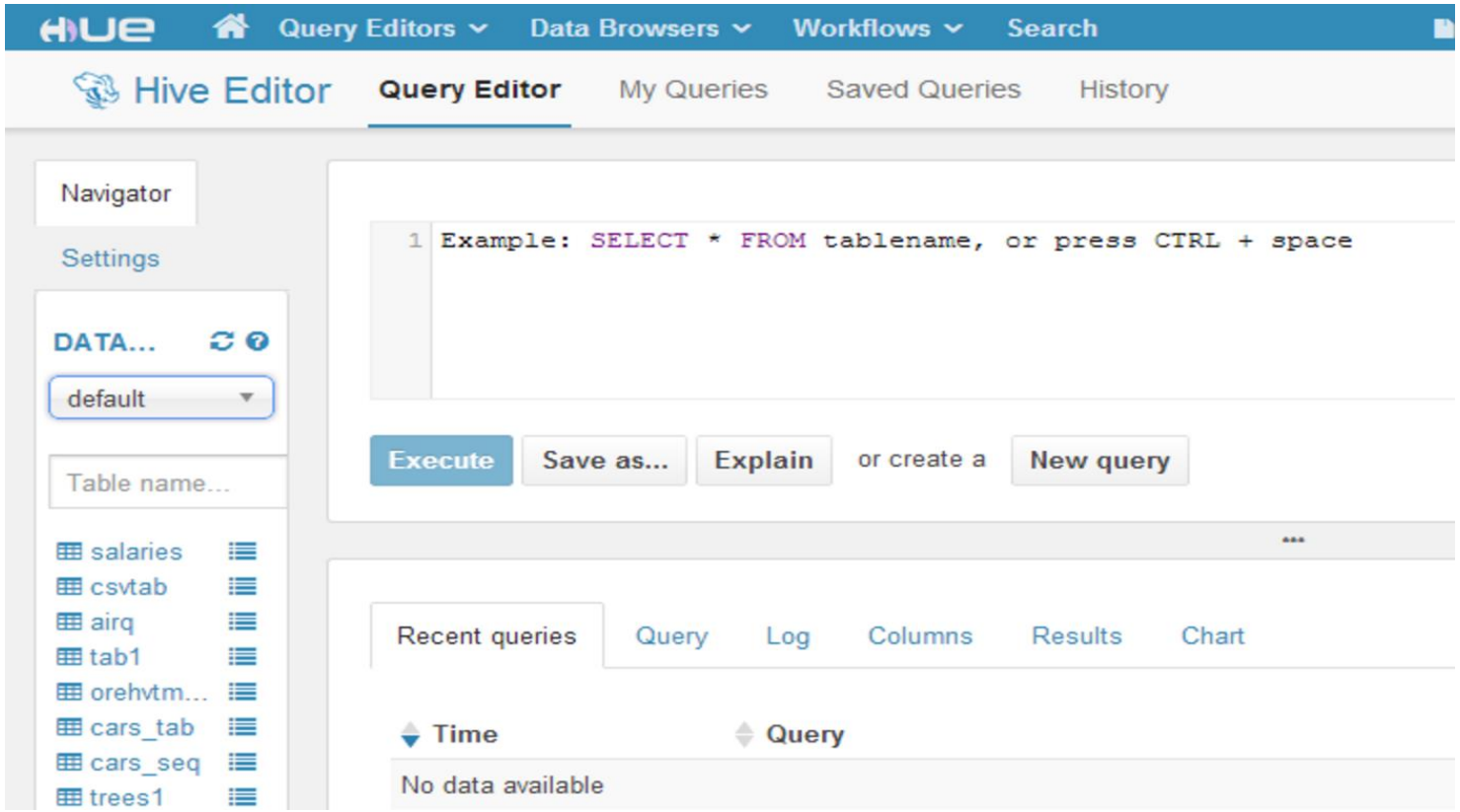
Apache Ambari

intuitive, easy-to-use Hadoop management web UI. Apache Ambari was donated by Hortonworks team. It's a powerful and nice interface for Hadoop and other typical applications from the Hadoop ecosystem.



Hadoop | Management Layer

is an open-source Web interface that supports Apache Hadoop and its ecosystem, licensed under the Apache v2 license



Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Sqoop
Data Exchange

Flume
Log Collector

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

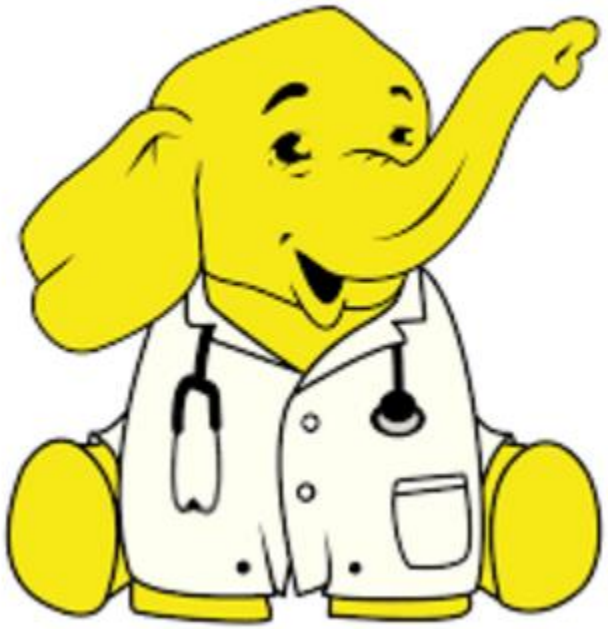
R Connectors
Statistics

Hive
SQL Query

Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

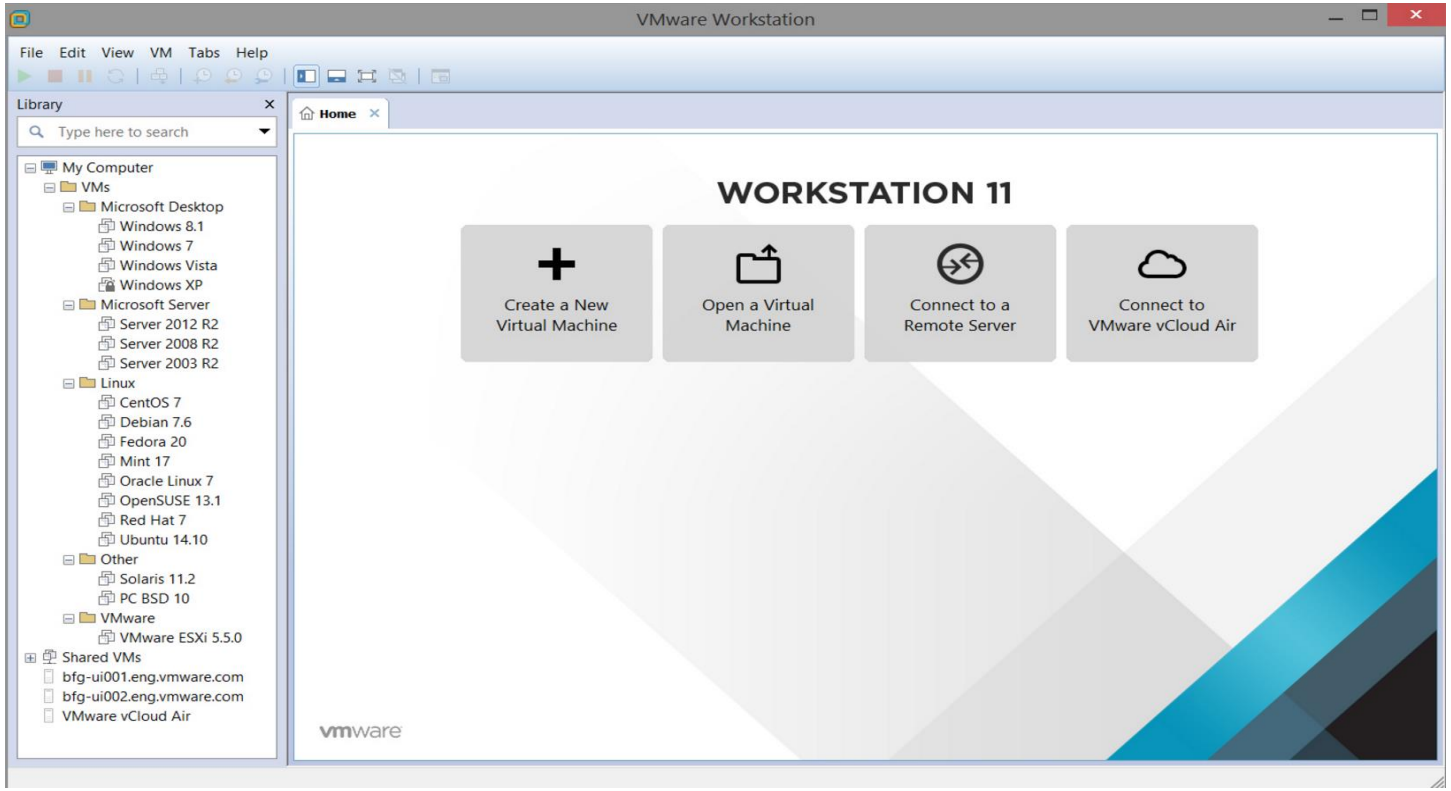
HDFS
Hadoop Distributed File System



Current Setup

Current Setup

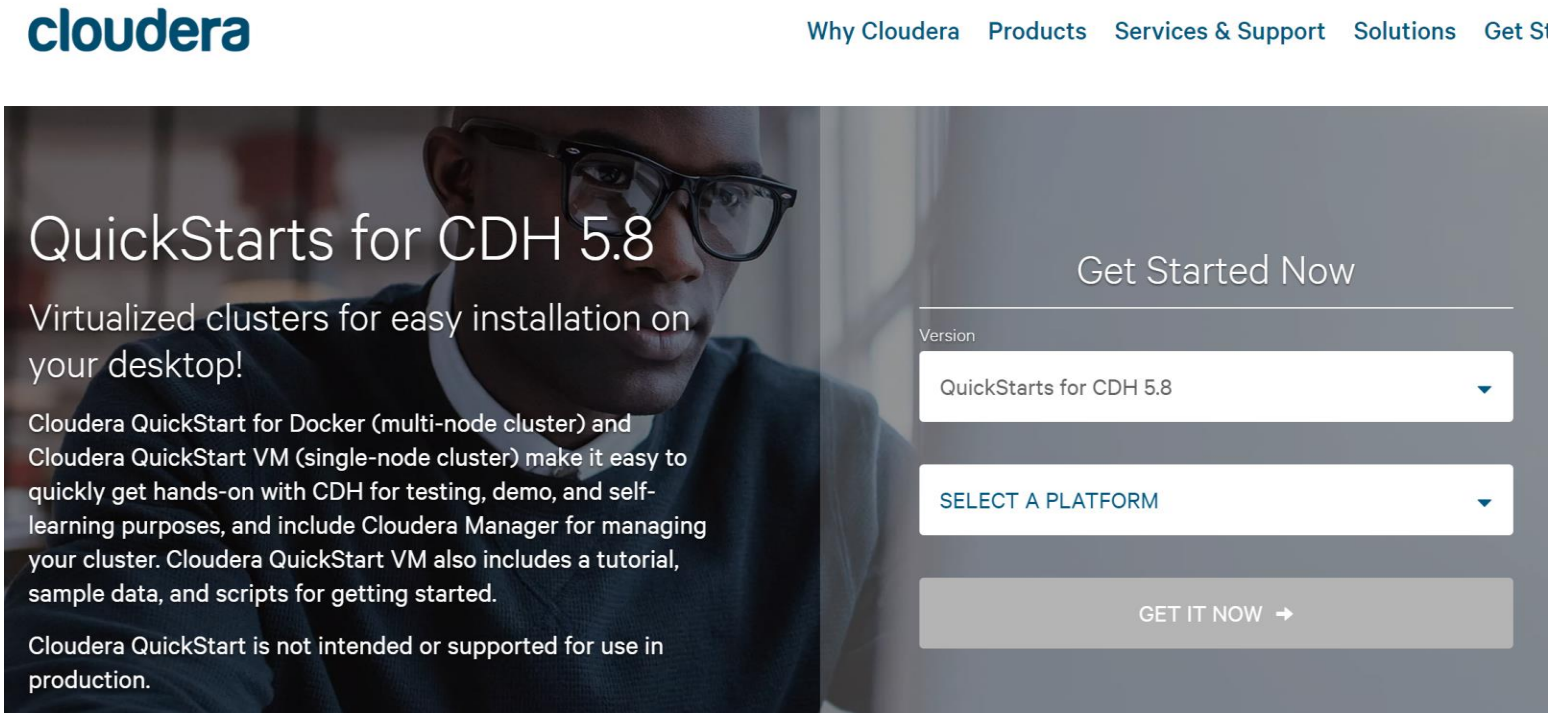
is a subsidiary of Dell Technologies, that provides cloud and virtualization software and services.



http://www.vmware.com/

Current Setup

The VM make it easy to quickly get hands-on with CDH for testing, demo, and self-learning purposes, and include Cloudera Manager for managing your cluster. Cloudera QuickStart VM also includes a tutorial, sample data, and scripts for getting started.

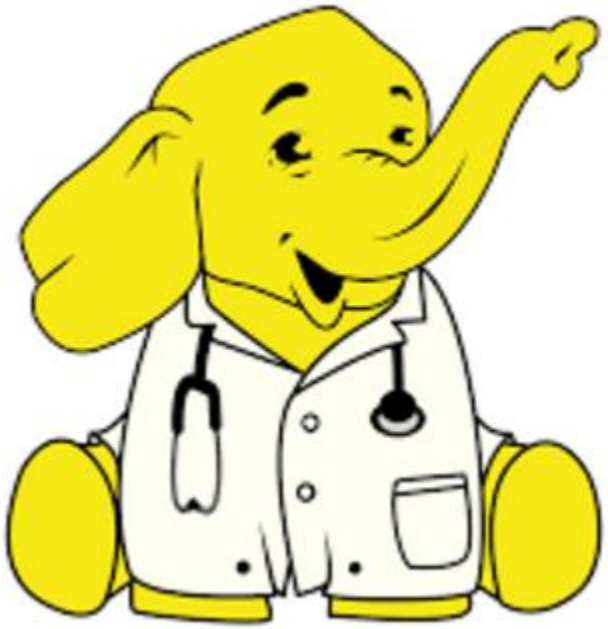


http://www.cloudera.com/downloads/quickstart_vms/5-8.html

Inside the VM:

Our RDBMS

Hadoop Storage



Use Case

The case:

Data Sources

File
Video

MySQL

RDBMS

A platform for manipulating data stored in HDFS via a high-level language called Pig Latin. It does data extractions, transformations and loading, and basic analysis in patch mode

A data warehousing and SQL-like query language that presents data in the form of tables. Hive programming is similar to database programming.

cloudera

open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop.

+ableau

Connect BI tools to the Hadoop cluster

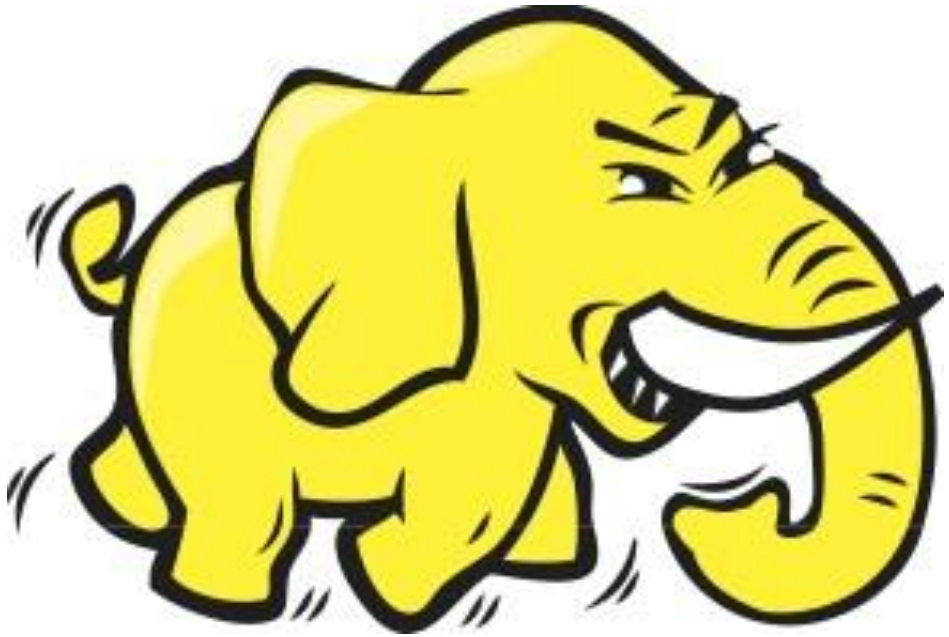
HDFS :

A platform for manipulating data stored in HDFS via a high-level language called Pig Latin. It does data extractions, transformations and loading, and basic analysis in patch mode

Cloudera CDH cluster

Basic Linux Commands

cat [filename]	Display file's contents to the standard output device (usually your monitor).
cd /directorypath	Change to directory.
chmod [options] mode filename	Change a file's permissions.
clear	Clear a command line screen/window for a fresh start.
cp [options] source destination	Copy files and directories.
ls [options]	List directory contents.
mkdir [options] directory	Create a new directory.
mv [options] source destination	Rename or move file(s) or directories.
pwd	Display the pathname for the current directory.
touch filename	Create an empty file with the specified name.
who [options]	Display who is logged on.



Demo

Questions