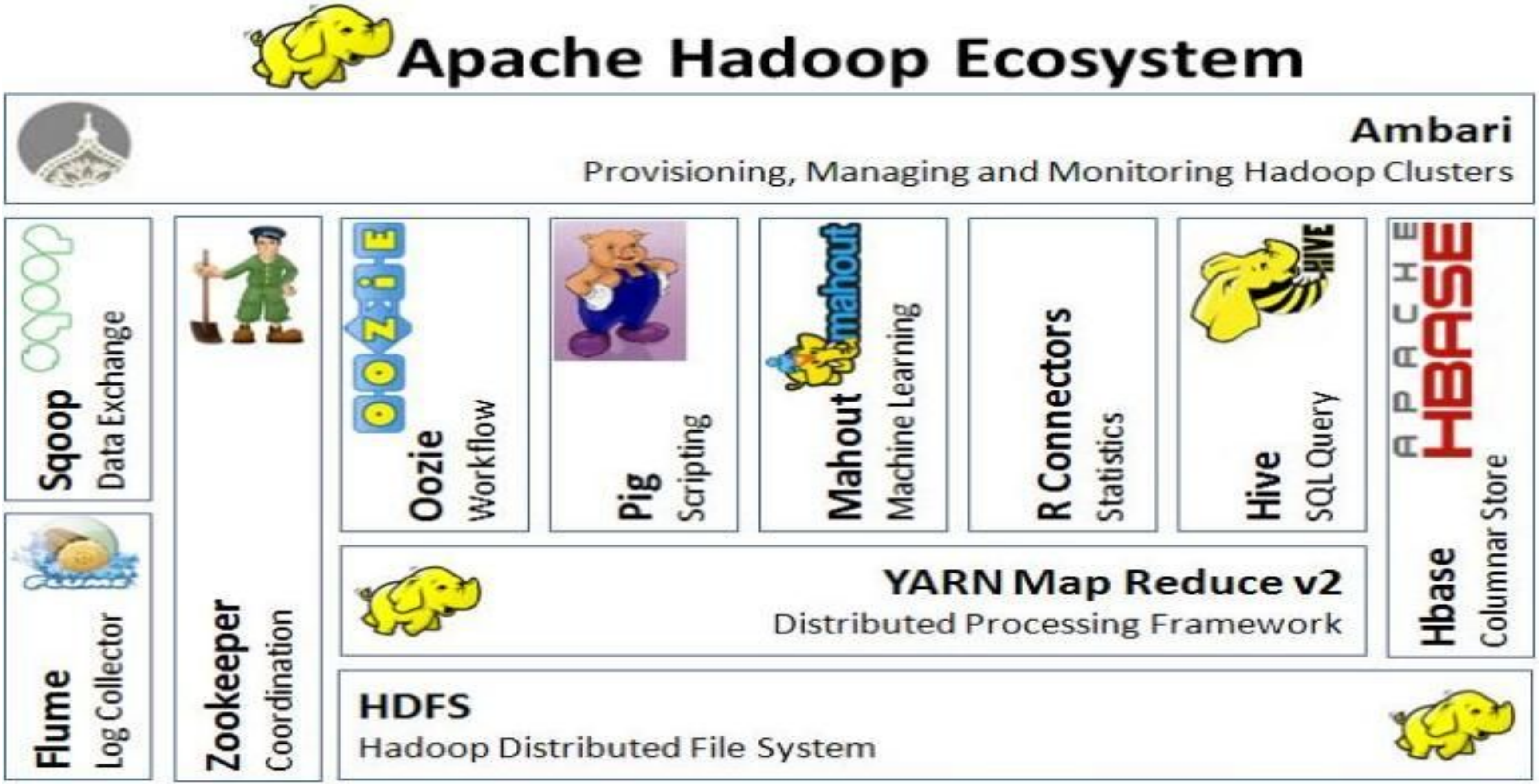
Intro to Hadoop Ecosystem

- Abhay Dandekar

Agenda

1. Hadoop Eco-system
2. Hadoop YARN
3. Zookeeper
4. Hbase
5. Hive
6. Pig
7. Sqoop
8. Flume
9. Oozie
10. Spark

Hadoop Eco-system



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

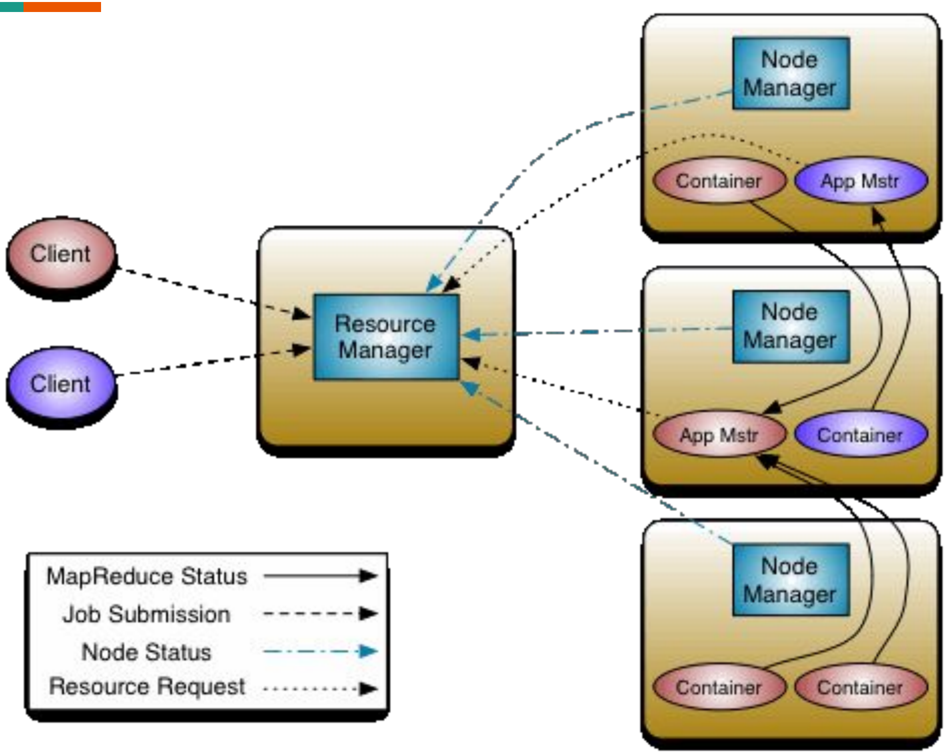
Hive
SQL Query

Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

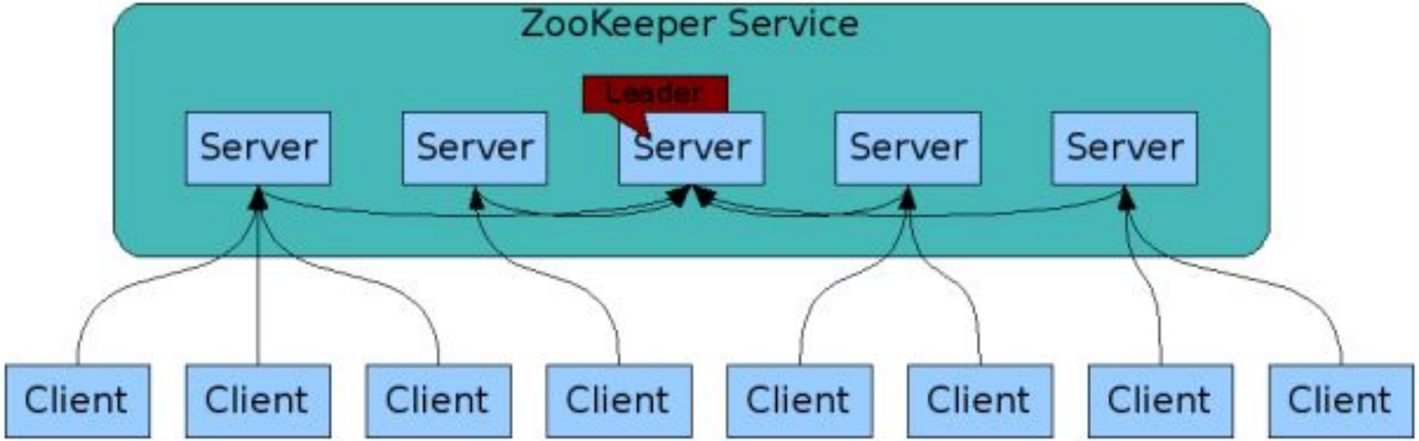
HDFS
Hadoop Distributed File System

YARN - Yet Another Resource Negotiator



1. RM
2. NM
3. AppMaster
4. YarnChild

Zookeeper



1. Leader
2. Quorum
3. Ensemble

Hbase

1. NoSQL
2. Data is in billions or atleast in millions
3. Built on top of HDFS
4. Works with ReST / Java Data Objects / Scala / Jython

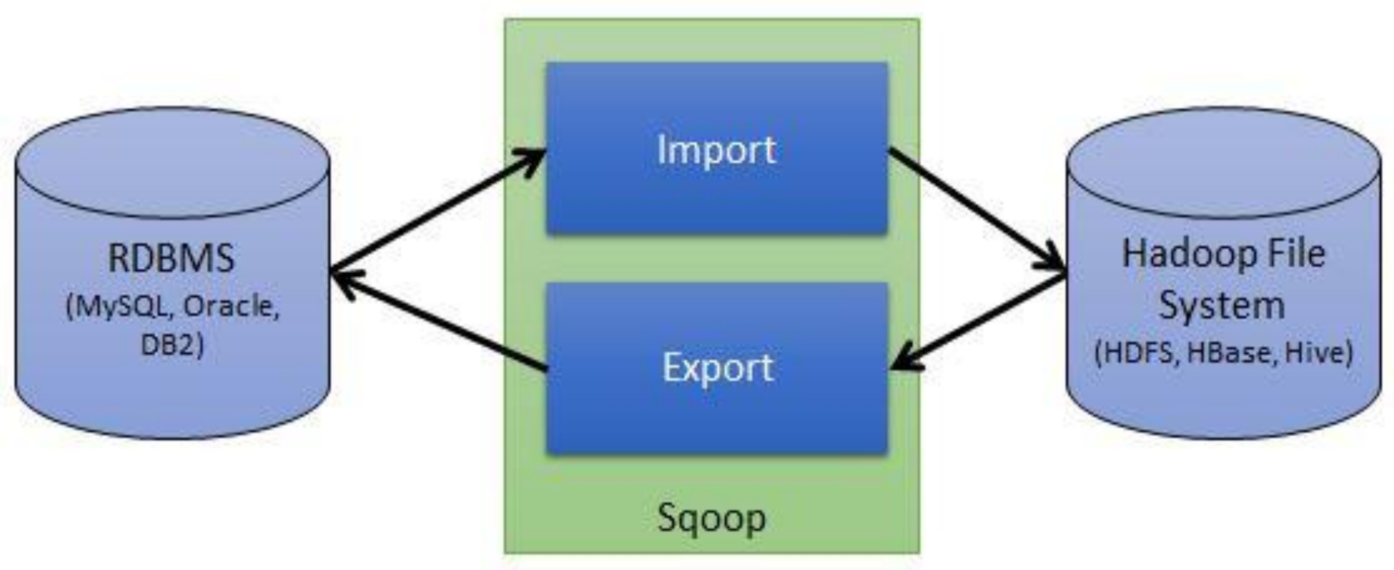
Hive

1. SQL over MapReduce
2. JDBC Connectivity
3. Inbuilt SQL Optimisation

Pig

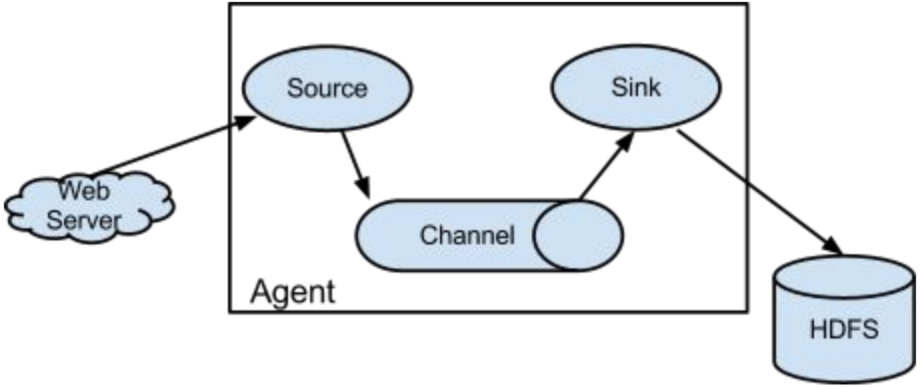
1. High level language for processing analytical workloads
2. Can help avoid SQLs
3. Can bring in heavy optimisation

Sqoop

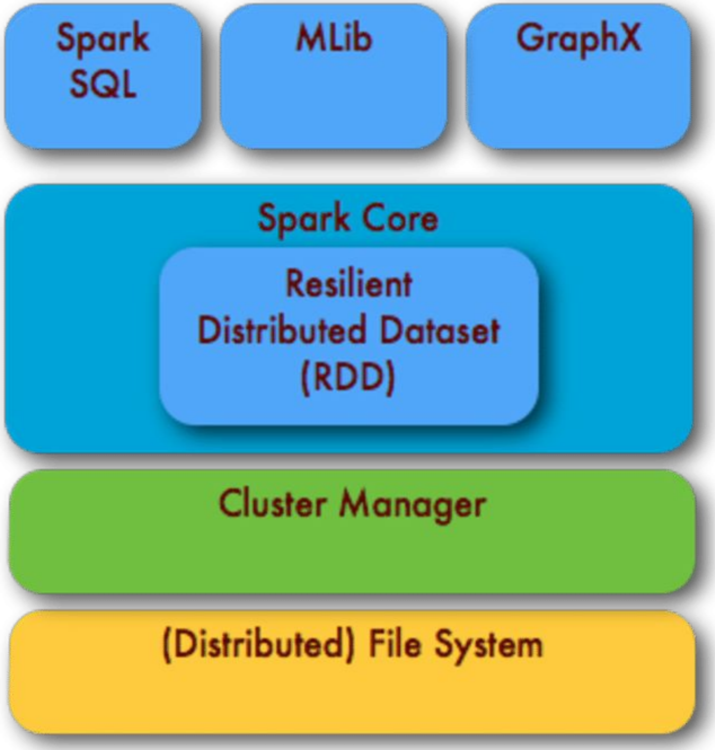

RDBMS
(MySQL, Oracle, DB2)
Import
Export
Sqoop
Hadoop File System
(HDFS, HBase, Hive)

Flume

1. Streaming data processing system
2. Collecting aggregating and moving large data across systems
3. Typically used in data ingestion



Spark



Spark SQL

MLib

GraphX

Spark Core

Resilient Distributed Dataset (RDD)

Cluster Manager

(Distributed) File System