CSCE 5300: Introduction to Big Data and Data Science

Lesson 1

Overview

Overview

- Evaluation Criteria
- Topics to be covered
- Installations
- In class Exercise

Grading Criteria

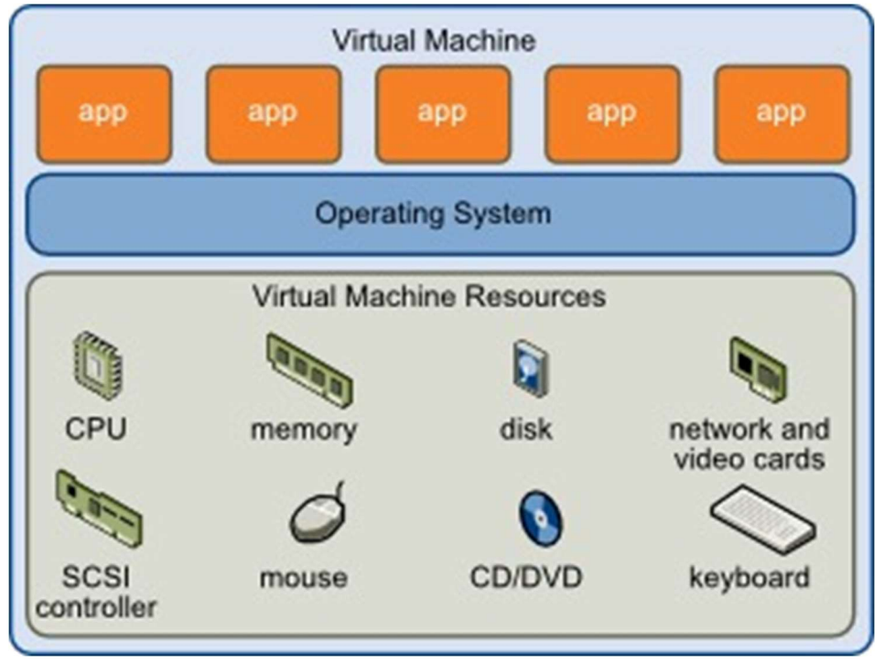
- 20% Quizzes (individual)
- 25% In-class Tasks
- 30% Project
- 25% Exam

Topics to be Covered

- Big Data Overview, Installations
- HDFS / Map Reduce / Big Data Applications
- Hadoop Dependent Query Based No SQL Database Hive
- Hadoop to SQL Parallel Transfer Engine: Sqoop
- Parallel Indexing: Solr & Lucene
- Independent Column Based No SQL Database: Cassandra
- Spark Programming with RDDs and applications
- Spark: Data Frames and SQL
- Machine Learning and Big Data Analytics Applications
- Data Visualization, Deep Learning Concepts
- Spark with RDD and streaming
- GraphX, GraphFrames, Graph Analytics Applications
- Parallel Computing

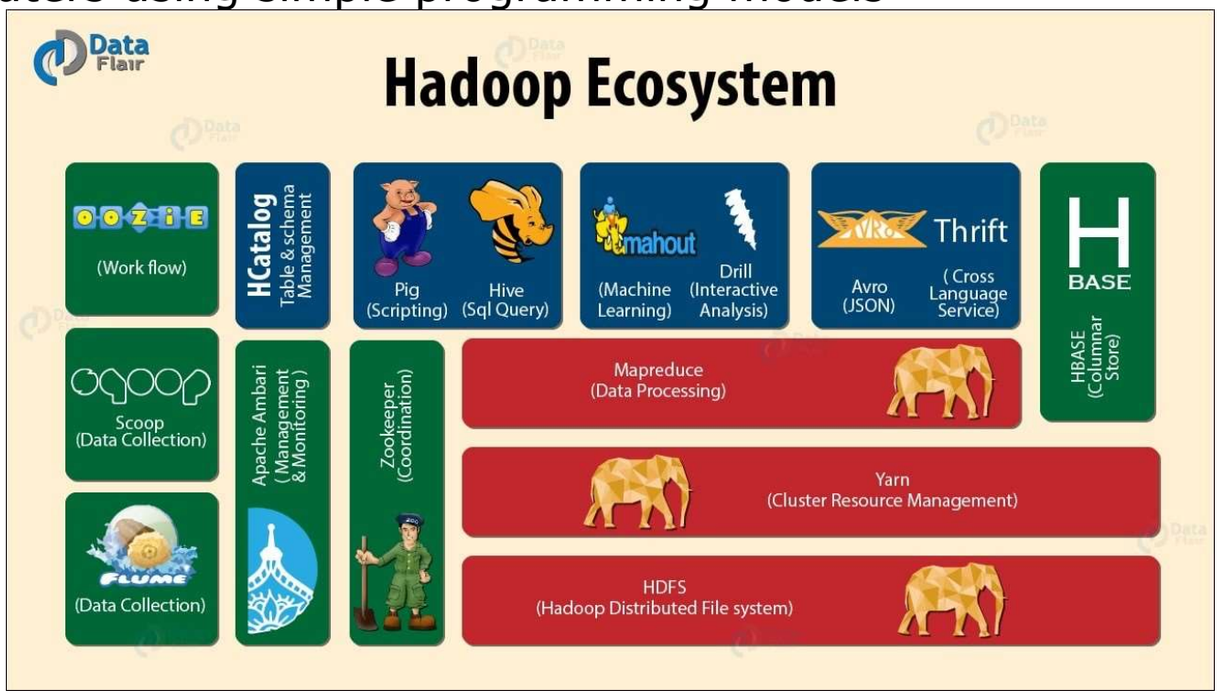
Cloudera
Virtual Machine

- In computing, a **virtual machine** (**VM**) is an emulation of a computer system
- Virtual machines are based on computer architectures and provide functionality of a physical computer.
- Their implementations may involve specialized hardware, software, or a combination

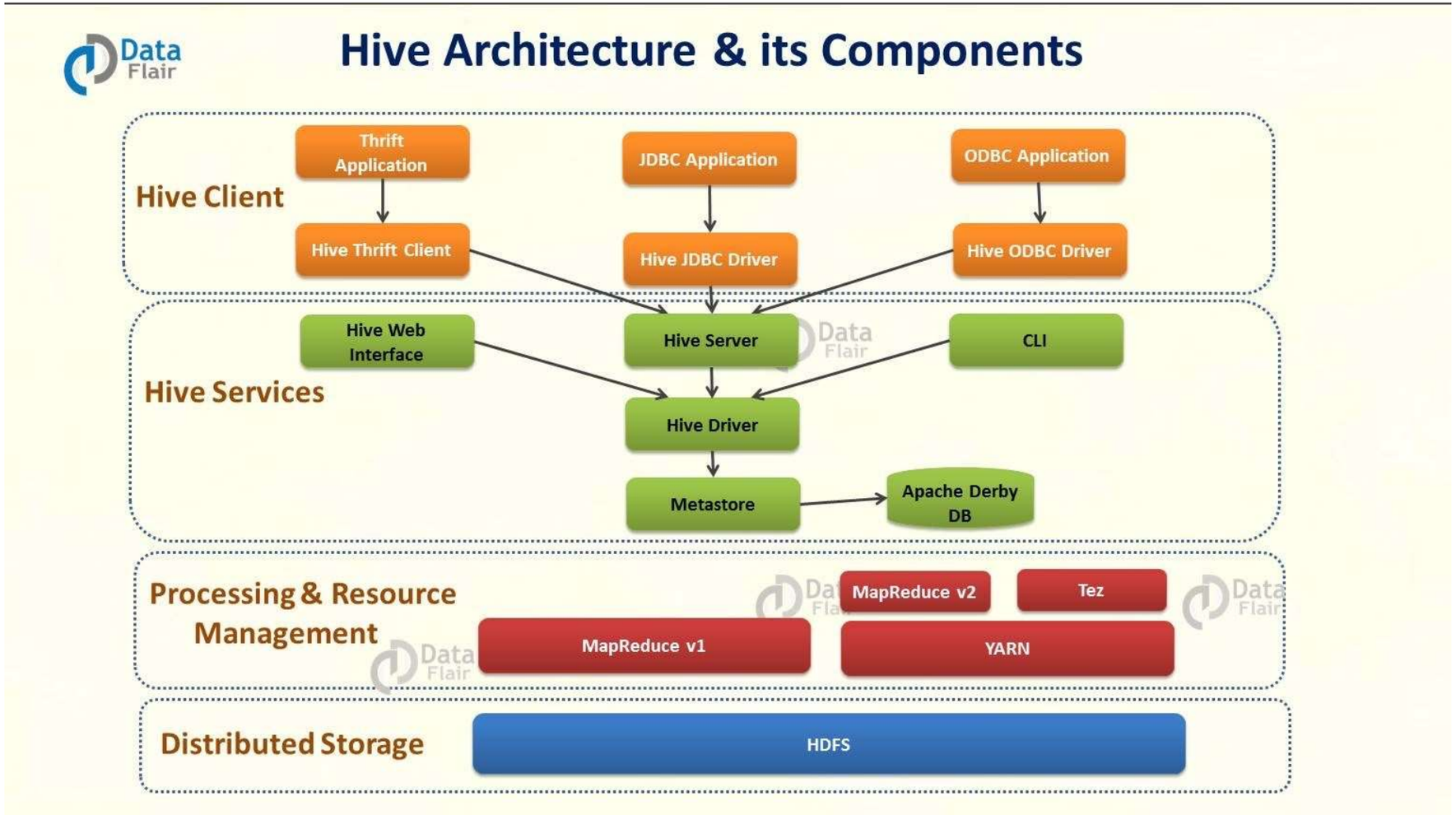


Hadoop Eco-system

A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models



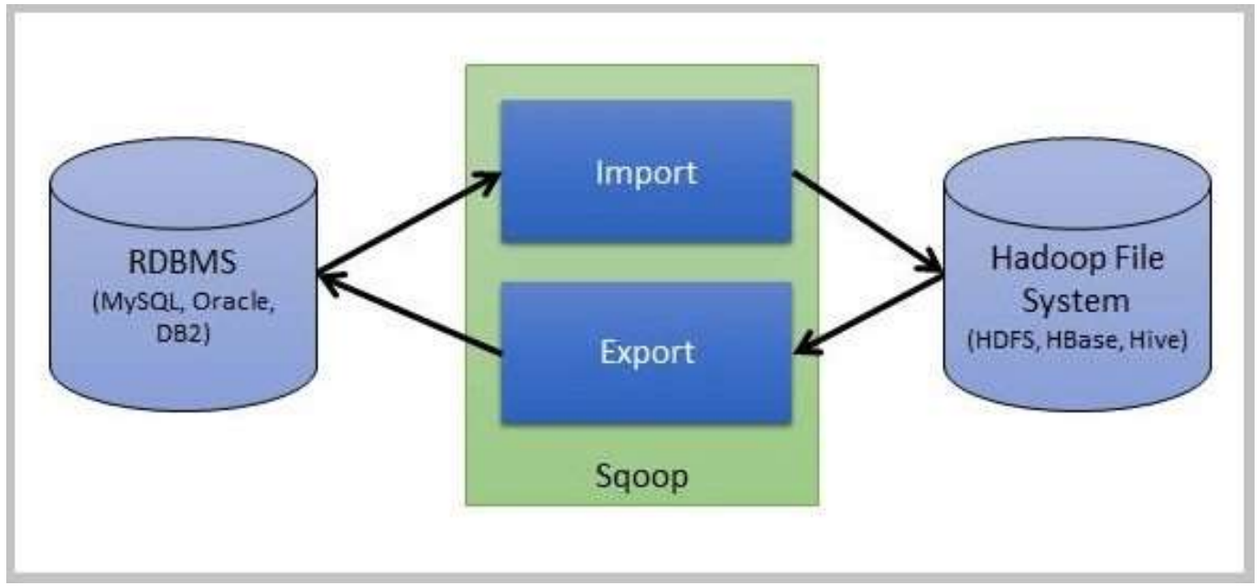
Source: https://data-flair.training/blogs/hadoop-ecosystem-components/



Source: https://data-flair.training/blogs/apache-hive-architecture/

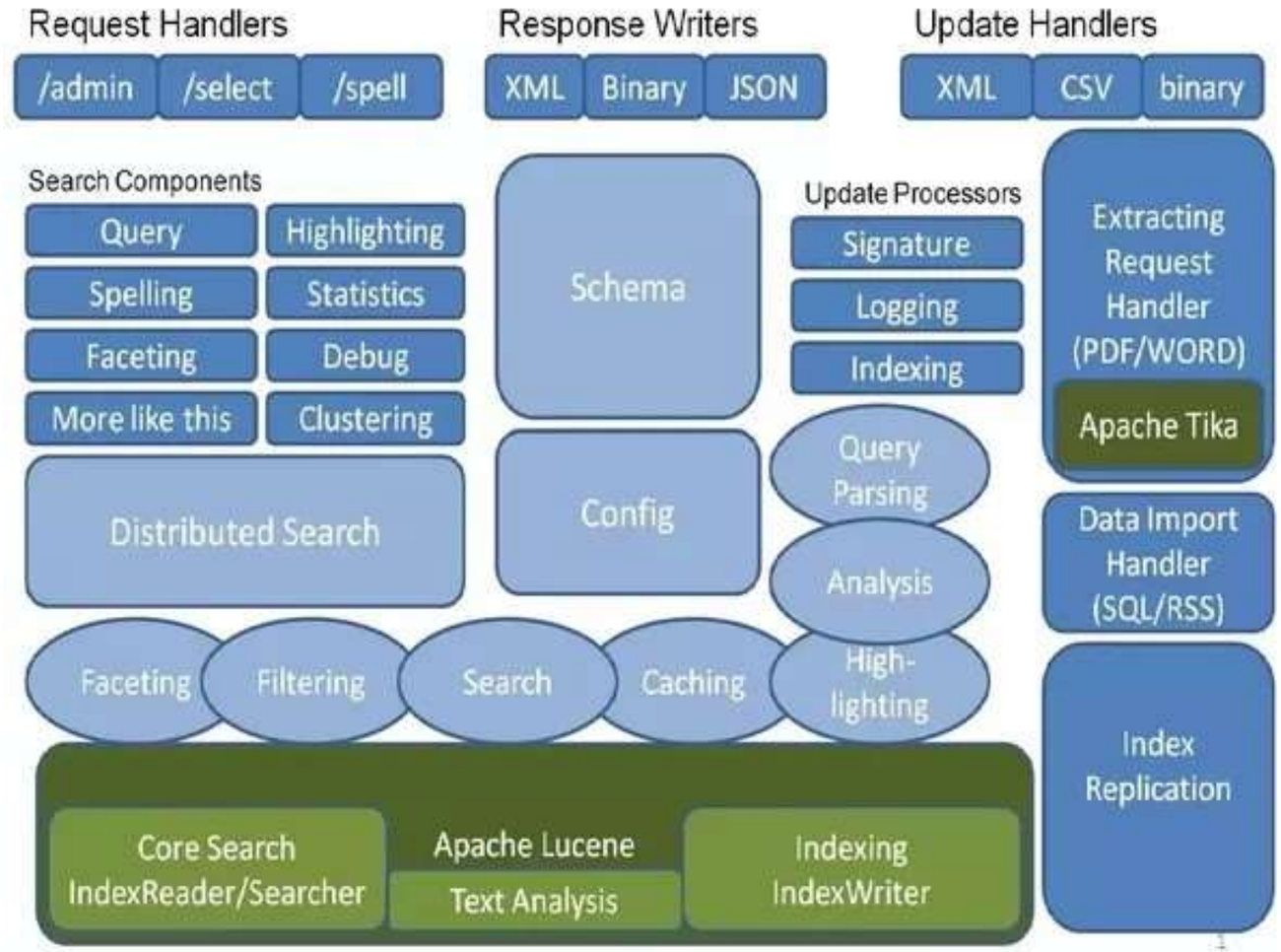
Sqoop

Application for transferring data between relational databases and Hadoop



Source: https://www.hdfstutorial.com/sqoop-architecture/

Lucene/Solr Architecture

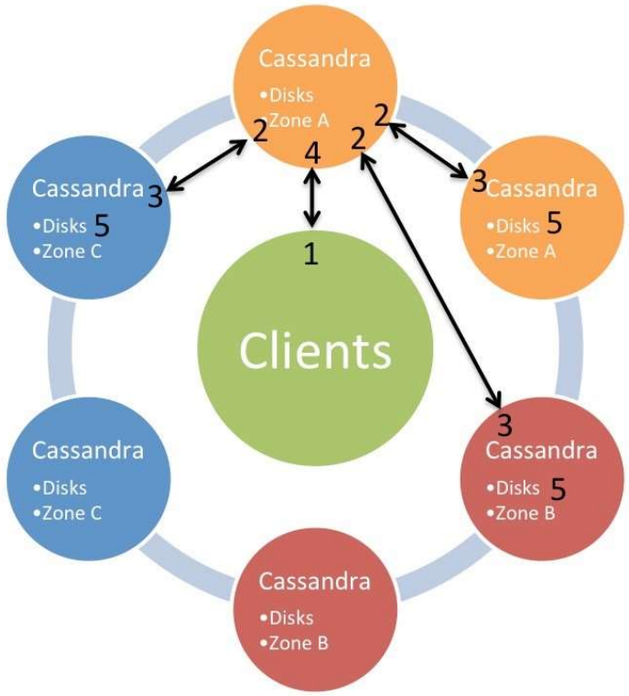


Source: https://www.quora.com/What-is-the-internal-architecture-of-Apache-solr

Cassandra Write Data Flows

Single Region, Multiple Availability Zone

1. Client Writes to any Cassandra Node
2. Coordinator Node replicates to nodes and Zones
3. Nodes return ack to coordinator
4. Coordinator returns ack to client
5. Data written to internal commit log disk



If a node goes offline, hinted handoff completes the write when the node comes back up.

Requests can choose to wait for one node, a quorum, or all nodes to ack the write

SSTable disk writes and compactions occur asynchronously

NETFLIX

Source: https://intellipaat.com/tutorial/cassandra-tutorial/brief-architecture-of-cassandra/