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

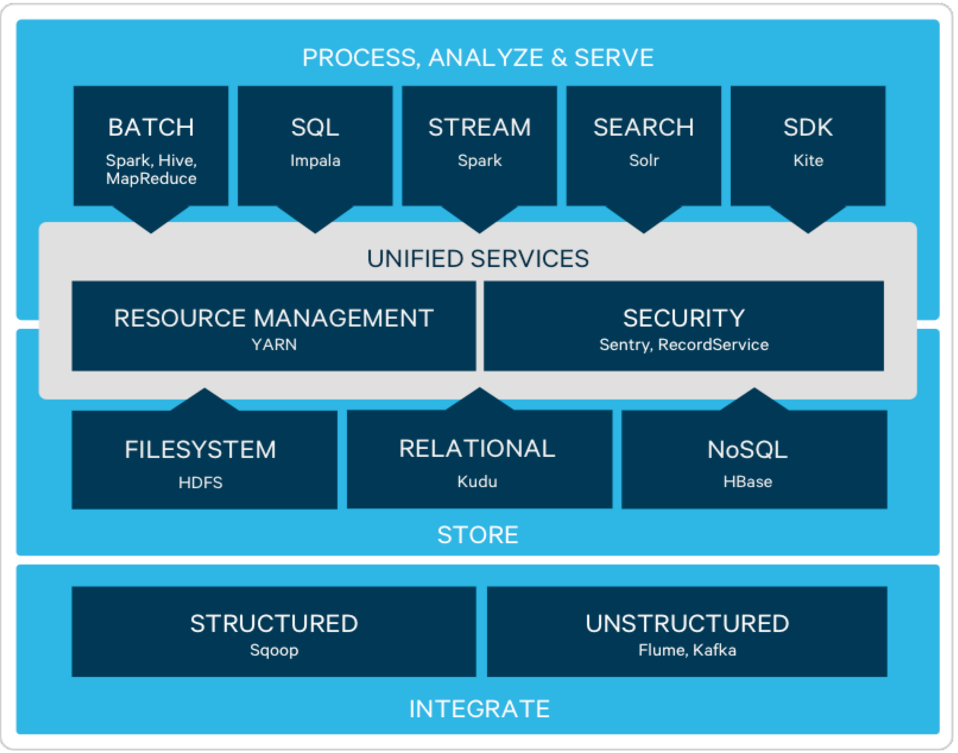
Evaluation Plan		
Exam	20%	Individual
Quiz (3-4)	15%	Individual
In class Exercise	30%	Individual
Project	30%	Team (3-4)
Paper	5%	Individual

Topics to be Covered

- Big Data Overview, Installations, Data Visualization
- HDFS / Map Reduce
- HDFS / Map Reduce and 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
- Spark Programming with RDDs and applications
- Spark: Data Frames and SQL
- Machine Learning and Big Data Analytics Applications
- Deep Learning Concepts
- Spark with RDD and streaming
- GraphX, GraphFrames, Graph Analytics Applications
- Parallel Computing

Cloudera

- A software platform for data engineering, data warehousing, machine learning and analytics that runs in the cloud or on premises.
- Cloudera started as a hybrid open-source Apache Hadoop distribution, CDH (Cloudera Distribution Including Apache Hadoop), that targeted enterprise-class deployments of that technology



PROCESS, ANALYZE & SERVE

BATCH
Spark, Hive, MapReduce

SQL
Impala

STREAM
Spark

SEARCH
Solr

SDK
Kite

UNIFIED SERVICES

RESOURCE MANAGEMENT
YARN

SECURITY
Sentry, RecordService

FILESYSTEM
HDFS

RELATIONAL
Kudu

NoSQL
HBase

STORE

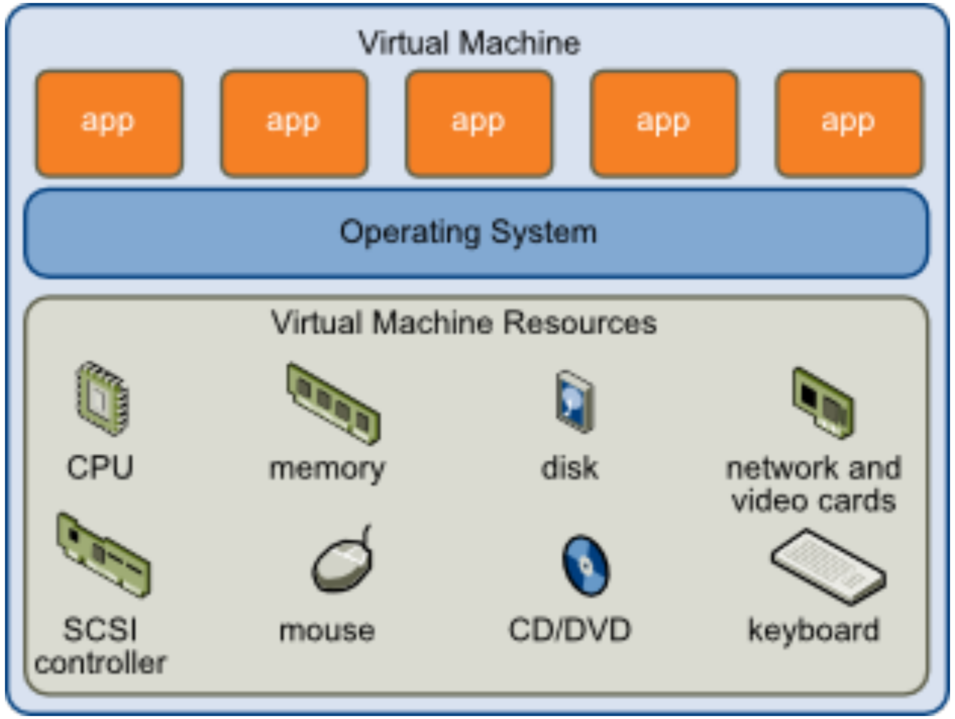
STRUCTURED
Sqoop

UNSTRUCTURED
Flume, Kafka

INTEGRATE

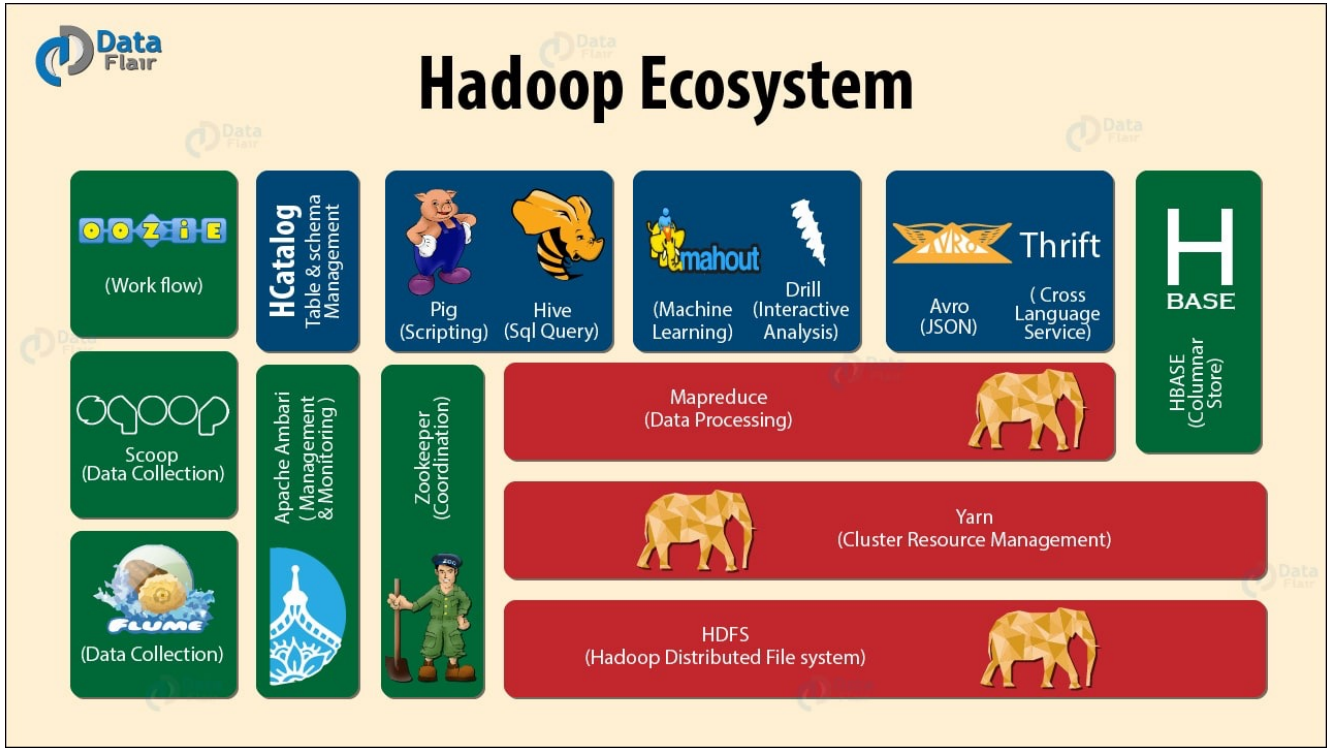
Virtual Machine

- 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

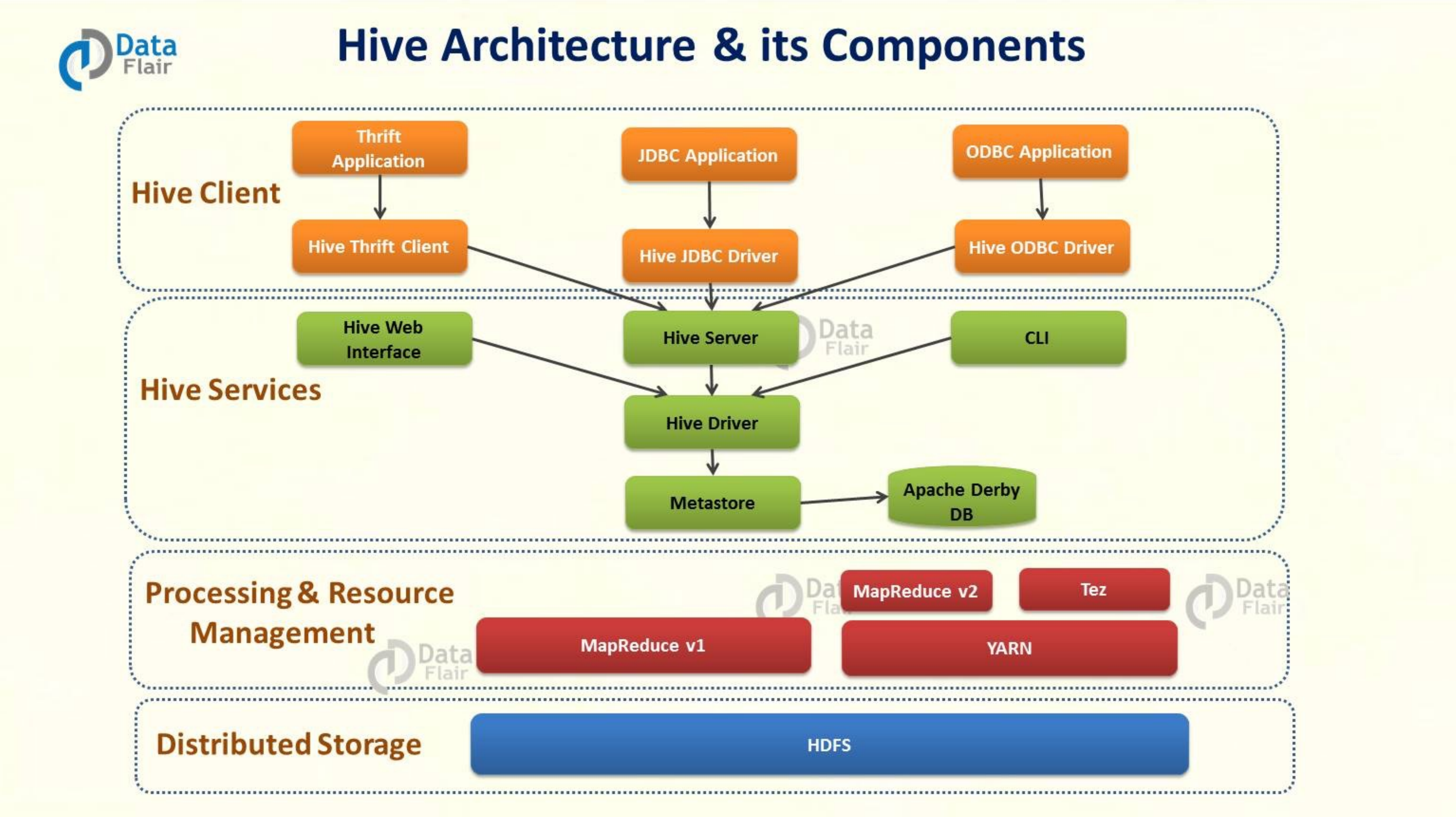


Hadoop Eco-system

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



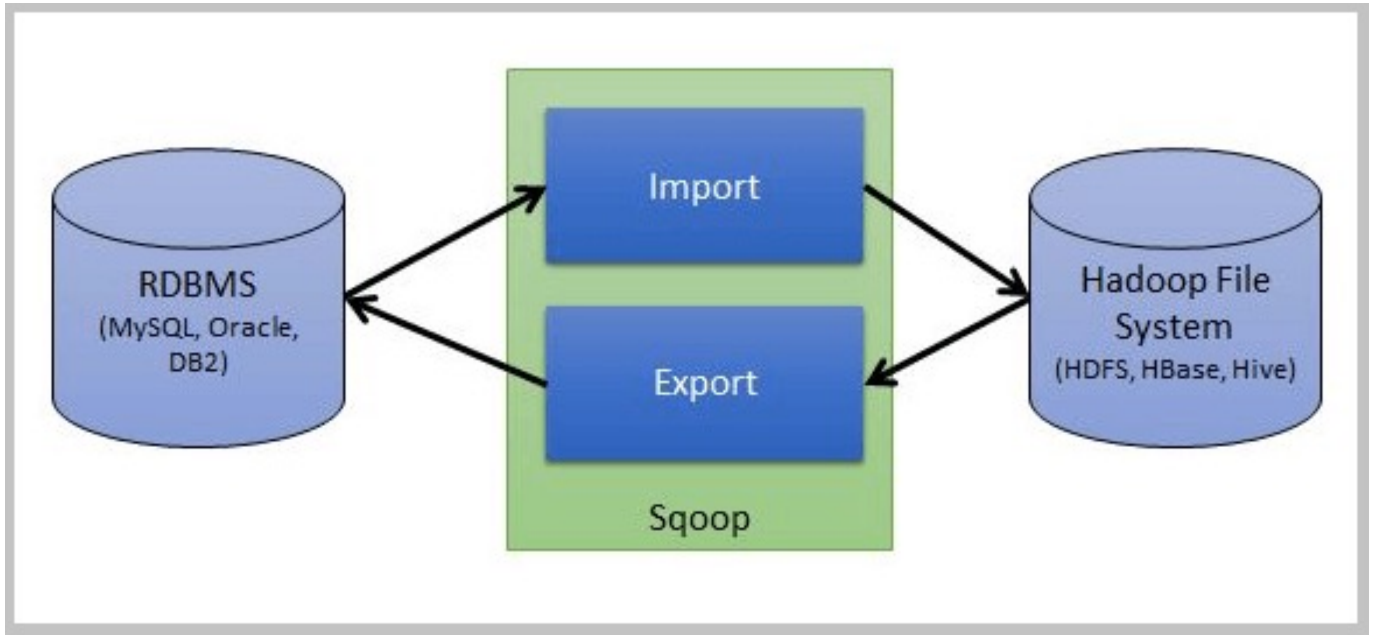
Source: https://data-flair.training/blogs/hadoop-ecosystem-components/



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

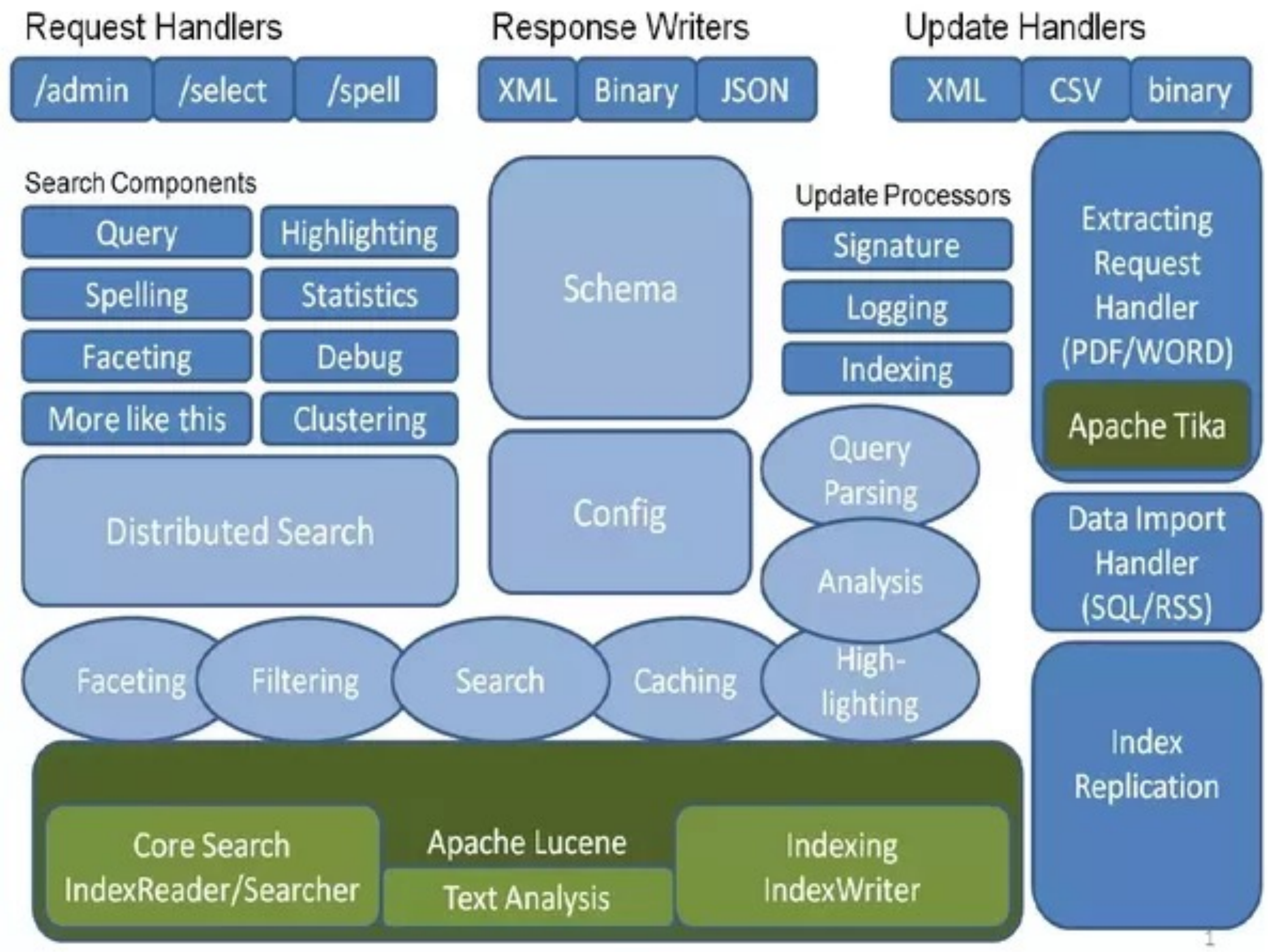
Sqoop

Application for transferring data between relational databases and Hadoop



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

Lucene/Solr Architecture



Request Handlers

/admin /select /spell

Response Writers

XML Binary JSON

Update Handlers

XML CSV binary

Search Components

Query

Highlighting

Spelling

Statistics

Faceting

Debug

More like this

Clustering

Distributed Search

Schema

Config

Update Processors

Signature

Logging

Indexing

Query Parsing

Analysis

High-lighting

Extracting Request Handler (PDF/WORD)

Apache Tika

Data Import Handler (SQL/RSS)

Index Replication

Faceting Filtering Search Caching

Apache Lucene

Core Search IndexReader/Searcher

Text Analysis

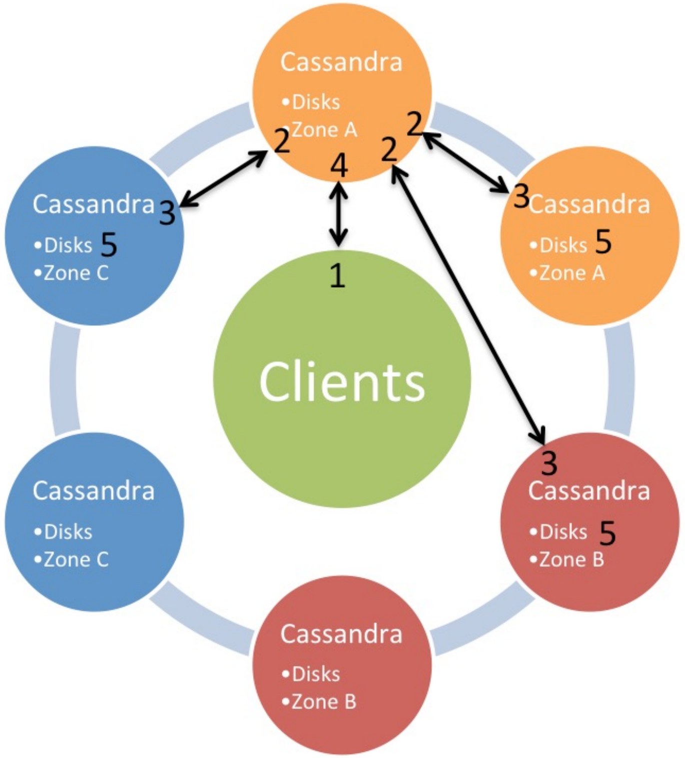
Indexing IndexWriter

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

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