Introduction

22 June 2020 22:35

Ground Rules

1) You will be muted for the entire session
2) Use can use chatbox for any questions and interactions
3) There will be a refreshment break for 10-15 minutes in the middle of the session.
4) Please be regular to these sessions
5) We have a heterogenous batch of learners with varying experience levels. Hence I request you to respect everyone's questions.
6) I will try to answer most of the questions in the session itself. But if someone has out of context or advanced questions , I will address them at the end of the session or beginning of next session.
7) Please complete your assignments on time

Quick look at LMS

Expectations of the participants

Course Outline

22 June 2020 22:50

1) Introduction to Big Data Hadoop and Spark
2) Introduction to Scala for Apache Spark
3) Functional Programming and OOPS concept in Scala
4) Deep Dive into Apache Spark Framework
5) Playing with Spark RDDS
6) Data Frames and Spark SQL
7) Machine Learning using Spark ML Lib
8) Deep Dive into Spark Mllib
9) Understanding Apache Kafka and Apache Flume
10) Apache Spark Streaming - Processing Multiple Batches
11) Apache Spark Streaming - Data Sources
12) In-Class Project

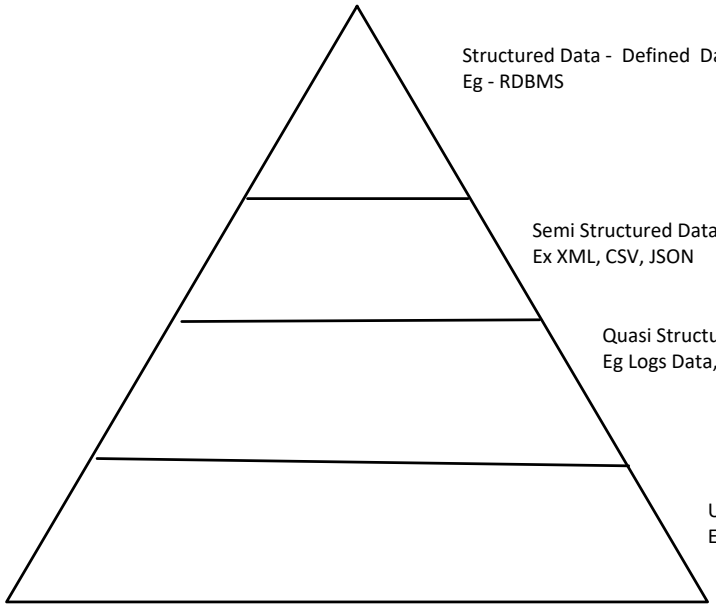
Topics for Today

22 June 2020 22:53

1. Structure of data today
2. Big Data
3. Transportation Services Use Case
4. Big Data Solution Architecture
5. Hadoop
6. Distributed File System
7. Hadoop Cluster Architecture
8. Block Replication
9. Rack Awareness
10. Hadoop Ecosystem
11. Yarn
12. Yarn Components

Structure of data Today

22 June 2020 23:01



Structured Data - Defined Data Type, format, structure,
Eg - RDBMS

Semi Structured Data - Textual Data with a Noticable Pattern, enable Parsing
Ex XML, CSV, JSON

Quasi Structured Data - Texual Data with erratic data format
Eg Logs Data, Web Click Stream Data

Unstructured Data - No Inherent Structure,
Eg Images and Videos

Big Data

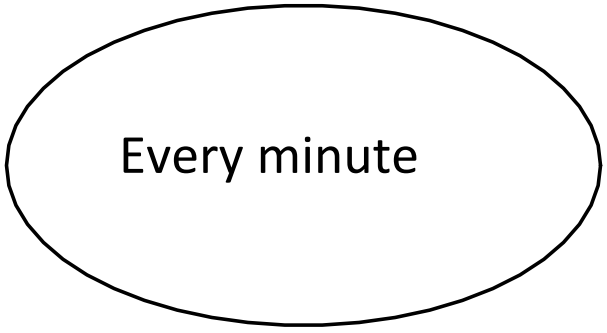
22 June 2020 23:15

Collection of data so large and complex that It becomes difficult to analyses with traditional tools

5.1 Million Facebook Posts

1.4 Million Tweets on Twitter

3.5 Million People post on Instagram



Volume - Size of data is huge
Variety - csv, json, txt, orc, parquet, avro, pdf, xml
Velocity - Speed of data generation
Value - Correct Meaning from the data
Veracity - Uncertainties and Inconsistencies

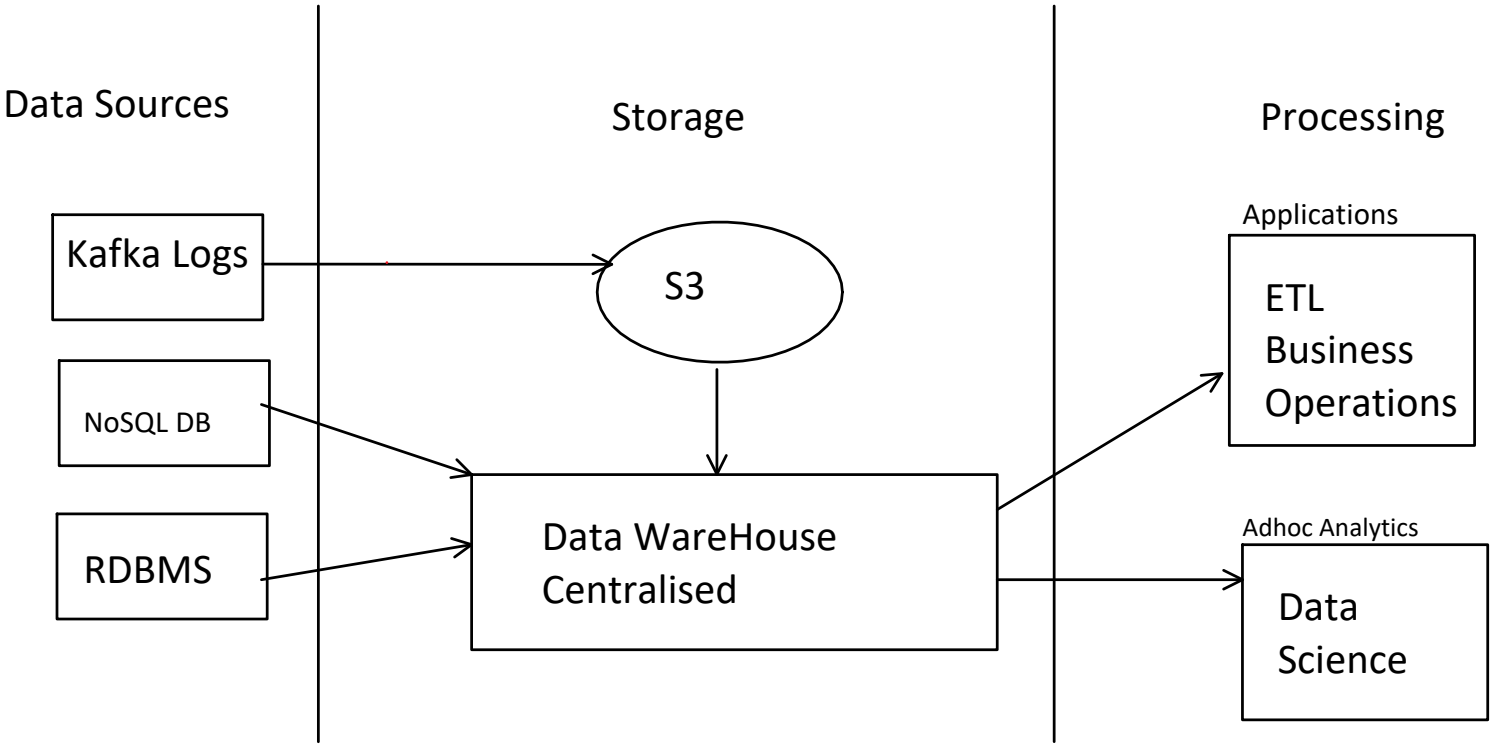
Common Scenario

Telecommunication - Network Utilization

Banking -
Ecommerce -

Transportation Services Use Case

22 June 2020 23:23



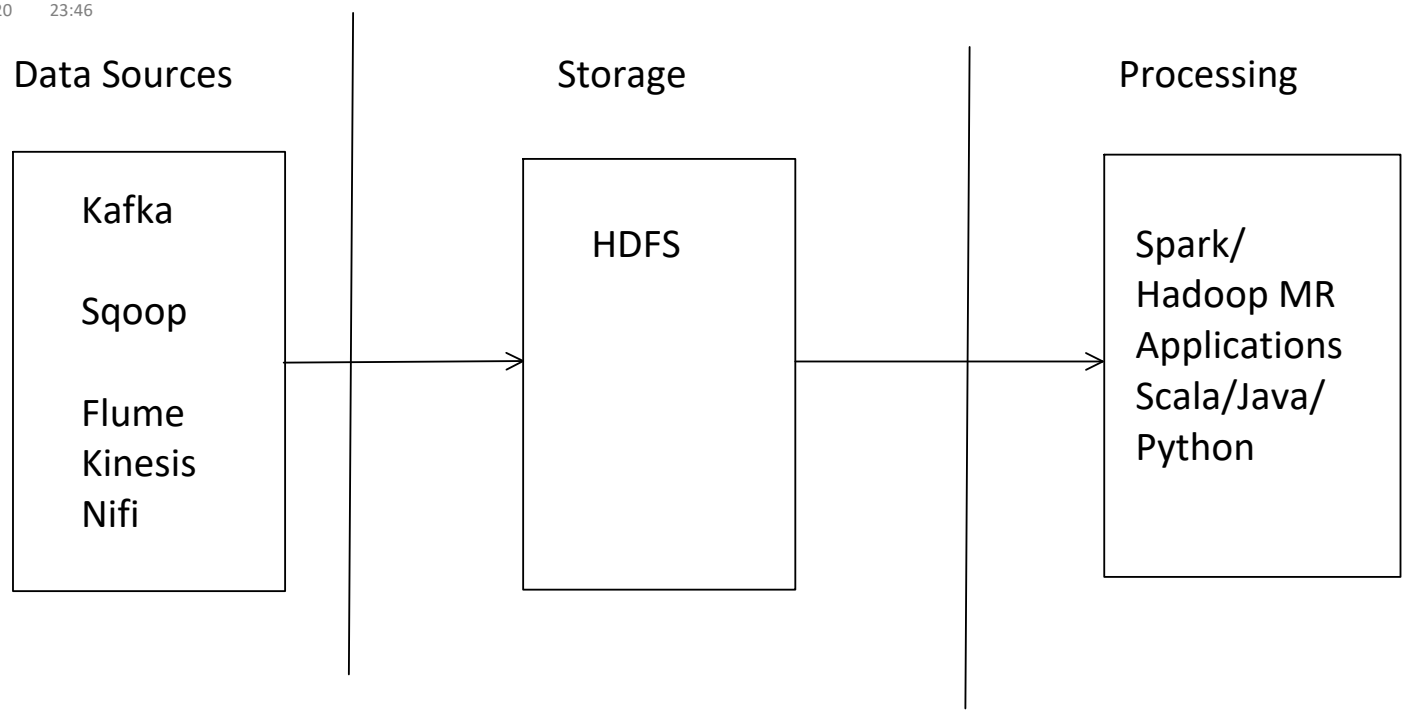
Big Machine

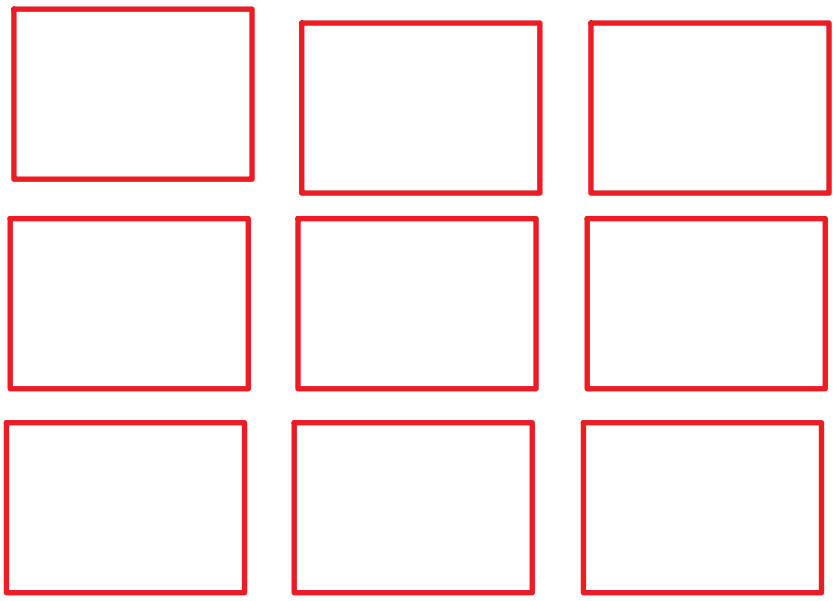
Data is Stored

Single Point of Failure
Scalability - Problematic
Processing Time is more

Big Data Solution Architecture

22 June 2020 23:46





Hadoop

22 June 2020 23:50

What is Hadoop?

Open Source Framework that allows distributed processing of large datasets

Features of Hadoop

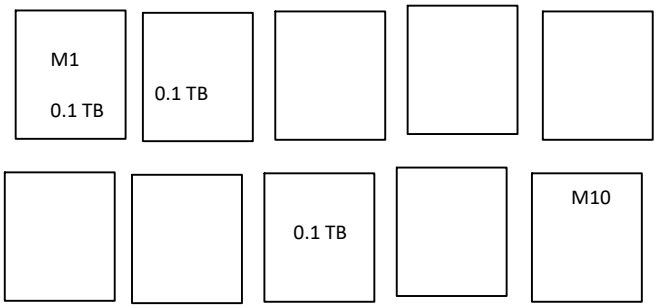
1) Open Source FrameWork
2) Scalable
3) Economic
4) Flexible

5) Fault Tolerant

Distributed File System

22 June 2020 23:56

4 I/o Channels
100 MB/s
1 TB
43 Minutes

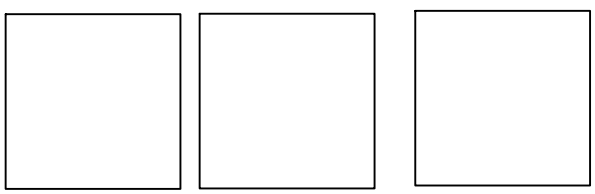


1 Machine
Time taken to Read
1 TB of data - 43 Minutes

10 Machines

Time Required ? minutes

Horizontal Scaling

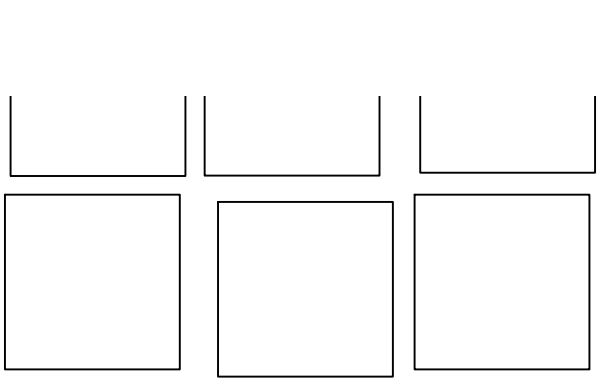


Current

4 Core CPU 8 GB Ram

Vertical Scaling

24 Core CPU
48 GB Ram



4 Core CPU 8 GB Ram
500 GB Hard Disk
100 MB/s

24 Core CPU
48 GB Ram
3 TB Hard Disk
100 MBPS i/o

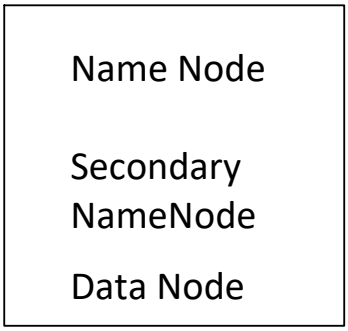
Hadoop Core Components

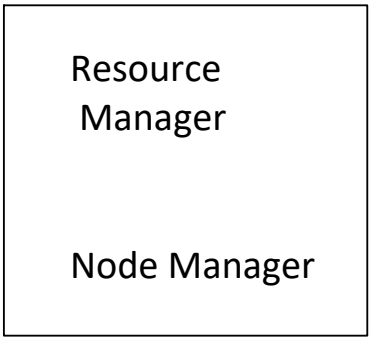
22 June 2020 23:58

HDFS - For Storage

Master Slave Architecture

Yarn - For Resource Management





HDFS - Hadoop Distributed File System

Name Node - Master

- Manages Data Nodes
- Records the metadata of data present in Data nodes

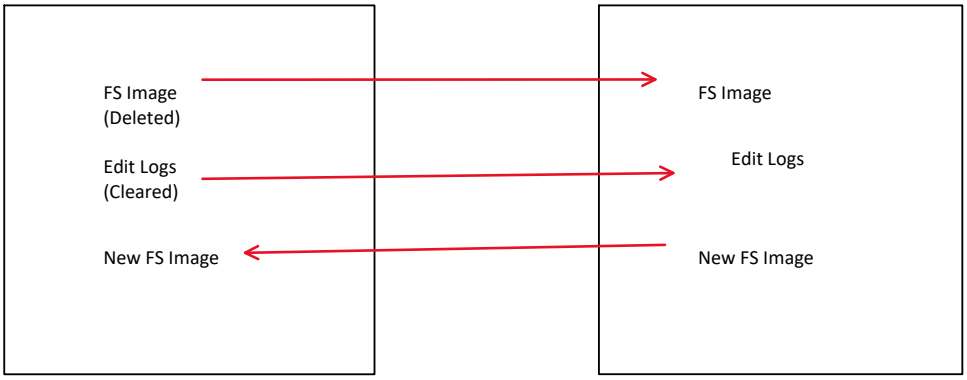
Data Node - Slave

- Actual data is stored
- Serve Read and Write Requests

Checkpointing

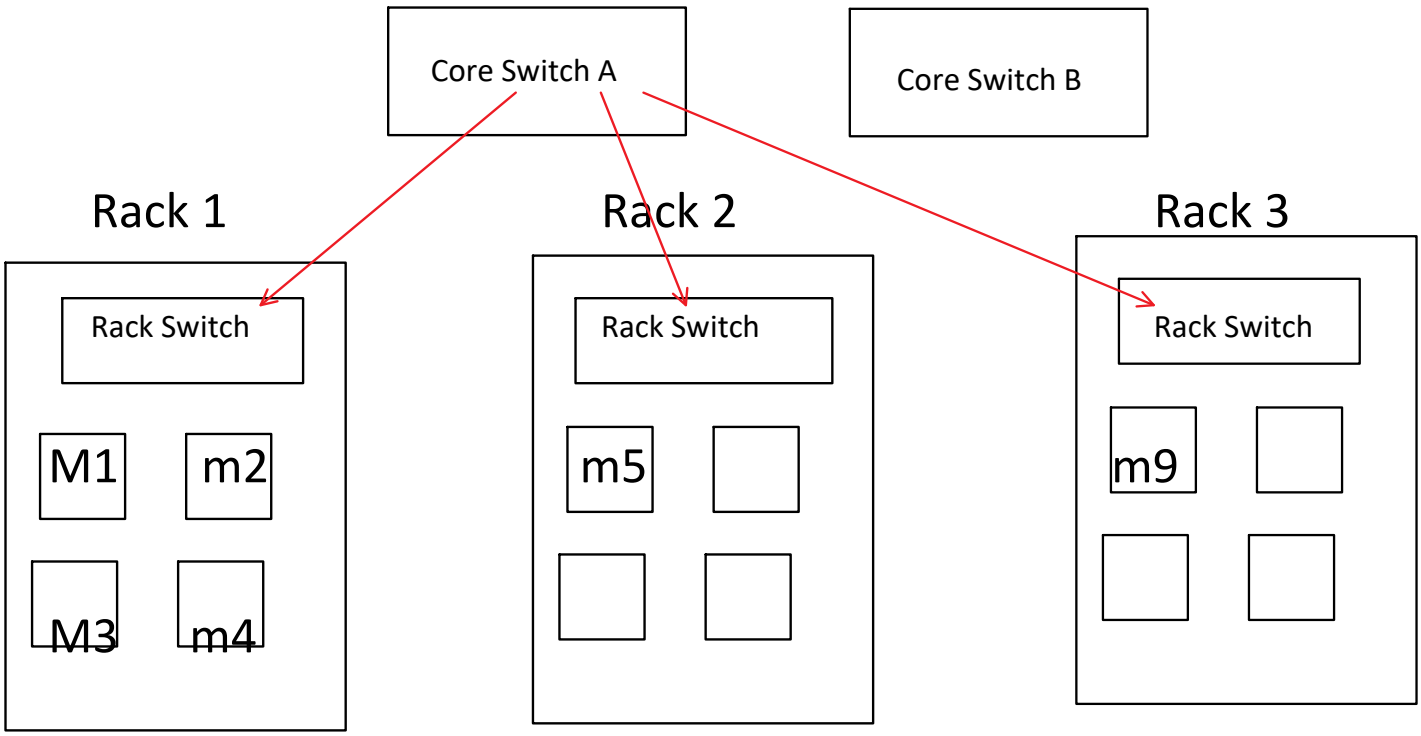
NameNode

Secondary NameNode


FS Image
(Deleted)
Edit Logs
(Cleared)
New FS Image
FS Image
Edit Logs
New FS Image

Hadoop Cluster Architecture

23 June 2020 00:11



Block Replication and Rack Awareness

23 June 2020 00:12

How data is stored on HDFS
Data is stored in form of Blocks
128 MB

248 MB File - 128 MB , 120 MB
514 MB - 4 - 128 MB , 2 MB

Is it safe to have just 1 copy of each block?

Replication Factor 3

RacKAwareness

Slide 42

Hadoop Ecosystem

23 June 2020 00:16

Slide 49/50

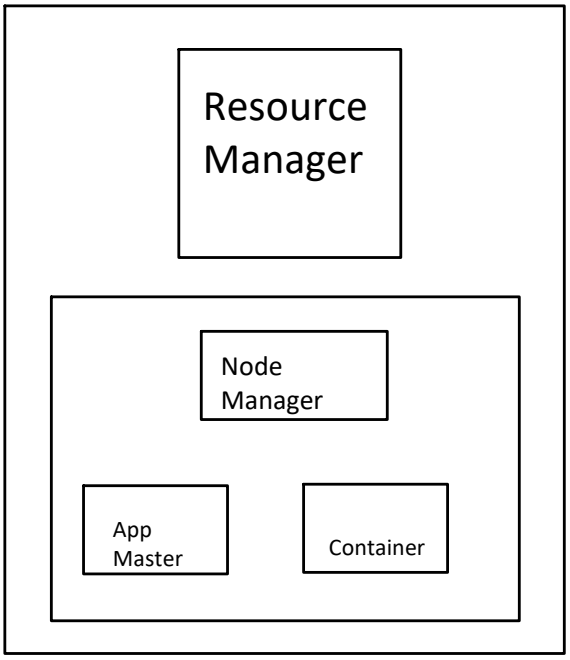
YARN Components

23 June 2020 23:50

Master Daemon , It manages all other daemons
Accepts Jobs Submission

Responsible for Containers, monitoring their resources usage

App Master
One per application
Negotiates Resources



Container

CPU,RAM
Actuals task run

Slide 56