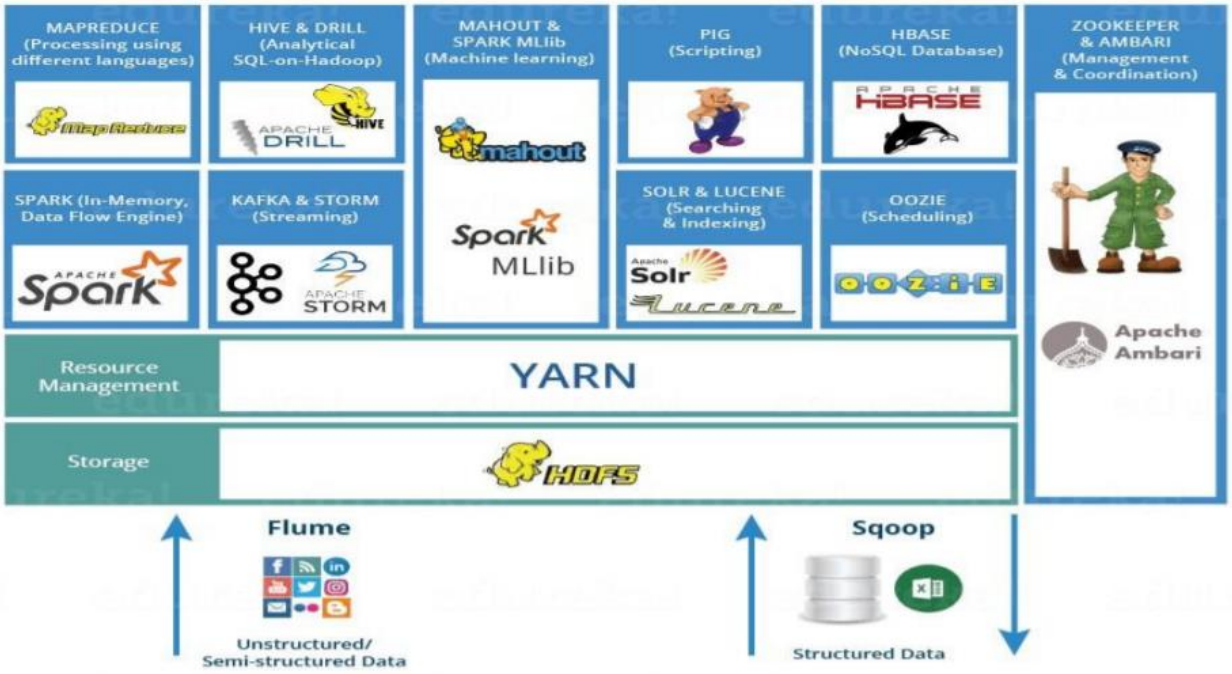
Practical-6

- **Task : Study of Hadoop Ecosystem.**

Hadoop Ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problems. You can consider it as a suite which encompasses a number of services (ingesting, storing, analysing and maintaining) inside it. Let us discuss and get a brief idea about how the services work individually and in collaboration.

- HDFS -> Hadoop Distributed File System
- YARN -> Yet Another Resource Negotiator
- MapReduce -> Data processing using programming
- Spark -> In-memory Data Processing
- PIG, HIVE-> Data Processing Services using Query (SQL-like)
- HBase -> NoSQL Database
- Mahout, Spark MLlib -> Machine Learning
- Apache Drill -> SQL on Hadoop
- Zookeeper -> Managing Cluster
- Oozie -> Job Scheduling
- Flume, Sqoop -> Data Ingesting Services
- Solr & Lucene -> Searching & Indexing
- Ambari -> Provision, Monitor and Maintain cluster

Below are the Hadoop components, that together form a Hadoop ecosystem, I will be covering each of them in this blog:



HDFS

- Hadoop Distributed File System is the core component or you can say, the backbone of Hadoop Ecosystem.
- HDFS is the one, which makes it possible to store different types of large data sets (i.e. structured, unstructured and semi structured data).
- HDFS has two core components, i.e. NameNode and DataNode.
    - The NameNode is the main node and it doesn't store the actual data. It contains metadata, just like a log file or you can say as a table of content.
      Therefore, it requires less storage and high computational resources.
    - On the other hand, all your data is stored on the DataNodes and hence it
      requires more storage resources. These DataNodes are commodity hardware
      (like your laptops and desktops) in the distributed environment. That's the
      reason, why Hadoop solutions are very cost effective

YARN

- Consider YARN as the brain of your Hadoop Ecosystem. It performs all your
- processing activities by allocating resources and scheduling tasks.
- It has two major components, i.e. ResourceManager and NodeManager.
    - ResourceManager is again a main node in the processing department.
    - NodeManagers are installed on every DataNode. It is responsible for execution of task on every single DataNode.

MAPREDUCE

- It is the core component of processing in a Hadoop Ecosystem as it provides the logic of processing.
- In a MapReduce program, Map() and Reduce() are two functions.
    - The Map function performs actions like filtering, grouping and sorting.
    - While Reduce function aggregates and summarizes the result produced by map function.
    - The result generated by the Map function is a key value pair (K, V) which acts as the input for Reduce function.

APACHE PIG

- PIG has two parts: Pig Latin, the language and the pig runtime, for the execution environment. You can better understand it as Java and JVM.
- The compiler internally converts pig latin to MapReduce. It produces a sequential set of MapReduce jobs, and that's an abstraction (which works like black box).
- PIG was initially developed by Yahoo.
- It gives you a platform for building data flow for ETL (Extract, Transform and Load), processing and analyzing huge data sets.

APACHE HIVE

- Facebook created HIVE for people who are fluent with SQL. Thus, HIVEmakes them feel at home while working in a Hadoop Ecosystem.
- Basically, HIVE is a data warehousing component which performs reading, writing and managing large data sets in a distributed environment using SQL-like interface.
- HIVE + SQL = HQL
- The query language of Hive is called Hive Query Language(HQL), which is very similar like SQL.

APACHE MAHOUT

- Mahout which is renowned for machine learning. Mahout provides an environment for creating machine learning applications which are scalable.
- It performs collaborative filtering, clustering and classification. Some people also consider frequent item set missing as Mahout's function.

APACHE SPARK

- Apache Spark is a framework for real time data analytics in a distributed computing environment.

- It executes in-memory computations to increase speed of data processing over Map-Reduce.
- It is 100x faster than Hadoop for large scale data processing by exploiting inmemory computations and other optimizations. Therefore, it requires high processing power than Map-Reduce.

APACHE HBASE

- HBase is an open source, non-relational distributed database. In other words, it is a NoSQL database.
- It supports all types of data and that is why, it's capable of handling anything and everything inside a Hadoop ecosystem.
- It is modelled after Google's BigTable, which is a distributed storage system designed to cope up with large data sets.

APACHE DRILL

- As the name suggests, Apache Drill is used to drill into any kind of data. It's an open source application which works with distributed environment to analyze large data sets.
- It is a replica of Google Dremel.
- It supports different kinds NoSQL databases and file systems, which is a powerful feature of Drill. For example: Azure Blob Storage, Google Cloud Storage, HBase, MongoDB, MapR-DB HDFS, MapR-FS, Amazon S3, Swift, NAS and local files.

APACHE ZOOKEEPER

- Apache Zookeeper is the coordinator of any Hadoop job which includes a combination of various services in a Hadoop Ecosystem.
- Apache Zookeeper coordinates with various services in a distributed environment.

APACHE OOZIE

- Consider Apache Oozie as a clock and alarm service inside Hadoop Ecosystem. For Apache jobs, Oozie has been just like a scheduler. It schedules Hadoop jobs and binds them together as one logical work.
- There are two kinds of Oozie jobs:
    - Oozie workflow: These are sequential set of actions to be executed. You can assume it as a relay race. Where each athlete waits for the last one to complete his part.
    - Oozie Coordinator: These are the Oozie jobs which are triggered when the data is made available to it.

APACHE FLUME

- Ingesting data is an important part of our Hadoop Ecosystem.
- The Flume is a service which helps in ingesting unstructured and semi-structured data into HDFS.
- It gives us a solution which is reliable and distributed and helps us in collecting, aggregating and moving large amount of data sets.
- It helps us to ingest online streaming data from various sources like network traffic, social media, email messages, log files etc. in HDFS.

APACHE SQOOP

- Now, let us talk about another data ingesting service i.e. Sqoop. The major difference between Flume and Sqoop is that:
- Flume only ingests unstructured data or semi-structured data into HDFS.
- While Sqoop can import as well as export structured data from RDBMS or Enterprise data warehouses to HDFS or vice versa.

APACHE SOLR & LUCENE

- Apache Solr and Apache Lucene are the two services which are used forsearching and indexing in Hadoop Ecosystem.
- Apache Lucene is based on Java, which also helps in spell checking.
- If Apache Lucene is the engine, Apache Solr is the car built around it. Solr is a complete application built around Lucene.
- It uses the Lucene Java search library as a core for search and full indexing.

APACHE AMBARI

- Ambari is an Apache Software Foundation Project which aims at making Hadoop ecosystem more manageable. It includes software for provisioning, managing and monitoring Apache Hadoop clusters.
- The Ambari provides:
    - Hadoop cluster provisioning:
    - Hadoop cluster management:
    - Hadoop cluster monitoring:

❖ **Task : Explore the Services provided for Big Data Analytics by the below Software Company**

a) Cloudera

- Data Hub

A powerful cloud service that radically simplifies building modern, missioncritical data-

driven applications with enterprise security, governance, scale, and control.
Key benefits & features

- Enable complex, multi-component workloads
- PaaS-like experience with flexibility
- Flexibility, choice, and control
- Machine learning meets mission critical
- True cloud-native architecture
- Robust orchestration and automation

- Data Warehouse

Deliver self-service analytics on massive amounts of verified data to thousands of users without compromising cost, speed, or security.
Key features

- Hybrid and multi-cloud
- Unprecedented scale and volume
- Self-service analytics
- Real-time analytics at scale
- Easy to use

- Machine Learning

Accelerate data driven decision making from research to production with a secure, scalable, and open platform for ML.
Key features

- Containerized ML workspaces
- SDX for training data & ML models
- CML workbench & bring your own IDE
- Machine learning experiments
- Complete MLOps toolset
- Cloud-native hybrid data architecture

- CDP Private Cloud

Deliver powerful data analytics and machine learning in minutes with 50% less data center infrastructure.
Benefits of CDP Private Cloud

- Reduce data center costs
- Consolidate your clusters
- Guarantee SLAS for key applications
- Quickly onboard applications
- Customize your environments

b) MapR

- HPE EZMERAL DATA FABRIC

Provide enterprise-wide global access to data and ensure the success of your datadriven

transformation initiatives -- with a unified data fabric across your data centers, multiple clouds, and the edge.

- HPE Ezmeral Container Platform

Software platform designed to run both cloud-native and non-cloud native applications in containers.

- HPE Ezmeral ML Ops

A software solution that extends the capabilities of the HPE Ezmeral Container Platform to support the entire machine learning lifecycle and implement DevOpslike processes to standardize machine learning workflows.

c) Hortonworks

- Hortonworks HDP

The HDP Sandbox makes it easy to get started with Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, Druid and Data Analytics Studio (DAS).

- Hortonworks DataFlow--Sandbox

It easy to get started with Apache NiFi, Apache Kafka, Apache Storm, and Streaming Analytics Manager (SAM).

❖ **Task : Consider data storage for "amazon products" data. Each product data, "ProData" which is in a file of size less than or equal to 64 MB. A data block stores the full file data for a product of 'ProData_idN", where N = 1 to 500.Data block default size is 64MB. Default Replication in Data nodes is 3.**

1. How the files of each product will be distributed at a Hadoop cluster? How many product data can be stored at one cluster? Assume that each rack has two DataNodes for processing each of 64 GB memory. Assume that cluster consists of 120 racks, and thus 240 DataNodes.

Ans:

Data block default size is 64mb. Each product data file size is less than or equal to 64mb. Therefore, for each product data one data block is sufficient. A data block is in DataNode.

Assume that each rack has two DataNodes for processing each of 64 GB memory. Each node can thus store 64GB/64MB = 1024 data blocks product data.

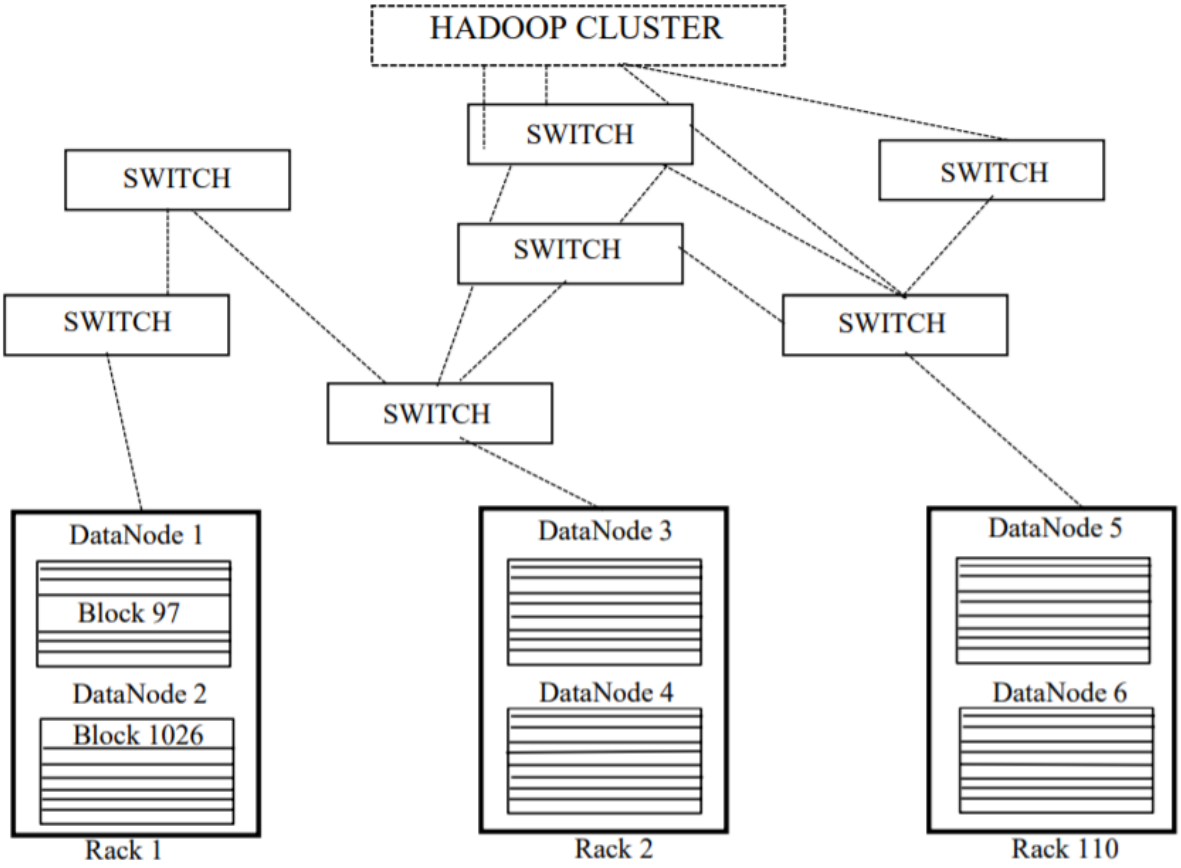
Each rack can thus store 2 * 64GB/64MB = 2048 data blocks.
Each data block default replicated three times in DataNode. Therefore data stored in cluster = 120 * 2048/3 = 81920.
Therefore, the maximum number of 81920 “ProData_idN” files can be distributed per cluster.

2. What is the total memory capacity of the cluster in TB? and each rack capacity in TB?

Ans:
Total memory capacity of the cluster = 120*128 = 15360GB = 15TB. Total memory capacity of each rack = 2*1024*64mb = 128GB

3. Show the distributed blocks for products with ID= 97 and 1026. Assumedefault replication in the Data Nodes = 3.

Ans:



4. What shall be the changes when a “ProData” file size ≤ 128 MB?

Ans:
Changes will be that each node will have half the number of data blocks.