Pune Institute of Computer Technology,

Dhankawadi, Pune - 43

Academic year: 2021-22

CASE STUDY ON

HEALTH-CARE SYSTEM WITH HADOOP ECOSYSTEM COMPONENTS

Vedant Bothikar - 31115

Shubham Chemate - 31118

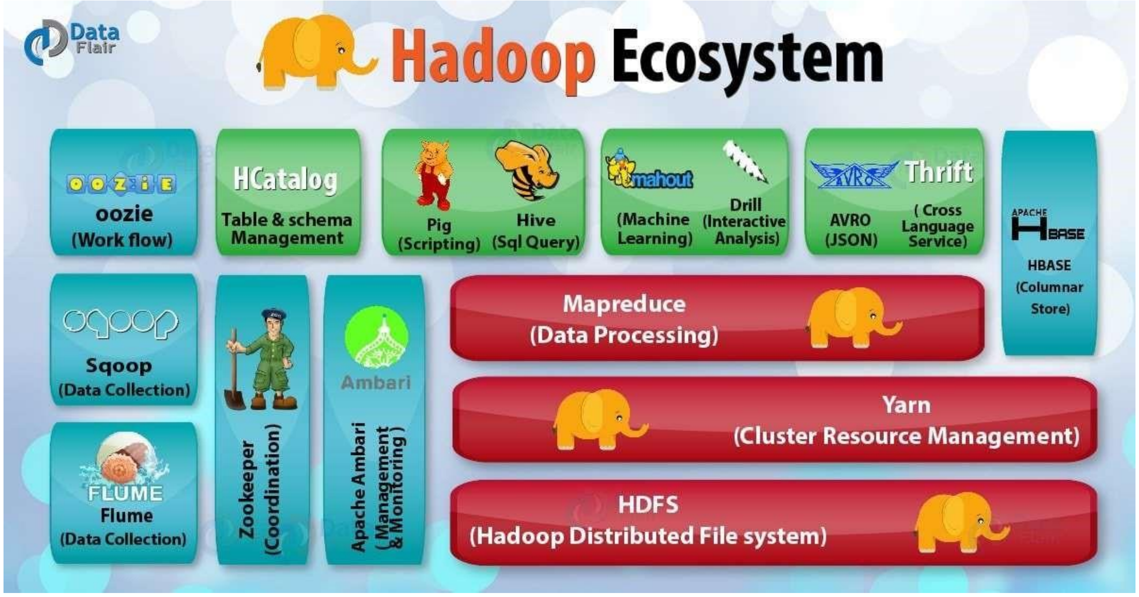
Under the guidance of

Prof. A. A. Chandorkar

DEPARTMENT OF COMPUTER ENGINEERING

Academic Year 2021-22

Hadoop Ecosystem Components



Hadoop Distributed File System

HDFS Components:

There are two major components of Hadoop HDFS- NameNode and DataNode. Let's now discuss these Hadoop HDFS Components-

i. NameNode

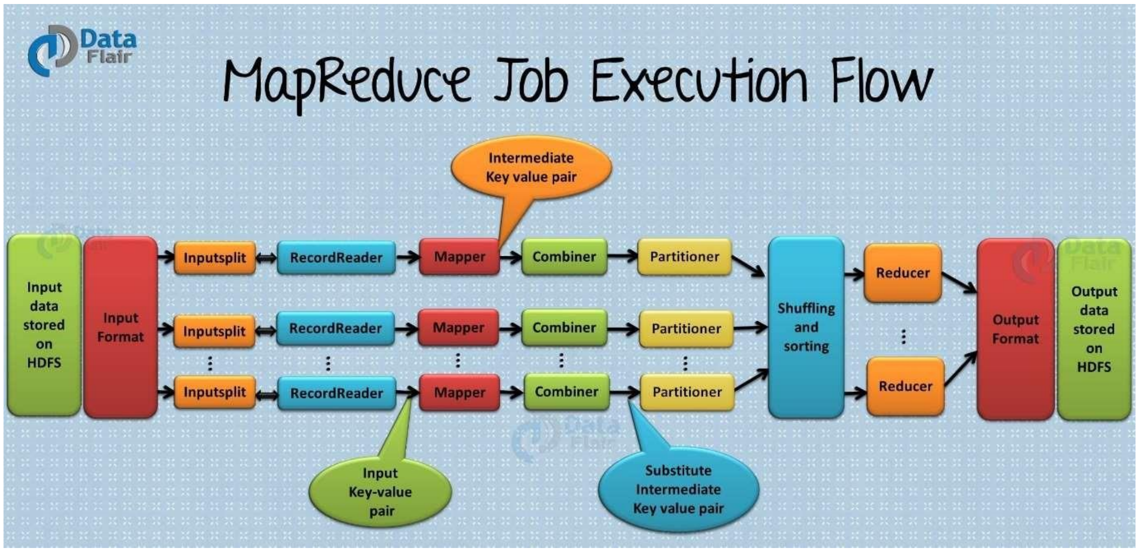
It is also known as Master node. NameNode does not store actual data or dataset. NameNode stores Metadata i.e. number of blocks, their location, on which Rack, which Datanode the data is stored and other details. It consists of files and directories.

ii. DataNode

It is also known as Slave. HDFS Datanode is responsible for storing actual data in HDFS. Datanode performs read and write operation as per the request of the clients. Replica block of Datanode consists of 2 files on the file system. The first file is for data and second file is for recording the block's metadata. HDFS Metadata includes checksums for data. At startup, each Datanode connects to its corresponding Namenode and does handshaking. Verification of namespace ID and software version of DataNode take place by handshaking. At the time of mismatch found, DataNode goes down automatically.

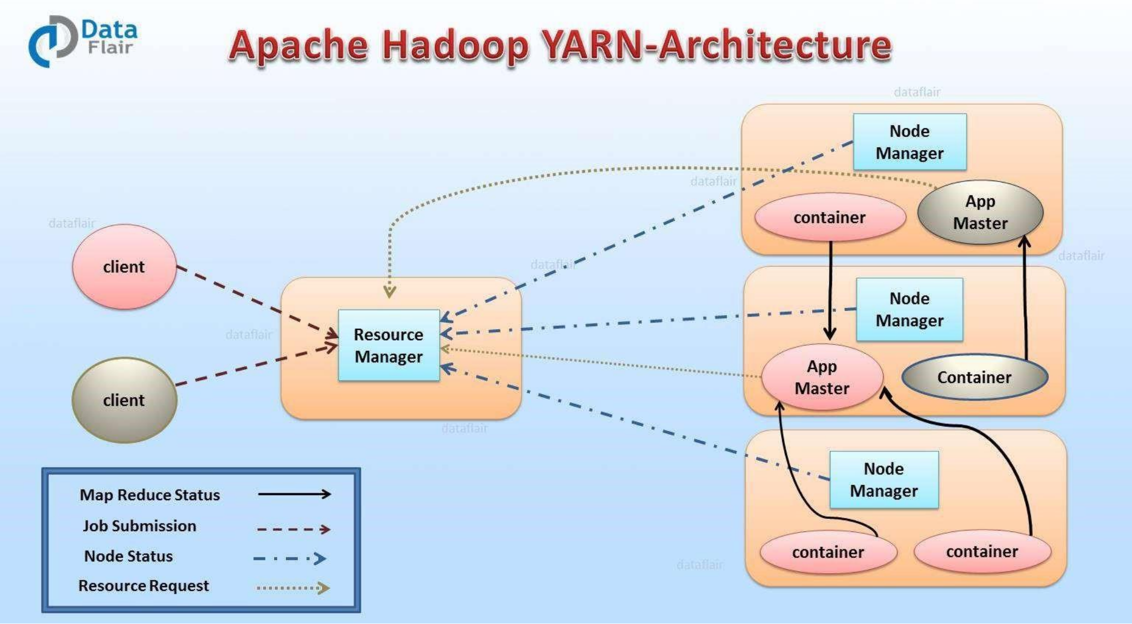
MapReduce

Hadoop MapReduce is the core Hadoop ecosystem component which provides data processing. MapReduce is a software framework for easily writing applications that process the vast amount of structured and unstructured data stored in the Hadoop Distributed File system.



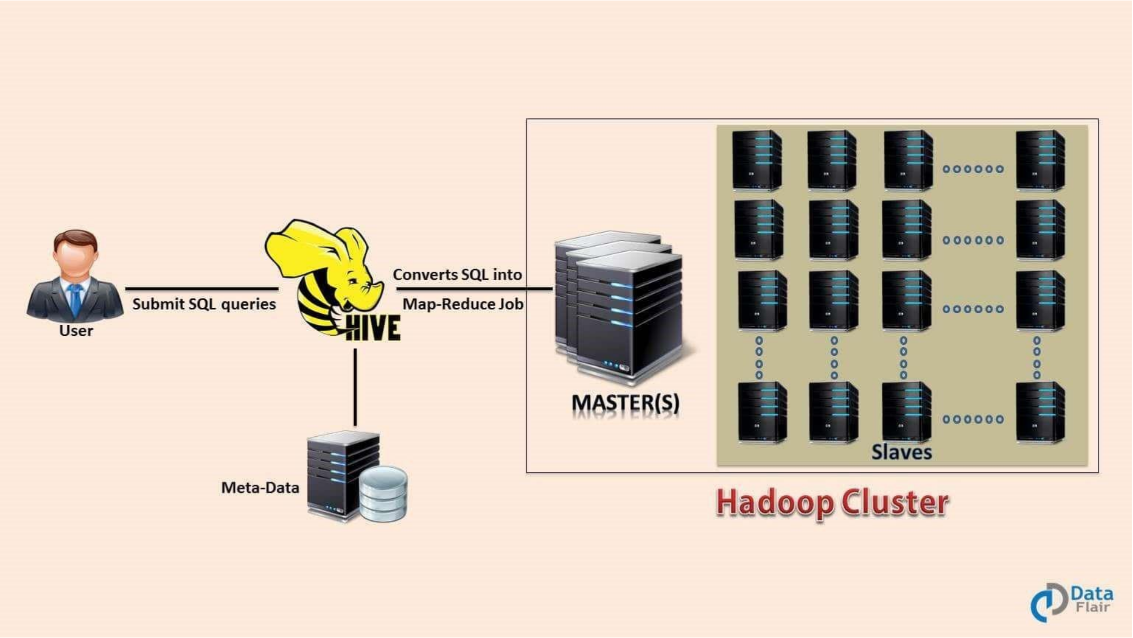
YARN

Hadoop YARN (Yet Another Resource Negotiator) is a Hadoop ecosystem component that provides the resource management. Yarn is also one the most important component of Hadoop Ecosystem. YARN is called as the operating system of Hadoop as it is responsible for managing and monitoring workloads. It allows multiple data processing engines such as real-time streaming and batch processing to handle data stored on a single platform.



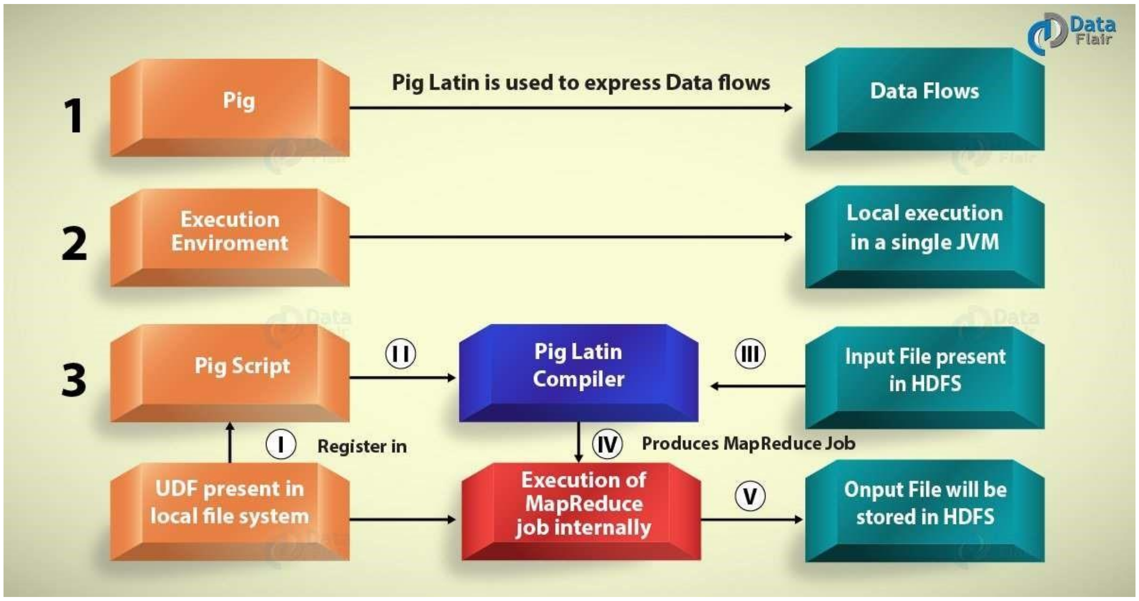
Hive

The Hadoop ecosystem component, Apache Hive, is an open source data warehouse system for querying and analyzing large datasets stored in Hadoop files. Hive do three main functions: data summarization, query, and analysis.



Pig

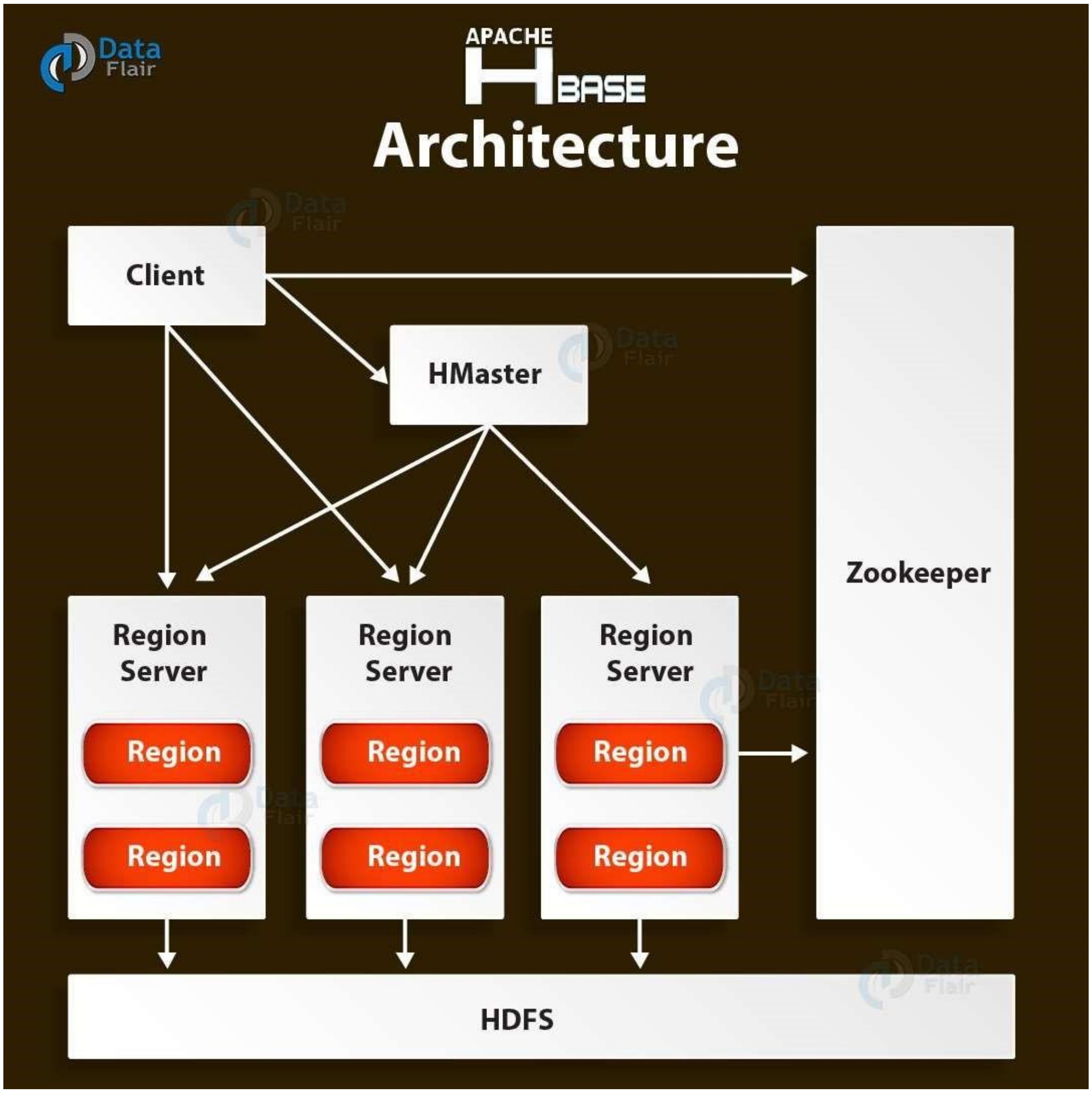
Apache Pig is a high-level language platform for analyzing and querying huge dataset that are stored in HDFS. Pig as a component of Hadoop Ecosystem uses PigLatin language. It is very similar to SQL. It loads the data, applies the required filters and dumps the data in the required format. For Programs execution, pig requires Java runtime environment.



Data Flair
1
Pig
Pig Latin is used to express Data flows
Data Flows
2
Execution Enviroment
Local execution in a single JVM
3
Pig Script
II
Pig Latin Compiler
III
Input File present in HDFS
I
Register in
IV
Produces MapReduce Job
UDF present in local file system
Execution of MapReduce job internally
V
Onput File will be stored in HDFS

HBase

Apache HBase is a Hadoop ecosystem component which is a distributed database that was designed to store structured data in tables that could have billions of row and millions of columns. HBase is scalable, distributed, and NoSQL database that is built on top of HDFS. HBase, provide real-time access to read or write data in HDFS.



Data Driven Digital Marketing - How Hadoop Can Help

Data leads to proper analysis, which in turn leads to more conversions. Every successful marketing strategy relies on data to get the desired results. In today's online world, people use multiple devices to access information and marketers need the right data in order to segment and implement cross-device strategies.

Hadoop - The New Force in the World of Big Data

Hadoop has come as a new force in the world of Big Data. It is worth noting that more than half of the Fortune 50 use Hadoop. The rising pressure of data overload is handled effectively by Hadoop. Companies wish to be data driven, which simply means having a unified view of the customer. As much of the data used by marketers are found in databases, and corporate houses invest a lot in big data warehouses - which are more appropriately referred to as RDBs (Relational Databases) - cheaper data processing is required.

Today’s digital marketing companies need an affordable data management platform that can support petabyte-scale data processing and real-time analytics. We have data routinely popping up in audio, video, images, social media, text, meta data, etc. Handling such a vast amount of data efficiently requires lots of hardware and processing power. Hadoop is the best fit considering this scenario because it uses industry standard hardware, it allows the data to be processed faster and more efficiently, and the cost of storage is cheaper than a relational data warehouse system. Large corporate houses like Facebook and Yahoo use Hadoop as a solution to process large sets of data.

How Hadoop Helps to Scale Processing?

The Apache Hadoop allows for the distributed processing of Big Data across clusters of computers. It consists of four different parts: Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN and Hadoop MapReduce. It splits files into large blocks of data and evenly distributes them across the nodes forming the cluster.

Cloudera CDH is just one example of a scalable and high-performance Hadoop environment. It is the only Hadoop solution that offers unified batch processing, interactive SQL, interactive search, and role-based access controls. The basic framework is built on the assumption that data easily gets in but it usually doesn’t get out. This helps to quickly and reliably get data into Hadoop, and then proceed to work on solo channel marketing efforts with customers directly from Hadoop.

Hadoop - The Feasible Solution to Data Processing

The crux is that companies need data-driven digital marketing. The Apache Hadoop is a feasible solution to solve complex big data processing problems for digital marketing companies. The more data we have, the better we can analyze; the better we analyze, the faster we make more appropriate decisions in order to generate more profit. You do the math.