ONLINEGUWAHATI™

A division of Irisidea Technologies Pvt Ltd (www.irisidea.com)

irisidea DATA TRANSFORMATION COMPANY

www.irisidea.com

Basic Concepts of Hadoop and its Eco System to process Big Data

A Sharp-Witted Reference Guide for Beginners

In a nutshell, Big Data and Hadoop

Big data can be termed as very large data sets that existing traditional software technologies are unable to store, process, mining, handling etc and eventually fails to uncover the insights with meaning of the underlying data.

The unit of data sets generally start from hundreds of petabyte to Zettabyte and even more than that. Due to exponential growth of digitalization and internet penetration, Big Data is getting generated from multiple sources with different variety of data format and stores in distributed storage layer for processing and analysis.

Hadoop is the most popular software framework developed by Apache community to store and process big data. This is a open source framework where java programming is language used. The entire Hadoop framework can be customized according to the need by downloading entire source code from Apache's repository. Commodity hardware can effectively be utilized to build multi node cluster and install Hadoop framework in order to process different format of extremely large data sets.

Chapter - 1	Basic of Hadoop
Chapter - 2	Hadoop Distributed File System (HDFS)
Chapter - 3	Map Reduce
Chapter - 4	Spark
Chapter - 5	Hadoop Development Environment
Chapter - 6	Hive
Chapter - 7	Flume
Chapter - 8	Hadoop Eco System

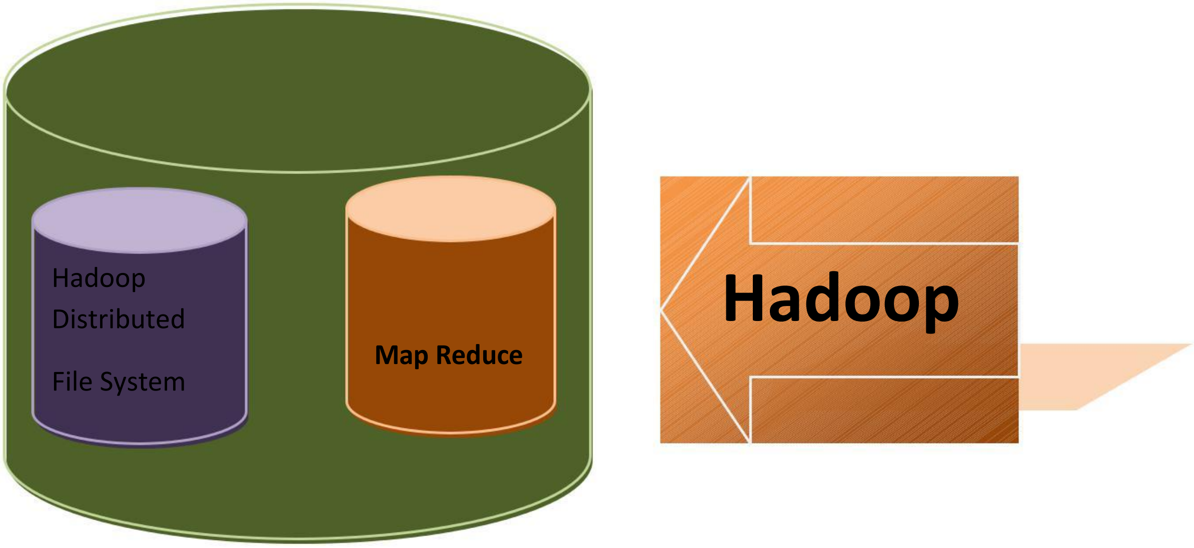
Chapter -1 Basic of Hadoop

Hadoop is a software framework/foundation developed by Apache community. The name Hadoop is not an acronym; it's a man-made. Hadoop was created by Doug Cutting who created Apache Lucene, a text search library. Hadoop has developed using Java programming language. The concept of Hadoop generated from the Google distributed file system , called GFS.

In April 2008, a world record was created to became the fasted system to sort an entire terabyte of data by Hadoop which was running on a 910-node cluster. Yahoo had extensively used Hadoop during that phase to developed their internet-scale search engine since it was need huge volume of data to conducted intricate analysis and testing.

Hadoop is a combination of two major component

- Storage layer (HDFS)
- Processing layer (Map Reduce)

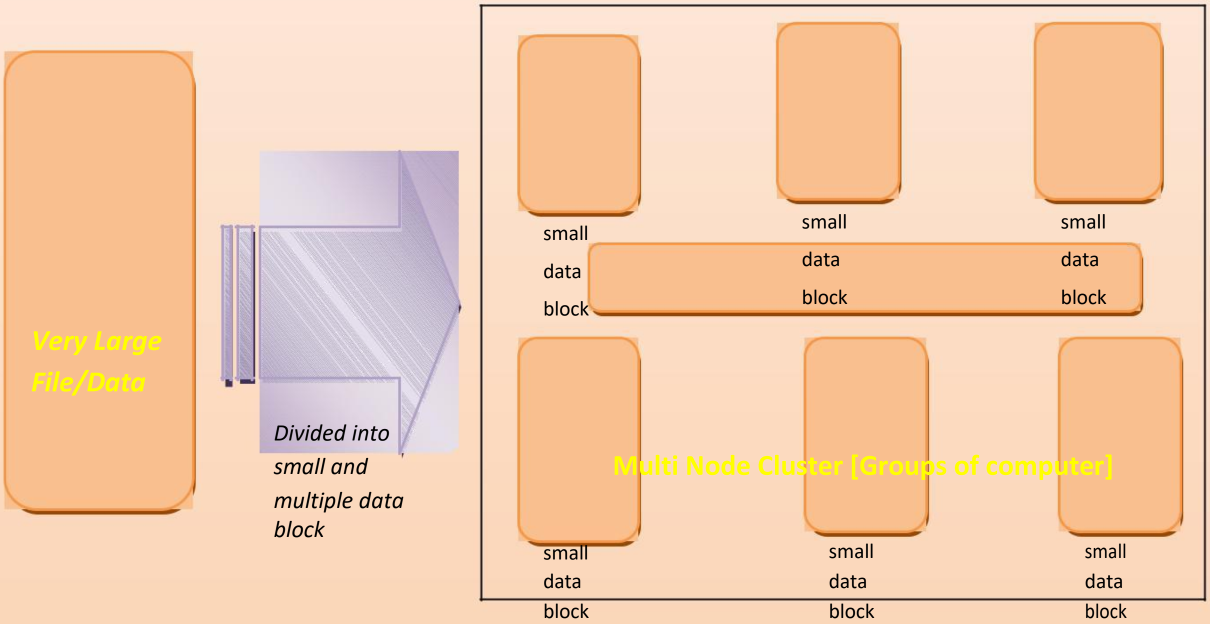


Chapter -2 Hadoop Distributed File System (HDFS)

Hadoop distributed file system (HDFS) is used to store data in a distributed manner across multiple node or computers in cluster. It is a container of holding very large data/files which we term as Big Data.

HDFS don't have any restriction to store data in terms of any format. As we know that due to exponential growth of digitalization around the globe, 80 percent data are now produced in unstructured format. Email, Blog, Messages etc are coming under unstructured data format category and can't be stored in available traditional storage system or database.

For analysis, mining, machine learning etc, unstructured data can be loaded in HDFS without any preprocessing which in fact is beyond imagination with available traditional Relational Database Management System (RDBMS), Data Warehousing System.



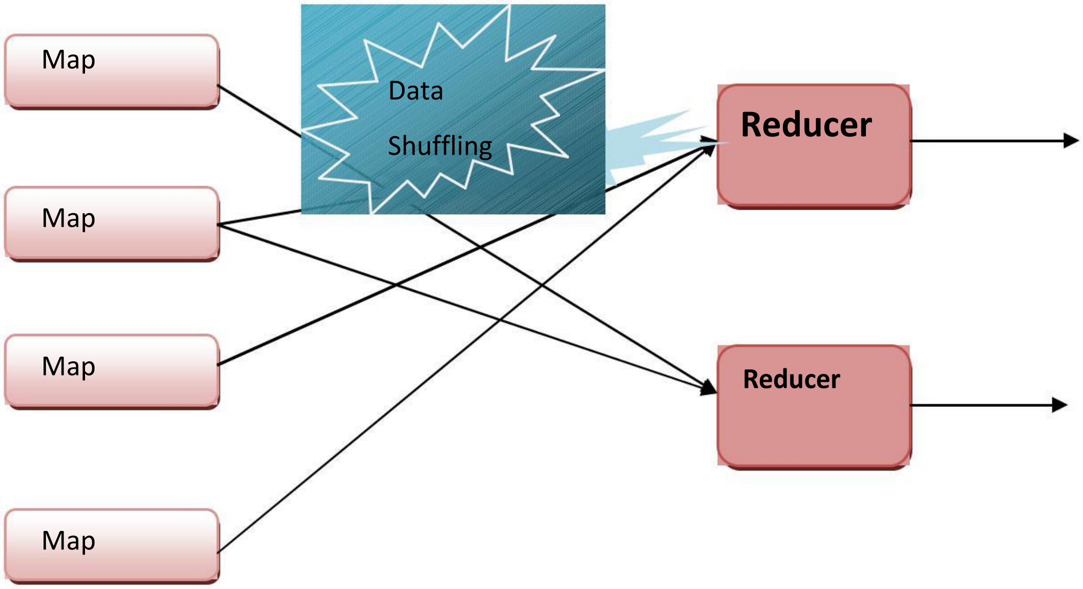
Chapter - 3 Map Reduce

Map Reduce belongs to processing layer software component in Hadoop. This framework can process very vast amount of data (multi-terabyte data-sets) in-parallel and in a distributed manner on large cluster (thousand node/computers).

Once huge volume of data files get spitted into multiple or small data blocks inside HDFS, Map Reduce starts processing each block in parallel to get the desire output.

Commodity hardware is sufficient to execute Map Reduce programming . It works by breaking the processing into two phases. Map phase as well as Reduce phase. Both the phases accepts input and emit output data in key value pair.

Data shuffling takes place over network on the entire cluster once Map phase starts emitting output and those will be input data to the Reducer phase. Reducer is responsible to produce the final output result and it will be stored again in HDFS.

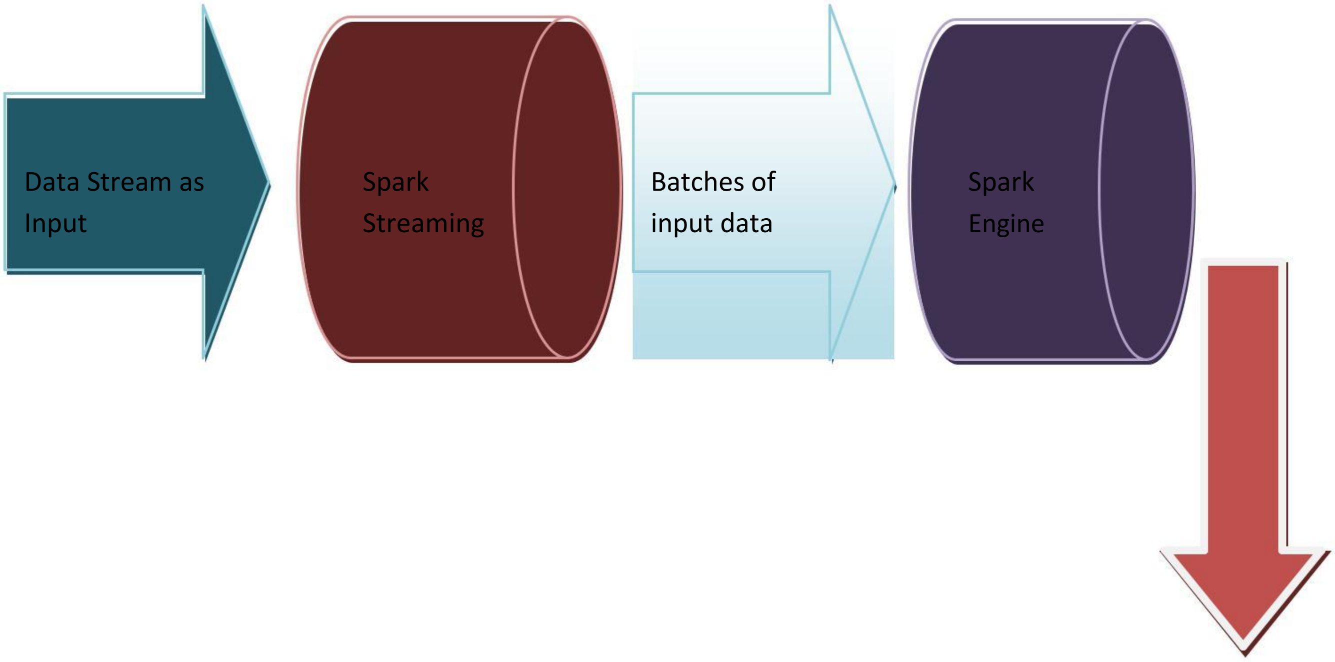


Chapter - 4 Spark

Open-source cluster-computing framework from Apache community. Originally developed at the University of California, Berkeley's AMPLab. Unlike Map Reduce in Hadoop, Spark cannot execute alone. ***Spark is not belongs to Hadoop framework.***

Apache's spark is famous to process data when it is on motion. It is excellent and very efficient framework to process and analyze real time streaming data. Advanced DAG execution engine that supports acyclic data flow and in-memory computing makes Spark very fast and general engine for large scale data processing.

Apache Spark requires a cluster manager and a distributed storage system. Hadoop distributed file system (HDFS) can be effectively utilized to store huge volume of data that Spark need.

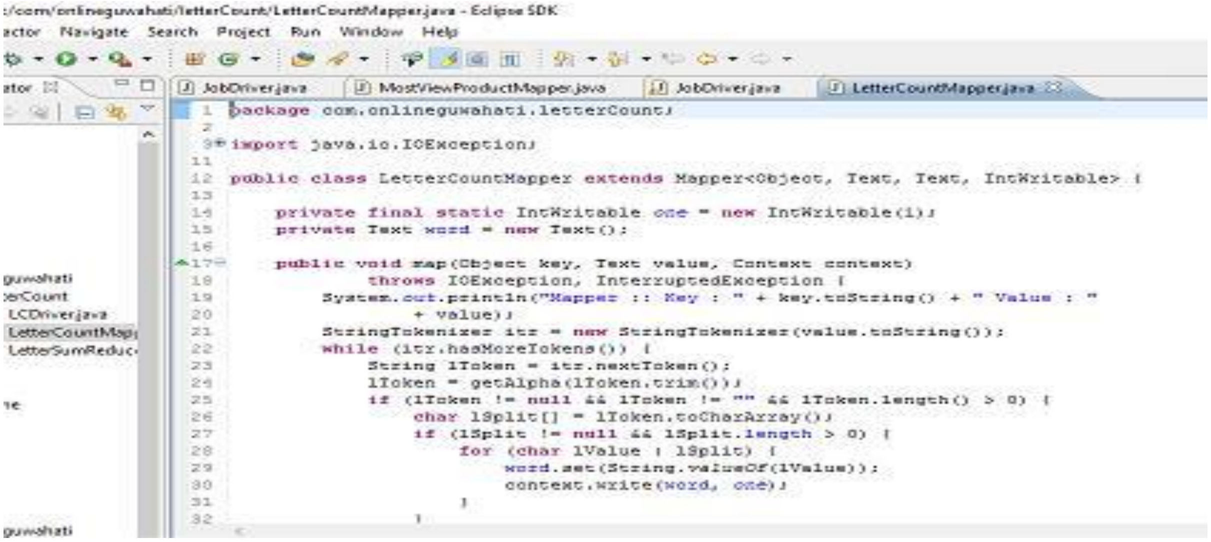


Chapter - 5 Hadoop Development Environment

There are multiple option to create Hadoop development environment. Using commodity hardware multi node cluster can be created and based on the volume of data, we can scale the cluster linearly. For installation and maintenances, a dedicated IT team is mandatory. This kind of environment typically needs to be provided by the organizations/companies who are involve commercially to process big data and helping other companies/organization to take strategic critical business decisions.

Yahoo, Facebook, Google, MapR, Pivotal etc. have their own development as well as production environment.

Leveraging cloud, we can have Hadoop development environment. This model is pay per use basis. Giant cloud server provider like Amazon, IBM, Microsoft's Azure etc offer the environment as a SAAS model (Software as a service). There are no constraints to scale up the cluster, resource allocation, choosing operating system etc in each node of the designed cluster. Besides, all the software licensing, hardware configurations, antivirus software etc are completely managed by the vendors. The subscribers have only to pay according to the use of services.



:/com/onlineguwahati/letterCount/LetterCountMapper.java - Eclipse SDK
actor Navigate Search Project Run Window Help
ator
JobDriver.java
MostViewProductMapper.java
JobDriver.java
LetterCountMapper.java
guwahati
erCount
LCDriver.java
LetterCountMap
LetterSumReduc
ne
guwahati

```
package com.onlineguwahati.letterCount;

import java.io.IOException;

public class LetterCountMapper extends Mapper<Object, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context)
            throws IOException, InterruptedException {
        System.out.println("Mapper :: Key : " + key.toString() + " Value : "
                + value);
        StringTokenizer itr = new StringTokenizer(value.toString());
        while (itr.hasMoreTokens()) {
            String lToken = itr.nextToken();
            lToken = getAlpha(lToken.trim());
            if (lToken != null && lToken != "" && lToken.length() > 0) {
                char lSplit[] = lToken.toCharArray();
                if (lSplit != null && lSplit.length > 0) {
                    for (char lValue : lSplit) {
                        word.set(String.valueOf(lValue));
                        context.write(word, one);
                    }
                }
```

For the students who are aspiring to learn and practice Hadoop, we can install and configure a development environment of single node Hadoop cluster where Map/Reduce programming can be practiced. We can effectively use this virtual machine on top of Windows operating system where Ubuntu 14.X can be installed and subsequently Hadoop binaries.

Chapter - 6 HIVE

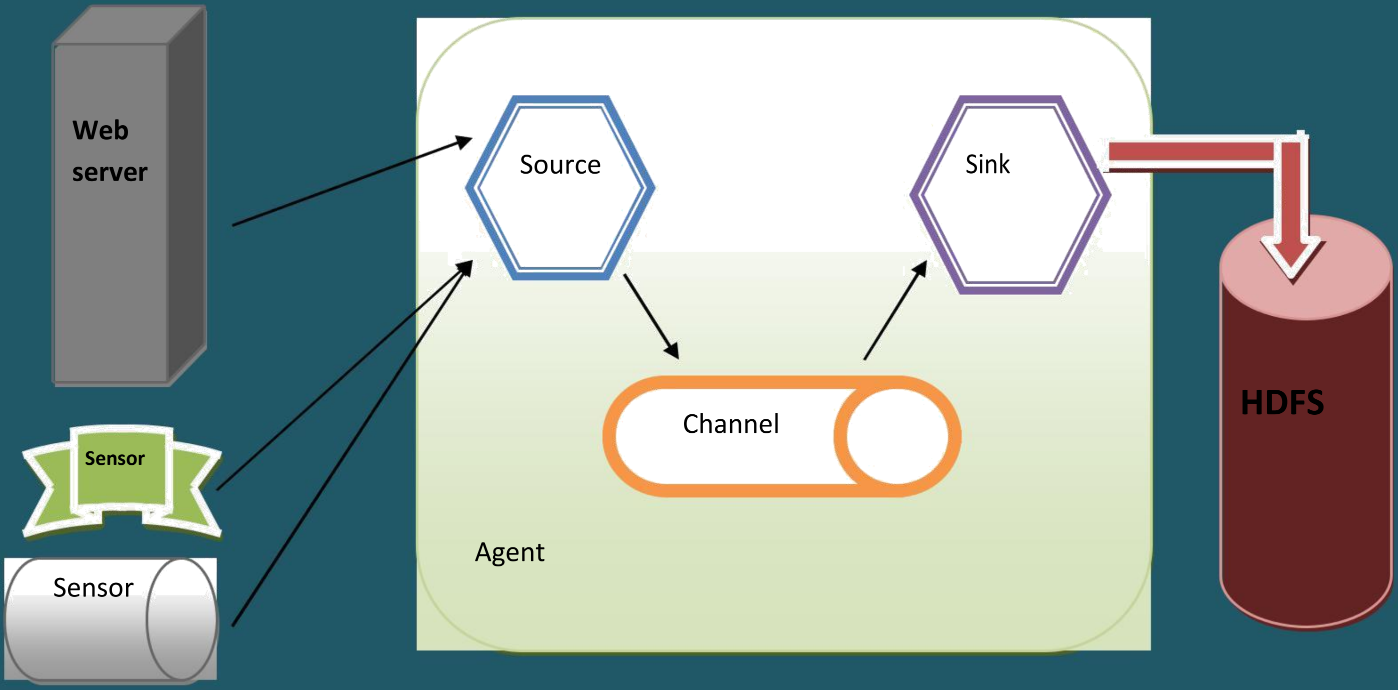
Hive can be considered as one of the important ingredients in the Information platform. Today's era being the data era, we should have the mechanism/software to facilitate reading, writing, and managing large datasets residing in distributed storage. Using SQL and Hive is the best option for it

Apache's Hive is a framework for data warehousing on top of Hadoop. It is designed for easy and effective data aggregation, ad-hoc querying and analysis of huge volumes of data. Hive is cost effective and addresses all the scalability requirements. It was initially developed by Facebook to run queries on the huge volume of data and HDFS was responsible to store all the data.

Hive is not designed for online transaction processing. It has language capabilities similar to SQL and it is called HQL (Hive Query Language). HQL supports many built-in functions and operators. HQL queries are internally converted to Map Reduce job and executed over HDFS in a distributed manner and produced the desired result. Hive has a flexibility to write and execute UDF (User Define Function).

Chapter - 7 Flume

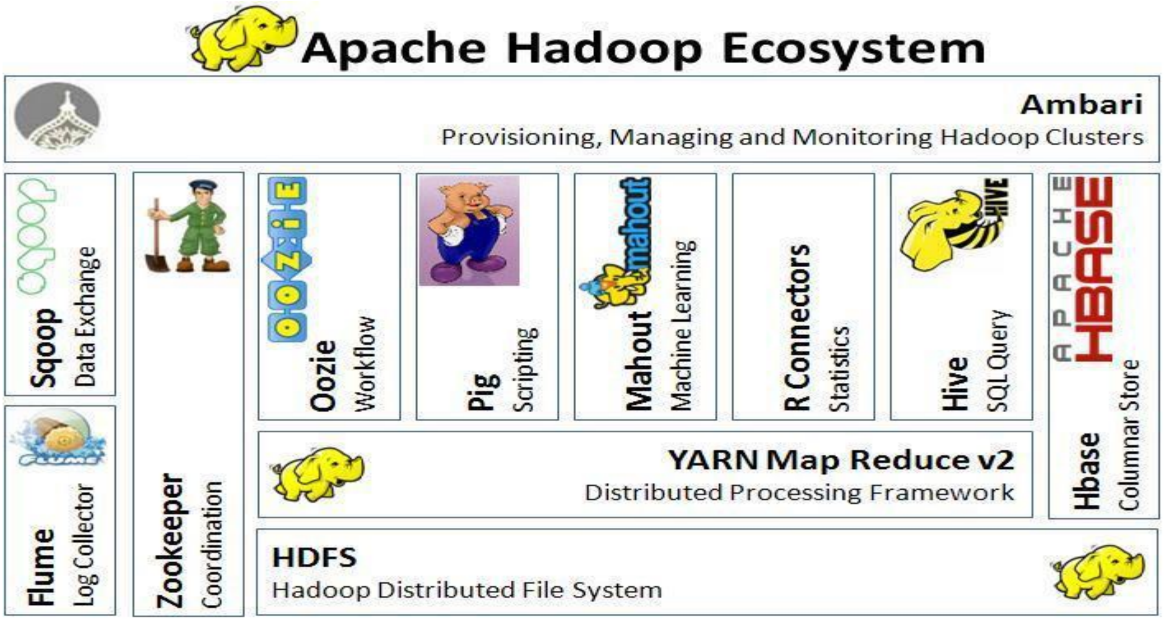
Apache's Flume is another important component which belongs to Hadoop eco system. After identification of Big Data generation source like web server log, various sensor's output or events etc, Flume can be connected to transfer those data continuously to HDFS for persistence in a distributed manner.



Flume is efficiently designed for moving bulk quantities of streaming data into HDFS. It is reliable, distributed and render service for collecting and aggregating log files generated in the web/application servers, network traffic data.

Besides, Flume can be used to transport data in any format like structured, semi-stuctrued and unstructred generated from the social media, email messages etc. Twitter streaming data can be ingested into HDFS using Flume.

Chapter - 8 Hadoop Eco System.



As represented in the above diagram, there are other components or tools consolidated in the Eco system those addresses different functionality.

Flume :- Collects log data and dump to HDFS

Sqoop:- Exchange data between HDFS and RDBMS, also in reverse direction.

Zookeeper:- To coordinating among the nodes in the cluster.

Oozie:- Workflow scheduler system to manage Apache Hadoop jobs

Pig:- A scripting programming tool developed by Yahoo.

Mahut:- Develop scalable machine learning algorithms

R Connectors:- A statistical tool

HBase:- NoSQL database and natively integrated with HDFS.

YARN:- Responsible to split up the functionalities of resource management and job scheduling/monitoring into separate daemons.

Ambari:- Provisioning, managing and monitoring component for Hadoop cluster