Practical No:1

Aim: Identify Any OSS AND CREATE DETAIL REPORT ABOUT IT.



1}Idea:

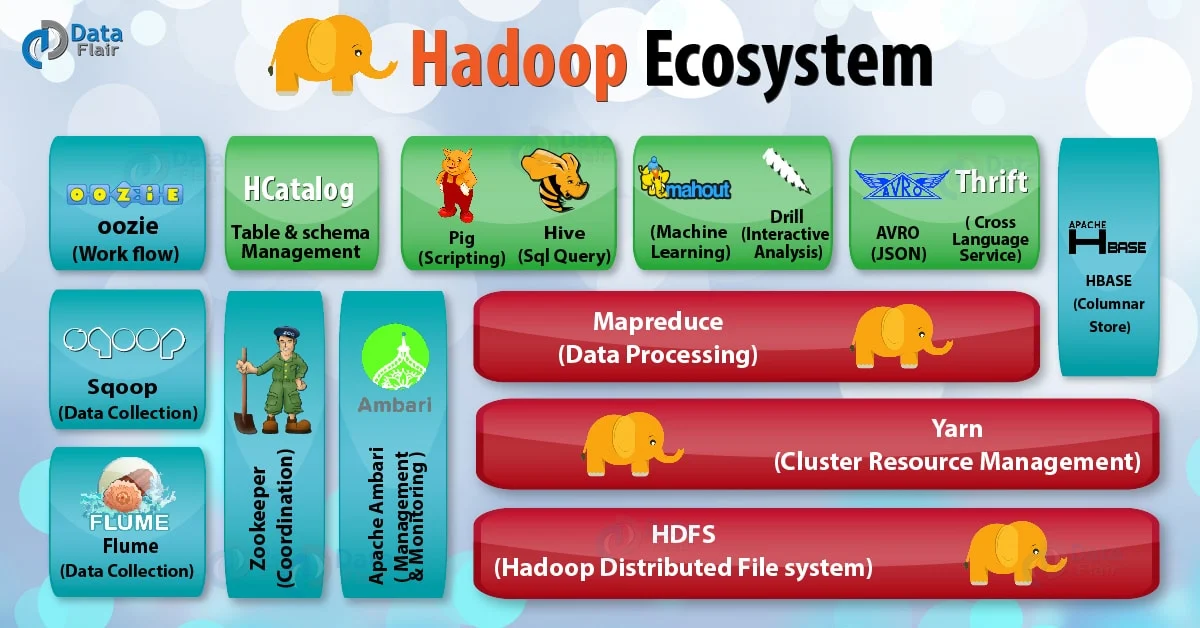
Apache Hadoop is **an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data**. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

According to its co-founders, [Doug Cutting](https://en.wikipedia.org/wiki/Doug_Cutting) and [Mike Cafarella](https://en.wikipedia.org/wiki/Mike_Cafarella), the genesis of Hadoop was the Google File System paper that was published in October 2003. This paper spawned another one from Google – "MapReduce: Simplified Data Processing on Large Clusters". Development started on the [Apache Nutch](https://en.wikipedia.org/wiki/Apache_Nutch) project, but was moved to the new Hadoop subproject in January 2006.[]](https://en.wikipedia.org/wiki/Apache_Hadoop#cite_note-19) Doug Cutting, who was working at [Yahoo!](https://en.wikipedia.org/wiki/Yahoo!) at the time, named it after his son's toy elephant. The initial code that was factored out of Nutch consisted of about 5,000 lines of code for HDFS and about 6,000 lines of code for MapReduce.

In March 2006, Owen O'Malley was the first committer to add to the Hadoop project; Hadoop 0.1.0 was released in April 2006 It continues to evolve through contributions that are being made to the project The very first design document for the Hadoop Distributed File System was written by Dhruba Borthakur in 2007.

2}What Problem does it solve?

Hadoop consists of four main modules:

* Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.
* Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.
* MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.
* Hadoop Common – Provides common Java libraries that can be used across all modules.
* 

Licencing model:

Apache Hadoop

|  |  |
| --- | --- |
| **Original author(s)** | **Doug Cutting, Mike Cafarella** |
| Written in | Java |
| Operating system | Cross-platform |
| Type | Distributed file system |
| License | Apache License 2.0 |

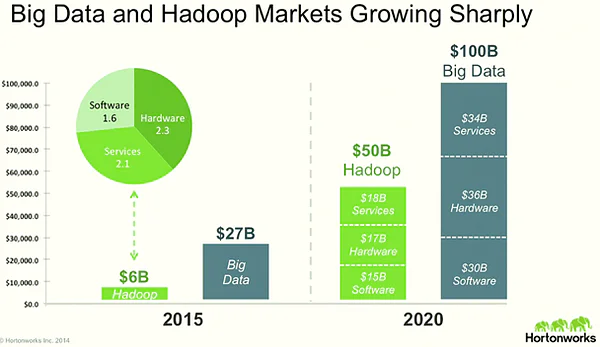
4}Intent Behind Making it opensource?

The Hadoop ecosystem is a comprehensive, well-organized arrangement that makes big data analysis simple and accurate. The Hadoop ecosystem comprises several software applications, each specializing in a specific task. However, while the entire ecosystem is a combination of software tools, each of the tools in itself is capable of doing a specialized job independently. This means that you can pick and choose the specific tools needed to fulfill your purpose — Hadoop is that flexible. Hadoop does not bind you by rules that compel you to use the software in a certain way. You can use the source code in any manner you like.

Let's take a look at an overview of how the Hadoop ecosystem works and also how it embraces the open-source principles along the way. Let's start with a basic definition of Hadoop. According to IBM, “Apache Hadoop is an open source software project that enables [distributed processing](https://www.techopedia.com/definition/3351/distributed-processing) of large [data sets](https://www.techopedia.com/definition/3348/data-set-ibm-mainframe) across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with very high degree of [fault tolerance](https://www.techopedia.com/definition/3362/fault-tolerance). Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the [application layer](https://www.techopedia.com/definition/6006/application-layer).”

5}Monetizing model:

Apache is a 501(c)3 charity, not a company.Instead, it makes its money from donations from the millions of users it has, especially bigger corporations. And this money funds the development of the Apache software and multiple other projects they produce.For those not familiar with the US tax system, donations made to a 501(c)3 charity are tax-deductible. Essentially, companies get to donate money and write it off and in turn benefit from the software produced.It’s a win-win situation for everyone



6}Popularity:

Apache Hadoop is the most popular and powerful big data tool, Hadoop provides the world’s most reliable storage layer. In this section of the features of Hadoop, let us discuss various key features of Hadoop.

1. Hadoop is Open Source

2. Hadoop cluster is Highly Scalable

3. Hadoop provides Fault Tolerance

4. Hadoop provides High Availability

5. Hadoop is very Cost-Effective

6. Hadoop is Faster in Data Processing

7. Hadoop is based on Data Locality concept

8. Hadoop provides Feasibility

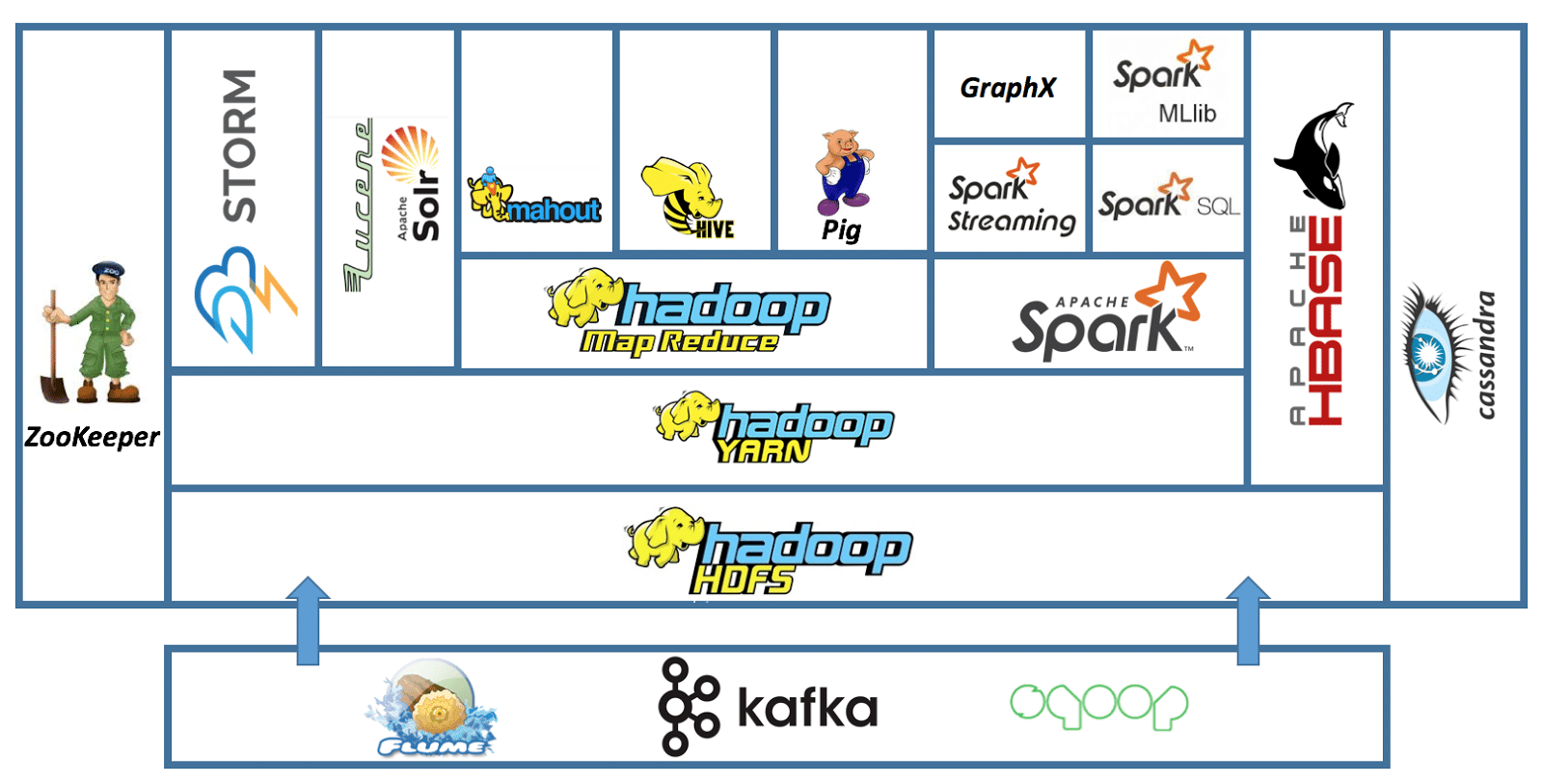
9. Hadoop is Easy to use

10. Hadoop ensures Data Reliability

Impact:

Since the crash of 2008, the importance of big data in finance has surged. The demand for big data has increased to a staggering proportion. This caused many industry experts to examine the connections between regulatory changes in finance and the need for the kind of solutions offered by Hadoop.

Business sectors are heavily impacted by big data. This include finance, [healthcare](https://www.chetu.com/healthcare.php), marketing and technology. Decisions are made based on both the quality as well as the quantity of information that is available at the time. Organizing information is an analytical process, and the difficulty of performing this function increases with volume. This particular problem leads developers to come up with a solution to address this specific issue.



**Hadoop developers recognize what is called the four V’s of big data:**

**Volume:** Around 2.5 quintillion bytes of data are generated every single day. This volume is a relatively recent phenomenon; around 90 percent of the data in the entire world was created in only a matter of years.

**Variety:** The type of data collected must be stored and analyzed in order to have any kind of meaning or application for businesses.

**Velocity:** This relates to the speed that is required of the hardware to store and analyze the date. This is a critical element for financial institutions involved in trading, for example.

**Veracity:** The content of the data must be analyzed in order to be applied. The veracity refers to the accuracy of the information, which requires complex processes to sort factual data considering the sheer volume of information involved.

