INTERNATIONALJOURNALOFRESEARCH SCIENCE& MANAGEMENT

A STUDY OF HADOOP ECOSYSTEM

Ms.Prachi Dhaulakhandi*
* Assistant Professor Dept. Of I.T.,C.T.Girls PG College, Kashipur(U.S.Nagar)

Keywords: e-learning, Hadoop Ecosystem

Abstract

Hadoop rose around 10 years ago with the aim to improve Yahoo's search engine. Hadoop was created by Doug Cutting and Mike Cafarella and was named after a toy elephant belonging to Doug Cutting's son. Also Hadoop has its origin in Apache Nutch, an open source web search engine.

Introduction

History

Hadoop rose around 10 years ago with the aim to improve Yahoo's search engine. Hadoop was created by Doug Cutting and Mike Cafarella and was named after a toy elephant belonging to Doug Cutting's son. Also Hadoop has its origin in Apache Nutch, an open source web search engine. In October 2003 Google released paper on the Google File System, a scalable distributed file system paper for large distributed data-intensive applications and in December 2004 they released paper on MapReduce, a programming model and an associated implementation for processing and generating large data sets. Doug Cutting and team realized that the Google was generalizing the processes into a framework that automated steps that they were doing manually and that prove very revelatory to them. Cutting and Cafarella over few months built up the underlying file systems and processing framework that became Hadoop and ported Nutch on top of it. [CESY11] [GIGA13]

Yahoo which was equally impressed with Google File System and MapReduce wanted to build open source technologies based on them. Then Cutting joined Yahoo and they spun out the storage and processing parts of Nutch to form Hadoop as an open source Apache Software Foundation project. Hadoop initially supported 5 to 20 nodes and slowly the Hadoop's scalability increased from single digit nodes to an ultimate thousands of nodes. Yahoo invested heavy resources into the project which made Hadoop technology to mature and slowly Hadoop rolled into power of analytics for various production applications. Later Yahoo merged its search and advertising into one unit with the help of Hadoop and it implemented some scheduling changes within Hadoop that tackle security and workflow. [GIGA13]

Definition

Hadoop is 100% open source framework that enables distributed parallel processing of huge amounts of data across standard servers that store and process the data. It is written in Java with some code in C and command line utilities written in Shell script. Hadoop consists of a storage part called Hadoop Distributed File System (HDFS) and a processing part called MapReduce. It distributes the files which are split into large chunks across nodes to process the data. This allows the data to be transferred faster and more efficiently than other conventional methods. Hadoop is widely used by companies like Facebook, Twitter, Linkedln, Yahoo!. [CESY11] [DRIV14]

Big data and need for Hadoop

Big data is simply a way of describing data which grew far beyond the ability of traditional systems to handle it and the problems that were unsolvable by traditional systems. The factors behind big data challenge can be categorized in terms of "the three Vs of big data". [IJSR14]

- *Volume:* Dozens of terabytes of data. The reasons behind these increasing volume of data are increase in automated processes, increase in interconnected systems and increase in number of people living online.
- *Variety:* This refers to the organization of data in unstructured, semi structured, and structured data category. Some datasets even include multi-structured data.
- *Velocity:* Some data that enters the organization has a limited time window before that data can be transformed and stored into a data warehouse for further analysis. The velocity challenge increases with high volume of data. [IJSR14]

So if the data storage has any of the above characteristics, then we can say we have a big data challenge. Hadoop is designed for above 3Vs. It is a technology tool that is suited for high volumes of data and data structures. But Hadoop does not fully resolve the problems with data velocity challenges as it does not deal with data in motion. [IJSR14]

Hadoop installation

Installation on Linux for Single Node Cluster

Supported Platforms

- GNU/Linux/Windows is supported as a development and production platform

Required Software

- Java 7 or late version of Java 6 must be installed (OpenJDK or JDK/JRE).
- ssh must be installed and sshd must be running

```
$ sudo apt-get install ssh
$ sudo apt-get install rsync
```

Installing Software

- Download Hadoop from Apache Download Mirrors
  http://www.trieuvan.com/apache/hadoop/common/
  Also check downloaded copy for tampering using GPG or SHA-256, details at *http://hadoop.apache.org/releases.html*

- Prepare to start Hadoop cluster
  - Unpack the downloaded Hadoop distribution. In the distribution, edit the file etc/hadoop/hadoop-env.sh to define some parameters as follows:

```
# set to the root of your Java installation
export JAVA_HOME=/usr/java/latest
```

```
$ bin/hadoop
```

Now Hadoop cluster can be started in one of the three supported modes:

- Local Standalone Mode
- Pseudo-Distributed Mode
- Fully-Distributed Mode

- Standalone Operation
  Hadoop is configured to run in a non-distributed mode by default as a single Java process.

- Psuedo-Distributed Operation where Hadoop runs on a single-node. Hadoop daemon runs as a separate Java process.
  - Configuration
    Use following

etc/hadoop/core-site.xml:

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

etc/hadoop/hdfs-site.xml:

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
```

```
</property>
</configuration>
```

- Setup passphraseless ssh

  ```
  $ ssh localhost
  ```

- Execution of job on MapReduce locally
  - Format Filesystem

    ```
    $ bin/hdfs namenode -format
    ```

  - Start NameNode daemon and DataNode daemon

    ```
    $ sbin/start-dfs.sh
    ```

  - Browse the web interface for the NameNode

    http://localhost:50070/
  - Make the HDFS directories

    ```
     $ bin/hdfs dfs -mkdir /user
    $ bin/hdfs dfs -mkdir /user/<username>
    ```

  - Copy the input files into the distributed filesystem

    ```
    $ bin/hdfs dfs -put etc/hadoop input
    ```

  - Run an example and check output from distributed filesystem

    ```
    $ bin/hadoop jar share/hadoop/mapreduce/Hadoop-
    mapreduce-examples-2.7.1.jar grep input output
    'dfs[a-z.]+'

     $ bin/hdfs dfs -get output output
    $ cat output/*

    $ bin/hdfs dfs -cat output/*
    ```

  - Stop the daemon

    ```
    $ sbin/stop-dfs.sh
    ```

Conclusion

Hadoop is powerful because it is extensible and it is easy to integrate. Its popularity is due to its ability to store, analyze and access large amounts of data. It is cost effective as it scales across clusters of commodity hardware. Hadoop is not actually a single product but instead a collection of several components. When all these components are merged, it makes Hadoop very user friendly. But there are some limitations which are not immediately obvious but should be considered. But implementing Hadoop gives users amazing amount of tools and resources that allow them to truly personalize their big data experience to fit to their business need.

References

[1] [CESY11] White, Tom. "Hadoop: The Definitive Guide" United States*:* 2nd edition, O'Reilly Media, October 2011

[2] http://ce.sysu.edu.cn/hope/UploadFiles/Education/2011/10/201110221516245419.pdf

[3] [DRIV14] DeRoos, Dirk. "Hadoop for Dummies" Canada: John Wiley & Sons, 2014.

[4] https://drive.google.com/file/d/0B6KicIDH0ZKqOUNwdjlOVXFhams/edit?pli=1

[5] [GIGA13] "The history of Hadoop: From 4 nodes to the future of data", March 2013, Last access: 2015/25/11

[6] https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/

[7] [IJSR14] Bhosale, Harshawardhan. & Gadekar, Devendra. "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications , Volume 4, Issue 10 , October 2014, Last access: 2015/25/11

[8] http://www.ijsrp.org/research-paper-1014/ijsrp-p34125.pdf