Experiment No.:06

Title: Installation and Configuration of Hadoop & Eucalyptus. Develop map reduce application using Hadoop cluster setup (Single node and multimode).

(Example URL patterns count and other)

Objectives: From this experiment, the student will be able to

- To make the students understand use of Hadoop & Eucalyptus.
- To learn the implement & develop applications in Hadoop cluster.

Theory:

Introduction:

Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are four major elements of Hadoop i.e. HDFS, MapReduce, YARN, and Hadoop Common.

Hadoop is designed to scale up from a single computer to thousands of clustered computers, with each machine offering local computation and storage. In this way, Hadoop can efficiently store and process large datasets ranging in size from gigabytes to petabytes of data.

Four modules comprise the primary Hadoop framework and work collectively to form the Hadoop ecosystem:

Hadoop Distributed File System (HDFS): As the primary component of the Hadoop ecosystem, HDFS is a distributed file system that provides high-throughput access to application data with no need for schemas to be defined up front.

Yet Another Resource Negotiator (YARN): YARN is a resource-management platform responsible for managing compute resources in clusters and using them to schedule users' applications. It performs scheduling and resource allocation across the Hadoop system.

MapReduce: MapReduce is a programming model for large-scale data processing. Using distributed and parallel computation algorithms, MapReduce makes it possible to carry over processing logic and helps to write applications that transform big datasets into one manageable set.

Hadoop Common: Hadoop Common includes the libraries and utilities used and shared by other Hadoop modules.

All Hadoop modules are designed with a fundamental assumption that hardware failures of individual machines or racks of machines are common and should be automatically handled in software by the framework. The Apache Hadoop MapReduce and HDFS components were originally derived from Google MapReduce and Google File System (GFS) papers.

Beyond HDFS, YARN, and MapReduce, the entire Hadoop open source ecosystem continues to grow and includes many tools and applications to help collect, store, process, analyze, and manage big data. These include Apache Pig, Apache Hive, Apache HBase, Apache Spark, Presto, and Apache Zeppelin.

What are the benefits of Hadoop?

Fault tolerance

In the Hadoop ecosystem, even if individual nodes experience high rates of failure when running jobs on a large cluster, data is replicated across a cluster so that it can be recovered easily should disk, node, or rack failures occur.

Cost control

Hadoop controls costs by storing data more affordably per terabyte than other platforms. Instead of thousands to tens of thousands of dollars per terabyte being spent on hardware, Hadoop delivers compute and storage on affordable standard commodity hardware for hundreds of dollars per terabyte.

Open source framework innovation

Hadoop is backed by global communities united around introducing new concepts and capabilities faster and more effectively than internal teams working on proprietary solutions. The collective power of an open source community delivers more ideas, quicker development, and troubleshooting when issues arise, which translates into a faster time to market.

MapReduce

MapReduce divides a task into small parts and assigns them to many computers. Later the results are collected at one place and integrated to form the result dataset.
The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pair).
The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into smaller set of tuples.

Steps:

1. Download jdk1.8 for Windows10 from link.
   https://www.oracle.com/java/technologies/downloads/#java8-windows
2. Download Hadoop from given link.
   https://hadoop.apache.org/
3. Download latest version 3.3.0
4. Now install Java SE development Kit 8.
5. Set destination directory to C:/Java and install it
6. Then merge this “Java” folder with the folder that is in program file named “java”.
7. Now we will set environment variable and path.
8. Go to edit environment variable by just search environment variable
9. We have to add new variable
10. Enter variable name and value(path of jdk bin folder)
11. Then edit path variable just by clicking path and add the same path here
12. Java has successfully installed.
13. Now we will install Hadoop

14. First we will extract Hadoop folder and move it to C: drive. Note: Ignore some warning after extraction.
15. Rename it just "hadoop" for convenience
16. Perform some configurations
17. Go to "hadoop" folder and edit 5 files from it
    - core-site.xml
    - hdfs-site.xml
    - mapred-site.xml
    - yarn-site.xml
    - hadoop.env.xml
18. Update core-site.xml with below mentioned code

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

19. Update mapred-site.xml with below mentioned code

```
  - For mapred-site.xml

<configuration>
 <property>
 <name>mapreduce.framework.name</name>
 <value>yarn</value>
 </property>
</configuration>
```

20. Update yarn-site.xml with below mentioned code

```
  -For yarn-site.xml

<configuration>
 <property>
 <name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
 <property>

<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
 </property>
</configuration>
```

21. Create a folder "data" under "C:/Hadoop"
    - Create folder "datanode" under "C:/Hadoop/data"
    - Create folder "namenode" under "C:/Hadoop/data"
    -
22. Update hdfs-site.xml with below mentioned code

```
  -For hdfs-site.xml


<configuration>
```

```
 <property>
<name>dfs.replication</name>
 <value>1</value>
 </property>
 <property>
 <name>dfs.namenode.name.dir</name>
 <value>C:\hadoop\data\namenode</value>
 </property>
 <property>
 <name>dfs.datanode.data.dir</name>
 <value>C:\hadoop\data\datanode</value>
 </property>
</configuration>
```

23. In “hadoop-env.cmd” we just update “JAVA_HOME” with JDK path
24. `Note : JAVA_HOME=C:\java\jdk1.8.0_271`
25. Path variable name “HADOOP_HOME” and value(Pth of bin folder inside hadoop folder)
26. Then edit path in system variable and add both “bin” and “sbin” path here
27. Noe the last step we are going to do is replace original “bin” folder with the bin folder which can download from description
28. Installation and configuration of Hadoop has successfully completed
29. Now we will test word-count task on Hadoop
30. Run “cmd” as Administration and connect to hadoop server using following commands.
    ```
     -> hdfs namenode -format
    ```
31. Now move to hadoop\sbin
    ```
    -> start all
    ```
32. Search “localhost:9870” for confirmation
33. Make input directory in HDFS `(hadoop fs -mkdir/input_dir)`
34. Make input text file in C: drive(you can take any sample text)
35. Copy the input text file in the input directory `(hadoop fs -put C:/input_file.txt /input_dir).`
36. Verify input_file.txt available in HDFS input directoty. `(hadoop fs -ls /input_dir/).`
37. You can verify content. (`hadoop dfs -cat /input_dir/input_file.txt)`
38. Now work for word-count.
39. `(hadoop jar C:/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jarwordcount /input_dir /output_dir).`
40. Verify output. (`hadoop dfs -cat /output_dir/*).`

References:

Download link for Java 8 kit: https://www.oracle.com/java/technologies/downloads/#java8

Download link for Hadoop: https://hadoop.apache.org/

Hadoop configuration codes:
https://drive.google.com/file/d/1M9FvC4U4kvc888zns8AyCmclNUSmXhcf/view

Commands for word count task: https://drive.google.com/file/d/1Hf0B-hlyDQM85dulB-DO2HZqANTTqYKQ/view

Run Wordcount Program on Hadoop-3.3.0 windows 10
https://www.youtube.com/watch?v=nsi4nVS16lc