Data Analytics Assignment 1

Hadoop

Introduction

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Two main components:

- HDFS (Hadoop Distributed File System): This is the storage component of Hadoop, which allows for the storage of large amounts of data across multiple machines. It is designed to work with commodity hardware, which makes it cost-effective.
- YARN (Yet Another Resource Negotiator): This is the resource management component of Hadoop, which manages the allocation of resources (such as CPU and memory) for processing the data stored in HDFS.

History of Hadoop

Hadoop was started with **Doug Cutting and Mike Cafarella** in the year 2002 when they both started to work on Apache Nutch project. **In 2003,** they came across a paper that described the architecture of Google's distributed file system, called **GFS (Google File System)** which was published by Google, for storing the large data sets. **In 2004,** Google published one more paper on the technique **MapReduce**, which was the solution of processing those large datasets. So **in 2006,** Doug Cutting joined Yahoo along with Nutch project. He wanted to provide the world with an open-source, reliable, scalable computing framework, with the help of Yahoo. **In 2007,** Yahoo successfully tested Hadoop on a 1000 node cluster and start using it.

In January of 2008, Yahoo released Hadoop as an open source project to ASF(Apache Software Foundation). And in July of 2008, Apache Software Foundation successfully tested a 4000 node cluster with Hadoop

 Java: Java 8 or higher is required.

 SSH: SSH must be enabled and configured (for communication between nodes in a cluster).

 RAM: Minimum 4GB of RAM (recommended 8GB or more).

 Disk Space: At least 500GB of disk space is recommended for data storage. **Network Requirements**

CPU: Dual-core or higher

- **Bandwidth**: High-speed network connectivity (Gigabit Ethernet recommended).
- **Latency**: Low-latency network for better performance.

Versions

. Hadoop 0.x (Initial Releases)

- **Hadoop 0.1.0 (2006)**
    - Initial release by Doug Cutting and Mike Cafarella.
    - Basic functionalities with MapReduce and HDFS

      Hadoop 1.0.0 (2011

Hadoop 2.x (YARN Introduction)

- **Hadoop 2.0.0-alpha (2012)**
    - Introduction of YARN (Yet Another Resource Negotiator) separating resource management and job scheduling.

- **Hadoop 2.2.0 (2013)**

- First stable release with YARN.
- Improved HDFS performance and federation.

. Hadoop 3.x (Enhanced Features)

- **Hadoop 3.0.0 (2017)**
    - Erasure Coding for HDFS to reduce storage overhead.
    - Support for more than 2 NameNodes (HDFS federation).

 Hadoop 3.1.0 (2018)

- Support for GPU scheduling in YARN.
- Improvements in YARN Timeline Service v.2.

 Hadoop 3.2.0 (2019)

- Native support for ARM architecture.
- Enhanced support for cloud storage (e.g., Amazon S3, Azure Blob Storage).

different versions of hadoop since start, till latest with imp points only for imp ones

ChatGPT

Here's a summary of the major Hadoop versions and their significant features or improvements:

1. Hadoop 0.x (Initial Releases)

- **Hadoop 0.1.0 (2006)**
    - Initial release by Doug Cutting and Mike Cafarella.
    - Basic functionalities with MapReduce and HDFS.

2. Hadoop 1.x (Stable)

- **Hadoop 1.0.0 (2011)**
    - First stable version.
    - Introduced HDFS Federation for better scalability.
    - JobTracker and TaskTracker architecture for MapReduce.

3. Hadoop 2.x (YARN Introduction)

- **Hadoop 2.0.0-alpha (2012)**
    - Introduction of YARN (Yet Another Resource Negotiator) separating resource management and job scheduling.
    - HDFS High Availability (HA) with automatic failover.
- **Hadoop 2.2.0 (2013)**
    - First stable release with YARN.
    - Improved HDFS performance and federation.
- **Hadoop 2.6.0 (2014)**
    - Enhanced YARN resource management.
    - Support for rolling upgrades.
- **Hadoop 2.7.0 (2015)**
    - Support for Windows and improved HDFS.
    - Introduction of NFSv3 access to HDFS.

4. Hadoop 3.x (Enhanced Features)

- **Hadoop 3.0.0 (2017)**
    - Erasure Coding for HDFS to reduce storage overhead.
    - Support for more than 2 NameNodes (HDFS federation).
    - Docker support in YARN.
- **Hadoop 3.1.0 (2018)**
    - Support for GPU scheduling in YARN.
    - Improvements in YARN Timeline Service v.2.
- **Hadoop 3.2.0 (2019)**

    - Native support for ARM architecture.
    - Enhanced support for cloud storage (e.g., Amazon S3, Azure Blob Storage).
- **Hadoop 3.3.0 (2020)**
    - Improved scalability and stability.
    - Enhanced support for containerized applications.
- **Hadoop 3.3.1 (2021)**
    - Various bug fixes and performance improvements.
- **Hadoop 3.3.2 (2021)**
    - Minor enhancements and critical bug fixes.

5. Hadoop 3.4.0 (2022)

- Latest stable release as of now.
- Continued improvements in performance, security, and compatibility.

Steps:

Install Java

Hadoop requires Java 8 or higher.

 Download and Install Java: Download the JDK from Oracle and install it.

 Set JAVA_HOME:

- Right-click on This PC or Computer on your desktop and select Properties.
- Click on Advanced system settings and then Environment Variables.
- Click New under System variables and add:
    - **Variable name:** JAVA_HOME
    - **Variable value:** C:\Program Files\Java\jdk-11.x.x (adjust path if necessary)

Download Hadoop:

Go to the Apache Hadoop releases page and download the latest stable release.

On Windows, extract the .tar.gz file using WinRAR or 7-Zip and move it to C:\hadoop.

 Right-click on This PC or Computer and select Properties.

 Click Advanced system settings and then Environment Variables.

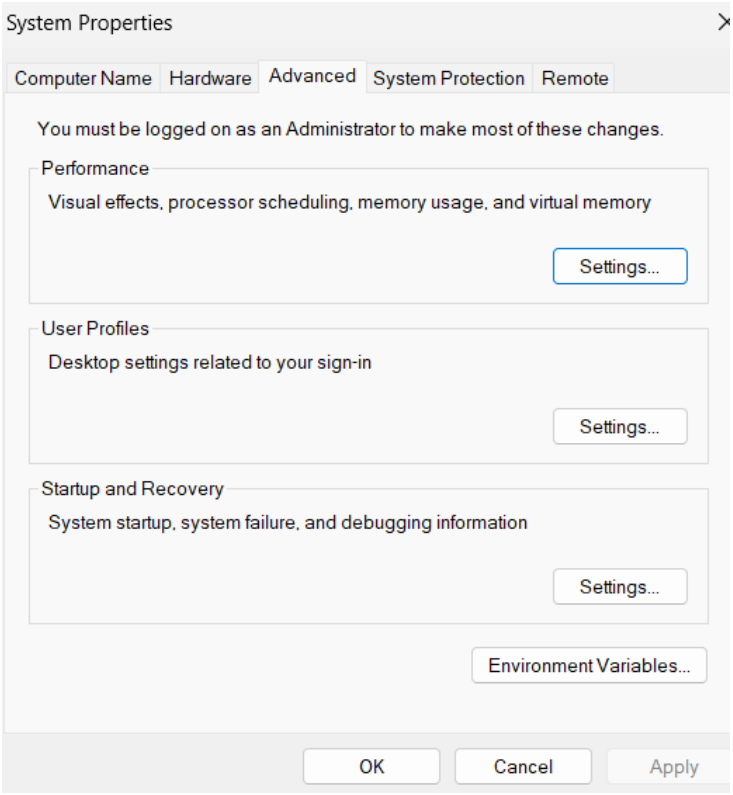
 Click New under System variables and add:

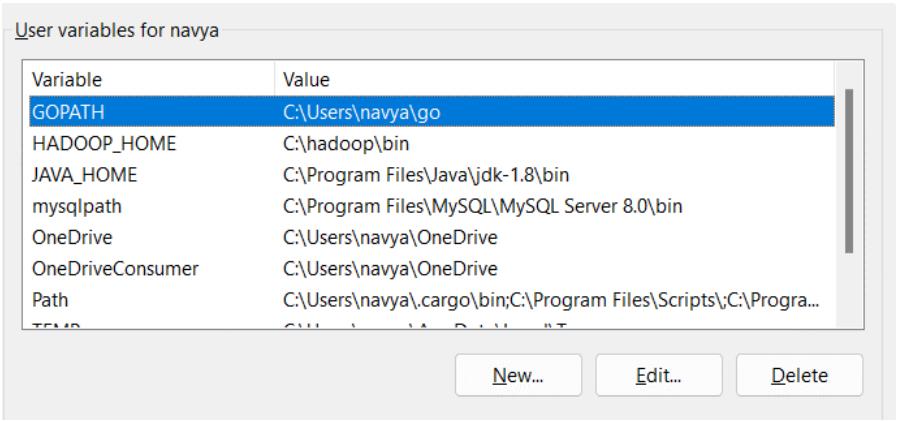
- **Variable name:** HADOOP_HOME

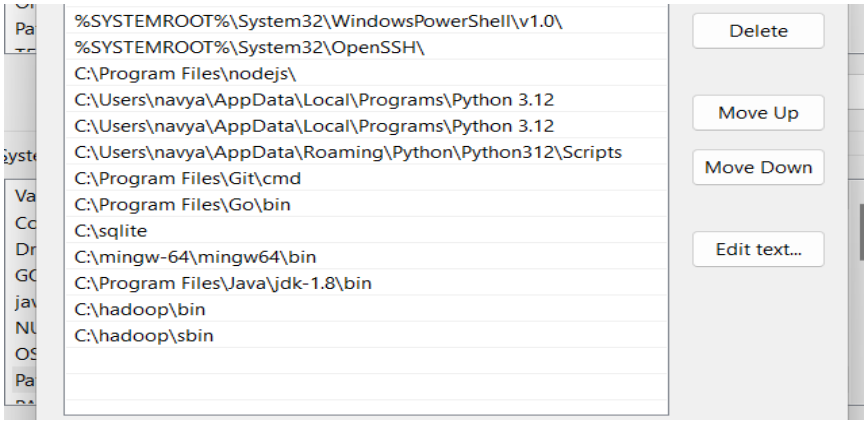
- **Variable value:** C:\hadoop
- Update Path variable and add %HADOOP_HOME%\bin.

Edit the configuration in the 4 files, add the <property> code.

Then go to command prompt and type the following







yarn-site.xml

C: > hadoop > etc > hadoop > yarn-site.xml

```xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>

</configuration>

```

mapred-site.xml

C: > hadoop > etc > hadoop > mapred-site.xml

```xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

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

```

httpfs-site.xml

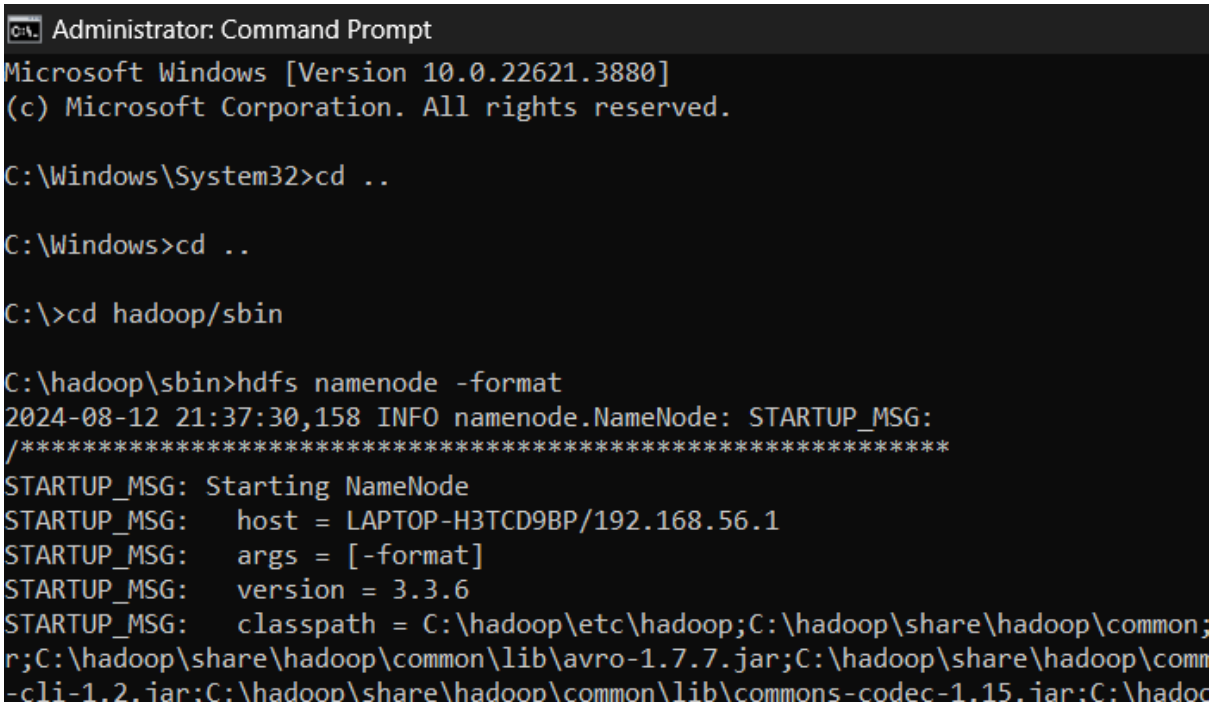
C: > hadoop > etc > hadoop > httpfs-site.xml

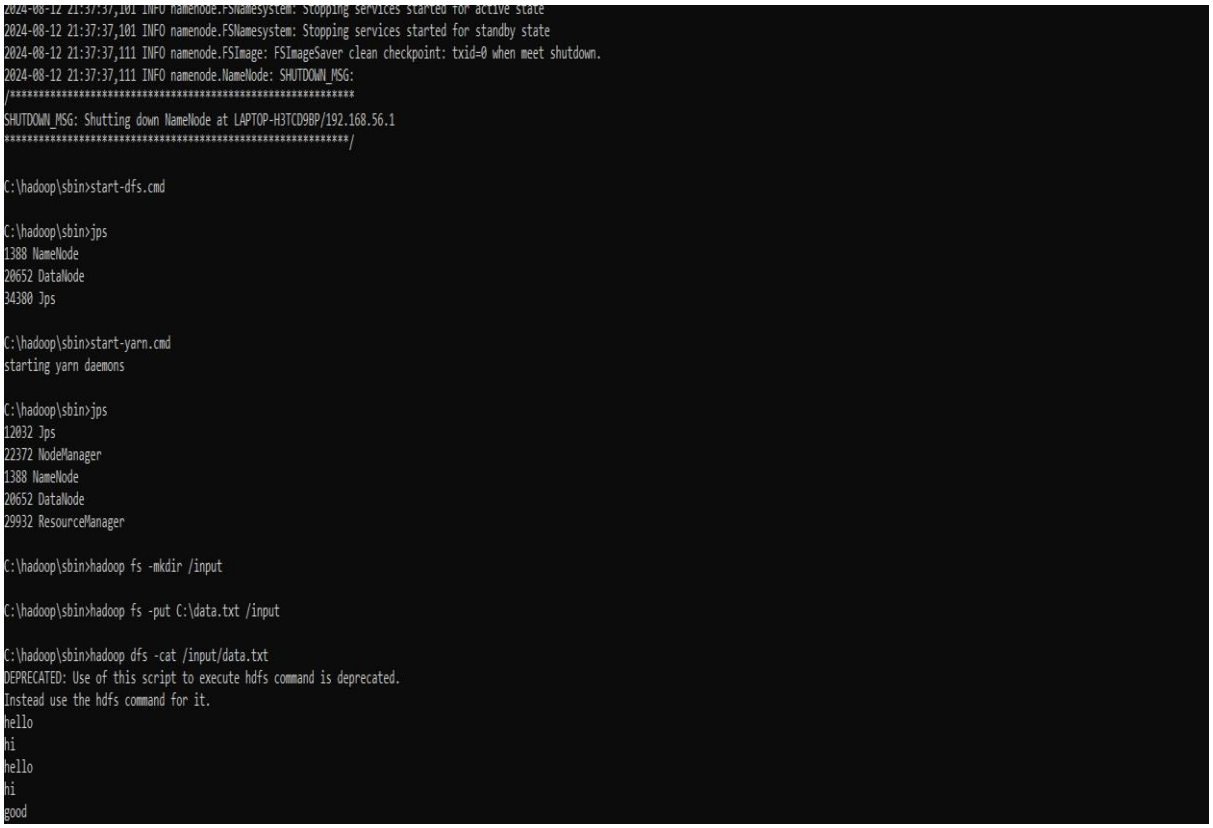
```
<!--
  You may obtain a copy of the License at

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License.
-->
<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>
```

core-site.xml

C: > hadoop > etc > hadoop > core-site.xml

```
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

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


Administrator: Command Prompt
Microsoft Windows [Version 10.0.22621.3880]
(c) Microsoft Corporation. All rights reserved.
C:\Windows\System32>cd ..
C:\Windows>cd ..
C:\>cd hadoop/sbin
C:\hadoop\sbin>hdfs namenode -format
2024-08-12 21:37:30,158 INFO namenode.NameNode: STARTUP_MSG:
/**
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = LAPTOP-H3TCD9BP/192.168.56.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.6
STARTUP_MSG: classpath = C:\hadoop\etc\hadoop;C:\hadoop\share\hadoop\common;
r;C:\hadoop\share\hadoop\common\lib\avro-1.7.7.jar;C:\hadoop\share\hadoop\comm
-cli-1.2.jar;C:\hadoop\share\hadoop\common\lib\commons-codec-1.15.jar;C:\hadoo


2024-08-12 21:37:37,101 INFO namenode.FSNamesystem: Stopping services started for active state
2024-08-12 21:37:37,101 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-08-12 21:37:37,111 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-08-12 21:37:37,111 INFO namenode.NameNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down NameNode at LAPTOP-H3TCD9BP/192.168.56.1
**/
C:\hadoop\sbin>start-dfs.cmd
C:\hadoop\sbin>jps
1388 NameNode
20652 DataNode
84380 Jps
C:\hadoop\sbin>start-yarn.cmd
starting yarn daemons
C:\hadoop\sbin>jps
12032 Jps
22372 NodeManager
1388 NameNode
20652 DataNode
29932 ResourceManager
C:\hadoop\sbin>hadoop fs -mkdir /input
C:\hadoop\sbin>hadoop fs -put C:\data.txt /input
C:\hadoop\sbin>hadoop dfs -cat /input/data.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
hello
hi
hello
hi
good

```
                Physical memory (bytes) snapshot=609239040
                Virtual memory (bytes) snapshot=834080768
                Total committed heap usage (bytes)=441450496
                Peak Map Physical memory (bytes)=363778048
                Peak Map Virtual memory (bytes)=423948288
                Peak Reduce Physical memory (bytes)=245460992
                Peak Reduce Virtual memory (bytes)=410136576
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=70
        File Output Format Counters
                Bytes Written=42

C:\hadoop\sbin>hadoop fs -cat /out/*
cat: `/out/*': No such file or directory

C:\hadoop\sbin>hadoop fs -cat /output/*
bye     1
day     1
good    2
hello   4
hi      3
morning 1

C:\hadoop\sbin>A
```

```
**********************************************************/

:\hadoop\sbin>start-dfs.cmd

:\hadoop\sbin>jps
388 NameNode
0652 DataNode
4380 Jps

:\hadoop\sbin>start-yarn.cmd
tarting yarn daemons

:\hadoop\sbin>jps
2032 Jps
2372 NodeManager
388 NameNode
0652 DataNode
9932 ResourceManager

:\hadoop\sbin>hadoop fs -mkdir /input
```

```
C:\hadoop\sbin>hadoop dfs -cat /input/data.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
hello
hi
hello
hi
good
morning
good day
bye
hello
hello
hi
C:\hadoop\sbin>hadoop fs -cat /output/*
bye     1
day     1
good    2
hello   4
hi      3
morning 1

C:\hadoop\sbin>
```