DATA ANALYTICS ASSIGNMENT

Apache Hadoop:

**Apache Hadoop** is an open-source framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models. It is designed to scale up from a single server to thousands of machines, each offering local computation and storage. Hadoop is widely used for big data analytics and supports a range of use cases, from data warehousing and ETL (Extract, Transform, Load) operations to machine learning and data mining.

* 1. **History of Hadoop:**

Hadoop originated from projects at Yahoo! and Apache Nutch, which were inspired by Google's papers on the Google File System (GFS) and MapReduce.

* **2003**: Google published the GFS paper, which described a distributed file system designed to handle large data sets using commodity hardware.
* **2004**: Google released the MapReduce paper, describing a programming model and associated implementation for processing and generating large datasets.
* **2006**: Doug Cutting and Mike Cafarella created Hadoop to support the Nutch search engine project.
* **2008**: Hadoop became a top-level Apache project, gaining wider adoption and further development.
* Since then, Hadoop has become a cornerstone of big data, with ongoing contributions and improvements from the open-source community and industry.
  1. **Versions of Hadoop:**

Hadoop has evolved through several major versions, each offering performance improvements, feature enhancements, and expanded support.

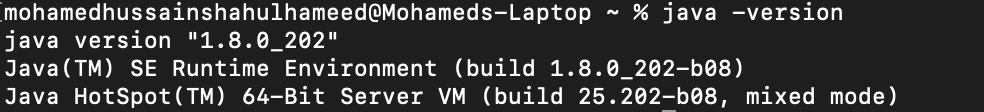
* **Hadoop 1.x**: The initial version, featuring MapReduce as the primary processing engine and HDFS for storage.
* **Hadoop 2.x**: Introduced YARN (Yet Another Resource Negotiator) which allowed for better resource management and enabled multiple processing models other than MapReduce to work on Hadoop clusters (like Apache Tez, Apache Spark).
* **Hadoop 3.x**: Introduced features such as erasure coding for more efficient storage, support for more than 2 Name Nodes, and containerization with Docker.
  1. **System Requirements for Hadoop (all OS):**

Hadoop can be installed on various operating systems like Windows, macOS, and Linux. Below are general system requirements:

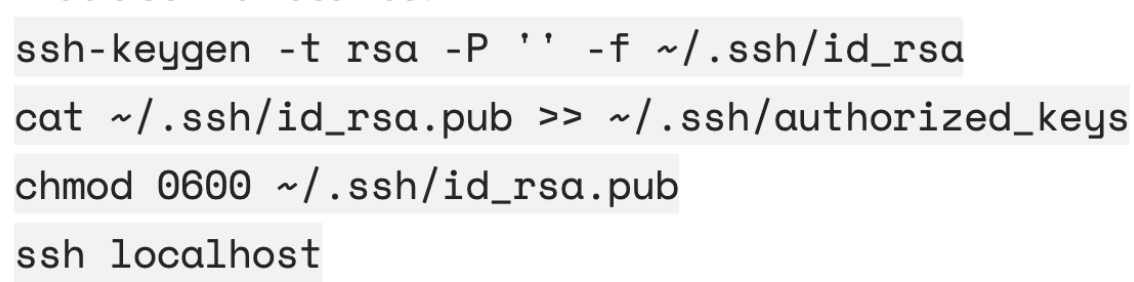
* **RAM**: Minimum 4 GB RAM (8 GB or more recommended).
* **CPU**: Multi-core processor (quad-core or higher recommended for faster processing).
* **Disk Space**: At least 20 GB of free disk space for storage (more for large datasets and better performance).
* **Java**: Requires Java Development Kit (JDK) 8 or higher.
* **Operating Systems**:
  + Linux (preferred OS for Hadoop clusters due to its stability and performance in distributed environments).
  + macOS (for development purposes).
  + Windows (not typically used for production clusters but supported in standalone mode).

1.4 Installation Steps on Macos:

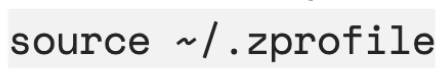
Step 1: Install Java 1.8 version



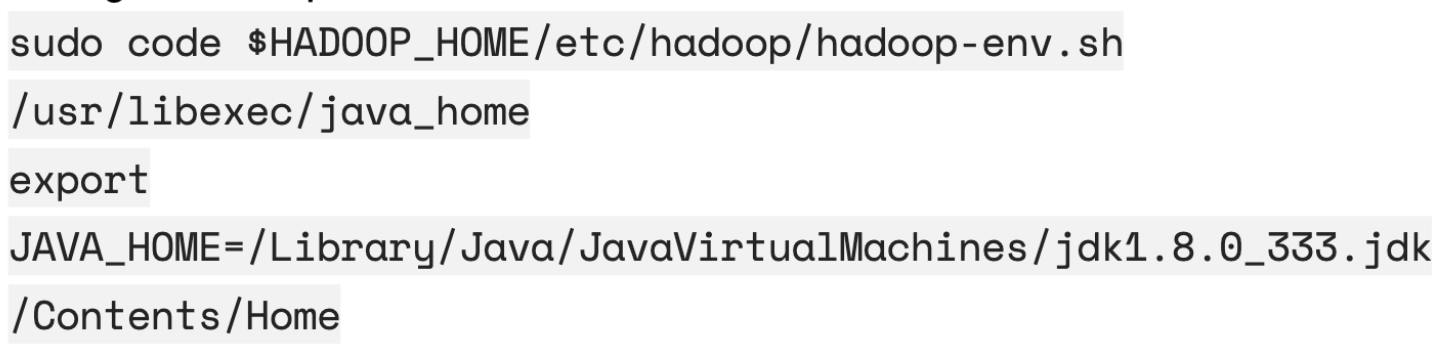
Step 2: Enable SSH for local host



Step 3: Download Hadoop and modify z-profile



Step 4: Configure Hadoop



Step 5: Edit core-site.xml

<property>  
 <name>fs.defaultFS</name>  
 <value>hdfs://localhost:9000</value>  
 <final>true</final>  
</property>

Step 6: Edit hdfs-site.xml

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

Step 7: Edit mapred-site.xml

<configuration>  
 <property>  
 <name>yarn.app.mapreduce.am.env</name>  
 <value>HADOOP\_MAPRED\_HOME=/opt/homebrew/opt/hadoop</value>  
 </property>  
 <property>  
 <name>mapreduce.map.env</name>  
 <value>HADOOP\_MAPRED\_HOME=/opt/homebrew/opt/hadoop</value>  
 </property>  
 <property>  
 <name>mapreduce.reduce.env</name>  
 <value>HADOOP\_MAPRED\_HOME=/opt/homebrew/opt/hadoop</value>  
 </property>  
</configuration>

Step 8: Edit yarn-site.xml

<configuration>  
 <property>  
 <name>yarn.nodemanager.aux-services</name>  
 <value>mapreduce\_shuffle</value>   
 </property>  
</configuration>

Step 9: Start Hadoop

Start-all.sh

* 1. Installation Screenshots

