# **Apache Hadoop and Hive Installation & Configuration**

## **1. Environment Setup**

### **VM and OS Selection:**

* Installed a fresh CentOS 7 VM.
* Configured network settings for internet access.
* Disabled firewall and SELinux:  
   sudo systemctl stop firewalld
* sudo systemctl disable firewalld
* sudo setenforce 0
* Updated CentOS repositories and installed necessary packages:  
   sudo yum update -y
* sudo yum install -y wget vim net-tools

### **Issues Faced in CentOS and Resolutions:**

* **Repository Errors:**
  + Encountered errors like Could not resolve host: mirrorlist.centos.org.
* Resolved by manually updating /etc/resolv.conf to use Google DNS:  
   echo "nameserver 8.8.8.8" | sudo tee /etc/resolv.conf
  + Updated and enabled required repositories manually.
* **Package Dependency Issues:**
  + PostgreSQL package conflicts resolved by manually installing required dependencies.

### **User and Group Setup:**

* Created a Hadoop user and group:  
   sudo groupadd hadoop
* sudo useradd -m -G hadoop hadoop
* sudo passwd hadoop
* Changed ownership of /opt to Hadoop user:  
   sudo chown -R hadoop:hadoop /opt

## **2. Hadoop Installation & Configuration**

### **Download and Extract Hadoop:**

* Downloaded Hadoop and extracted it to /opt/hadoop
* Set environment variables in ~/.bashrc:  
   export HADOOP\_HOME=/opt/hadoop
* export PATH=$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$PATH
* export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop
* Reloaded ~/.bashrc:  
   source ~/.bashrc

### **Configured Hadoop:**

* Updated core-site.xml:  
   <property>
* <name>fs.defaultFS</name>
* <value>hdfs://cmg.hadoop.com:9000</value>
* </property>
* <property>
* <name>hadoop.proxyuser.deco.hosts</name>
* <value>\*</value>
* </property>
* <property>
* <name>hadoop.proxyuser.deco.groups</name>
* <value>\*</value>
* </property>
* Updated hdfs-site.xml, mapred-site.xml, yarn-site.xml.
* Formatted NameNode:  
   hdfs namenode -format
* Started Hadoop services:  
   start-dfs.sh
* start-yarn.sh
* Verified services using:  
   jps

## **3. PostgreSQL Installation & Hive Metastore Configuration**

### **PostgreSQL Installation:**

* Installed PostgreSQL:  
   sudo yum install -y postgresql-server postgresql-contrib
* Initialized PostgreSQL database:  
   sudo postgresql-setup initdb
* Started PostgreSQL service:  
   sudo systemctl start postgresql
* sudo systemctl enable postgresql

### **Hive Metastore Configuration:**

* Created a Hive Metastore database and user in PostgreSQL:  
   CREATE DATABASE metastore;
* CREATE USER hive WITH PASSWORD 'hivepassword';
* GRANT ALL PRIVILEGES ON DATABASE metastore TO hive;
* Updated hive-site.xml:  
   <property>
* <name>javax.jdo.option.ConnectionURL</name>
* <value>jdbc:postgresql://localhost:5432/metastore</value>
* </property>
* <property>
* <name>hive.server2.thrift.port</name>
* <value>10000</value>
* </property>
* <property>
* <name>hive.metastore.uris</name>
* <value>thrift://localhost:9083</value>
* </property>
* <property>
* <name>hive.server2.enable.doAs</name>
* <value>true</value>
* <description>
* Setting this property to true will have HiveServer2 execute
* Hive operations as the user making the calls to it.
* </description>
* </property>
* Ran SchemaTool to initialize the Metastore schema:  
   schematool -dbType postgres -initSchema

## **4. Hive Installation & Troubleshooting**

### **Hive Installation:**

* Extracted Hive to /opt/hive
* Set environment variables in ~/.bashrc:  
   export HIVE\_HOME=/opt/hive
* export PATH=$HIVE\_HOME/bin:$PATH
* Started Hive Metastore:  
   hive --service metastore &
* Started HiveServer2:  
   hive --service hiveserver2 &
* Encountered **impersonation issue**:  
   User: hadoop is not allowed to impersonate anonymous
* (To be resolved in next steps)

## **5. Sqoop Setup**

* Installed Sqoop.
* Configured MySQL/PostgreSQL connectors.
* Verified connectivity using:  
   sqoop list-databases --connect jdbc:postgresql://localhost:5432/ --username postgres --password postgres

## **6. Pending Issues & Next Steps**

* Resolve HiveServer2 connection issue.
* Configure impersonation for Hive.
* Test Sqoop with real data transfers.
* Document any additional configurations and troubleshooting steps.
* Review CentOS repo issues and resolve package dependency problems.