Local Hadoop Installation

Version 1.0

[**Scope**](#_heading=h.3znysh7) **3**

[**Installations Steps**](#_heading=h.2et92p0) **4**

[HDFS Installation](#_heading=h.tyjcwt) 4

[YARN Installation](#_heading=h.3dy6vkm) 6

[Pig Installation](#_heading=h.1t3h5sf) 7

[Sqoop Installation](#_heading=h.4d34og8) 8

[Hive Installation](#_heading=h.2s8eyo1) 9

[Spark Installation](#_heading=h.17dp8vu) 10

[Flume Installation](#_heading=h.3rdcrjn) 11

[HBase Installation ( Pseudo Distributed )](#_heading=h.6wmi29kkdwvt) 12

[Kafka Installation](#_heading=h.5xs2ilg4ldnz) 13

# 

# 

# Scope

This document is designed as a guide to local installation of various hadoop components on your local.

At actual production installations, a distribution like Cloudera Distribution / Hontonworks Distribution are preferred. Local installations are preferred only for local development work.

# Installations Steps

## HDFS Installation

Download the tarball from the download location mentioned below.

Link : [Apache\_link\_for\_v3.3.1](https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz)

Follow the steps below for installation.

1. Untar the downloaded jar
   1. $ tar -xvzf <YOUR\_DOWNLOADED\_TAR\_BALL\_LOCATION>
2. Export HADOOP\_HOME=<YOUR\_UNTARRED\_LOCATION>
3. Edit the configuration files for namenode dir and datanode dir
   1. Add the following entries to core-site.xml

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

* 1. Add the following entries to hdfs-site.xml:

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:///home/abhay/DBDA\_HOME/DATA/HDFS/namenode</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>file:///home/abhay/DBDA\_HOME/DATA/HDFS/datanode</value>

</property>

1. If this is your first installation, format HDFS using the following command
   1. cd $HADOOP\_HOME/bin
   2. ./hdfs namenode -format
2. Start HDFS using the startup script
   1. cd $HADOOP\_HOME
   2. sbin/start-dfs.sh
3. Your HDFS should startup
4. Check the installation using the following command
   1. cd $HADOOP\_HOME/bin
   2. ./hdfs dfs -ls /
   3. ./hdfs dfs -put ../README.txt /
   4. Access the UI from <http://localhost:9870/>
5. Stop HDFS
   1. cd $HADOOP\_HOME
   2. sbin/stop-dfs.sh

## YARN Installation

Download the tarball from the download location mentioned below.

Link : [Apache\_link\_for\_v3.3.1](https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz)

Follow the steps below for installation.

1. Untar the downloaded jar
   1. $ tar -xvzf <YOUR\_DOWNLOADED\_TAR\_BALL\_LOCATION>
2. Export HADOOP\_HOME=<YOUR\_UNTARRED\_LOCATION>
3. Edit the configuration files for YARN and Map-Reduce.
   1. Edit etc/hadoop/yarn-site.xml

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.env-whitelist</name>

<value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PREPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_HOME,PATH,LANG,TZ,HADOOP\_MAPRED\_HOME</value>

</property>

* 1. Edit etc/hadoop/mapred-site.xml  
      <property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

<property>

<name>mapreduce.application.classpath</name>

<value>$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/\*:$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/lib/\*</value>

</property>

1. Start YARN using the following script
   1. $ sbin/start-yarn.sh

## Pig Installation

Download the tarball from the download location mentioned below.

Link : [Inet-Link](https://dlcdn.apache.org/pig/latest/pig-0.17.0.tar.gz)

Follow the steps below for installation.

1. Untar the downloaded jar
   1. $ tar -xvzf <YOUR\_DOWNLOADED\_TAR\_BALL\_LOCATION>
2. Export HADOOP\_HOME=<YOUR\_HADOOP\_LOCATION>
3. Export PATH=$PATH:<YOUR\_UNTAR\_LOCATION>/bin

## Sqoop Installation

Download the tarball from the download location mentioned below.

Link : [Inet-Link](http://archive.apache.org/dist/sqoop/1.4.2/sqoop-1.4.2.bin__hadoop-2.0.0-alpha.tar.gz)

Follow the steps below for installation.

1. Untar the downloaded jar
   1. $ tar -xvzf <YOUR\_DOWNLOADED\_TAR\_BALL\_LOCATION>
2. Export HADOOP\_HOME=<YOUR\_HADOOP\_LOCATION>
3. Export SQOOP\_HOME=<YOUR\_UNTAR\_LOCATION>
4. Export PATH=$PATH:<YOUR\_UNTAR\_LOCATION>/bin
5. Store all your JDBC jars (e.g mysql-connector-<version>.jar) under $SQOOP\_HOME/lib

## Hive Installation

Download the tarball from the download location mentioned below.

Link : [Inet-Link](https://dlcdn.apache.org/hive/hive-2.3.9/apache-hive-2.3.9-bin.tar.gz)

Follow the steps below for installation.

1. Untar the downloaded jar
   1. $ tar -xvzf <YOUR\_DOWNLOADED\_TAR\_BALL\_LOCATION>
2. Export HADOOP\_HOME=<YOUR\_HADOOP\_LOCATION>
3. Export HIVE\_HOME=<YOUR\_HIVE\_UNTARRED\_LOCATION>
4. Export PATH=$PATH:<YOUR\_UNTAR\_LOCATION>/bin
5. Create HDFS dir using the following commands
   1. $HADOOP\_HOME/bin/hadoop fs -mkdir -p /user/hive/warehouse
   2. $HADOOP\_HOME/bin/hadoop fs -mkdir -p /tmp
6. Provide updated permissions using the following commands
   1. $HADOOP\_HOME/bin/hadoop fs -chmod g+w /user/hive/warehouse
   2. $HADOOP\_HOME/bin/hadoop fs -chmod g+w /tmp
7. Initialise the metastore DB using the schema\_tool
   1. cd $HIVE\_HOME # ENSURE YOU ARE ALWAYS IN HIVE\_HOME
   2. $HIVE\_HOME/bin/schematool -dbType derby -initSchema
8. Start you HIVE prompt
   1. cd $HIVE\_HOME # ENSURE YOU ARE ALWAYS IN HIVE\_HOME
   2. ./bin/hive
9. Test your hive prompt. Execute the following on hive prompt
   1. show databases # THIS COMMAND SHOULD NOT ERROR OUT

Note : Recommended tutorial : <https://cwiki.apache.org/confluence/display/Hive/GettingStarted#GettingStarted-InstallationandConfiguration>

Steps :

## Spark Installation

Download the tarball from the download location mentioned below.

Link : [Inet-Link](https://www.apache.org/dyn/closer.lua/spark/spark-3.2.0/spark-3.2.0-bin-hadoop3.2.tgz)

Follow the steps below for installation.

1. Untar the downloaded jar
   1. $ tar -xvzf <YOUR\_DOWNLOADED\_TAR\_BALL\_LOCATION>
2. Export HADOOP\_HOME=<YOUR\_HADOOP\_LOCATION>
3. Export SPARK\_HOME=<YOUR\_UNTAR\_LOCATION>
4. Start master using the following command
   1. $ sbin/start-master.sh
5. Start slave using the following command
   1. $ sbin/start-worker.sh spark://<your\_master\_host>:7077
   2. Note : 7077 is the default port, it may vary across versions

## Flume Installation

Download the tarball from the download location mentioned below.

Link : [Inet-Link](http://www.apache.org/dyn/closer.lua/flume/1.9.0/apache-flume-1.9.0-bin.tar.gz)

Follow the steps below for installation.

1. Untar the downloaded jar
   1. $ tar -xvzf <YOUR\_DOWNLOADED\_TAR\_BALL\_LOCATION>
2. Export FLUME\_HOME=<YOUR\_UNTAR\_LOCATION>

## HBase Installation ( Pseudo Distributed )

Download the tarball from the download location mentioned below.

Link : [Inet-Link](https://dlcdn.apache.org/hbase/3.0.0-alpha-2/hbase-3.0.0-alpha-2-bin.tar.gz)

Follow the steps below for installation.

* Ensure your JAVA\_HOME is already set ( in /etc/environment )
* Extract the tarball using the following command:
  1. tar -xvzf <YOUR\_HBASE\_TARBALL>
* export HBASE\_HOME=<YOUR\_HBASE\_EXTRACTED\_LOCATION>
* Delete the existing properties in hbase-site.xml.
* Add the following in hbase-site.xml:

<property>

<name>hbase.cluster.distributed</name>

<value>true</value>

</property>

<property>

<name>hbase.rootdir</name>

<value>hdfs://localhost:9000/hbase</value>

</property>

* Run the following command
  1. cd $HBASE\_HOME
  2. ./bin/start-hbase.sh
* Run $jps command. You should see HMaster, HRegionserver, HQuorumPeer running.
* Run the following command to see if hbase dir is populated. If startup is successful, you will see some content in the hbase dir.
  1. hdfs dfs -ls /hbase

## Kafka Installation

Download the tarball from the download location mentioned below.

Link : [Inet-Link](https://www.apache.org/dyn/closer.cgi?path=/kafka/3.0.0/kafka_2.13-3.0.0.tgz)

Follow the steps below for installation.

* Extract the tarball using the following command:
  1. tar -xvzf <YOUR\_KAFKA\_TARBALL>
* export KAFKA\_HOME=<YOUR\_KAFKA\_EXTRACTED\_LOCATION>
* Ensure HBase is down ( esp Quorumpeer )
* Start your zookeeper. Run the following
  1. cd KAFKA\_HOME
  2. mkdir logs
  3. bin/zookeeper-server-start.sh config/zookeeper.properties > logs/zk.log 2>&1 &
  4. bin/kafka-server-start.sh config/server.properties > logs/kafka.log 2>&1 &
* Test your installation
  1. From 1st Terminal and check the topic creation
     1. bin/kafka-topics.sh --create --partitions 1 --replication-factor 1 --topic quickstart-events --bootstrap-server localhost:9092
     2. bin/kafka-topics.sh --describe --topic quickstart-events --bootstrap-server localhost:9092
     3. bin/kafka-console-producer.sh --topic quickstart-events --bootstrap-server localhost:9092
        1. Enter values in the terminal
  2. From 2nd Terminal, start a consumer and see if things are being consumed
     1. bin/kafka-console-consumer.sh --topic quickstart-events --from-beginning --bootstrap-server localhost:9092
* To stop the servers run the following:
  1. bin/kafka-server-stop.sh
  2. bin/zookeeper-server-stop.sh