Tez Installation Guide

For this project, we will be installing and configuring three systems: Hadoop, Hive, and Tez. I have experienced some version incompatibility so it is best to use the versions I have tested on.

Hadoop ⇒ Version 2.6.0
Hive ⇒ Version 0.13.1
Tez ⇒ Version 0.4.1

1) [SKIP IF YOU HAVE CLONED THE TEZ REPO] Download the systems.

To download Hadoop: http://hadoop.apache.org/releases.html#Download
To download Hive: https://hive.apache.org/downloads.html
To download Tez: http://tez.apache.org/releases/index.html

2) Set up Hadoop on Pseudo-Distributed Mode.

It is easiest to follow the instructions detailed here:
http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html#YARN_on_Single_Node

Tips:

i) for the two paths you need to set - `JAVA_HOME` and `HADOOP_PREFIX`, you can follow the documentation in `path_configs.txt`.

ii) if you cloned the Tez repo, all the configs for `etc/hadoop/core-site.xml`, `etc/hadoop/hdfs-site.xml`, `etc/hadoop/mapred-site.xml`, and `etc/hadoop/yarn-site.xml` should be set correctly.

Verify that your Hadoop setup is correct. You can do this by running HDFS and YARN. The following links should be running `http://localhost:50070/` (Name Node) and `http://localhost:8088/` (Resource Manager)

3) Set up Hive.

It is easiest to follow the instructions detailed here:
https://cwiki.apache.org/confluence/display/Hive/GettingStarted#GettingStarted-RunningHive

Tips:

i) you need to hav `HIVE_HOME` and `HADOOP_HOME` set.

ii) you also need to make sure you add `HIVE_HOME` to your `PATH`.

Verify that Hive is installed correctly by running some sample commands. To enter your Hive console, run the command: `bin/hive`.

4) Set up Tez paths.

If you did not clone the repo, and downloaded Tez separately, you will need to:
1. unpack the zip
2. build the jars with command: `mvn package -DskipTests -Dtar`
3. explode the tar from `tez-dist/target/tez-0.4.1-incubating.tar.gz`
4. this should result in `/tez-0.4.1-incubating`, which is what you will get if you cloned the repo

At this point, you will need to set up some paths, specifically, `TEZ_INSTALL_DIR`, `TEZ_CONF_DIR`, `TEZ_JARS`, `HIVE_AUX_JARS_PATH`. The exports for these paths are available at path_configs.txt as well.

```
export TEZ_INSTALL_DIR='[PATH_TO_TEZ_DIR]/tez-0.4.1-incubating'
export TEZ_CONF_DIR=$TEZ_INSTALL_DIR/conf
export TEZ_JARS=$(echo "$TEZ_INSTALL_DIR"/*.jar | tr ' ' ':'):$(echo
"$TEZ_INSTALL_DIR"/lib/*.jar | tr ' ' ':')
export HIVE_AUX_JARS_PATH="${TEZ_JARS}:${TEZ_CONF_DIR}"
```

5) Update Hadoop Classpath.

Add the following export statement to your `.bash_profile`/`.bashrc` as well. This adds the TEZ paths to your `HADOOP_CLASSPATH`.

```
export HADOOP_CLASSPATH =
"${TEZ_CONF_DIR}:${TEZ_JARS}/*:${TEZ_JARS}/lib/*:${HADOOP_CLASSPATH}"
```

6) Set up HDFS.

Create directory `/apps/tez-0.4.1` on HDFS. This is where all your Tez jars will go. You will also need to the change the permissions of this directory. The commands are:

Command: `bin/hdfs dfs -mkdir -p /apps/tez-0.4.1`
Command: `bin/hdfs dfs -chmod g+w /apps/tez-0.4.1`

You also need to make a `/user` directory on HDFS with your [name]. For me that is `/user/jaywang`.

Command: `bin/hdfs dfs -mkdir -p /user/jaywang`

7) Copy Tez jars to HDFS.

The commands are:

Command: `bin/hdfs dfs -copyFromLocal $TEZ_INSTALL_DIR/*` `/apps/tez-0.4.1`
Command: `bin/hdfs dfs -copyFromLocal $TEZ_INSTALL_DIR/lib/*` `/apps/tez-0.4.1`

8) Update mapred-site.xml.

The `mapred-site.xml` file should be located in `/hadoop-2.6.0/etc/hadoop/`. Note that this has already been done if you cloned the repo.

Under `mapreduce.framework.name`, change the value from `yarn` to `yarn-tez`.

9) Create a tez-site.xml.

Create a `tez-site.xml` file in your `$TEZ_CONF_DIR`. Note that this has already been done if you cloned the repo.

File contents should be:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>
  <name>tez.lib.uris</name>
  <value>${fs.default.name}/apps/tez-0.4.1/</value>
</property>

</configuration>
```

That's it! Now, to run Hive on Tez, make sure to set the execution engine to Tez and run the Hive job as you would normally. There is another guide called Tez Tutorial Guide that walks through how to run Hive on Tez once the installation and setup is complete.