CS-839 NoSQL Systems
Dr. Vinu Venugopal, Divyam Agrawal, Aniket Mitra & Nandakishore S Menon

Tutorial Sheet #1

This tutorial sheet aims to help you in setting up your working environment on a local computer (such as a laptop or PC) and to run your first WordCount example in a local installation of Apache Hadoop 3.3.1.

Set Up Your Local Working Environment

1. Download and install **Java 8** (JDK 1.8) by choosing the respective installer binaries from: `http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html`

   - Linux users can directly download the `jdk-<version>-linux-x64.tar.gz` package, unpack the tar archive, and install the JDK with: `tar -zxvf jdk-<version>-linux-x64.tar.gz`
   - Windows users have to run the installation binaries and then add the `bin` folder with the `java` binary of their JDK installation to the front of their `PATH` variable under the *Windows System Variables.*
   - MacOS users can simply click on the downloaded `.dmg` file and install the application.

   Make sure that your `JAVA_HOME` environment variable points to your installation directory of Java 8.

2. Download and install Eclipse (Eclipse IDE for Java Developers) by choosing the respective installer binaries from: `http://www.eclipse.org/downloads/packages/`

   - Linux users can install **Eclipse** using `sudo tar -xzvf <path_to_the_downloaded_package>`
   - Windows users simply unzip the downloaded file into a directory of your choice. There is no need to run any installer.
   - MacOS users can simply double-click the downloaded `.dmg` file and follow the screen instructions to install Eclipse.

3. Download and install **Hadoop 3.3.1** by choosing a respective mirror with the `hadoop-3.3.1.tar.gz` binaries from: `http://hadoop.apache.org/releases.html`

   Unpack the tar archive and make sure that your `HADOOP_HOME` environment variable points to your installation directory of Hadoop. Additionally verify that your `PATH` environment variable contains the `bin` directory in which your `hadoop` binary is located.

Run the WordCount Examples in Apache Hadoop

1. Download the `Wikipedia-50-ARTICLES.tar.gz` file from Moodle and extract the tar archive into a local directory.
2. Download the `WordCount.java` class from LMS and put them into a new Java project in your Eclipse IDE for Java.
3. Import the `hadoop-common-3.3.1.jar` and `hadoop-mapreduce-client-core-3.3.1.jar` libraries from your local Hadoop installation (located in the `./share/hadoop/common` and `./share/hadoop/mapreduce` subdirectories of your `HADOOP_HOME` installation directory) as external jars into the Java build path of your Eclipse project.
4. Export your Java project into a single jar file called `WordCount.jar` from your Eclipse IDE to a local folder.
5. Run the WordCount example by typing

   ```
   hadoop jar WordCount.jar WordCount <input_directory> <output_directory>
   ```

   where `<input_directory>` is the directory containing your Wikipedia articles and `<output_directory>` is a new directory that will contain the output of your WordCount example.