Project Artifacts for Reproducibility

Softwares:

Apache Hadoop	https://hadoop.apache.org/
Apache Hive	https://hive.apache.org/
Java JDK	https://www.oracle.com/java/technologies/downloads/
Apache Hadoop Documentation	https://hadoop.apache.org/docs/stable/

Code:

Starting Apache Hadoop:

```
start-all.sh
start-dfs.sh                    # this will start namenode, datanode and
namenode
start-yarn.sh                   # this will start node manager and resource
manager
Jps                  # check running daemons
Check localhost web UI
Namenode Health: http://localhost:9870
Hadoop UI: http://localhost:8088/cluster/nodes
```

Stoping Hadoop:

```
stop-yarn.sh, stop-dfs.sh; alternatively stop-all.sh
```

Starting Apache Hive:

Creating Hive warehouse:

```
hdfs dfs -mkdir -p /user/hive/warehouse
hdfs dfs -chmod g+w /user/hive/warehouse
```

Initiate the Derby database to store metadata:

```
$HIVE_HOME/bin/schematool -dbType derby -initSchema

cd $HIVE_HOME/bin
hive               #starts hive
```

Creating a database and table in Hive:

```
CREATE DATABASE student_details;
USE student_detail;
CREATE TABLE IF NOT EXISTS student(Student_Name STRING, Student_No
INT, Student_Mark FLOAT) ROW FORMAT DELIMITED FIELDS TERMI-
NATED BY ',';
INSERT INTO TABLE student VALUES
('student1',1,'100'),('student2',2,'0'),('student3',3,'90');
SELECT * FROM student;
```