

Department of Computing

Academic Year 2022/2023

Infrastructure for Big Data – Assessment 3

**STUDENT NAME: M.S.Shruthi**

**STUDENT NUMBER: C00290801**

**COURSE NAME: Master of Science in Data Science**

**DEPARTMENT: Department of Computing and Data Science**

**SUBMITTED TO: Michael Gleeson**

**DATE OF SUBMISSION: December 23rd 2022.**

**Description of ScreenCast:**

Initially, I set up 3 VM manually ,Apache Hadoop ,MAP Reduce ,YARN and Hive .

In the video, I have shown 3 VM java version, Hadoop version and hostname.

Format name node in master. Brought all services up in master Node.

Verified through Java processing status JPS. Name node and data node are displayed in master and slave nodes respectively.

**Map-reduce job process:**

Created sample file in master and slave nodes. Copied to HDFS ,appended the 3 files.

Executed map-reduce job on the appended file. Map-reduce executed successfully .

**HIVE:**

Taken sample csv file from Kaggle dataset by name “diabetes.csv”. Uploaded to HDFS .

Created table for diabetes with columns corresponding to the CSV File. Loaded into HIVE database. Retrieved the records.

Steps to be performed:

**Data transaction:**

start-all.sh

jps

Hadoop fs -mkdir /hive\_data

Hadoop fs -copyFromLocal diabetes.csv /hive\_data

hadoop fs -ls /hive\_data

**hive**

show tables;

create table diabetes (Pregnancies int,Glucose int,BloodPressure int,SkinThickness int,Insulin int,BMI double,Diabetes double,Age int,Outcome int)row format delimited fields terminated by ',';

load data inpath '/testing/diabetes.csv' into table diabetes;

select count(\*) from diabetes where Pregnancies=7;

select count(\*) from diabetes;

select count(\*) from diabetes where Glucose=180;

**Map reduce Jobs:**

Hadoop fs -mkdir /map\_reduce

sudo vi Unimaster.txt

Hadoop fs -copyFromLocal Unimaster.txt /mapreduce

sudo vi Unislave1.txt, hadoop fs -appendToFile Unislave1.txt /mapreduce/Unimaster.txt

sudo vi Unislave2.txt, hadoop fs -appendToFile Unislave2.txt /mapreduce/Unimaster.txt

hadoop fs -cat /mapreduce/Unimaster.txt, hadoop fs -ls /mapreduce

**Command to execute mapreduce:**

cd /usr/local/hadoop/share/hadoop/mapreduce

hadoop jar hadoop-mapreduce-examples-3.3.1.jar wordcount /mapreduce/Unimaster.txt /mapreduce/Unimaster\_output

To delete directory in HDFS : hdfs dfs -rm -R /employee

**Vagrant:**

I have automated through vagrant building 3 VM. Installed java,Hadoop, configured the xml files. Executed name node format and java process status is displaying in name node.