**JAVA**

**$ javac --version**

**$ sudo apt install default-jdk**

**$ sudo apt install default-jre**

**$ javac --version**

**$ cd <1MS23SCS/SCN\*\*>**

**Hadoop**

**$ tar xvzf hadoop-3.4.0.tar.gz**

**$ cd hadoop-3.4.0/**

**$ gedit bash.sh**

**export JAVA\_HOME=$(readlink -f $(which javac) | awk 'BEGIN {FS="/bin"} {print $1}')**

**export PATH=$(echo $PATH):$(pwd)/bin**

**export CLASSPATH=$(hadoop classpath)**

**$ source bash.sh**

**$ hadoop**

**PIG:**

**$ hadoop**

**$ tar -xvf pig-0.17.0.tar.gz**

**$ cd pig-0.17.0/**

**$ gedit bash.sh**

**export PIG\_INSTALL=$(pwd)**

**export PATH=$PATH:$(pwd)/bin**

**$ source bash.sh**

**$ pig -version**

**$ pig**

**Write Pig Latin scripts for Crop Production Dataset.**

1. **Calculate total production of each crop.**
2. **Find the average production per year for each crop.**
3. **Filter all crops grown in ‘Karnataka’**
4. **Calculate the total area used for each crop in the year 2010.**

crop\_prod = LOAD 'Datasets/crop\_production.csv' USING PigStorage(',') AS (State\_Name:chararray, District\_Name:chararray, Crop\_Year:int, Season:chararray, Crop:chararray, Area:float, Production:float);

DESCRIBE crop\_prod;

total\_production = GROUP crop\_prod BY Crop;

sum\_production = FOREACH total\_production GENERATE group AS Crop, SUM(crop\_prod.Production) AS Total\_Production;

DUMP sum\_production;

grouped\_by\_crop\_year = GROUP crop\_prod BY (Crop, Crop\_Year);

average\_production = FOREACH grouped\_by\_crop\_year GENERATE group.Crop AS Crop, group.Crop\_Year AS Crop\_Year, AVG(crop\_prod.Production) AS Avg\_Production;

DUMP average\_production;

specific\_state = FILTER crop\_prod BY State\_Name == ‘Karnataka’;

unique\_crops = GROUP specific\_state BY Crop;

DUMP unique\_crops;

specific\_year = FILTER crop\_prod BY Crop\_Year == 2010;

total\_area = GROUP specific\_year BY Crop;

sum\_area = FOREACH total\_area GENERATE group AS Crop, SUM(specific\_year.Area) AS Total\_Area;

DUMP sum\_area;

**Write Pig Latin scripts for Olympic Athletes and Hosts Datasets.**

1. **Filter athletes participated in the “Tokyo 2020” games.**
2. **Filter the games held in “China”.**
3. **Group games by season and count the number of games in each session.**
4. **Filter games that occurred after the year 2000.**

athletes = LOAD 'olympic\_athletes.csv' USING PigStorage(',') AS (athlete\_url: chararray, athlete\_full\_name: chararray, games\_participations: int, first\_game: chararray, athlete\_year\_birth: float, athlete\_medals: chararray, bio: chararray);

hosts = LOAD 'olympic\_hosts.csv' USING PigStorage(',') AS (game\_slug: chararray, game\_end\_date: chararray, game\_start\_date: chararray, game\_location: chararray, game\_name: chararray, game\_season: chararray, game\_year: int);

DESCRIBE athletes;

DESCRIBE hosts;

tokyo\_2020\_athletes = FILTER athletes BY first\_game == ‘Tokyo 2020’;

DUMP tokyo\_2020\_athletes;

games\_in\_china = FILTER hosts BY game\_location == 'China';

DUMP games\_in\_china;

grouped\_by\_season = GROUP hosts BY game\_season;

counted\_by\_season = FOREACH grouped\_by\_season GENERATE group AS game\_season, COUNT(hosts) AS num\_games;

DUMP counted\_by\_season;

games\_after\_2000 = FILTER hosts BY game\_year > 2000;

DUMP games\_after\_2000;