**ValidselecjICS 411: Big Data Storage**

**,**

**Homework Assignment 1**

Due: See Syllabus

**Objective:**

The objective of this assignment is to review data management using relational database management systems. Basically, you will design, create, and write SQL queries using MySQL DBMS.

**Requirements:**

In this assignment, you are asked to design and create a Weather database that includes weather reports on wind and temperature that were collected eight different stations.

Creating the database and importing data:

* The data to be loaded in the database is provided to you in a zipped file called “assignment1\_data.zip”. When you unzip the file, you will get the following 3 files (location.csv,temperature.csv, and wind.csv). Open each file and familiarize yourself with the data format. The data in these files is interpreted as follows:
  + location.csv: station name, latitude, longitude
  + wind.csv: station name, year, month, wind speed
  + temperature.csv: station name, year, month, temperature
* Create database tables to hold the data given in the three files. Make sure to identify primary keys and foreign keys as appropriate.
* Load the data from the given data files into the database tables. Using the Cloudera VM, store the data files under the directory ‘/home/cloudera/Assignment1/data/’. In MySQL you can load data using the following syntax:

mysql>load data infile '/home/cloudera/Assignment1/data/location.csv'

>into table Location

>fields terminated by ','

>lines terminated by '\n';

* If you get an error from MySQL that the file cannot be read, you can change the file permissions as follows: browse to the directory including the file using the file browser, right click on file name, choose ‘Properties’ and make sure all permissions are set to be ‘Read and Write’.

**SQL Queries:**

For each question below, write one or more SQL query to find the required output.

1. Produce a list of station name, year, month, wind speed, temperature.

SELECT Wind.station\_name, Wind.year, Wind.month, wind\_speed, temperature

FROM Wind, Temperature

WHERE Wind.station\_name = Temperature.station\_name and Wind.year = Temperature.year and Wind.month = Temperature.month;

1. For each station, find the total number of valid wind reports and the total number of valid temperature reports. (Note: do not count NULL as a valid report).

SELECT station\_name, SUM(case when wind\_speed != 0 then 1 else 0 end) as Valid\_Wind, SUM(case when temperature!= 0 then 1 else 0 end) as Valid\_Temp

FROM Wind\_Temp\_View

GROUP BY station\_name;

1. For each station, find the total number of wind reports and the total number of temperature reports in each year. From the output, identify stations that did not report a valid reading every month.

select station\_name, year,

sum(case when wind\_speed != 0 then 1 else 0 end) as Valid\_Wind\_Report,

sum(case when temperature != 0 then 1 else 0 end) as Valid\_Temp\_Report

from Wind\_Temp\_View

group by station\_name, year;

1. Find how many wind speed reports are collected in each month? How many temperature reports are collected each month?

SELECT station\_name, month, SUM(case when wind\_speed != 0 then 1 else 0 end) as Valid\_Wind, SUM(case when temperature != 0 then 1 else 0 end) as Valid\_Temp

FROM Wind\_Temp\_View

GROUP BY station\_name, month;

1. For each station, find the first year at which the station started to report wind speeds and the first year at which the station started to report temperature readings.

SELECT station\_name, min(year)

FROM Wind\_Temp\_View

GROUP BY station\_name;

1. Find the coldest and hottest temperatures for each station.

SELECT station\_name as Station, min(temperature) as MIN, max(temperature) as MAX

FROM Temperature

GROUP BY station\_name;

1. What is the maximum and minimum temperatures reported in 2000? What are the stations that reported these maximum and minimum temperatures?

SELECT station\_name, min(temperature) as Min, max(temperature) as Max

FROM Temperature

WHERE year = 2000;

1. What is the average wind speed at 90-degree latitude? (Note: do not look at the data to find what stations are located at 90-degree latitude, however, you have to use 90 in your query)

SELECT Location.station\_name, AVG(wind\_speed)

FROM Wind, Location

WHERE latitude = 90 and Wind.station\_name = Location.station\_name;

1. The name of the weather station at the South Pole is called Clean Air, because very little man-made pollution can be found there. Find out what temperatures were recorded at the South Pole such that the output is sorted by temperatures. When is it Summer and when it is winter in South pole?

select station\_name,year,month, temperature

from Wind\_Temp\_View

where station\_name = 'Clean\_Air'

order by temperature ASC;

From looking at the data my guess for when it is summer is between November – February and winter is from May – September .

1. For each station, find the maximum, minimum, and average temperature reported in a certain month of the year. Comment on when do you think it is summer and when it is winter for each station.

SELECT station\_name,

MAX(temperature) AS MAX\_Temp,

MIN(temperature) AS MIN\_Temp,

AVG(temperature) AS AVG\_Temp

FROM Wind\_Temp\_View

WHERE month = 'jun'

GROUP BY station\_name;

**What to submit:**

Submit only one .sql script (file) that includes SQL statements to:

1. create the database
2. create the tables
3. load tables with data
4. answers to all the 10 queries.
5. Whenever needed, write a comment with each query to answer the question asked on the output of that query.

Make sure that your scripts runs on MySQL without giving any errors. You can test your script on MySQL as follows:

* Write all your sql commands in a file and save with extensions ‘.sql’ (e.g, assignment1.sql)
* Assume, you saved your sql script under the directory ‘/home/cloudera/Assignment1,’. Then you can run the script using the following command:

mysql> source /home/cloudera/Assignment1/assignment1.sql