**M2 Test Submission**

1. **Data Cleaning**

**Commands used in Pig:**

* Source = LOAD '/home/ektashah/heathrow.txt' USING PigStorage('\t') AS (year : int, month :int, tmax : int, tmin:int, af:chararray, rain : int, sun\_hours : chararray);
* rowData = FILTER Source by year IS NOT NULL;
* Datavals = FOREACH rowData GENERATE year, month, tmax, tmin, REPLACE(af,'---','') as FrostDays,rain,REPLACE(sun\_hours,'---','') AS Sunshine\_Hours
* Datavals1 = FOREACH rowData GENERATE year, month, tmax, tmin, af,rain,REPLACE(sun\_hours,'Provisional','') AS Sunhours
* Datavals2 = FOREACH Datavals1 GENERATE year, month, tmax, tmin, af,rain,RTRIM(REPLACE(Sunhours,'#','')) AS Sunhours1
* STORE Datavals2 INTO 'exam/scrubbedweather' USING PigStorage(' ');

1. **Convert Celsius into Fahrenheit - Pig**

**Python code - convert\_weather.py**

@outputSchema("c\_weather: {(year : int, month :int, tmax : float, tmin:float, af:float, rain : float, sun\_hours : float)}")

def fahrenheit(weather):

year, month, tmax, tmin, af, rain, sun\_hours = weather.split(' ')

tmax = float(9.0/5)\*float(tmax)+32

tmin = float(9.0/5)\*float(tmin)+32

return int(year), int(month), tmax, tmin, af, rain, sun\_hours

**Commands for Pig:**

* REGISTER 'exam/convert\_weather.py' USING jython AS convert\_weather;
* rowData = LOAD 'exam/scrubbedweather' as (weather:chararray);
* ConvWeather = FOREACH rowData GENERATE FLATTEN(convert\_weather.fahrenheit(weather));
* STORE ConvWeather INTO 'exam/convertedweather' USING PigStorage(' ');

1. **Hive**

**Create table in hive**

Drop table if exists weather;

create external table weather(

year int,

month int,

tmax float,

tmin float,

af float,

rain float,

sun\_hours float)

row format delimited fields terminated by ' '

Stored as textfile location '/home/ektashah/convertedweather';

**Python Code : convertmm.py**

#!/usr/bin/env python

import sys

import string

while True:

line=sys.stdin.readline()

if not line:

break

row=string.strip(line, "\n")

year, month, tmax, tmin, af, rain, sun\_hours = string.split(row, "\t")

rain = float(rain)/25.4

print "\t".join([year, month, tmax, tmin, af, str(rain),sun\_hours])

**Hive:**

* add file /home/ektashah/convertmm.py;
* SELECT TRANSFORM(year, month, tmax, tmin, af, rain, sun\_hours)

USING 'python convertmm.py'

as(year int, month int, tmax float, tmin float, af float, rain float, sun\_hours float)

from weather;