BIG DATA ANALYTICS

UNIT-II
PIG

- An Example
- Generating Examples

A simple example by writing the program to calculate the maximum recorded temperature by year for the weather dataset in Pig Latin

---

$ bash

hduser072@adminitlab5-OptiPlex-3050:~$ cd Desktop

hduser072@adminitlab5-OptiPlex-3050:~/Desktop$ cd RMMPIGBDA

bash: cd: RMMPIGBDA: No such file or directory

hduser072@adminitlab5-OptiPlex-3050:~/Desktop$ cd RMMBDAPIG

hduser072@adminitlab5-OptiPlex-3050:~/Desktop/RMMBDAPIG$ ls

bigfile pig_1666863504164.log
derby.log pig-java-code
HadoopBook.pdf pigsample
hive-all 'Pig step wise execution instructions'
'hive queries-rmm' pigwordcount.pig
isGood.jar sample.txt
pig_1666849438515.log scirptcode
pig_1666860663135.log upperconverter.jar
pig_1666862859987.log wordcount.txt

hduser072@adminitlab5-OptiPlex-3050:~/Desktop/RMMBDAPIG$ pig -x local

```
grunt>

grunt> records = LOAD 'sample.txt' USING PigStorage(' ') AS (year:Chararray,temperature:int,quality:int);


grunt> dump records;

(1950,0,1)
(1950,22,1)
(1950,-11,1)
(1949,111,1)
(1949,78,1)

grunt>

grunt> filtered_records = FILTER records BY temperature!=9999 AND(quality==0 OR quality==1 OR quality==5 OR quality==9);

grunt> DUMP filtered_records;

(1950,0,1)
(1950,22,1)
(1950,-11,1)
(1949,111,1)
(1949,78,1)

grunt>

grunt> dump filtered_records;

(1950,0,1)
(1950,22,1)
(1950,-11,1)
(1949,111,1)
(1949,78,1)
```

grunt> grouped_records = GROUP filtered_records BY year;

grunt> DUMP grouped_records;

(1949,{(1949,78,1),(1949,111,1)})
(1950,{(1950,-11,1),(1950,22,1),(1950,0,1)})

grunt>

grunt> max_temp = FOREACH grouped_records GENERATE group,MAX(filtered_records.temperature);

grunt> dump max_temp;

(1949,111)
(1950,22)

grunt> describe records;

records: {year: chararray,temperature: int,quality: int}

grunt> describe filtered_records;

filtered_records: {year: chararray,temperature: int,quality: int}

grunt> describe grouped_records;

grouped_records: {group: chararray,filtered_records: {(year: chararray,temperature: int,quality: int)}}

grunt> describe max_temp;

max_temp: {group: chararray,int}

Generating Examples

In this example, we've used a small sample dataset with just a handful of rows to make it easier to follow the data flow and aid debugging. Creating a cut-down dataset is an art, as ideally it should be rich enough to cover all the cases to exercise your queries (the completeness property), yet small enough to make sense to the programmer (the conciseness property). Using a random sample doesn't work well in general because join and filter operations tend to remove all random data, leaving an empty result, which is not illustrative of the general data flow?

With the ILLUSTRATE operator, Pig provides a tool for generating a reasonably complete and concise sample dataset. Here is the output from running ILLUSTRATE on our dataset

```
grunt> ILLUSTRATE max_temp;
-------------------------------------------------------------------------------
| records    | year:chararray    | temperature:int    | quality:int    |
-------------------------------------------------------------------------------
|            | 1949              | 111                | 1              |
|            | 1949              | 78                 | 1              |
|            | 1949              | 0                  | 0              |
|            | 1949              | 9999               | 1              |
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
| filtered_records    | year:chararray    | temperature:int    | quality:int    |
-------------------------------------------------------------------------------
|                     | 1949              | 111                | 1              |
|                     | 1949              | 78                 | 1              |
|                     | 1949              | 0                  | 0              |
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
| grouped_records | group:chararray | filtered_records:bag{:tuple(year:chararray,temperature:int,quality:int)} |
-------------------------------------------------------------------------------
|                 | 1949            | {}                                                                      |
|                 | 1949            | {}                                                                      |
-------------------------------------------------------------------------------
-----------------------------------------------
| max_temp    | group:chararray    | :int    |
-----------------------------------------------
|             | 1949               | 111     |
-----------------------------------------------
grunt>
```