Experiment No: 3

Pig Latin

Pig Latin is a data flow language used by Apache Pig to analyze the data in Hadoop.

Latin Data Types

Scalar types: Int, float, double, chararray, bytearray

Complex types: tuple, bag, map

Apache Pig Run Modes

Apache Pig executes in two modes: Local Mode and MapReduce Mode.

Local : pig -x local

MapReduce: pig -x mapreduce

it will take you into a grunt shell.

bda.txt

```
1991	33	Mumbai
1991	38	Pune
1991	38	Pune
1991	36	Delhi
1990	55	Pune
1990	55	Delhi
1996	65	Mumbai
1996	36	Pune
2020	66	Mumbai
2020	52	Pune
2025	55	Mumbai
2025	62	Delhi
2025	62	Pune
2025	62	Pune
```

```
pig -x local
```

Various Operators in Apache PIG

LOAD: load the text file data from the file system. [local / HDFS]

```
my_bag = LOAD ‘/home/dbit/Desktop/bda.txt’ as  (year:int, temp:int, city:chararray) ;

describe my_bag;

dump my_bag;
```

```
my_bag1 = LOAD ‘/home/dbit/Desktop/bda.txt’ as  (year, temp, city) ;

describe my_bag1;
```

```
my_bag2 = LOAD ‘/home/dbit/Desktop/bda.txt’ ;
describe my_bag2;
```

DISTINCT Operator: is used to remove duplicate tuples in a relation

```
my_bag = LOAD ‘/home/dbit/Desktop/bda.txt’ as  (year:int, temp:int, city:chararray) ;

dump my_bag;

distinct_result = DISTINCT my_bag;

dump distinct_result;
```

FILTER Operator: filter tuples on some condition

```
temp_55 =  FILTER my_bag  by temp>55;
dump temp_55;

city_pune = FILTER my_bag  by city==’Pune’;
dump city_pune;
```

Group Operator: is used to group the data in one or more relations

```
group_city = GROUP my_bag by city;
dump group_city;
```

LIMIT Operator: is used to limit the number of output tuples

```
limit_bag = LIMIT my_bag  5;
dump limit_bag;
```

ORDER BY Operator: sorts a relation based on one or more fields

```
order_bag = ORDER  my_bag by temp ;
dump order_bag;

order_bag5 = ORDER  my_bag by temp DESC;
dump order_bag5;
```

PigStorage: Storing output of PIG in Local file system:

```
store order_bag5 into '/home/dbit/Desktop/order.txt' using PigStorage(';');
```

Exercises for Students:

Q. Load the following data.txt file (on Local file system)

Sam	Mumbai	66
Jim	Pune	77
Tom	Pune	55
Herry	Mumbai	65
Sony	Delhi	51
Sia	Pune	65
Sia	Pune	65
Sia	Pune	65
Tina	Mumbai	53
Jia	Delhi	76
Sara	Delhi	65

Q. Remove duplicate tuples and put in a bag say “result1” and show its content.

Q. Sort the data in ascending and descending both and put the sorted data in the bag “result2” and “result3” respectively.

Q. put the first seven tuples in the bag “result4” and show its content.

Q. Show all the details of the students who scored more than 60 marks in “result5”.

Q. Group the students according to their city and store it in the bag saying “result6”.

Q. Show which student scored maximum and which student scored minimum marks in “result7”

NOTE: Store all outputs [result1 to result7] of above queries in Local file system.