Big Data and Hadoop Developer Social Media Project

By Pravin Wagh

Title	**Analyse data set from Stack Exchange**
Case Study	We need to study and analyse the dataset provided by the Social Media Co. Stack Exchange.
Dataset	The Dataset is provided along with the Project.
Analysis Objective	• Top 10 most commonly used tags in this data set. • Average time to answer questions. • Number of questions which got answered within 1 hour. • Tags of questions which got answered within 1 hour.
Attributes	qid Unique question id i User id of questioner qs Score of the question qt Time of the question (in epoch time) tags a comma-separated list of the tags. qvc Number of views of this question qac Number of answers for this question Aid Unique answer id j User id of answerer as Score of the answer at Time of the answer (in epoch time)
Data Structure	Unstructured

Top 10 most commonly used tags in this data set.

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_dataset_answers1.csv'
            USING PigStorage('_')
            AS
            (qid:chararray,i:chararray,qs:chararray,qt:chararray,tags:chararray,qvc:chararray,
             qac:chararray,aid:chararray,j:chararray,as:chararray,at:chararray);
            generate_tags = FOREACH social_data GENERATE tags;
            token_tags = FOREACH generate_tags GENERATE TOKENIZE(tags);
            format_tags = FOREACH token_tags GENERATE FLATTEN($0) AS tagged;
            group_tags = GROUP format_tags BY tagged;
            count_tags = FOREACH group_tags GENERATE group, COUNT(format_tags) as
                         calccount;
            sort_tags = ORDER count_tags BY calccount DESC;
            top10_tags = LIMIT sort_tags 10;
DUMP top10_tags;
```

Result

```
(1238479830,176)(1242829327,138) (1240545634,102) (1239779339,99) (1237529231,95)
(1241094622,93) (1240352042,92) (1237350979,85) (1242941717,81) (1236696722,76)
```

Average time to answer questions.

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_dataset_answers1.csv'
            USING PigStorage('_')
            AS
            (qid:chararray,i:chararray,qs:chararray,qt:chararray,tags:chararray,qvc:chararray,
             qac:chararray,aid:chararray,j:chararray,as:chararray,at:long);
            GroupQid = GROUP social_data BY qid;
            AvgTime = FOREACH GroupQid GENERATE group, social_data.qid as qid,
                      AVG(social_data.at) as averageTime;
            CalAvgTime = FOREACH AvgTime GENERATE qid,
                         ToDate((long)averageTime*1000) as averageAnsTime;
DUMP CalAvgTime;
```

Result

```
({("1)},1970-01-01T00:00:02.000Z) ({("2)},1970-01-01T00:00:00.000Z) ({("3)},1970-01-
01T00:00:03.000Z) ({("4)},1970-01-01T00:00:18.000Z) ({("5)},1970-01-01T00:00:04.000Z)
({("6)},1970-01-01T00:00:06.000Z) ({("7)},1970-01-01T00:00:01.000Z) ({("8)},1970-01-
01T00:00:12.000Z) ({("9)},1970-01-01T00:00:01.000Z) ({(22)},1970-01-01T00:00:01.000Z)
({(25)},1970-01-01T00:00:01.000Z) ({(40)},1970-01-01T00:00:01.000Z) ({(41)},1970-01-
01T00:00:00.000Z) ({(42)},1970-01-01T00:00:01.000Z)........
```

Number of questions which got answered within 1 hour.

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_dataset_answers1.csv'
                    USING PigStorage('_')
                    AS
                    (qid:chararray,i:chararray,qs:chararray,qt:chararray,tags:chararray,qvc:chararray,
                     qac:chararray,aid:chararray,j:chararray,as:chararray,at:chararray);
                    generate_qid = FOREACH social_data GENERATE qid as q,
                                          ToDate((long)at*1000) as time;
                    qid_gethour = FOREACH generate_qid GENERATE q as q, time as time,
                                         GetHour(time) as hour;
                    qid_hourless1 = FILTER qid_gethour by hour <= 1;
DUMP qid_hourless1;
```

Result

```
("1,1970-01-01T00:00:02.000Z,0) ("2,1970-01-01T00:00:00.000Z,0) ("3,1970-01-
01T00:00:03.000Z,0) ("4,1970-01-01T00:00:18.000Z,0) ("5,1970-01-01T00:00:04.000Z,0)
("6,1970-01-01T00:00:06.000Z,0) ("7,1970-01-01T00:00:01.000Z,0) ("8,1970-01-
01T00:00:12.000Z,0) ("9,1970-01-01T00:00:01.000Z,0) ("10,1970-01-01T00:00:08.000Z,0)
("11,1970-01-01T00:00:01.000Z,0) ("12,1970-01-01T00:00:03.000Z,0) ("13,1970-01-
01T00:00:05.000Z,0) ("14,1970-01-01T00:00:00.000Z,0) …….
```

Tags of questions which got answered within 1 hour

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_dataset_answers1.csv'
                USING PigStorage('_')
                AS
                (qid:chararray,i:chararray,qs:chararray,qt:chararray,tags:chararray,qvc:chararray,
                 qac:chararray,aid:chararray,j:chararray,as:chararray,at:chararray);
                generate_tags = FOREACH social_data GENERATE tags, qid as q,
                                    ToDate((long)at*1000) as time;
                hourly_tags = FOREACH generate_tags GENERATE TOKENIZE(tags), q as q,
                                 GetHour(time) as hour;
                flatten_tags = FOREACH hourly_tags GENERATE FLATTEN($0) AS tag, q as q,
                                hour as hour;
                hourlessOne = FILTER flatten_tags by hour <= 1;
                Order_tags = ORDER hourlessOne by tag;
DUMP Order_tags;
```

Result

```
(1235000081,"1,0) (1235000081,"2,0) (1235000140,"3,0) (1235000140,"4,0)
(1235000140,"5,0) (1235000140,"6,0) (1235000140,"7,0) (1235000140,"8,0)
(1235000140,"9,0) (1235000140,"10,0) (1235000140,"11,0) (1235000140,"12,0)
(1235000140,"13,0) (1235000140,"14,0) (1235000140,"15,0) ....
```