

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	<ul style="list-style-type: none"> • 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	<table> <tr><td>qid</td><td>Unique question id</td></tr> <tr><td>i</td><td>User id of questioner</td></tr> <tr><td>qs</td><td>Score of the question</td></tr> <tr><td>qt</td><td>Time of the question (in epoch time)</td></tr> <tr><td>tags</td><td>a comma-separated list of the tags.</td></tr> <tr><td>qvc</td><td>Number of views of this question</td></tr> <tr><td>qac</td><td>Number of answers for this question</td></tr> <tr><td>Aid</td><td>Unique answer id</td></tr> <tr><td>j</td><td>User id of answerer</td></tr> <tr><td>as</td><td>Score of the answer</td></tr> <tr><td>at</td><td>Time of the answer (in epoch time)</td></tr> </table>	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)
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)
```

Pig Script

Menu

Learning on Simplilearn

Pig

Learning on Simplilearn

+

< > ↺ 🗖

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

0 📶 ❤️ ☁️ ❌ 📶

pravin18in_gmail

My Scripts

Query history

Date	Pig Script
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Page 1 of 1.

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_datase
t_answers1.csv' USING PigStorage('_') AS (qid:chararray,i:chararray,qs:
chararray,qt:chararray,tags:chararray,qvc:chararray,qac:chararray,aid:c
hararray,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 c
alccount;
sort_tags = ORDER count_tags BY calccount DESC;
top10_tags = LIMIT sort_tags 10;
DUMP top10_tags;
```

Close

Windows Taskbar: Windows logo, Task View, Edge, Pro..., Calculator, Un..., Un..., Pig..., Inb..., Ap..., Sp..., Inc..., pra..., W Pr..., P Big..., Un..., Desktop, 4:33 PM 2/9/2017, ENG, 1

Results

Menu Learning on Simplilearn Pig Learning on Simplilearn

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

pravin18in_gmail

My Scripts Query history

Date	Pig Script
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:16	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Page 1 of 1

Results

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

Close

Windows Taskbar: Big... Unti... Pig... Inb... Apa... Spe... Incr... prav... Proj... Big... Desktop 4:20 PM 2/9/2017

Logs

Menu

Learning on Simplilearn

Pig

Learning on Simplilearn

+

< > ↺ 🗖

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

0 ❤ ☁ ❌ 📶

pravin18in_gmail

My Scripts

Query history

Date	Pig Script
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Page 1 of 1.

Logs

s://cloudlabns/tmp/temp-509551022/tmp-1834993195,

Input(s):
Successfully read 263540 records (24806129 bytes) from: "/user/pravin18in_gmail/socialMedia/Project3_dataset_answers1.csv"

Output(s):
Successfully stored 10 records (192 bytes) in: "hdfs://cloudlabns/tmp/temp-509551022/tmp-1834993195"

Counters:
Total records written : 10
Total bytes written : 192
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1485517424020_6319 -> job_1485517424020_6320,
job_1485517424020_6320 -> job_1485517424020_6322,
job_1485517424020_6322 -> job_1485517424020_6323,
job_1485517424020_6323

Close

Windows Taskbar: 🪟 🕒 📁 📧 Pro... 🧮 🖨 📄 🌐 Un... 🔄 Pig... 🔍 Inb... 🔍 Ap... 🔍 Sp... 🔍 Inc... 🔍 pra... 📄 Pr... 📄 Big... 🗑 Un... Desktop 🔊 🔊 🔊 ENG 4:31 PM 2/9/2017 🗨 1

Average time to answer questions.

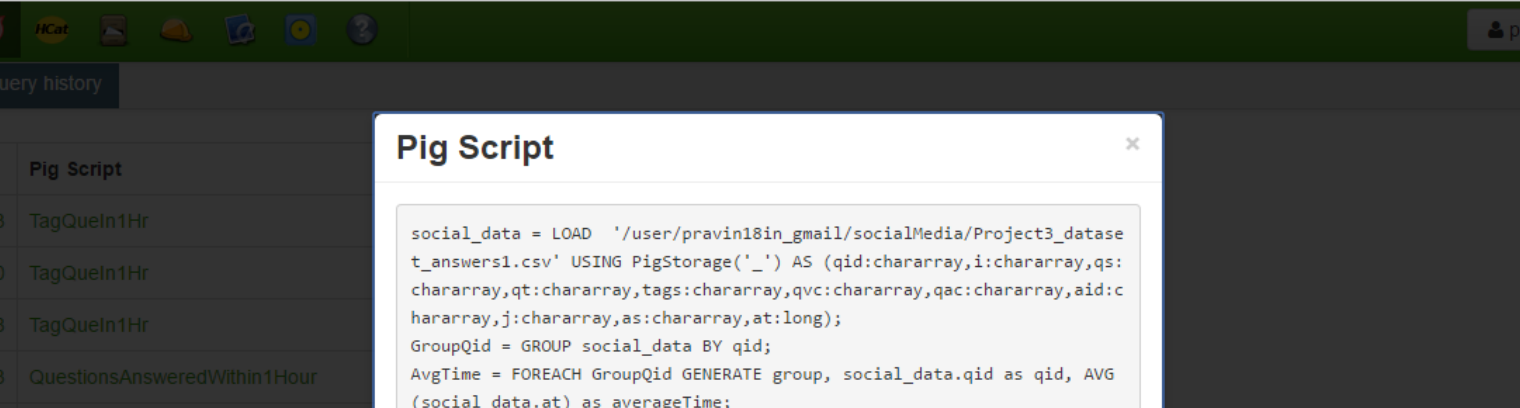
Pig Script

```
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).....
```

Pig Script



The screenshot shows the Amazon EMR console interface. A modal window titled "Pig Script" is open, displaying a Pig script. The background shows a table of query history with columns "Date" and "Pig Script".

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_datase
t_answers1.csv' USING PigStorage('_') AS (qid:chararray,i:chararray,qs:
chararray,qt:chararray,tags:chararray,qvc:chararray,qac:chararray,aid:c
hararray,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*100
0) as averageAnsTime;
DUMP CalAvgTime;
```

The background table shows the following data:

Date	Pig Script
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Results

Menu Learning on Simplilearn Pig Learning on Simplilearn +

< > ↺ ☰ http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/2193/ 0 x ♥ ☁ x ⬇

My Scripts Query history

The Job job_1485517424020_6373 has been started successfully.
You can always go back to [Query History](#) for results after the run.

```
(({"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)
(({43},1970-01-01T00:00:01.000Z)
(({44},1970-01-01T00:00:08.000Z)
(({45},1970-01-01T00:00:05.000Z)
(({46},1970-01-01T00:00:00.000Z)
(({73},1970-01-01T00:00:00.000Z)
(({74},1970-01-01T00:00:02.000Z)
(({ "10"},1970-01-01T00:00:08.000Z)
(({ "11"},1970-01-01T00:00:01.000Z)
(({ "12"},1970-01-01T00:00:03.000Z)
(({ "13"},1970-01-01T00:00:05.000Z)
(({ "14"},1970-01-01T00:00:00.000Z)
(({ "15"},1970-01-01T00:00:00.000Z)
```

Windows Taskbar: File Explorer, Pro..., Calculator, Chrome, Un..., Pig..., Inb..., Ap..., Sp..., Inc..., pra..., W Pr..., P Big..., Un..., Desktop, 4:43 PM 2/9/2017, ENG

Logs

Menu

Learning on Simplilearn

Pig

Learning on Simplilearn

+

< > ↺ 🗖

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

0 ❤ ☁ ❌ 📶

pravin18in_gmail

My Scripts

Query history

Date	Pig Script
09.02.2017 03:10	SocialMediaAverageTimetoAnswerQue
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Page 1 of 1.

Logs

Input(s):

Successfully read 263540 records (24806129 bytes) from: "/user/pravin18in_gmail/socialMedia/Project3_dataset_answers1.csv"

Output(s):

Successfully stored 263540 records (7243683 bytes) in: "hdfs://cloud1abns/tmp/temp262833470/tmp99676778"

Counters:

Total records written : 263540

Total bytes written : 7243683

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job_1485517424020_6374

2017-02-09 11:11:26,535 [main] INFO org.apache.hadoop.yarn.client.ap

Close

Windows Taskbar: Windows logo, Task View, Edge, Pro..., Calculator, Chrome, Un..., Pig..., Inb..., Ap..., Sp..., Inc..., pra..., W Pr..., P Big..., Un..., Desktop, 4:44 PM 2/9/2017, ENG, 1

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) .....
```

Pig Script

Menu

Learning on Simplilearn

Pig

Learning on Simplilearn

+

< > ↺ 🗖

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

0 🔒 ❤️ ☁️ ❌ 📶

pravin18in_gmail

My Scripts

Query history

Date	Pig Script
09.02.2017 03:16	SocialMediaAverageTimetoAnswerQue
09.02.2017 03:10	SocialMediaAverageTimetoAnswerQue
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQuestions
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Page 1 of 1.

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_datase
t_answers1.csv' USING PigStorage('_') AS (qid:chararray,i:chararray,qs:
chararray,qt:chararray,tags:chararray,qvc:chararray,qac:chararray,aid:c
hararray,j:chararray,as:chararray,at:long);
generate_qid = FOREACH social_data GENERATE qid as q, ToDate((long)at*1
000) as time;
qid_gethour = FOREACH generate_qid GENERATE q as q, time as time, GetHo
ur(time) as hour;
qid_hourless1 = FILTER qid_gethour by hour <= 1;
DUMP qid_hourless1;
```

Close

Windows Taskbar: Proj..., Unti..., Pig..., Inb..., Pig..., Rae..., Incr..., prav..., Proj..., Big..., Desktop, 4:56 PM 2/9/2017

Results

Menu Learning on Simplilearn Pig Learning on Simplilearn +

< > ↺ ☰ http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/2195/ 0 x ♥ ☁ x ↻

My Scripts Query history

The Job job_1485517424020_6377 has been started successfully.
You can always go back to Query History for results after the run.

```
("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)
("15,1970-01-01T00:00:00.000Z,0)
("16,1970-01-01T00:00:03.000Z,0)
("17,1970-01-01T00:00:00.000Z,0)
("18,1970-01-01T00:00:01.000Z,0)
("19,1970-01-01T00:00:00.000Z,0)
("20,1970-01-01T00:00:02.000Z,0)
("21,1970-01-01T00:00:00.000Z,0)
(22,1970-01-01T00:00:01.000Z,0)
("23,1970-01-01T00:00:03.000Z,0)
("24,1970-01-01T00:00:01.000Z,0)
(25,1970-01-01T00:00:01.000Z,0)
("26,1970-01-01T00:00:06.000Z,0)
```

Windows Taskbar: Proj... Unti... Pig ... Inb... Pig ... Rae... Incr... prav... Proj... Big... Desktop 5:01 PM 2/9/2017 ENG

Logs

Menu

Learning on Simplilearn

Pig

Learning on Simplilearn

+

< > ↺ 🗖

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

0 ❤ ☁ ❌ 📶

pravin18in_gmail

My Scripts

Query history

Date	Pig Script
09.02.2017 03:16	SocialMediaAverageTimetoAnswerQue
09.02.2017 03:10	SocialMediaAverageTimetoAnswerQue
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQue
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Page 1 of 1.

Logs

7275279/tmp290988009,

Input(s):
Successfully read 263540 records (24806129 bytes) from: "/user/pravin18in_gmail/socialMedia/Project3_dataset_answers1.csv"

Output(s):
Successfully stored 258587 records (6590397 bytes) in: "hdfs://cloudl abns/tmp/temp-987275279/tmp290988009"

Counters:
Total records written : 258587
Total bytes written : 6590397
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1485517424020_6341

Close

Windows Taskbar: Win, Task View, Edge, Proj..., Calculator, File Explorer, Chrome, Un..., Pig ..., Inb..., Pig ..., Rae..., Incr..., prav..., Proj..., Big..., Desktop, 4:57 PM 2/9/2017, ENG, Notification

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) ....
```

Pig Script

Menu

Learning on Simplilearn

Pig

Learning on Simplilearn

+

<

>

↺

🗖

🌐

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

0

🔒

🔖

☁

❌

📶

📁

🐱

🐱

🐱

🐱

🐱

🐱

🐱

pravin18in_gmail

My Scripts

Query history

Date	Pig Script
09.02.2017 03:29	QuestionsAnsweredWithin1Hour
09.02.2017 03:16	SocialMediaAverageTimetoAnswerQu
09.02.2017 03:10	SocialMediaAverageTimetoAnswerQu
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQu
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQu
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQu
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Page 1 of 1.

Pig Script

```
social_data = LOAD '/user/pravin18in_gmail/socialMedia/Project3_datase
t_answers1.csv' USING PigStorage('_') AS (qid:chararray,i:chararray,qs:
chararray,qt:chararray,tags:chararray,qvc:chararray,qac:chararray,aid:c
hararray,j:chararray,as:chararray,at:long);
generate_tags = FOREACH social_data GENERATE tags, qid as q, ToDate((lo
ng)at*1000) as time;
hourly_tags = FOREACH generate_tags GENERATE TOKENIZE(tags), q as q, Ge
tHour(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;
```

Close

Results

Menu Learning on Simplilearn Pig Learning on Simplilearn +

< > ↺ ☰ http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/2196/ 0 x ♥ ☁ x ⬇

My Scripts Query history

The Job job_1485517424020_6385 has been started successfully.
You can always go back to Query History for results after the run.

```
(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")
(1235000140,"16,0")
(1235000140,"17,0")
(1235000140,"18,0")
(1235000369,"19,0")
(1235000369,"20,0")
(1235000369,"21,0")
(1235000377,"22,0")
(1235000414,"23,0")
(1235000414,"24,0")
(1235000427,"25,0")
```

Windows Taskbar: Proj... Unti... Pig ... Inb... Pig ... : Saa... Incr... prav... Proj... Big... Desktop 5:14 PM 2/9/2017

Logs

Menu

Learning on Simplilearn

Pig

Learning on Simplilearn

+

< > ↺ ☰

http://ec2-34-194-195-248.compute-1.amazonaws.com:8000/pig/query_history/#

0 ❤️ ☁️ ❌ 📶

pravin18in_gmail

My Scripts

Query history

Date	Pig Script
09.02.2017 03:42	TagQueIn1Hr
09.02.2017 03:29	QuestionsAnsweredWithin1Hour
09.02.2017 03:16	SocialMediaAverageTimetoAnswerQu
09.02.2017 03:10	SocialMediaAverageTimetoAnswerQu
09.02.2017 02:23	TagQueIn1Hr
09.02.2017 02:20	TagQueIn1Hr
09.02.2017 02:18	TagQueIn1Hr
09.02.2017 01:53	QuestionsAnsweredWithin1Hour
09.02.2017 01:34	SocialMediaAverageTimetoAnswerQu
09.02.2017 01:32	SocialMediaAverageTimetoAnswerQu
09.02.2017 01:29	SocialMediaAverageTimetoAnswerQu
09.02.2017 01:15	Social Media Project
09.02.2017 01:13	Social Media Project

Logs

```
job_1485517424020_6388 1 1 4 4 4 4
6 6 6 6 Order_tags ORDER_BY
hdfs://cloudlabns/tmp/temp1067670054/tmp-1045129771,

Input(s):
Successfully read 263540 records (24806129 bytes) from: "/user/pravin
18in_gmail/socialMedia/Project3_dataset_answers1.csv"

Output(s):
Successfully stored 258587 records (7107571 bytes) in: "hdfs://cloudl
abns/tmp/temp1067670054/tmp-1045129771"

Counters:
Total records written : 258587
Total bytes written : 7107571
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1485517424020_6386 -> job_1485517424020_6387,
job_1485517424020_6387 -> job_1485517424020_6388
```

Close

Windows Taskbar: Proj..., Unti..., Pig..., Inb..., Pig..., : Saa..., Incr..., prav..., Proj..., Big..., Desktop, 5:16 PM 2/9/2017