-------------------------------------------------------------------------------------------------------------------------------

Queries to Load Data from Bucket into Spark sql

-------------------------------------------------------------------------------------------------------------------------------

CREATE EXTERNAL TABLE tags

(id BIGINT,

tag\_name STRING,

count BIGINT,

excerpt\_post\_id BIGINT,

wiki\_post\_id BIGINT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset2/Tags';

-------------------------------------------------------------------------------------------------------------------------------

CREATE EXTERNAL TABLE votes(

id BIGINT,

creation\_date TIMESTAMP,

post\_id BIGINT,

vote\_type\_id BIGINT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset2/Votes/';

-------------------------------------------------------------------------------------------------------------------------------

CREATE EXTERNAL TABLE post\_history(

id BIGINT,

creation\_date TIMESTAMP,

post\_id BIGINT,

post\_history\_type\_id BIGINT,

revision\_guid STRING,

user\_id BIGINT,

text STRING,

comment STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset2/PostHistory/';

-------------------------------------------------------------------------------------------------------------------------------CREATE EXTERNAL TABLE comments(

id BIGINT,

text STRING,

creation\_date TIMESTAMP,

post\_id BIGINT,

user\_id BIGINT,

user\_display\_name STRING,

score BIGINT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset2/Comments';

-------------------------------------------------------------------------------------------------------------------------------

CREATE EXTERNAL TABLE badges(

id BIGINT,

name STRING,

date TIMESTAMP,

user\_id BIGINT,

class BIGINT,

tag\_based BOOLEAN)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset2/Badges';

-------------------------------------------------------------------------------------------------------------------------------

CREATE EXTERNAL TABLE users(

id INTEGER,

display\_name STRING,

about\_me STRING,

age STRING,

creation\_date TIMESTAMP,

last\_access\_date TIMESTAMP,

location STRING,

reputation INTEGER,

up\_votes INTEGER,

down\_votes INTEGER,

views INTEGER,

profile\_image\_url STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset2/Users/';

-------------------------------------------------------------------------------------------------------------------------------

CREATE EXTERNAL TABLE posts\_answers(

id BIGINT,

title STRING,

body STRING,

accepted\_answer\_id STRING,

answer\_count STRING,

comment\_count BIGINT,

community\_owned\_date TIMESTAMP,

creation\_date TIMESTAMP,

favorite\_count STRING,

last\_activity\_date TIMESTAMP,

last\_edit\_date TIMESTAMP,

last\_editor\_display\_name STRING,

last\_editor\_user\_id BIGINT,

owner\_display\_name STRING,

owner\_user\_id BIGINT,

parent\_id BIGINT,

post\_type\_id BIGINT,

score BIGINT,

tags STRING,

view\_count STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset2/PostsAnswers';

-------------------------------------------------------------------------------------------------------------------------------

CREATE EXTERNAL TABLE posts\_questions(

id BIGINT,

title STRING,

body STRING,

accepted\_answer\_id BIGINT,

answer\_count BIGINT,

coment\_count BIGINT,

community\_owned\_date TIMESTAMP,

creation\_date TIMESTAMP,

favorite\_count BIGINT,

last\_activity\_date TIMESTAMP,

last\_edit\_date TIMESTAMP,

last\_editor\_display\_name STRING,

last\_editor\_user\_id BIGINT,

owner\_display\_name STRING,

owner\_user\_id BIGINT,

parent\_id STRING,

post\_type\_id BIGINT,

score BIGINT,

tags STRING,

view\_count BIGINT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 'gs://cloud-cs-643/dataset1/PostsQuestions';

-------------------------------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------------------

Main Queries to Generate Data Insights

-------------------------------------------------------------------------------------------------------------------------------

1) INSIGHT 5.

Select sf\_table.tag,ny\_table.lang\_count,sf\_table.lang\_count,sf\_table.lang\_count-ny\_table.lang\_count as diff from (select pq.tag as tag ,count(\*) as lang\_count from (select id,explode(split(posts\_questions1.tags,'\\|')) as tag,owner\_user\_id from posts\_questions1) as pq join users1 as u on u.id=pq.owner\_user\_id where ( location like '%San Francisco%' or location like '%SF%' ) group by pq.tag) as sf\_table join (select pq.tag as tag ,count(\*) as lang\_count from (select id,explode(split(posts\_questions1.tags,'\\|')) as tag,owner\_user\_id from posts\_questions1) as pq join users1 as u on u.id=pq.owner\_user\_id where ( location like '%New York%') group by pq.tag) as ny\_table on ny\_table.tag=sf\_table.tag order by ny\_table.lang\_count desc limit 100;

-------------------------------------------------------------------------------------------------------------------------------

2) INSIGHT 4.

select ny.lang as language ,ny.count as ny\_count ,sf.count as sf\_count,bang.count as bangalore\_count , london.lang as london\_count from (

select lang, count(\*) as count from (

select pq.tags,"java" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where ( location like '%San Francisco%') and pq.tags like '%java%'

union all

select pq.tags, "python" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where ( location like '%San Francisco%') and pq.tags like '%python%'

union all

select pq.tags, "javascript" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where ( location like '%San Francisco%') and pq.tags like '%javascript%'

union all

select pq.tags, "sql" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where ( location like '%San Francisco%') and pq.tags like '%sql%'

union all

select pq.tags, "c#" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where ( location like '%San Francisco%') and pq.tags like '%c#%'

union all

select pq.tags, "android" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%San Francisco%') and pq.tags like '%android%'

union all

select pq.tags, "html" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%San Francisco%') and pq.tags like '%html%'

union all

select pq.tags, "jquery" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where ( location like '%San Francisco%') and pq.tags like '%jquery%'

union all

select pq.tags, "c++" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%San Francisco%') and pq.tags like '%c++%'

union all

select pq.tags, "css" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%San Francisco%') and pq.tags like '%css%')

group by lang) as ny join

(select lang, count(\*) as count from (

select pq.tags,"java" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%java%'

union all

select pq.tags, "python" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%python%'

union all

select pq.tags, "javascript" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%javascript%'

union all

select pq.tags, "sql" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%sql%'

union all

select pq.tags, "c#" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%c#%'

union all

select pq.tags, "android" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%android%'

union all

select pq.tags, "html" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%html%'

union all

select pq.tags, "jquery" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%jquery%'

union all

select pq.tags, "c++" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%c++%'

union all

select pq.tags, "css" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%New York%') and pq.tags like '%css%')

group by lang) as sf on ny.lang=sf.lang join

(select lang, count(\*) as count from (

select pq.tags,"java" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%java%'

union all

select pq.tags, "python" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%python%'

union all

select pq.tags, "javascript" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%javascript%'

union all

select pq.tags, "sql" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%sql%'

union all

select pq.tags, "c#" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%c#%'

union all

select pq.tags, "android" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%android%'

union all

select pq.tags, "html" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%html%'

union all

select pq.tags, "jquery" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%jquery%'

union all

select pq.tags, "c++" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%c++%'

union all

select pq.tags, "css" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%Bangalore%') and pq.tags like '%css%')

group by lang) as bang on ny.lang=bang.lang join

(select lang, count(\*) as count from (

select pq.tags,"java" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%java%'

union all

select pq.tags, "python" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%python%'

union all

select pq.tags, "javascript" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%javascript%'

union all

select pq.tags, "sql" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%sql%'

union all

select pq.tags, "c#" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%c#%'

union all

select pq.tags, "android" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%android%'

union all

select pq.tags, "html" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%html%'

union all

select pq.tags, "jquery" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%jquery%'

union all

select pq.tags, "c++" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%c++%'

union all

select pq.tags, "css" as lang FROM users1 as u join posts\_questions1 as pq on u.id=pq.owner\_user\_id where (location like '%London%') and pq.tags like '%css%')

group by lang) as london on ny.lang=london.lang

-------------------------------------------------------------------------------------------------------------------------------

3) INSIGHT-2

SELECT REGEXP\_EXTRACT(tags, ‘cloud’ ) AS Tag,

EXTRACT(YEAR FROM creation\_date) AS Year,

COUNT(\*) AS Number\_of\_Questions\_related\_to\_cloud,

SUM(IF(answer\_count > 0, 1, 0)) AS Number\_Questions\_with\_Answers

FROM

posts\_questions

GROUP BY

Tag, Year

HAVING

Year > 2008 AND Year <= 2019 AND Tag is not NULL

ORDER BY

Year;

-------------------------------------------------------------------------------------------------------------------------------

4) INSIGHT-3

SELECT User\_Tenure,

COUNT(1) AS Num\_Users,

ROUND(AVG(reputation)) AS Avg\_Reputation,

ROUND(AVG(num\_badges1)) AS Avg\_Num\_Badges

FROM (

SELECT users.id ,

ROUND(datediff(current\_date(),to\_date(users.creation\_date))/365) AS user\_tenure,

users.reputation AS reputation,

SUM(IF(badges1.user\_id IS NULL, 0, 1)) AS num\_badges

FROM users

LEFT JOIN badges1

ON users.id = badges1.user\_id

GROUP BY users.id, user\_tenure, reputation, num\_badges

)

GROUP BY User\_Tenure

ORDER BY User\_Tenure;

-------------------------------------------------------------------------------------------------------------------------------

5). INSIGHT-1

with ok as(

with ok2 as(SELECT

MOnth\_of\_Year,

Year,

COUNT(1) AS Num\_Questions,

SUM(answered\_in\_1h) AS Num\_Answered\_in\_1H,

ROUND(100 \* SUM(answered\_in\_1h) / COUNT(1),1) AS Percent\_Answered\_in\_1H

FROM

(

SELECT

q.id AS question\_id,

EXTRACT(MONTH FROM q.creation\_date) AS MOnth\_of\_Year,

EXTRACT(YEAR FROM q.creation\_date) AS Year,

MAX(IF(a.parent\_id IS NOT NULL AND

(UNIX\_SECONDS(a.creation\_date)-UNIX\_SECONDS(q.creation\_date))/(60\*60) <= 1, 1, 0)) AS answered\_in\_1h

FROM

posts\_questions1 as q

LEFT JOIN

posts\_answers1 as a

ON q.id = a.parent\_id

GROUP BY question\_id, MOnth\_of\_Year, Year;

)

GROUP BY

MOnth\_of\_Year,Year

ORDER BY

Year)

SELECT

MOnth\_of\_Year,

Year,

Num\_Questions,

Num\_Answered\_in\_1H,

Percent\_Answered\_in\_1H

FROM

ok2

where (Num\_questions in (select Num\_questions from ok2 where Year=2009 order by Num\_questions desc limit 3 ) and Year=2009 )

union all

SELECT

MOnth\_of\_Year,

Year,

Num\_Questions,

Num\_Answered\_in\_1H,

Percent\_Answered\_in\_1H

FROM

ok2

where (Num\_questions in (select Num\_questions from ok2 where Year=2010 order by Num\_questions desc limit 3 ) and Year=2010 )

union all

SELECT

MOnth\_of\_Year,

Year,

Num\_Questions,

Num\_Answered\_in\_1H,

Percent\_Answered\_in\_1H

FROM

ok2

where (Num\_questions in (select Num\_questions from ok2 where Year=2017 order by Num\_questions desc limit 3 ) and Year=2017 )

union all

SELECT

MOnth\_of\_Year,

Year,

Num\_Questions,

Num\_Answered\_in\_1H,

Percent\_Answered\_in\_1H

FROM

ok2

where (Num\_questions in (select Num\_questions from ok2 where Year=2018 order by Num\_questions desc limit 3 ) and Year=2018 )

)

SELECT

MOnth\_of\_Year,

Year,

Num\_Questions,

Num\_Answered\_in\_1H,

Percent\_Answered\_in\_1H

FROM

ok

where (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2008) and Year=2008) or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2009) and Year=2009)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2010) and Year=2010)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2011) and Year=2011)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2012) and Year=2012)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2013) and Year=2013)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2014) and Year=2014) or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2015) and Year=2015)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2016) and Year=2016)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2017) and Year=2017)or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2018) and Year=2018) or (Percent\_Answered\_in\_1H =(select max(Percent\_Answered\_in\_1H) from ok where Year=2019) and Year=2019)