StackOverFlow Data Analysis

Creating and Loading data into tables

Creating comments Table:

CREATE TABLE comment (id int, userid int)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED AS textfile LOCATION '/user/praveenkum79edu/StackOverFlow_Comment_Analysis';

LOAD DATA INPATH '/user/praveenkum79edu/SatckOverFlow_Data/comments.csv' OVERWRITE INTO TABLE comment;



Creating Post Table:

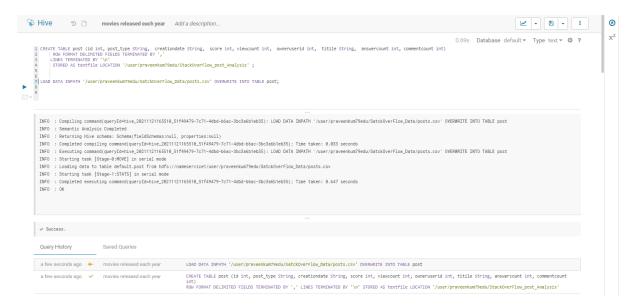
CREATE TABLE post (id int, post_type String, creationdate String, score int, viewcount int, owneruserid int, titile String, answercount int, commentcount int)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED AS textfile LOCATION '/user/praveenkum79edu/StackOverFlow_post_Analysis';

LOAD DATA INPATH '/user/praveenkum79edu/SatckOverFlow_Data/posts.csv' OVERWRITE INTO TABLE post;



Creating Posttype Table:

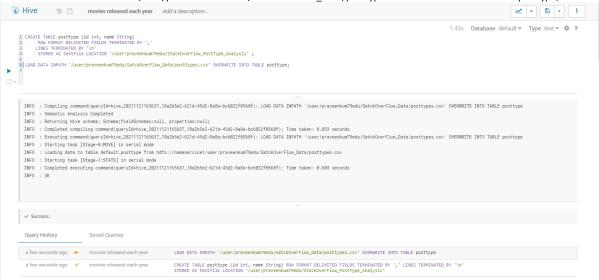
CREATE TABLE posttype (id int, name String)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED AS textfile LOCATION '/user/praveenkum79edu/StackOverFlow_PostType_Analysis';

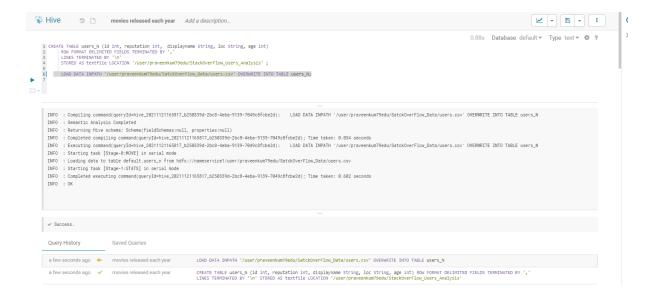
LOAD DATA INPATH '/user/praveenkum79edu/SatckOverFlow_Data/posttypes.csv' OVERWRITE INTO TABLE posttype;



Creating users Table:

CREATE TABLE users_N (id int, reputation int, displayname String, loc String, age int)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS textfile LOCATION '/user/praveenkum79edu/StackOverFlow_Users_Analysis';

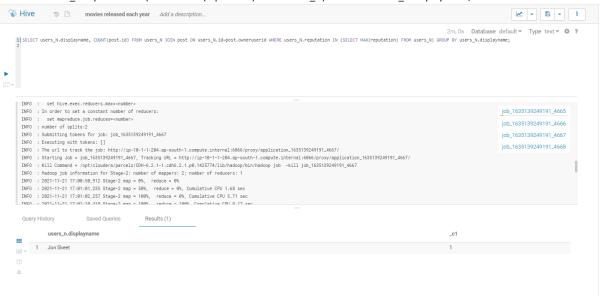
LOAD DATA INPATH '/user/praveenkum79edu/SatckOverFlow_Data/users.csv' OVERWRITE INTO TABLE users_N;



Problem Statement

Find the display name and no. of posts created by the user who has got maximum reputation

SELECT users_N.displayname, COUNT(post.id) FROM users_N JOIN post ON users_N.id=post.owneruserid WHERE users_N.reputation IN (SELECT MAX(reputation) FROM users_N) GROUP BY users_N.displayname;



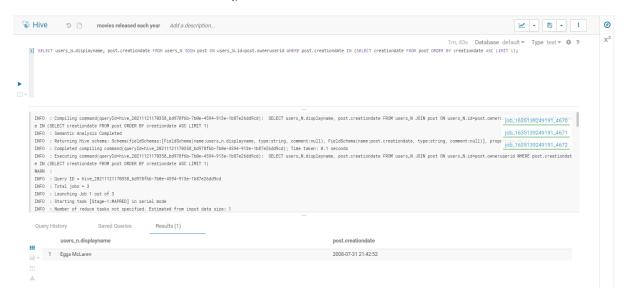
Find the average age of users on the Stack Overflow site.

SELECT CAST(AVG(age) AS INT) FROM users_N;



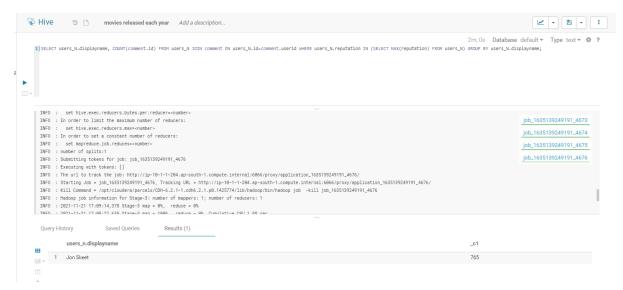
Find the display name of user who posted the oldest post on Stack Overflow (in terms of date).

SELECT users_N.displayname, post.creationdate FROM users_N JOIN post ON users_N.id=post.owneruserid WHERE post.creationdate IN (SELECT creationdate FROM post ORDER BY creationdate ASC LIMIT 1);



Find the display name and no. of comments done by the user who has got maximum reputation

SELECT users_N.displayname, COUNT(comment.id) FROM users_N JOIN comment ON users_N.id=comment.userid WHERE users_N.reputation IN (SELECT MAX(reputation) FROM users_N) GROUP BY users_N.displayname;



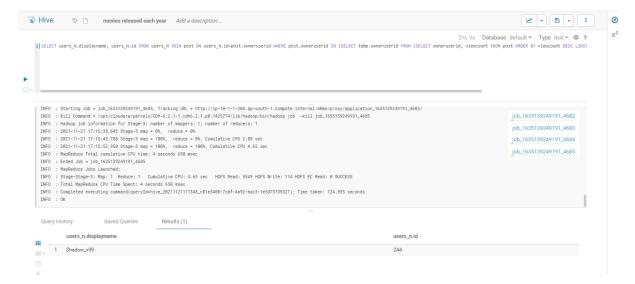
Find the display name of user who has created maximum no. of posts on Stack Overflow.

SELECT users_N.displayname, users_N.id FROM users_N JOIN post ON users_N.id=post.owneruserid WHERE post.owneruserid IN (SELECT temp.owneruserid FROM (SELECT owneruserid, COUNT(id) AS cnt FROM post GROUP BY owneruserid ORDER BY cnt DESC LIMIT 1) temp) group by users_N.displayname, users_N.id;



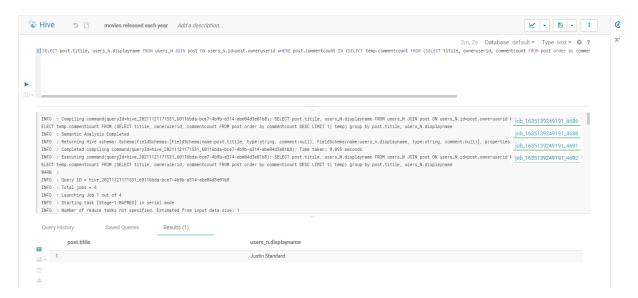
Find the owner name and id of user whose post has got maximum no. of view counts so far

SELECT users_N.displayname, users_N.id FROM users_N JOIN post ON users_N.id=post.owneruserid WHERE post.owneruserid IN (SELECT temp.owneruserid FROM (SELECT owneruserid, viewcount FROM post ORDER BY viewcount DESC LIMIT 1) temp) group by users_N.displayname, users_N.id;



Find the title and owner name of post who has got maximum no. of Comment count

SELECT post.titile, users_N.displayname FROM users_N JOIN post ON users_N.id=post.owneruserid WHERE post.commentcount IN (SELECT temp.commentcount FROM (SELECT titile, owneruserid, commentcount FROM post order by commentcount DESC LIMIT 1) temp) group by post.titile, users N.displayname;



Find the location which has maximum no of Stack Overflow users

SELECT tmpusr.loc FROM (SELECT loc, COUNT(id) AS cid FROM users_N WHERE loc != " GROUP BY loc ORDER BY cid DESC LIMIT 1) tmpusr;



Find the total no. of answers, posts, comments created by Indian users.

SELECT COUNT(comment.id) AS Comments from comment JOIN users_N on comment.userid=users_N.id Where users_N.loc='India' UNION

Select count(post.id) As Posts from post join users_N on post.owneruserid=users_N.id Where users_N.loc='India' UNION SELECT COUNT(temppost.id) Answers FROM (SELECT id, owneruserid FROM post WHERE post_type IN (SELECT id FROM posttype WHERE name='Answer')) temppost join users_N on temppost.owneruserid=users_N.id Where users_N.loc='India';

