POC: Stack Overflow

Problem Statement

1. Find the display name and no. of posts created by the user who has got maximum reputation.
2. Find the average age of users on the Stack Overflow site.
3. Find the display name of user who posted the oldest post on Stack Overflow (in terms of date).
4. Find the display name and no. of comments done by the user who has got maximum reputation.
5. Find the display name of user who has created maximum no. of posts on Stack Overflow.
6. Find the display name of user who has commented maximum no. of times on Stack Overflow.
7. Find the owner name and id of user whose post has got maximum no. of view counts so far.
8. Find the title and owner name of the post which has maximum no. of Answer count.
9. Find the title and owner name of post who has got maximum no. of Comment count.
10. Find the location which has maximum no of Stack Overflow users.
11. Find the total no. of answers, posts, comments created by Indian users.

**Dataset**

**Input format** - .csv

**Dataset Description-**

Comments.csv

id

userid

Posts.csv

id

post\_type

creationdate

score

viewcount

owneruserid

title

answercount

commentcount

PostTypes.csv

id

name

Users.csv

id

reputation

displayname

loc

age

**Algorithm to be followed:**

Step 1: Analysis of dataset and its attributes.

Step 2: Upload the datasets in MySQL.

Step 3: Using Sqoop load all the data into HDFS.

Step 4: Create table in Hive as per the attributes of the datasets provided and load the data into it.

Step 5: Execute the problem statements using SparkSQL.

Step 6: Store the same SparkSQL output in a Hive External Table.

**POC Flow chart:**