Creating the database named as ‘stackoverflow’

CREATE DATABASE stackoverflow LOCATION '/Users/chandana/Documents/Hadoop/hdfs\_tmp';

Using the database stackoverflow

use stackoverflow;

Creating the necessary tables: Tags, Answers, Questions

CREATE TABLE TAGS (id INT, Tags string) row format delimited fields terminated by ',';

LOAD DATA INPATH '/Users/chandana/Documents/Hadoop/hdfs\_tmp/Tags.csv' OVERWRITE INTO TABLE Tags;

CREATE TABLE answers (id INT, owner\_user\_id INT,creation\_date string, parent\_id INT, score INT) row format delimited fields terminated by ',';

LOAD DATA INPATH '/Users/chandana/Documents/Hadoop/hdfs\_tmp/Answers.csv' OVERWRITE INTO TABLE answers;

CREATE TABLE questions (id INT, owner\_user\_id INT, creation\_date string, score INT, Title string) row format delimited fields terminated by ',';

LOAD DATA INPATH '/Users/chandana/Documents/Hadoop/hdfs\_tmp/Questions.csv' OVERWRITE INTO TABLE questions;

1. How many unique questions are there in the dataset?

select count (distinct id) from questions;

O/P: 118978

2. How many users have asked at least 1 question?

select count (distinct owner\_user\_id) from questions;

O/P: 40712

3. How many users have written more than 1 answer?

Select count(\*) as count\_of\_users from (select count (owner\_user\_id) as no\_of\_users from answers group by owner\_user\_id having count (owner\_user\_id) >1) my\_new\_table;

o/p: 5280

4. Find the top 5 questions based on the descending order of the question score

select title,score from questions order by score desc limit 5;

"What does the ""yield"" keyword do?" 5524

What is a metaclass in Python? 3219

How do I check whether a file exists using Python? 2729

Does Python have a ternary conditional operator? 2655

Calling an external command in Python 2312

5. How many average questions are asked per day?

select To\_Date(creation\_date) from answers;

-----Check

Select sum(no\_of\_questions\_asked) as total\_questions, count (distinct To\_Date(creation\_date)) as total\_no\_of\_days, sum(no\_of\_questions\_asked)/count (distinct To\_Date(creation\_date)) as avg\_questions from (Select count (id) as no\_of\_questions\_asked, creation\_date from questions group by creation\_date) my\_new\_table;

O/P: 118978 1479 80.4448951994591

6. How many questions do not have any answer at all? (number of  
questions with no answers)

select count(distinct id) from questions where not exists (select parent\_id from answers where answers.parent\_id=questions.id);

O/P: 94877

7. How many average answers are there per question?

Select sum(no\_of\_answers) as total\_answers,count (\*) as total\_no\_questions, sum (no\_of\_answers)/count(\*) as avg\_answers from (select count (id) as no\_of\_answers, parent\_id from answers group by parent\_id) my\_new\_table;

o/p: 68522 24101 2.843118542799054

8.  How many average answers are there per question, excluding the  
questions without any answer?

9. How many questions are there on python syntax?

select count(tags) from tags where tags like '%python%'; 704197

select count(tags) from tags where tags like '%syntax%'; 1904

select count(tags) from tags where tags like '%python%' and tags like '%syntax%';

0

Select count ([questions.id](http://questions.id)) as no\_of\_questions from questions where id in (select id from tags python join tags syntax on python.id = syntax.id where python.tags = ‘python’ and syntax.tags = ‘syntax’);