HADOOP Assignment 2- Solutions

Document Authorization

Document name	HADOOP Assignment 2- Solutions.docx
Author(s)	Vanesa Lopez Garcia, Sabrina Mirtcheva, Daan Pelt, Moritz Steinbrecher, Hsiu-Chi Liu, Andrew Rizk, Pravat Ranjan Pasayat
Department	Group O-2-4, MBD-2018, IE HST
Last modified by	
Last modified date	20-Nov-2018
Authorization type	Baselined
Version	1.0

CONTENTS

1. Introduction

Use crypto database

```
use crypto;
```

Create an external table crypto.team_d over the files located in /user/flume/tweets/

```
create external table team_d (
id bigint,
lang string,
place struct<name:string,country_code:string>,
entities struct<media:array<struct<id:bigint,media_url:string>>>,
`user` struct<id:bigint,geo_enabled:boolean,followers_count:int>
)
row format serde 'org.apache.hive.hcatalog.data.JsonSerDe'
stored as textfile
location '/user/flume/tweets';
```

(1) Write a query that returns the total number of tweets in table crypto.team_d

```
select count(*) from team_d;
```

(2 & 3) Create a managed table crypto.team_d_parquet with same structure as crypto.team_d but stored in format PARQUET and insert all rows from crypto.team_d into crypto.team_d_parquet

```
create table team_d_parquet
stored as Parquet as
select * from team_d;
```

(4) Write a query that returns the total number of tweets in table crypto.team_d_parquet

```
select count(*) from team_d_parquet;
```

(5) Verify that both tables contain the same number of tweets and how much different are the query response times.

Both contains same number of rows (5300).

Time taken (count from team_d_parquet table): 0.048 seconds
Time taken (count from team_d table): 23.345 seconds

(6) Write a query that returns users with geolocation enabled in table crypto.team_d_parquet

```
select `user` from team_d_parquet where `user`.geo_enabled = true;
```

(7) Write a query that returns the number of tweets per language in table crypto.team_d_parquet

```
select lang, count(*) from team_d_parquet group by lang;
```

(8) Write a query that returns the top 10 users with more followers in table crypto.team_d_parquet

```
select `user` from team_d_parquet order by `user`.followers_count desc limit 10;
```

(9) Write a query that returns the geoname latitude,longitude and timezone by joining crypto.geonames and crypto.team_d_parquet

```
select a.id, b.latitude, b.longitude, b.timezone from team_d_parquet as a left join geonames as b on
upper(a.place.name) = upper(b.name) and a.place.country_code = b.country_code;
```

(10) Write a query that returns the average number of media elements in a tweet from table crypto.team_d_parquet

```
select count(entities.media)/count(*) from team_d_parquet;
```

(11) Write a query that returns the top 10 websites whose media contents are being shared from table crypto.team_d_parquet

```
CREATE TEMPORARY MACRO website(url string) parse_url(url, 'HOST');

select website(exp.url), count(*) as total from team_d_parquet d lateral view
explode(entities.media.media_url) exp as url group by website(exp.url) order by total desc limit 10;
```