Project 05 - Big Data Social Network

The *Big Data Analytics Company* wants to import data from its social network database, which has the following logical schema:

```
USER(email, fullname, age, followers_count)
POST(post_id, category, date, user_email)
LIKES(user_email, post_id)
```

Problem 1 (Sqoop - Hive)

Design a distributed datawarehouse for importing data from production database. The company is interested in incremental imports of information regarding users and posts and so, data should be efficiently stored with respect to the extraction date.

Problem 2 (HBase - Map Reduce)

Design a `MapReduce` job that joins *post.csv* and *likes.csv* into an `HBase` table (`posts`) with the following schema:

- Row key: `date` and `post_id`
- Family `info`: stores `category` and `user_email`
- Family `likes`: users that likes the post

Design a `MapReduce` job that reads from the previous `HBase` table and computes:

- For each user u and category c, the count of posts with category c made by u
- For each user u and category c, the average number of likes obtained by posts with category c made by u

Problem 3 (Spark)

The company is interested in predicting users' age by looking at common interests among users. In particular, starting from the *users.csv* dataset and computed metrics, compute:

- The most frequent category among the posts of the user u
- The category with the highest average of likes among posts of the user u

Resulting schema should be:

```
USER(email, age, followers_count, cat_max_post_count,cat_max_avg_post_likes)
```

Build a regression model to estimate the age of a user starting from statics about user posts. Choose the best model with the best hyperparmeters configuration.