Скачала датасет и загрузила его в свою домашнюю директорию HDFS:

```
taya — student4_1@manager:~/kaggle_task — ssh student4_1@manager.novalocal -i ~/private_key_hadoop —...
[student4_1@manager kaggle_task]$ hdfs dfs -mkdir /user/student4_1/RAW_recipes
[student4_1@manager kaggle_task]$ hdfs dfs -copyFromLocal ./RAW_recipes.csv /user/student4_1/RAW_recipes
[student4_1@manager kaggle_task]$ hdfs dfs -mkdir /user/student4_1/RAW_interactions
[student4_1@manager kaggle_task]$ hdfs dfs -copyFromLocal ./RAW_interactions.csv /user/student4_1/RAW_inte
ractions
[student4_1@manager kaggle_task]$ hdfs dfs -mkdir /user/student4_1/interactions_validation
[student4_1@manager kaggle_task]$ hdfs dfs -copyFromLocal ./interactions_validation.csv /user/student4_1/i
nteractions_validation
[student4_1@manager kaggle_task]$ hdfs dfs -mkdir /user/student4_1/interactions_train
[student4_1@manager kaggle_task]$ hdfs dfs -copyFromLocal ./interactions_train.csv /user/student4_1/intera
ctions_train
[student4_1@manager kaggle_task]$ hdfs dfs -mkdir /user/student4_1/interactions_test
[student4_1@manager kaggle_task]$ hdfs dfs -copyFromLocal ./interactions_test.csv /user/student4_1/interac
tions_test
[student4_1@manager kaggle_task]$ hdfs dfs -ls /user/student4_1
Found 8 items
drwx------   - student4_1 student4_1          0 2020-06-08 18:45 /user/student4_1/.Trash
drwxr-xr-x   - student4_1 student4_1          0 2020-06-08 18:42 /user/student4_1/PP_recipes
drwxr-xr-x   - student4_1 student4_1          0 2020-06-08 16:34 /user/student4_1/PP_users
drwxr-xr-x   - student4_1 student4_1          0 2020-06-08 18:48 /user/student4_1/RAW_interactions
drwxr-xr-x   - student4_1 student4_1          0 2020-06-08 18:47 /user/student4_1/RAW_recipes
drwxr-xr-x   - student4_1 student4_1          0 2020-06-08 18:51 /user/student4_1/interactions_test
drwxr-xr-x   - student4_1 student4_1          0 2020-06-08 18:50 /user/student4_1/interactions_train
drwxr-xr-x   - student4_1 student4_1          0 2020-06-08 18:49 /user/student4_1/interactions_validation
```

Создала базу данных student4_1 в HIVE и EXTERNAL таблицы внутри базы данных:

```
1  create database student4_1
2
3  CREATE EXTERNAL TABLE `student4_1.PP_users`(
4  `u` string COMMENT 'from deserializer',
5  `techniques` string COMMENT 'from deserializer',
6  `items` string COMMENT 'from deserializer',
7  `n_items` string COMMENT 'from deserializer',
8  `ratings` string COMMENT 'from deserializer',
9  `n_ratings` string COMMENT 'from deserializer')
10 ROW FORMAT SERDE
11   'org.apache.hadoop.hive.serde2.OpenCSVSerde'
12 STORED AS INPUTFORMAT
13   'org.apache.hadoop.mapred.TextInputFormat'
14 OUTPUTFORMAT
15   'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
16 LOCATION
17   '/user/student4_1/PP_users'
18 TBLPROPERTIES (
19   'serialization.null.format'='',
20   'skip.header.line.count'='1',
21   'transient_lastDdlTime'='1591631777')
22
23 CREATE EXTERNAL TABLE `student4_1.PP_recipes`(
24 id int,
25 i int,
26 name_tokens string,
27 ingredient_tokens string,
28 steps_tokens string,
29 techniques string,
30 calorie_level int,
31 ingredient_ids string
32 )
33 ROW FORMAT SERDE
```

Query History | Saved Queries | Results (7)

	tab_name
1	interactions_test
2	interactions_train
3	interactions_validation
4	pp_recipes
5	pp_users
6	raw_interactions
7	raw_recipes

Запросы:

```
167 select count(id) from student4_1.pp_recipes group by calorie_level limit 5;
```

Query History | Saved Queries | Results (3+)

	_c0
1	69699
2	63255
3	45311

```
170 select name, steps
171 from student4_1.RAW_interactions i
172 join
173 student4_1.RAW_recipes r
174 on i.user_id = r.id
175 where steps is not null
176 limit 30;
```

Query History | Saved Queries | Results (30)

	name	steps
1	use whatever kind of cheese you like. (i normally throw in whatever leftover odd bits are in the fridge	'roasted re
2	once upon a time	have a tast
3	this recipe is making corn bread from scratch	'salt'