NoSQL Systems - IMT2021041 (Bhavil)

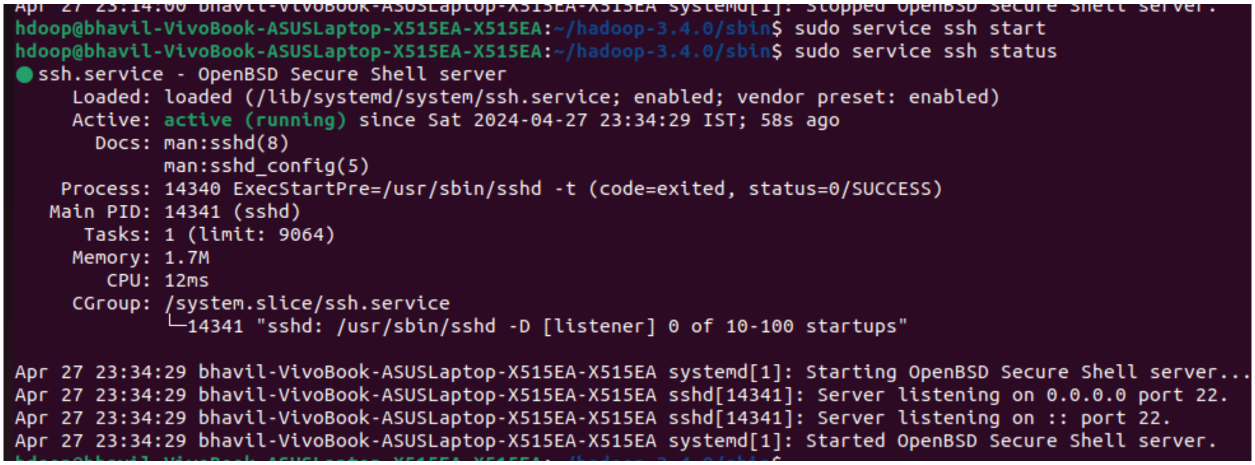
Assignment - 3

Apache Pig and Hive

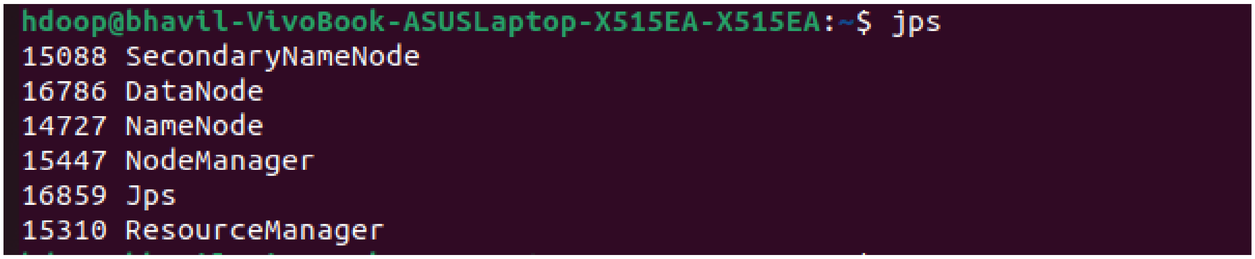
Introduction

Here, I installed Hive and Pig as told in the videos provided by sir.

The readme file has instructions to run them.



```
Apr 27 23:14:00 bhavil-VivoBook-ASUSLaptop-X515EA-X515EA systemd[1]: Stopped OpenBSD Secure Shell server.
hdoop@bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:~/hadoop-3.4.0/sbin$ sudo service ssh start
hdoop@bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:~/hadoop-3.4.0/sbin$ sudo service ssh status
● ssh.service - OpenBSD Secure Shell server
     Loaded: loaded (/lib/systemd/system/ssh.service; enabled; vendor preset: enabled)
     Active: active (running) since Sat 2024-04-27 23:34:29 IST; 58s ago
       Docs: man:sshd(8)
             man:sshd_config(5)
    Process: 14340 ExecStartPre=/usr/sbin/sshd -t (code=exited, status=0/SUCCESS)
   Main PID: 14341 (sshd)
      Tasks: 1 (limit: 9064)
     Memory: 1.7M
        CPU: 12ms
     CGroup: /system.slice/ssh.service
             └─14341 "sshd: /usr/sbin/sshd -D [listener] 0 of 10-100 startups"

Apr 27 23:34:29 bhavil-VivoBook-ASUSLaptop-X515EA-X515EA systemd[1]: Starting OpenBSD Secure Shell server...
Apr 27 23:34:29 bhavil-VivoBook-ASUSLaptop-X515EA-X515EA sshd[14341]: Server listening on 0.0.0.0 port 22.
Apr 27 23:34:29 bhavil-VivoBook-ASUSLaptop-X515EA-X515EA sshd[14341]: Server listening on :: port 22.
Apr 27 23:34:29 bhavil-VivoBook-ASUSLaptop-X515EA-X515EA systemd[1]: Started OpenBSD Secure Shell server.
```



```
hdoop@bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:~$ jps
15088 SecondaryNameNode
16786 DataNode
14727 NameNode
15447 NodeManager
16859 Jps
15310 ResourceManager
```

Effectively, we have to have hadoop running with the help of ssh localhost.

Only then Pig and Hive will work.

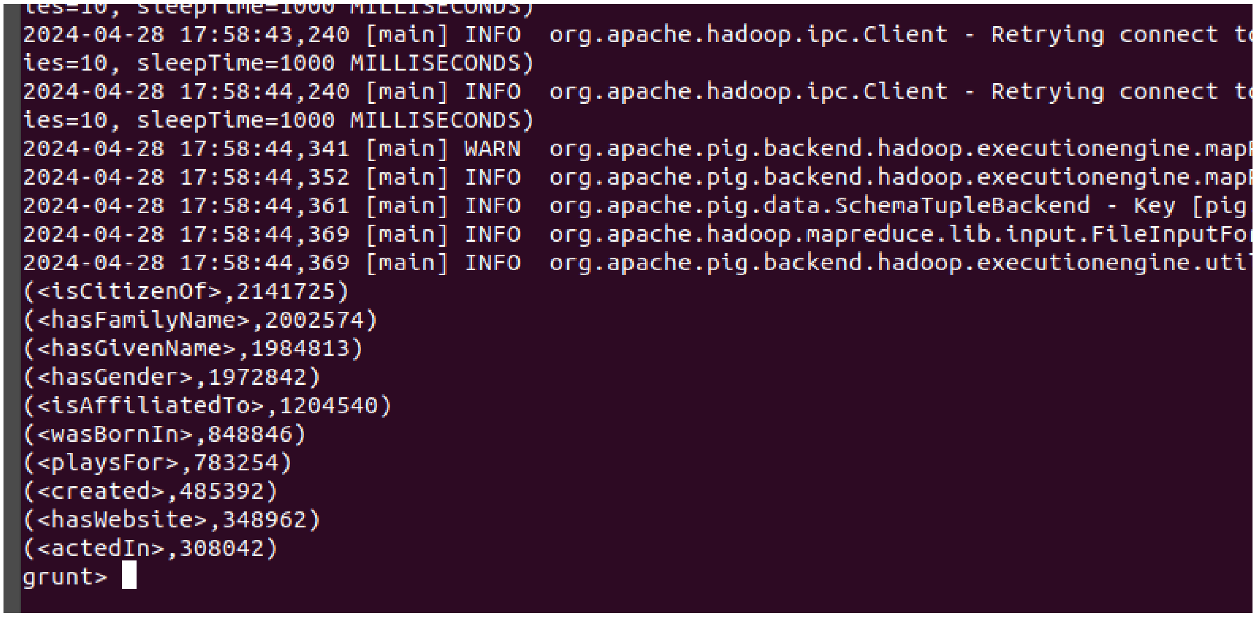
Otherwise, we will keep getting network problem.

PIG and PIG Latin

Part-A-1

The Pig logic starts by organizing the data into a schema with three fields: Subject, Predicate, and Object. Next, it filters the data to retain only the Predicate field due to memory constraints encountered when attempting to keep all fields. Then, the records are grouped based on the Predicate field. For each group, the goal is to determine the count of all members in the original table corresponding to that Predicate. Finally, the groups are sorted, and the top three results are returned, utilizing the ORDER BY and LIMIT operations to achieve this.

*In my pig's grunt, I had this thing, that grunt wasn't able to connect for 10 mins, then suddenly, it reconnected successfully, and ran my command.

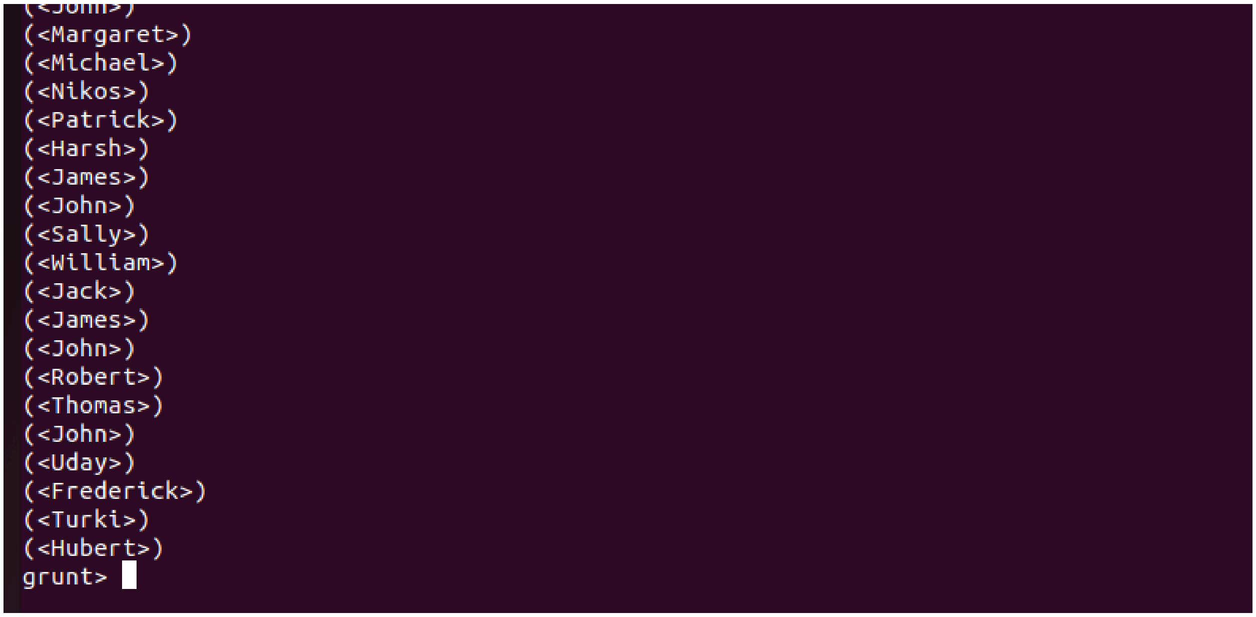


```
ies=10, sleepTime=1000 MILLISECONDS)
2024-04-28 17:58:43,240 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to
ies=10, sleepTime=1000 MILLISECONDS)
2024-04-28 17:58:44,240 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to
ies=10, sleepTime=1000 MILLISECONDS)
2024-04-28 17:58:44,341 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapR
2024-04-28 17:58:44,352 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapR
2024-04-28 17:58:44,361 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig
2024-04-28 17:58:44,369 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFor
2024-04-28 17:58:44,369 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util
(<isCitizenOf>,2141725)
(<hasFamilyName>,2002574)
(<hasGivenName>,1984813)
(<hasGender>,1972842)
(<isAffiliatedTo>,1204540)
(<wasBornIn>,848846)
(<playsFor>,783254)
(<created>,485392)
(<hasWebsite>,348962)
(<actedIn>,308042)
grunt>
```

Part-A-2

The Pig logic begins by filtering the data to retain only records with specific predicates, resulting in dataset R1. Next, it further filters the data to isolate records related to the "livesInPredicate," forming dataset R2. R2 is then grouped based on the subject's name. For each group in R2, the count of "livesIn" clauses associated with each name is calculated. Subsequently, R2 is filtered to include only names with a count greater than one. Following this, a join operation is performed between R1 and R2, where the object field of R1 is extracted. Finally, the resulting dataset is printed as the answer.

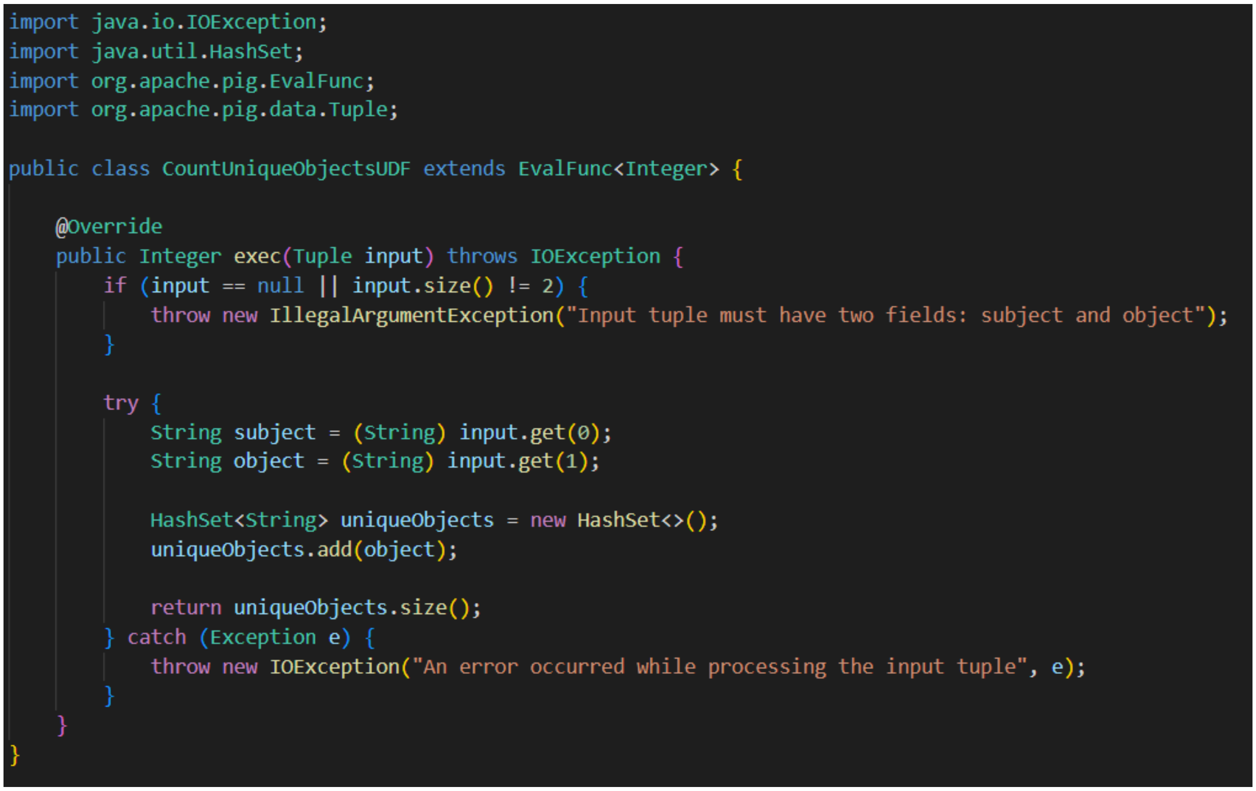
(a lot of names were there, filled my entire terminal, so couldnt get them)



```
(<John>)
(<Margaret>)
(<Michael>)
(<Nikos>)
(<Patrick>)
(<Harsh>)
(<James>)
(<John>)
(<Sally>)
(<William>)
(<Jack>)
(<James>)
(<John>)
(<Robert>)
(<Thomas>)
(<John>)
(<Uday>)
(<Frederick>)
(<Turki>)
(<Hubert>)
grunt>
```

Part-B

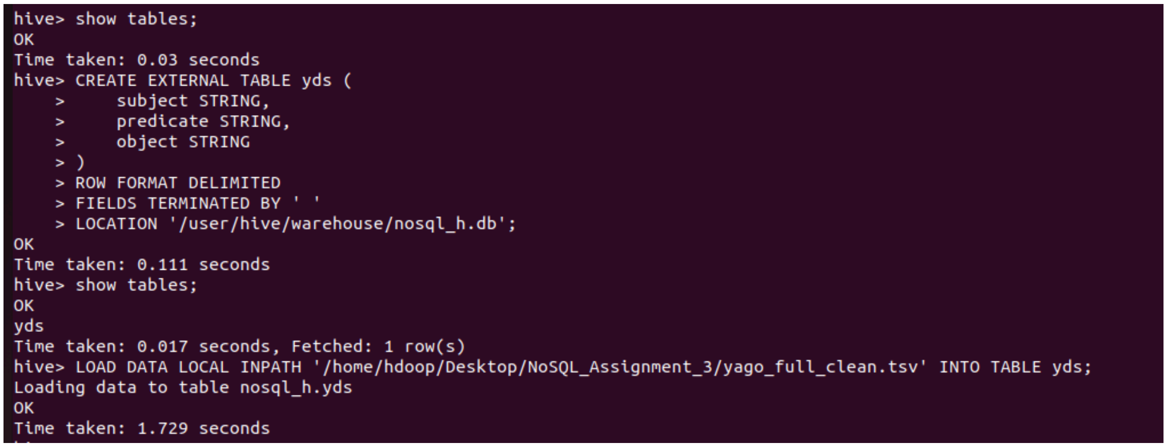
Here, we have to write the java code, make a jar file out of it, and register the jar file as a function in pig, and then use it as a regular operator in it.



```java
import java.io.IOException;
import java.util.HashSet;
import org.apache.pig.EvalFunc;
import org.apache.pig.data.Tuple;

public class CountUniqueObjectsUDF extends EvalFunc<Integer> {

    @Override
    public Integer exec(Tuple input) throws IOException {
        if (input == null || input.size() != 2) {
            throw new IllegalArgumentException("Input tuple must have two fields: subject and object");
        }

        try {
            String subject = (String) input.get(0);
            String object = (String) input.get(1);

            HashSet<String> uniqueObjects = new HashSet<>();
            uniqueObjects.add(object);

            return uniqueObjects.size();
        } catch (Exception e) {
            throw new IOException("An error occurred while processing the input tuple", e);
        }
    }
}
```

Hive and HiveQL

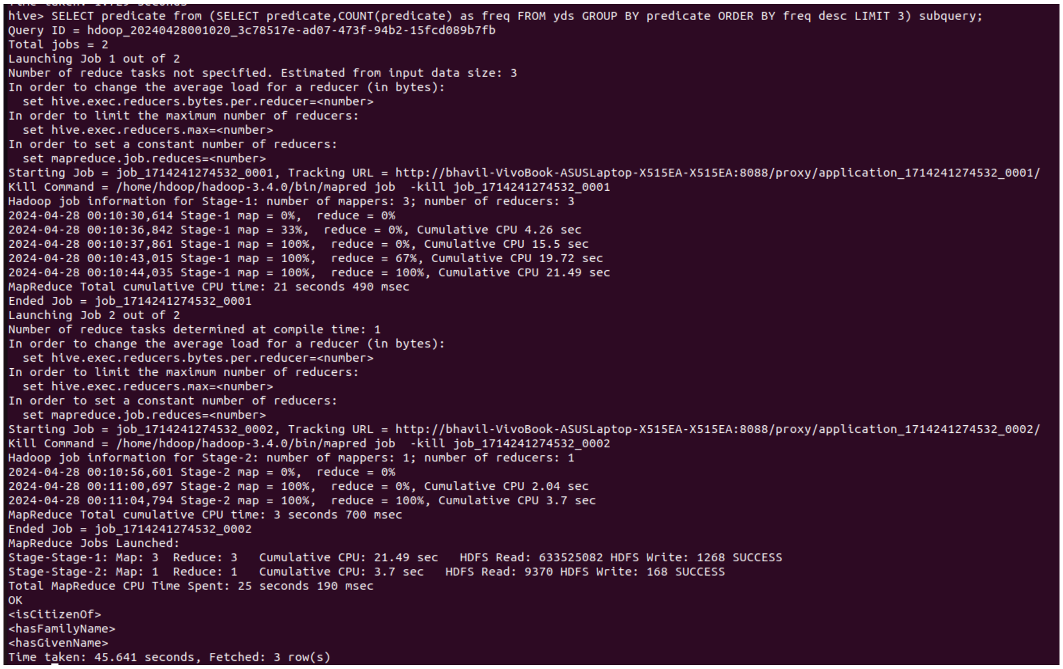
Creating the database, and loading the data to it.



```
hive> show tables;
OK
Time taken: 0.03 seconds
hive> CREATE EXTERNAL TABLE yds (
    >     subject STRING,
    >     predicate STRING,
    >     object STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ' '
    > LOCATION '/user/hive/warehouse/nosql_h.db';
OK
Time taken: 0.111 seconds
hive> show tables;
OK
yds
Time taken: 0.017 seconds, Fetched: 1 row(s)
hive> LOAD DATA LOCAL INPATH '/home/hdoop/Desktop/NoSQL_Assignment_3/yago_full_clean.tsv' INTO TABLE yds;
Loading data to table nosql_h.yds
OK
Time taken: 1.729 seconds
```

Part-A-1

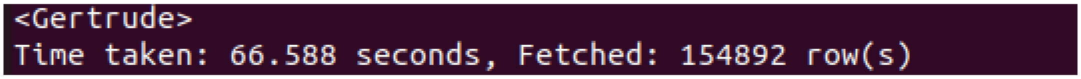
Group the data in the relation by the predicate field. For each unique predicate, calculate the frequency (count) of occurrences and alias it as freq. Order the grouped data by the frequency (freq) in descending order. Limit the result to the top 3 rows. Select only the predicate field from the result.



```
hive> SELECT predicate from (SELECT predicate,COUNT(predicate) as freq FROM yds GROUP BY predicate ORDER BY freq desc LIMIT 3) subquery;
Query ID = hdoop_20240428001020_3c78517e-ad07-473f-94b2-15fcd089b7fb
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 3
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714241274532_0001, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714241274532_0001/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714241274532_0001
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3
2024-04-28 00:10:30,614 Stage-1 map = 0%,  reduce = 0%
2024-04-28 00:10:36,842 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 4.26 sec
2024-04-28 00:10:37,861 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.5 sec
2024-04-28 00:10:43,015 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 19.72 sec
2024-04-28 00:10:44,035 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 21.49 sec
MapReduce Total cumulative CPU time: 21 seconds 490 msec
Ended Job = job_1714241274532_0001
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714241274532_0002, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714241274532_0002/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714241274532_0002
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-04-28 00:10:56,601 Stage-2 map = 0%,  reduce = 0%
2024-04-28 00:11:00,697 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.04 sec
2024-04-28 00:11:04,794 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.7 sec
MapReduce Total cumulative CPU time: 3 seconds 700 msec
Ended Job = job_1714241274532_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 21.49 sec   HDFS Read: 633525082 HDFS Write: 1268 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.7 sec   HDFS Read: 9370 HDFS Write: 168 SUCCESS
Total MapReduce CPU Time Spent: 25 seconds 190 msec
OK
<isCitizenOf>
<hasFamilyName>
<hasGivenName>
Time taken: 45.641 seconds, Fetched: 3 row(s)
```

Part-A-2

Filter the relation to retain only records with the predicate <hasGivenName>, aliased as table1. Filter the relation to retain only records with the subject found in a subquery result where subjects have more than one occurrence of the predicate <livesIn>, aliased as table2. Perform an inner join between table1 and table2 based on the subject field. Select only the object field from the result.



```
<Gertrude>
Time taken: 66.588 seconds, Fetched: 154892 row(s)
```

Part-B-1

Had to set these first for both partitioning and bucketing:

set hive.auto.convert.sortmerge.join = true;

set hive.optimize.bucketmapjoin = true;

set hive.optimize.bucketmapjoin.sortedmerge = true;

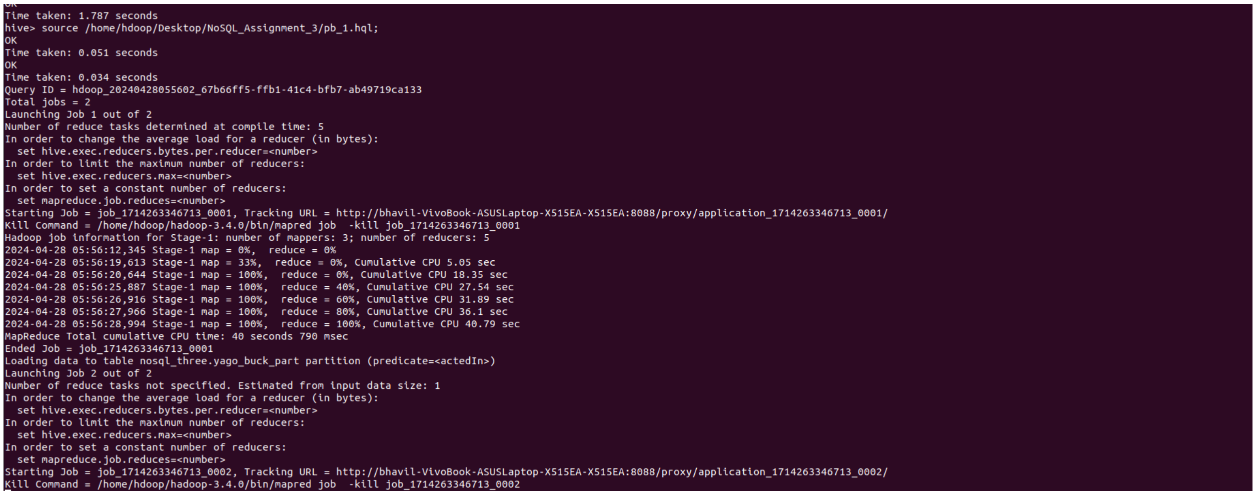
Partitioned by the predicate(string) and made 5 buckets.

Made an external table called yago, and put it inside hive warehouse. Created a subject and object table, and put it inside hive warehouse.

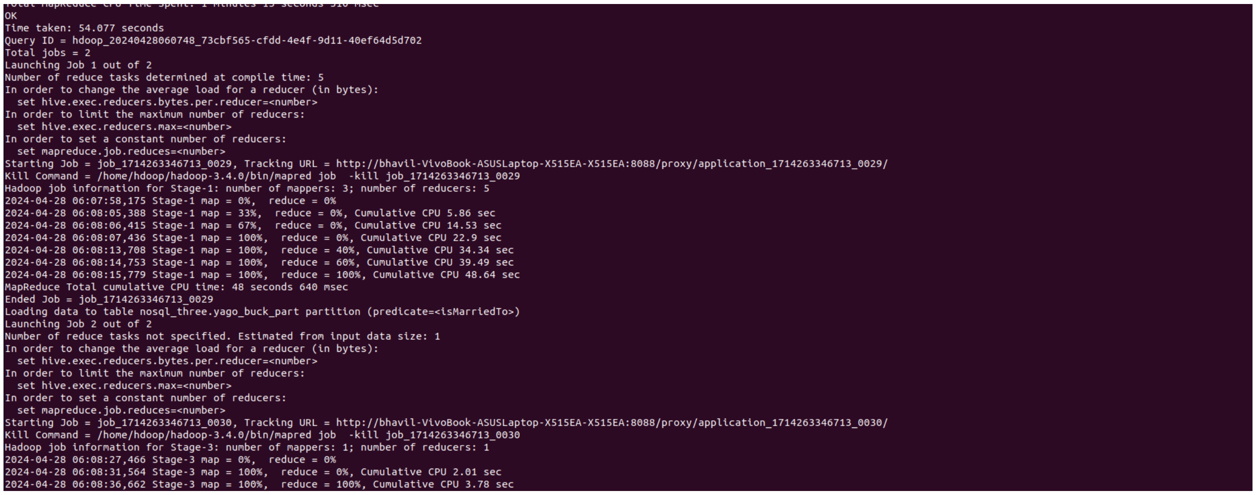
Note that, in the warehouse, we will be having the database, inside the db, we will be putting these tables.

Partitioned for all 29 predicates.

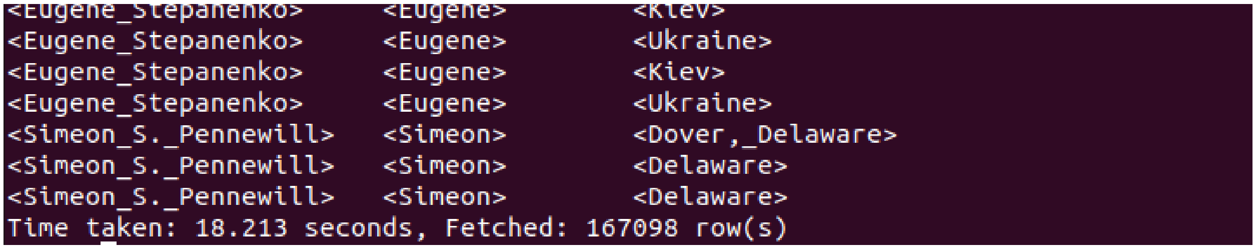
This Hive query retrieves data from the yago_buck_part table, specifically selecting subjects and objects where the predicate is "<hasGivenName>" and "<livesIn>" respectively. It then joins these selections based on the subject. In other words, it identifies individuals who have been given a name (the subject) and where they live (the object), linking these details together. The resulting output provides a list of individuals along with their given names and corresponding places of residence.



```
OK
Time taken: 1.787 seconds
hive> source /home/hdoop/Desktop/NoSQL_Assignment_3/pb_1.hql;
OK
Time taken: 0.051 seconds
OK
Time taken: 0.034 seconds
Query ID = hdoop_20240428055602_67b66ff5-ffb1-41c4-bfb7-ab49719ca133
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks determined at compile time: 5
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714263346713_0001, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714263346713_0001/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714263346713_0001
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 5
2024-04-28 05:56:12,345 Stage-1 map = 0%,  reduce = 0%
2024-04-28 05:56:19,613 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 5.05 sec
2024-04-28 05:56:20,644 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 18.35 sec
2024-04-28 05:56:25,887 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 27.54 sec
2024-04-28 05:56:26,916 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 31.89 sec
2024-04-28 05:56:27,966 Stage-1 map = 100%,  reduce = 80%, Cumulative CPU 36.1 sec
2024-04-28 05:56:28,994 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 40.79 sec
MapReduce Total cumulative CPU time: 40 seconds 790 msec
Ended Job = job_1714263346713_0001
Loading data to table nosql_three.yago_buck_part partition (predicate=<actedIn>)
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714263346713_0002, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714263346713_0002/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714263346713_0002
```



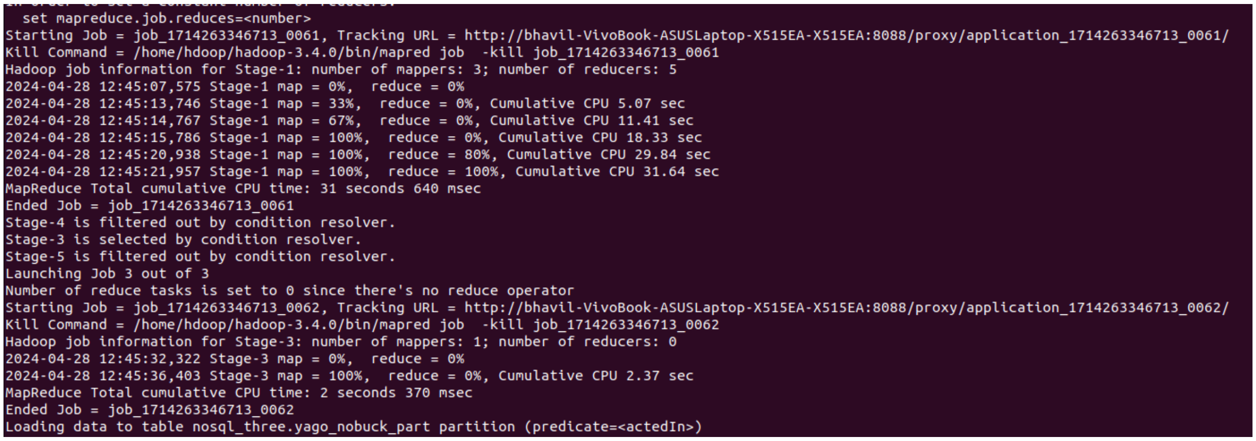
```
Total MapReduce CPU Time Spent: 1 minutes 15 seconds 310 msec
OK
Time taken: 54.077 seconds
Query ID = hdoop_20240428060748_73cbf565-cfdd-4e4f-9d11-40ef64d5d702
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks determined at compile time: 5
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714263346713_0029, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714263346713_0029/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714263346713_0029
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 5
2024-04-28 06:07:58,175 Stage-1 map = 0%,  reduce = 0%
2024-04-28 06:08:05,388 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 5.86 sec
2024-04-28 06:08:06,415 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 14.53 sec
2024-04-28 06:08:07,436 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 22.9 sec
2024-04-28 06:08:13,708 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 34.34 sec
2024-04-28 06:08:14,753 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 39.49 sec
2024-04-28 06:08:15,779 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 48.64 sec
MapReduce Total cumulative CPU time: 48 seconds 640 msec
Ended Job = job_1714263346713_0029
Loading data to table nosql_three.yago_buck_part partition (predicate=<isMarriedTo>)
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714263346713_0030, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714263346713_0030/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714263346713_0030
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2024-04-28 06:08:27,466 Stage-3 map = 0%,  reduce = 0%
2024-04-28 06:08:31,564 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.01 sec
2024-04-28 06:08:36,662 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.78 sec
```



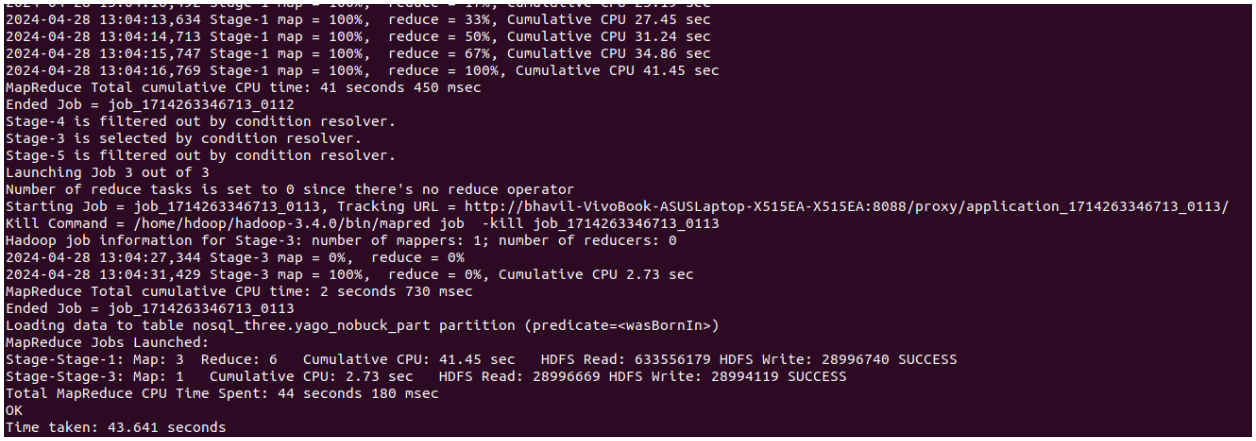
```
<Eugene_Stepanenko>     <Eugene>        <Kiev>
<Eugene_Stepanenko>     <Eugene>        <Ukraine>
<Eugene_Stepanenko>     <Eugene>        <Kiev>
<Eugene_Stepanenko>     <Eugene>        <Ukraine>
<Simeon_S._Pennewill>   <Simeon>        <Dover,_Delaware>
<Simeon_S._Pennewill>   <Simeon>        <Delaware>
<Simeon_S._Pennewill>   <Simeon>        <Delaware>
Time taken: 18.213 seconds, Fetched: 167098 row(s)
```

Part-B-2

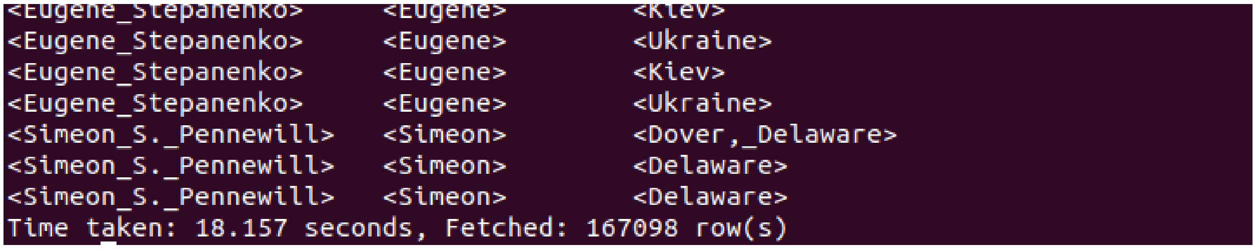
Effectively, here I don’t have to write the statement of bucketing, everything is same as before. Even the select statement is similar (new noBuck_part table instead of buck_part).



```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714263346713_0061, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714263346713_0061/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714263346713_0061
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 5
2024-04-28 12:45:07,575 Stage-1 map = 0%,  reduce = 0%
2024-04-28 12:45:13,746 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 5.07 sec
2024-04-28 12:45:14,767 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 11.41 sec
2024-04-28 12:45:15,786 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 18.33 sec
2024-04-28 12:45:20,938 Stage-1 map = 100%,  reduce = 80%, Cumulative CPU 29.84 sec
2024-04-28 12:45:21,957 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 31.64 sec
MapReduce Total cumulative CPU time: 31 seconds 640 msec
Ended Job = job_1714263346713_0061
Stage-4 is filtered out by condition resolver.
Stage-3 is selected by condition resolver.
Stage-5 is filtered out by condition resolver.
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1714263346713_0062, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714263346713_0062/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714263346713_0062
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-04-28 12:45:32,322 Stage-3 map = 0%,  reduce = 0%
2024-04-28 12:45:36,403 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.37 sec
MapReduce Total cumulative CPU time: 2 seconds 370 msec
Ended Job = job_1714263346713_0062
Loading data to table nosql_three.yago_nobuck_part partition (predicate=<actedIn>)
```



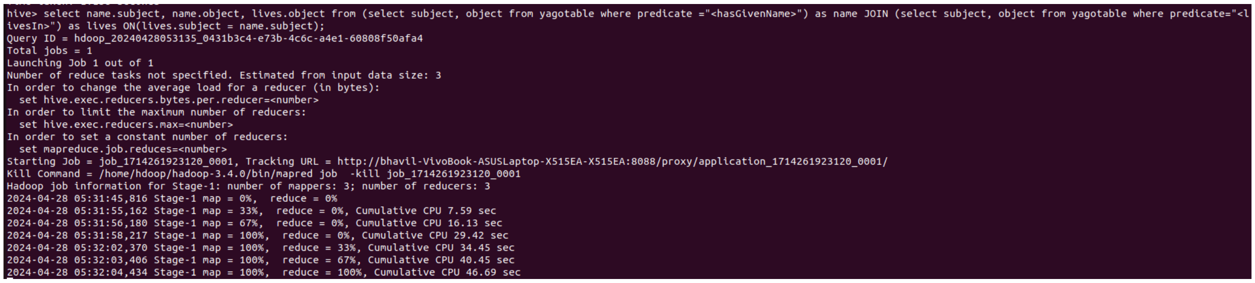
```
2024-04-28 13:04:10,492 Stage-1 map = 100%,  reduce = 17%, Cumulative CPU 23.19 sec
2024-04-28 13:04:13,634 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 27.45 sec
2024-04-28 13:04:14,713 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 31.24 sec
2024-04-28 13:04:15,747 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 34.86 sec
2024-04-28 13:04:16,769 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 41.45 sec
MapReduce Total cumulative CPU time: 41 seconds 450 msec
Ended Job = job_1714263346713_0112
Stage-4 is filtered out by condition resolver.
Stage-3 is selected by condition resolver.
Stage-5 is filtered out by condition resolver.
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1714263346713_0113, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714263346713_0113/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714263346713_0113
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-04-28 13:04:27,344 Stage-3 map = 0%,  reduce = 0%
2024-04-28 13:04:31,429 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.73 sec
MapReduce Total cumulative CPU time: 2 seconds 730 msec
Ended Job = job_1714263346713_0113
Loading data to table nosql_three.yago_nobuck_part partition (predicate=<wasBornIn>)
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 6   Cumulative CPU: 41.45 sec   HDFS Read: 633556179 HDFS Write: 28996740 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 2.73 sec   HDFS Read: 28996669 HDFS Write: 28994119 SUCCESS
Total MapReduce CPU Time Spent: 44 seconds 180 msec
OK
Time taken: 43.641 seconds
```



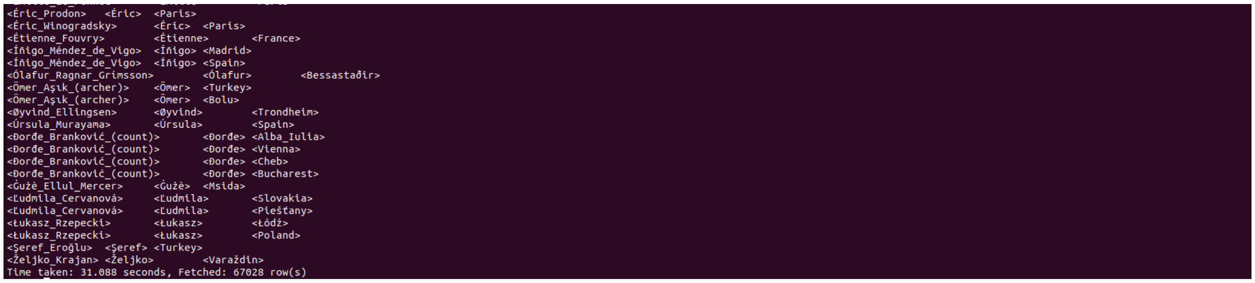
```
<Eugene_Stepanenko>     <Eugene>        <Kiev>
<Eugene_Stepanenko>     <Eugene>        <Ukraine>
<Eugene_Stepanenko>     <Eugene>        <Kiev>
<Eugene_Stepanenko>     <Eugene>        <Ukraine>
<Simeon_S._Pennewill>   <Simeon>        <Dover,_Delaware>
<Simeon_S._Pennewill>   <Simeon>        <Delaware>
<Simeon_S._Pennewill>   <Simeon>        <Delaware>
Time taken: 18.157 seconds, Fetched: 167098 row(s)
```

Part-B-3

Here, we won’t have to run the source file for 29 predicates partitioning at all. Just run the select statement directly.



```
hive> select name.subject, name.object, lives.object from (select subject, object from yagotable where predicate ="<hasGivenName>") as name JOIN (select subject, object from yagotable where predicate="<l
ivesIn>") as lives ON(lives.subject = name.subject);
Query ID = hdoop_20240428053135_0431b3c4-e73b-4c6c-a4e1-60808f50afa4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 3
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1714261923120_0001, Tracking URL = http://bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:8088/proxy/application_1714261923120_0001/
Kill Command = /home/hdoop/hadoop-3.4.0/bin/mapred job  -kill job_1714261923120_0001
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3
2024-04-28 05:31:45,816 Stage-1 map = 0%,  reduce = 0%
2024-04-28 05:31:55,162 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 7.59 sec
2024-04-28 05:31:56,180 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 16.13 sec
2024-04-28 05:31:58,217 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 29.42 sec
2024-04-28 05:32:02,370 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 34.45 sec
2024-04-28 05:32:03,406 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 40.45 sec
2024-04-28 05:32:04,434 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 46.69 sec
```



```
<Éric_Prodon>   <Éric>  <Paris>
<Éric_Winogradsky>      <Éric>  <Paris>
<Étienne_Fouvry>        <Étienne>       <France>
<Íñigo_Méndez_de_Vigo>  <Íñigo> <Madrid>
<Íñigo_Méndez_de_Vigo>  <Íñigo> <Spain>
<Ólafur_Ragnar_Grímsson>        <Ólafur>        <Bessastaðir>
<Ömer_Aşık_(archer)>    <Ömer>  <Turkey>
<Ömer_Aşık_(archer)>    <Ömer>  <Bolu>
<Øyvind_Ellingsen>      <Øyvind>        <Trondheim>
<Úrsula_Murayama>       <Úrsula>        <Spain>
<Đorđe_Branković_(count)>       <Đorđe> <Alba_Iulia>
<Đorđe_Branković_(count)>       <Đorđe> <Vienna>
<Đorđe_Branković_(count)>       <Đorđe> <Cheb>
<Đorđe_Branković_(count)>       <Đorđe> <Bucharest>
<Ġużè_Ellul_Mercer>     <Ġużè>  <Msida>
<Ľudmila_Cervanová>     <Ľudmila>       <Slovakia>
<Ľudmila_Cervanová>     <Ľudmila>       <Piešťany>
<Łukasz_Rzepecki>       <Łukasz>        <Łódź>
<Łukasz_Rzepecki>       <Łukasz>        <Poland>
<Şeref_Eroğlu>  <Şeref> <Turkey>
<Željko_Krajan> <Željko>        <Varaždin>
Time taken: 31.088 seconds, Fetched: 67028 row(s)
```