Data Ingestion Using Scoop And Data Analysis Using Hive

DATA INGESTION FROM RDS TO HDFS USING SQOOP

1. Sqoop import command *(See, sqoop/importFromRDSToHDFS.sqoop)*

   Note

   ✓ The table **Key_indicator_districtwise** available in RDS has some **NULL** values in some of the columns.

   ✓ Using the scoop command, the **NULL** values are replaced with **NA** for all String based columns and **\N** for all non-string based columns while importing data into HDFS.

   ✓ This is to make sure the **NULL** value is not written to the HDFS data.

2. Command to see the list of imported data *(See, hdfs/viewData.hdfs)*

EXTERNAL TABLE CREATION IN HIVE AND LOADING INGESTED DATA

1. Command to create the external table *(See, hive/createHiveTable.hql)*

   Note

   ✓ A database named **India_Annual_Health_Survey_2012_13_DB** is created. All the tables pertaining to this project will be created in this database.

   ✓ An external table named **IAHS_2012_13** is created with 645 columns. This table will be used as a master repository of data.

2. Command to load the ingested data into the external table *(See, hive/loadDataInHiveTable.hql)*

3. Queries to verify that the ingestion is correctly accomplished

   3.1 Query to count the total number of rows of data fetched from RDS using MySQL Workbench and from Hive using Hue

   MySQL Workbench *(See, sql/verificationQuery1.sql)*
   Hue *(See, hive/verificationQuery1.hql)*

   3.2 Query to select the top 10 rows and first 8 columns of the data fetched from RDS using MySQL Workbench and from Hive using Hue

   MySQL Workbench *(See, sql/verificationQuery2.sql)*
   Hue *(See, hive/verificationQuery2.hql)*

Note

- ✓ The above listed 02 queries and their results across the RDBMS table **Key_indicator_districtwise** and the HIVE table **IAHS_2012_13** should show that the data is correctly imported from RDS to HDFS using sqoop.
- ✓ Later, the same imported data is correctly ingested into the HIVE table **IAHS_2012_13.**

SUBSET SCHEMA CREATION IN HIVE TO SUPPORT ANALYSIS

1. Columns used in the subset schema
   ID
   State_Name
   State_District_Name
   AA_Households_Total
   AA_Population_Total
   CC_Sex_Ratio_All_Ages_Total
   LL_Total_Fertility_Rate_Total
   YY_Under_Five_Mortality_Rate_U5MR_Total_Person

2. Storage format used [Benchmark the performance before finalizing the storage format to be used. Create one schema using default format and one in any other format such as ORC for the columns to be used. Insert data into both the tables created. Compare the runtimes of the following queries and decide which format to be used.
   - ✓ *select count(*) from <Table Name>;*
   - ✓ *select State_Name, count(*) from <Table Name> group by State_Name;*
   - ✓ *select * from <Table Name> where State_Name = 'Uttar Pradesh';]*

Note

In point **03** below,

- ✓ A subset table named **IAHS_2012_13_TEXT** is created with default TEXT format.
- ✓ The subset table contains selected **08** columns.
- ✓ The data is ingested into this table from the master table **IAHS_2012_13**

In point **04** below,

- ✓ A subset table named **IAHS_2012_13_ORC** is created with ORC format.
- ✓ The subset table contains selected **08** columns.
- ✓ The data is ingested into this table from the master table **IAHS_2012_13**

In point **05** below,

- ✓ **03** sets of queries are executed against both the tables (reference, point 3 and 4) and their execution time is noted.
- ✓ On examining the execution time for all the **03** set of queries, it is observed that the queries executed on the table with ORC format has lower execution time in comparision to the execution time of queries executed on the table with default TEXT format.
- ✓ The difference in execution time of queries is marginal as the data set is small in size.
- ✓ The difference in execution time will increase for a voluminous production size data set.
- ✓ Based on the benchmarking performed for all the **03** queries, I have choosen the ORC format to be used for this project.
- ✓ Additionally, I have also used the compression algorithm SNAPPY with the ORC format as opposed to the non-compressed way of storing data with the default TEXT format.

  The data stored in compressed format saves on disk space which is again helpful when the size of the data set is voluminous.

3. Create and insert command for the default format *(See, hive/createAndInsertDefaultFormat.hql)*

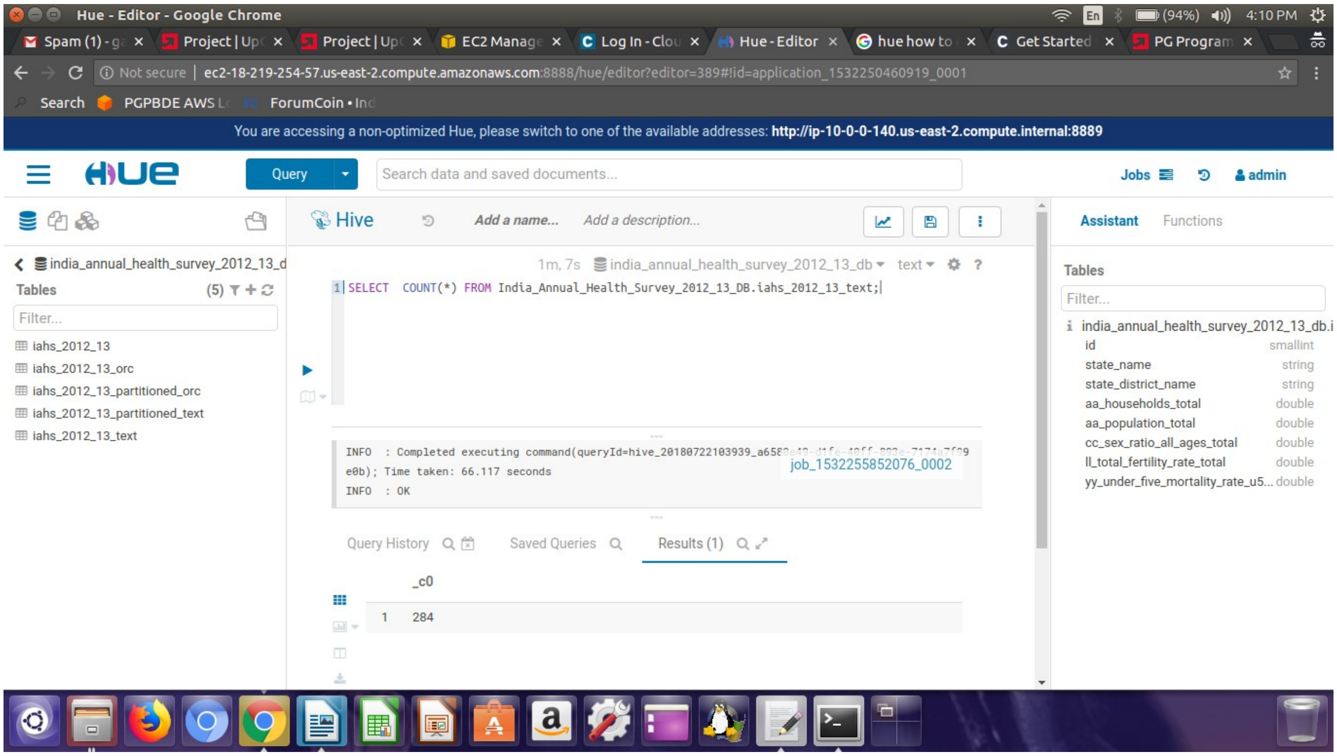
4. Create and insert command for the formats such as ORC *(See, hive/createAndInsertORCFormat.hql)*

5. Screenshot of runtimes against each query given above for the default format as well as for the formats such as ORC

TEXT FORMAT

SELECT COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text;

Time Taken: ***66.117 seconds***



Hue - Editor - Google Chrome
ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=389#!id=application_1532250460919_0001
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889
Query
Jobs
admin
Hive
Add a name...
Add a description...
Assistant
Functions
india_annual_health_survey_2012_13_d
Tables (5)
iahs_2012_13
iahs_2012_13_orc
iahs_2012_13_partitioned_orc
iahs_2012_13_partitioned_text
iahs_2012_13_text
1m, 7s
india_annual_health_survey_2012_13_db
text
Query History
Saved Queries
Results (1)
job_1532255852076_0002

```
SELECT  COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text;
```

```
INFO  : Completed executing command(queryId=hive_20180722103939_a6580e49-d1fe-40ff-893e-7174a7f99
e0b); Time taken: 66.117 seconds
INFO  : OK
```

	_c0
1	284

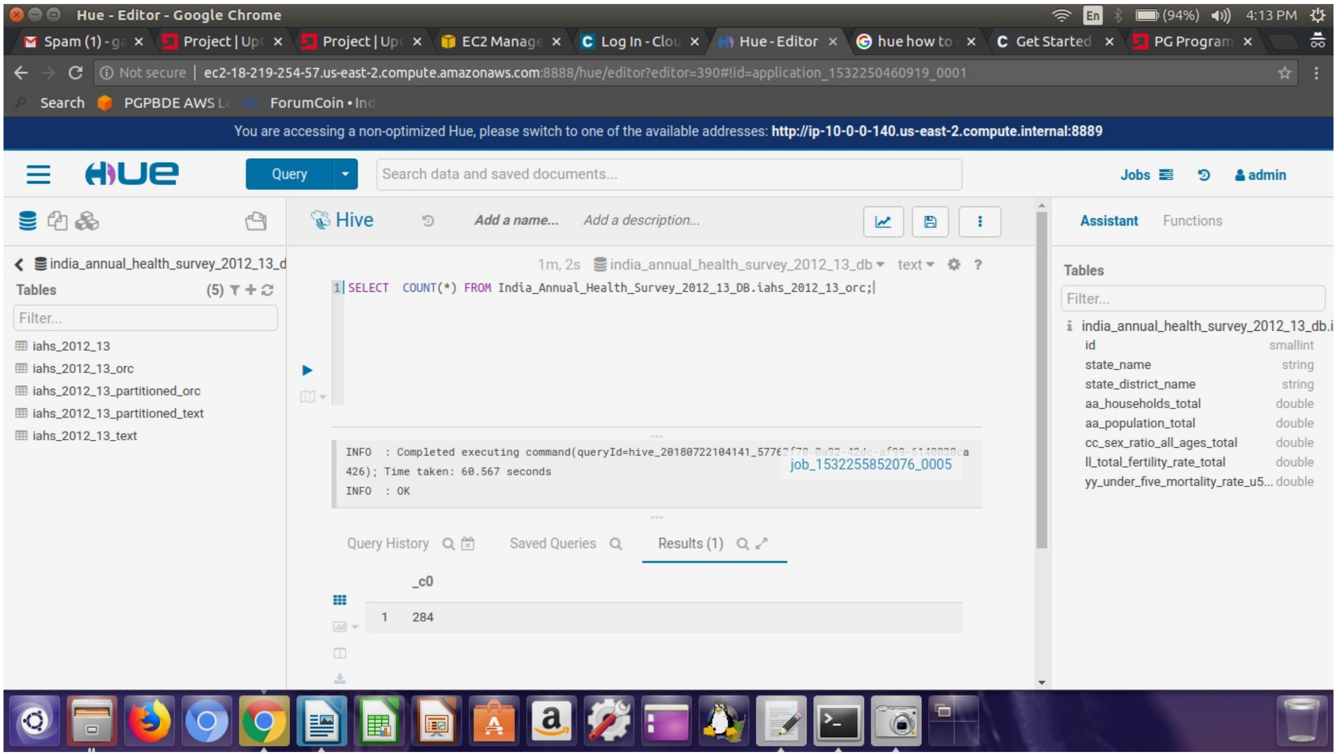
india_annual_health_survey_2012_13_db.i

Column	Type
id	smallint
state_name	string
state_district_name	string
aa_households_total	double
aa_population_total	double
cc_sex_ratio_all_ages_total	double
ll_total_fertility_rate_total	double
yy_under_five_mortality_rate_u5...	double

ORC FORMAT

SELECT COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc;

Time Taken: ***60.657 seconds***



You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

india_annual_health_survey_2012_13_d

Tables (5)

- iahs_2012_13
- iahs_2012_13_orc
- iahs_2012_13_partitioned_orc
- iahs_2012_13_partitioned_text
- iahs_2012_13_text

```
SELECT  COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc;
```

```
INFO  : Completed executing command(queryId=hive_20180722104141_57762f70-8a92-42dc-af99-6140030ca
426); Time taken: 60.567 seconds
INFO  : OK
```

job_1532255852076_0005

	_c0
1	284

Tables

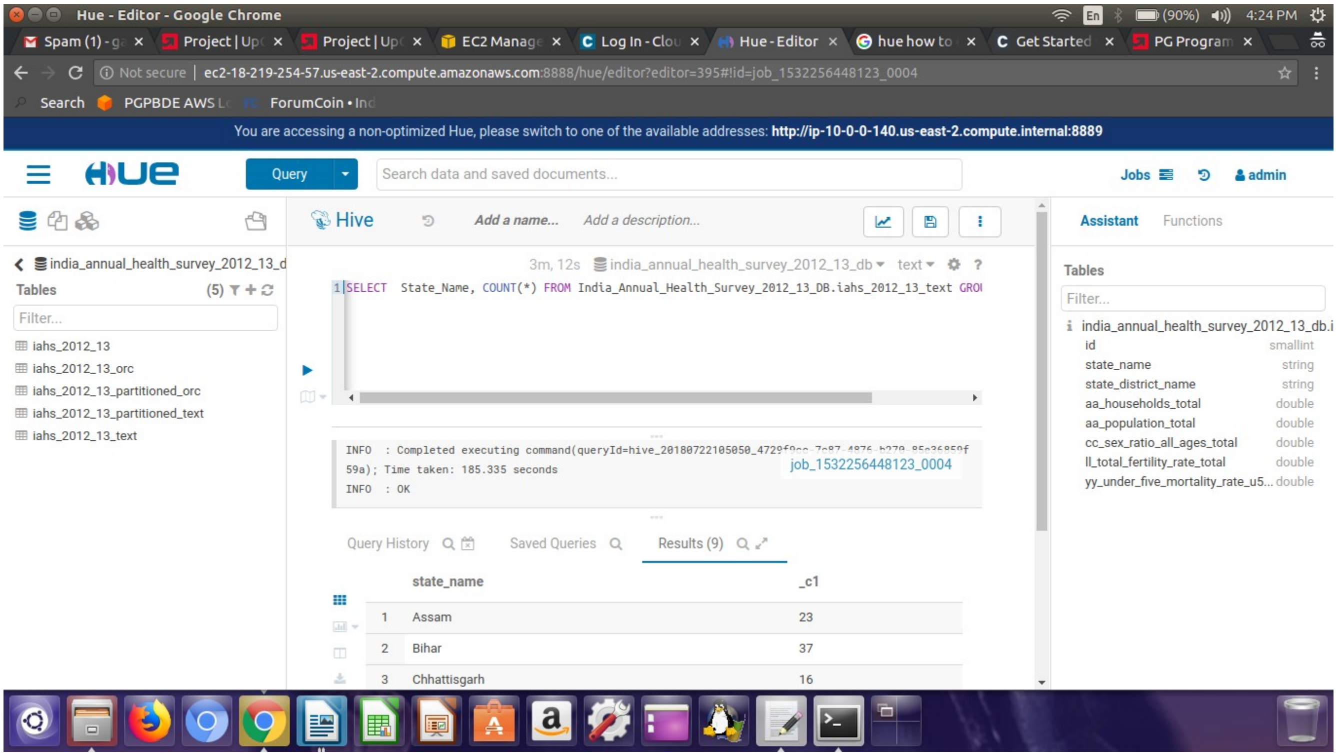
india_annual_health_survey_2012_13_db.i

id	smallint
state_name	string
state_district_name	string
aa_households_total	double
aa_population_total	double
cc_sex_ratio_all_ages_total	double
ll_total_fertility_rate_total	double
yy_under_five_mortality_rate_u5...	double

TEXT FORMAT

SELECT State_Name, COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text GROUP BY State_Name;

Time Taken: ***185.335 seconds***



Hue - Editor - Google Chrome
Not secure | ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=395#!id=job_1532256448123_0004
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889
Query
Search data and saved documents...
Jobs
admin
Hive
Add a name...
Add a description...
3m, 12s
india_annual_health_survey_2012_13_db
text
Assistant
Functions
Query History
Saved Queries
Results (9)

india_annual_health_survey_2012_13_d

Tables (5)

- iahs_2012_13
- iahs_2012_13_orc
- iahs_2012_13_partitioned_orc
- iahs_2012_13_partitioned_text
- iahs_2012_13_text

```
SELECT  State_Name, COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text GROU
```

```
INFO  : Completed executing command(queryId=hive_20180722105050_4729f0cc-7c87-4876-b270-85e36859f
59a); Time taken: 185.335 seconds
INFO  : OK
```

job_1532256448123_0004

	state_name	_c1
1	Assam	23
2	Bihar	37
3	Chhattisgarh	16

Tables

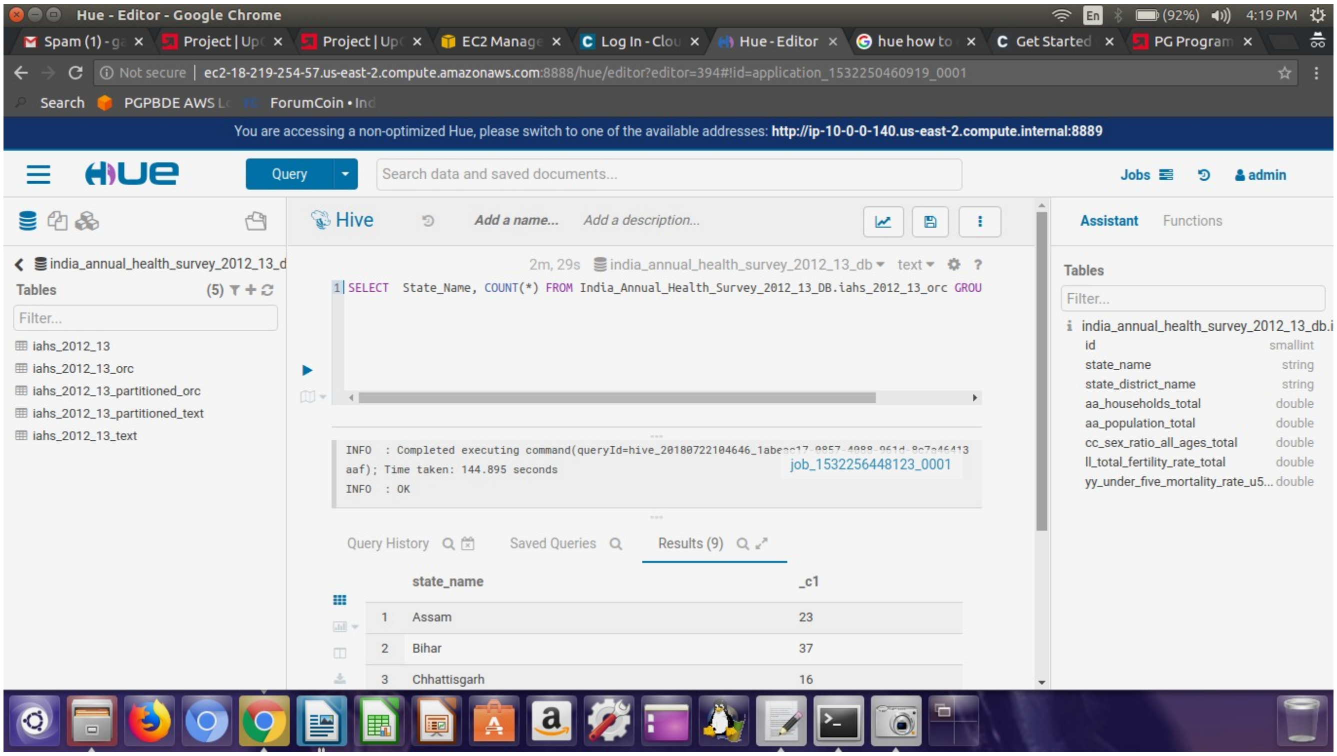
india_annual_health_survey_2012_13_db.i

id	smallint
state_name	string
state_district_name	string
aa_households_total	double
aa_population_total	double
cc_sex_ratio_all_ages_total	double
ll_total_fertility_rate_total	double
yy_under_five_mortality_rate_u5...	double

ORC FORMAT

SELECT State_Name, COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc GROUP BY State_Name;

Time Taken: ***144.896 seconds***



Hue - Editor - Google Chrome
ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=394#!id=application_1532250460919_0001
Hive
Add a name...
Add a description...
2m, 29s
india_annual_health_survey_2012_13_db
Query History
Saved Queries
Results (9)

You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889

Tables (5)

- iahs_2012_13
- iahs_2012_13_orc
- iahs_2012_13_partitioned_orc
- iahs_2012_13_partitioned_text
- iahs_2012_13_text

```
SELECT  State_Name, COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc GROU
```

```
INFO  : Completed executing command(queryId=hive_20180722104646_1abeac17-0857-4088-961d-8c7a46413
aaf); Time taken: 144.895 seconds
INFO  : OK
```

job_1532256448123_0001

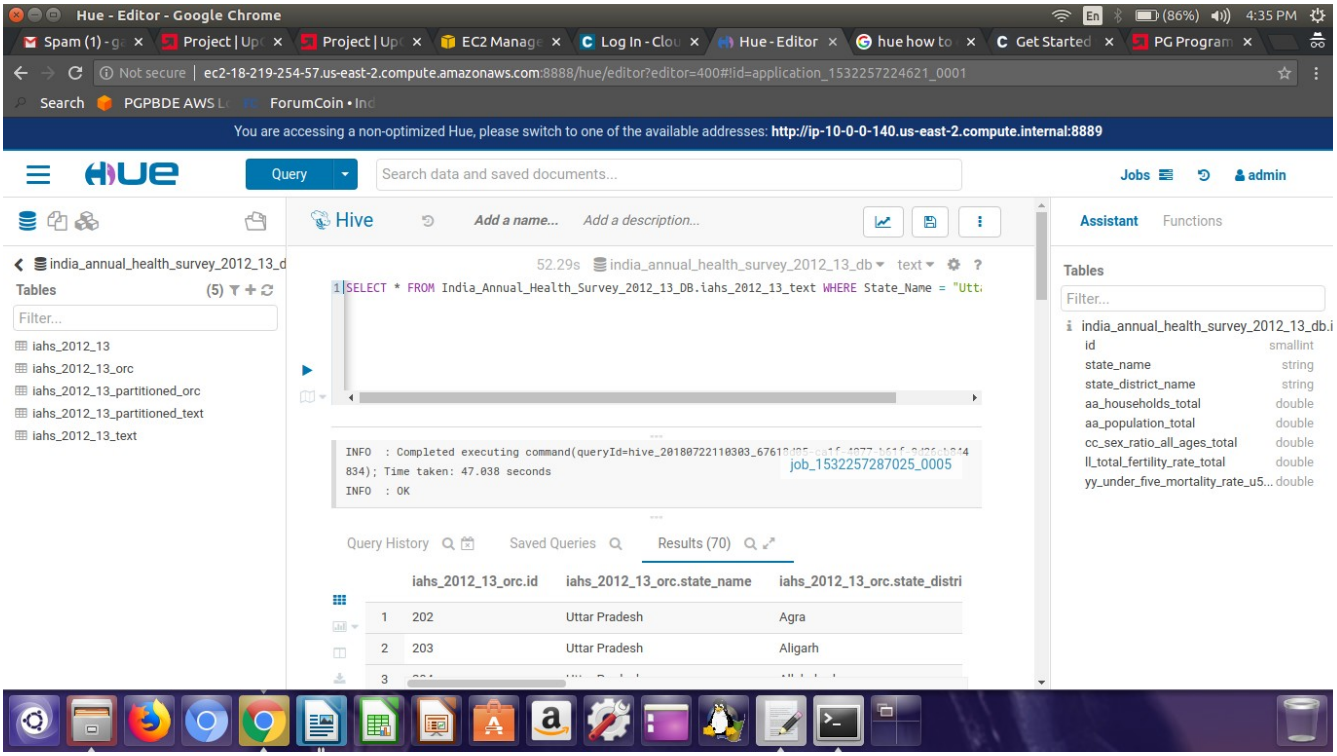
	state_name	_c1
1	Assam	23
2	Bihar	37
3	Chhattisgarh	16

india_annual_health_survey_2012_13_db.i	
id	smallint
state_name	string
state_district_name	string
aa_households_total	double
aa_population_total	double
cc_sex_ratio_all_ages_total	double
ll_total_fertility_rate_total	double
yy_under_five_mortality_rate_u5...	double

TEXT FORMAT

*SELECT * FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text WHERE State_Name = “Uttar Pradesh”;*

Time Taken: ***47.038 seconds***


Hue - Editor - Google Chrome
ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=400#!id=application_1532257224621_0001
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889
Query
Search data and saved documents...
Jobs
admin
Hive
Add a name...
Add a description...
52.29s
india_annual_health_survey_2012_13_db
text
india_annual_health_survey_2012_13_d
Tables (5)
iahs_2012_13
iahs_2012_13_orc
iahs_2012_13_partitioned_orc
iahs_2012_13_partitioned_text
iahs_2012_13_text
Assistant
Functions
Tables
Query History
Saved Queries
Results (70)

```
SELECT * FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text WHERE State_Name = "Utt
```

```
INFO  : Completed executing command(queryId=hive_20180722110303_67618d05-ca1f-4077-b61f-9d26cb844
834); Time taken: 47.038 seconds
INFO  : OK
```

job_1532257287025_0005

india_annual_health_survey_2012_13_db.i

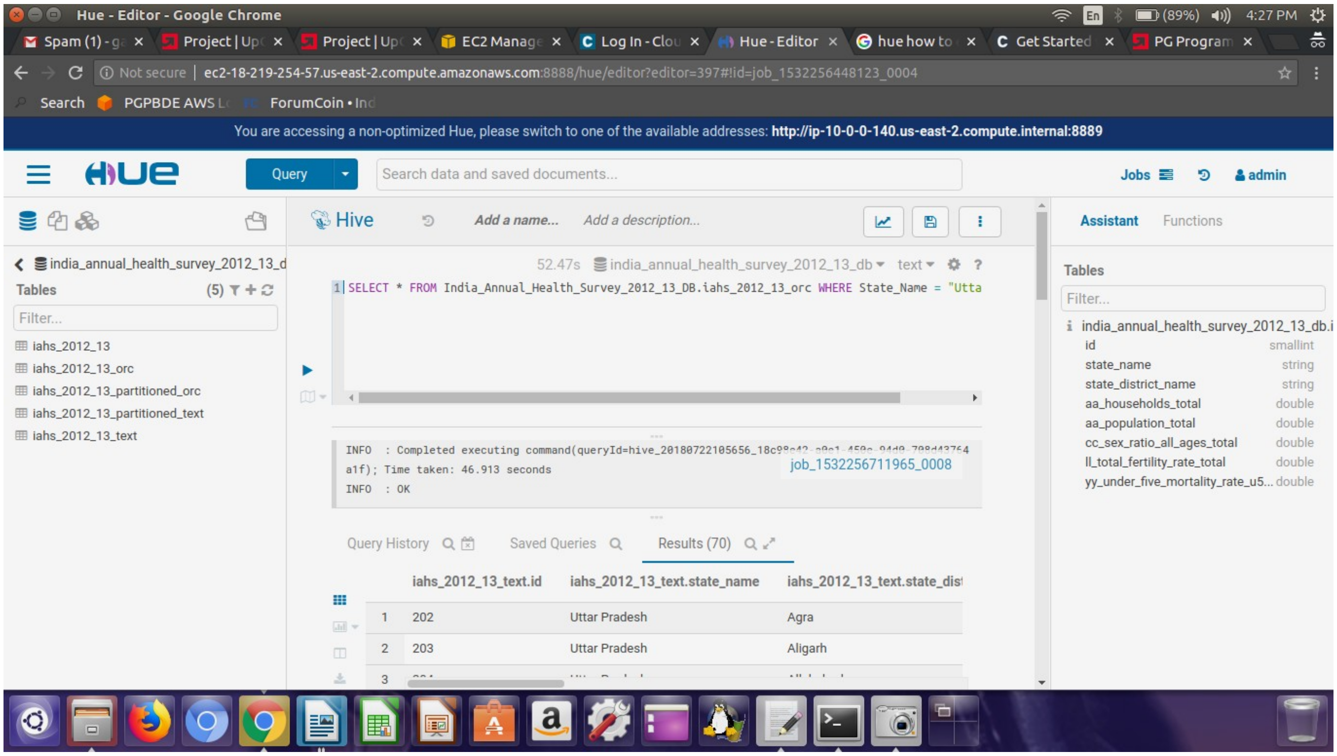
id	smallint
state_name	string
state_district_name	string
aa_households_total	double
aa_population_total	double
cc_sex_ratio_all_ages_total	double
ll_total_fertility_rate_total	double
yy_under_five_mortality_rate_u5...	double

	iahs_2012_13_orc.id	iahs_2012_13_orc.state_name	iahs_2012_13_orc.state_distri
1	202	Uttar Pradesh	Agra
2	203	Uttar Pradesh	Aligarh
3	[illegible]	[illegible]	[illegible]

ORC FORMAT

*SELECT * FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc WHERE State_Name = “Uttar Pradesh”;*

Time Taken: ***46.913 seconds***



Hue - Editor - Google Chrome
Not secure | ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=397#!id=job_1532256448123_0004
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889
Query
Search data and saved documents...
Jobs
admin
Hive
Add a name...
Add a description...
Assistant
Functions
52.47s
india_annual_health_survey_2012_13_db
text
Query History
Saved Queries
Results (70)

Tables (5)

- iahs_2012_13
- iahs_2012_13_orc
- iahs_2012_13_partitioned_orc
- iahs_2012_13_partitioned_text
- iahs_2012_13_text

```
SELECT * FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc WHERE State_Name = "Utta
```

```
INFO  : Completed executing command(queryId=hive_20180722105656_18c98c42-a0e1-450e-94d0-708d43764
a1f); Time taken: 46.913 seconds
INFO  : OK
```

job_1532256711965_0008

	iahs_2012_13_text.id	iahs_2012_13_text.state_name	iahs_2012_13_text.state_dist
1	202	Uttar Pradesh	Agra
2	203	Uttar Pradesh	Aligarh
3			

india_annual_health_survey_2012_13_db.i

id	smallint
state_name	string
state_district_name	string
aa_households_total	double
aa_population_total	double
cc_sex_ratio_all_ages_total	double
ll_total_fertility_rate_total	double
yy_under_five_mortality_rate_u5...	double

6. Create and insert command for the partition table for analyses 1 & 2. The partition table should be created using the table created above. *(See, hive/createAndInsertORCFormatPartitioned.hql)*

   Note

   For analyses 1 and 2,

   ✔ A partitioned table named **IAHS_2012_13_PARTITIONED_ORC_FORMAT** is created with ORC format.

   ✔ The data into this table is ingested from the master table **IAHS_2012_13**.

   ✔ This table will be used only for writing queries for analyses 1 and 2.

   For analyses 3, 4 and 5, the non-partitioned ORC format table **IAHS_2012_13_ORC** will be used.

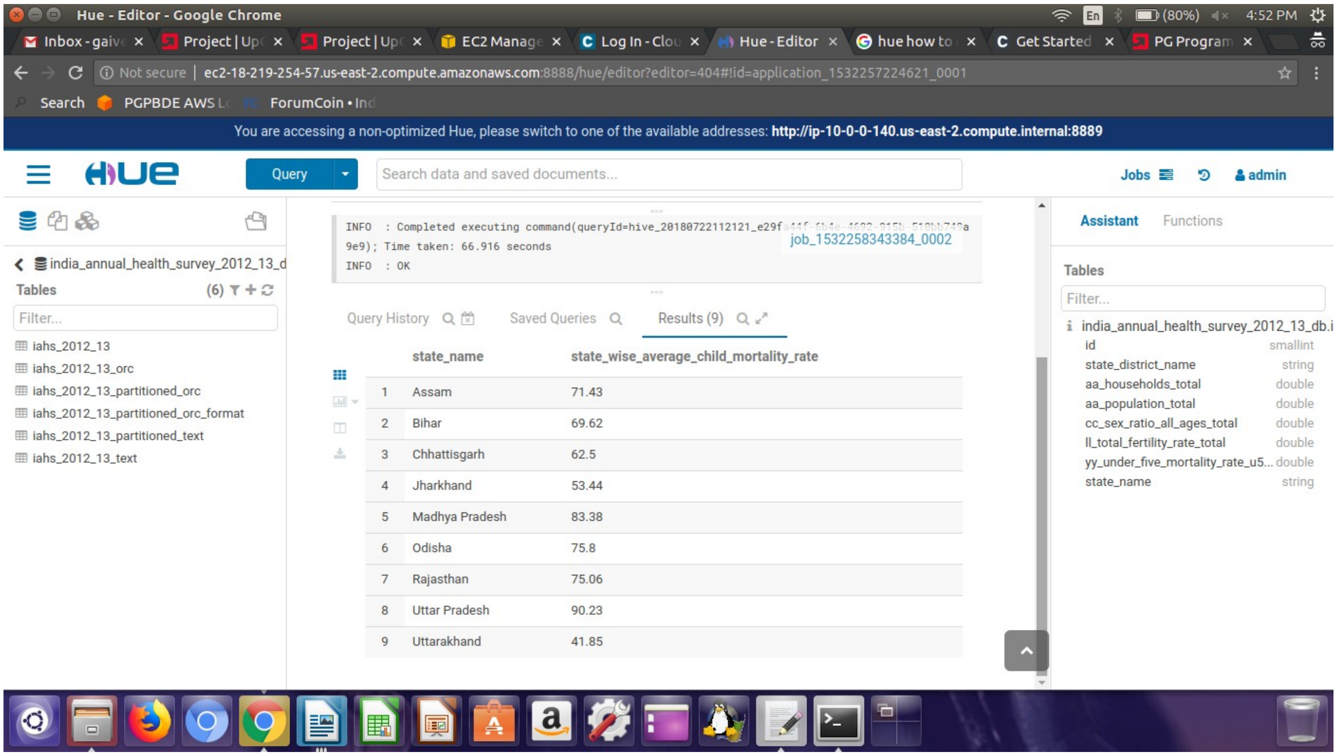
QUERY ANALYSIS, RESULT AND CHART

1. State wise child mortality rate

Query

*SELECT State_Name,
ROUND(AVG(YY_Under_Five_Mortality_Rate_U5MR_Total_Person),2) AS State_Wise_Average_Child_Mortality_Rate
FROM
India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_partitioned_orc_format GROUP BY State_Name;*

Screenshot of the result


Hue - Editor - Google Chrome
Inbox - gaiv
Project | Up
Project | Up
EC2 Manage
Log In - Clou
Hue - Editor
hue how to
Get Started
PG Program
Not secure | ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=404#!id=application_1532257224621_0001
Search
PGPBDE AWS L
ForumCoin • Ind
Query
Search data and saved documents...
Jobs
admin
india_annual_health_survey_2012_13_d
Tables (6)
Filter...
iahs_2012_13
iahs_2012_13_orc
iahs_2012_13_partitioned_orc
iahs_2012_13_partitioned_orc_format
iahs_2012_13_partitioned_text
iahs_2012_13_text
Query History
Saved Queries
Results (9)
Assistant
Functions
Tables
Filter...
india_annual_health_survey_2012_13_db.i

You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

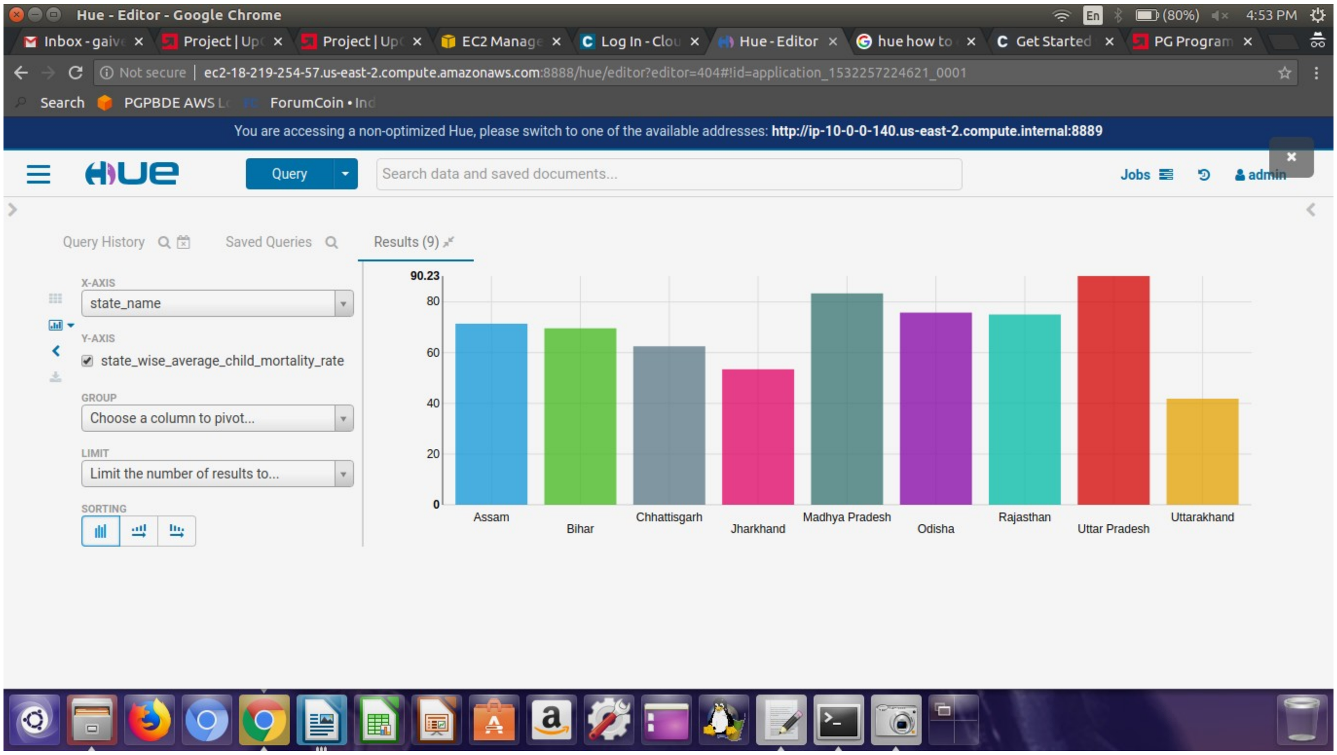
```
INFO  : Completed executing command(queryId=hive_20180722112121_e29fa44f-6b4e-4692-915b-510bb749a
9e9); Time taken: 66.916 seconds
INFO  : OK
```

job_1532258343384_0002

	state_name	state_wise_average_child_mortality_rate
1	Assam	71.43
2	Bihar	69.62
3	Chhattisgarh	62.5
4	Jharkhand	53.44
5	Madhya Pradesh	83.38
6	Odisha	75.8
7	Rajasthan	75.06
8	Uttar Pradesh	90.23
9	Uttarakhand	41.85

id	smallint
state_district_name	string
aa_households_total	double
aa_population_total	double
cc_sex_ratio_all_ages_total	double
ll_total_fertility_rate_total	double
yy_under_five_mortality_rate_u5...	double
state_name	string

Chart



You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

X-AXIS
state_name

Y-AXIS
☑ state_wise_average_child_mortality_rate

GROUP
Choose a column to pivot...

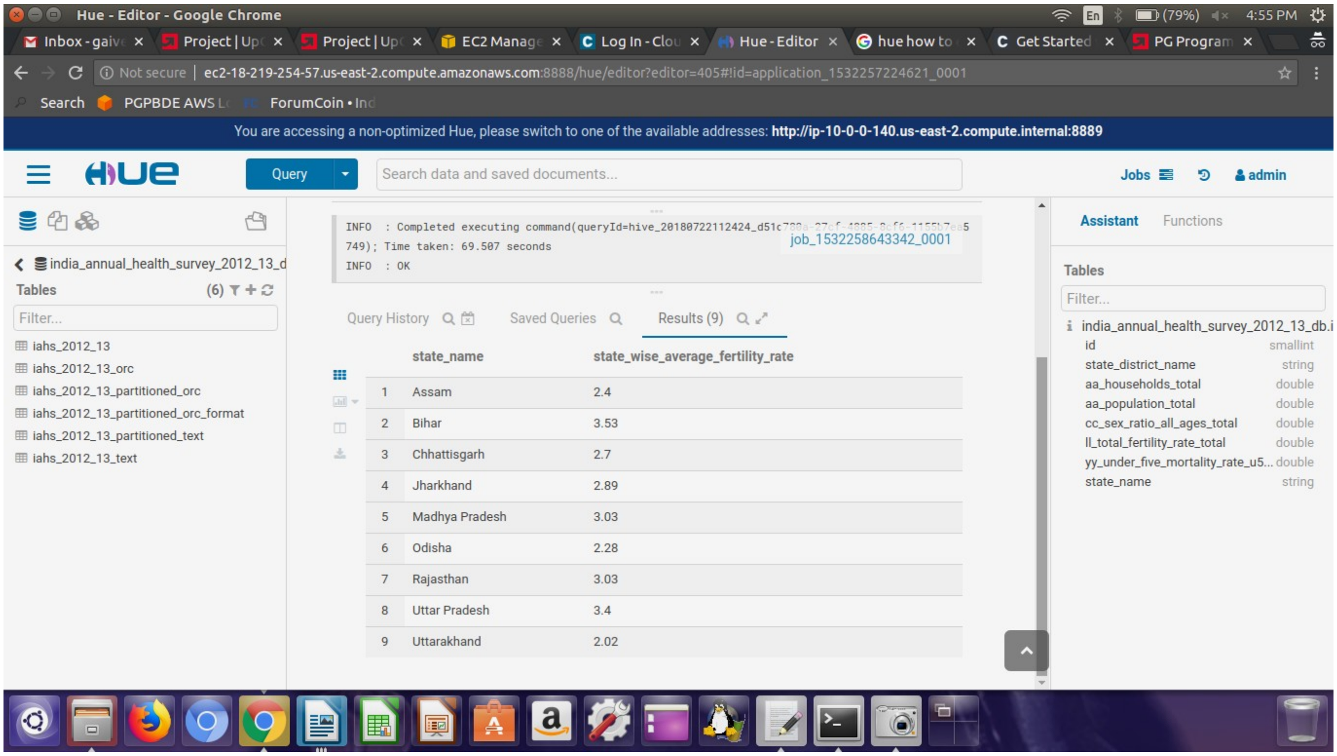
LIMIT
Limit the number of results to...

2. State wise fertility rate

Query

```
SELECT State_Name,
ROUND(AVG(LL_Total_Fertility_Rate_Total),2) AS
State_Wise_Average_Fertility_Rate
FROM
India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_partitioned_
orc_format GROUP BY State_Name;
```

Screenshot of the result



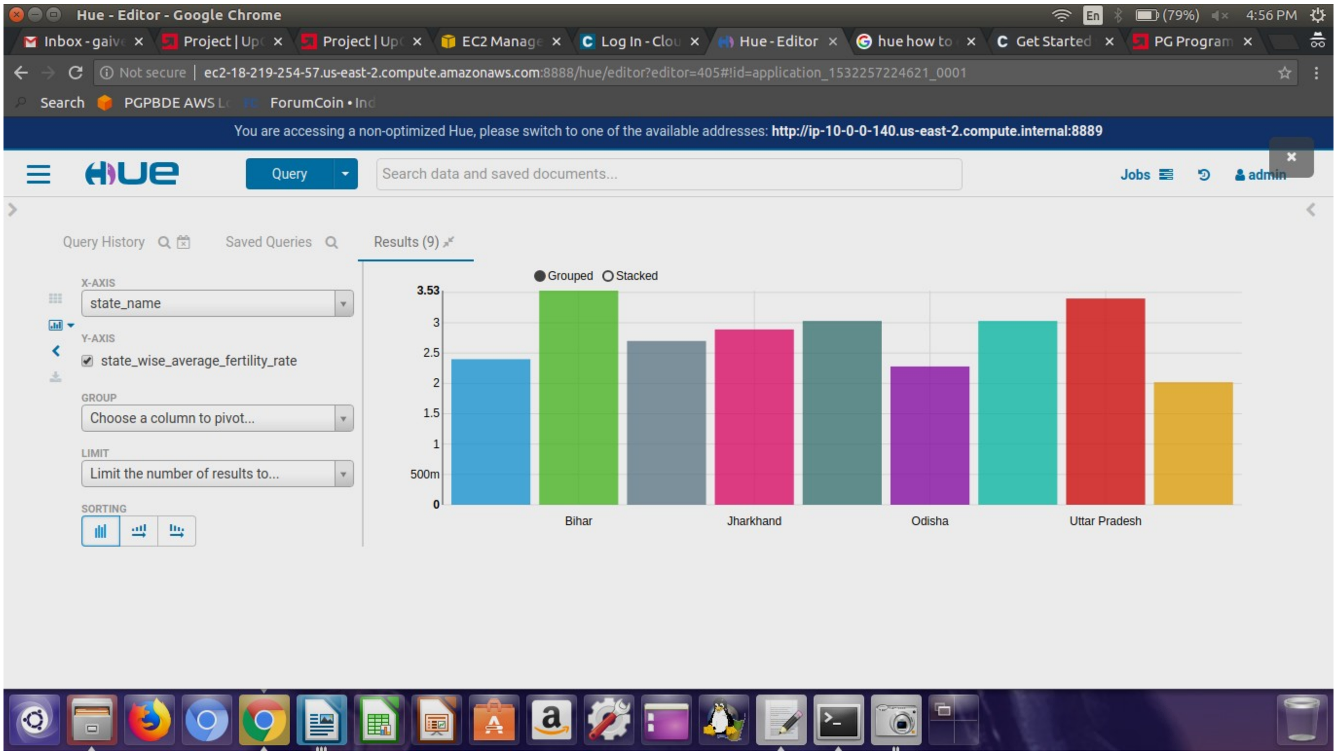
```
INFO  : Completed executing command(queryId=hive_20180722112424_d51c780a-27cf-4885-8cf6-1155b7ea5
749); Time taken: 69.507 seconds
INFO  : OK
```

job_1532258643342_0001

	state_name	state_wise_average_fertility_rate
1	Assam	2.4
2	Bihar	3.53
3	Chhattisgarh	2.7
4	Jharkhand	2.89
5	Madhya Pradesh	3.03
6	Odisha	2.28
7	Rajasthan	3.03
8	Uttar Pradesh	3.4
9	Uttarakhand	2.02

Hue - Editor - Google Chrome
ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=405#!id=application_1532257224621_0001
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889
Query
Search data and saved documents...
Jobs
admin
india_annual_health_survey_2012_13_d
Tables (6)
iahs_2012_13
iahs_2012_13_orc
iahs_2012_13_partitioned_orc
iahs_2012_13_partitioned_orc_format
iahs_2012_13_partitioned_text
iahs_2012_13_text
Query History
Saved Queries
Results (9)
Assistant
Functions
Tables
india_annual_health_survey_2012_13_db.i
id smallint
state_district_name string
aa_households_total double
aa_population_total double
cc_sex_ratio_all_ages_total double
ll_total_fertility_rate_total double
yy_under_five_mortality_rate_u5... double
state_name string

Chart



Hue - Editor - Google Chrome

ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=405#!id=application_1532257224621_0001

You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

X-AXIS: state_name

Y-AXIS: ☑ state_wise_average_fertility_rate

GROUP: Choose a column to pivot...

LIMIT: Limit the number of results to...

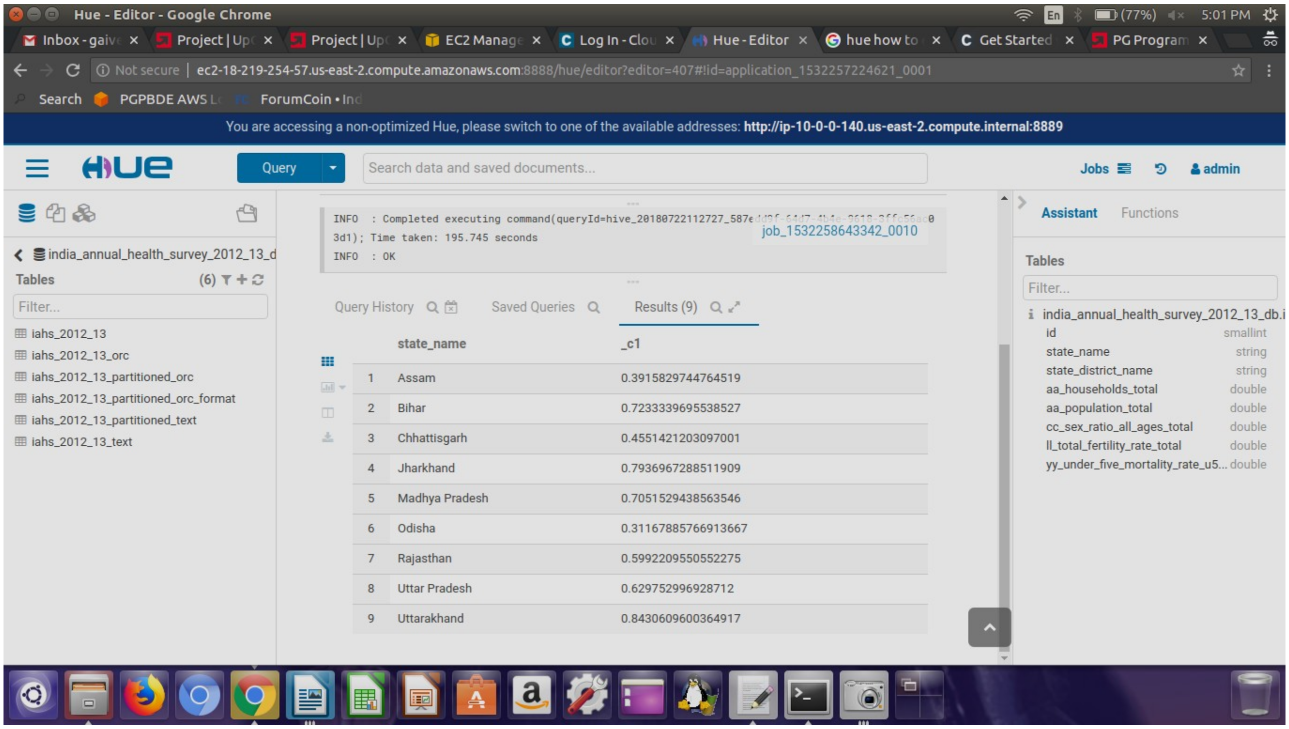
SORTING

3. Does high fertility correlate with high child mortality?

   Query

```
SELECT State_Name,
CORR(YY_Under_Five_Mortality_Rate_U5MR_Total_Person,
LL_Total_Fertility_Rate_Total)
FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc
GROUP BY State_Name;
```

   Screenshot of the result

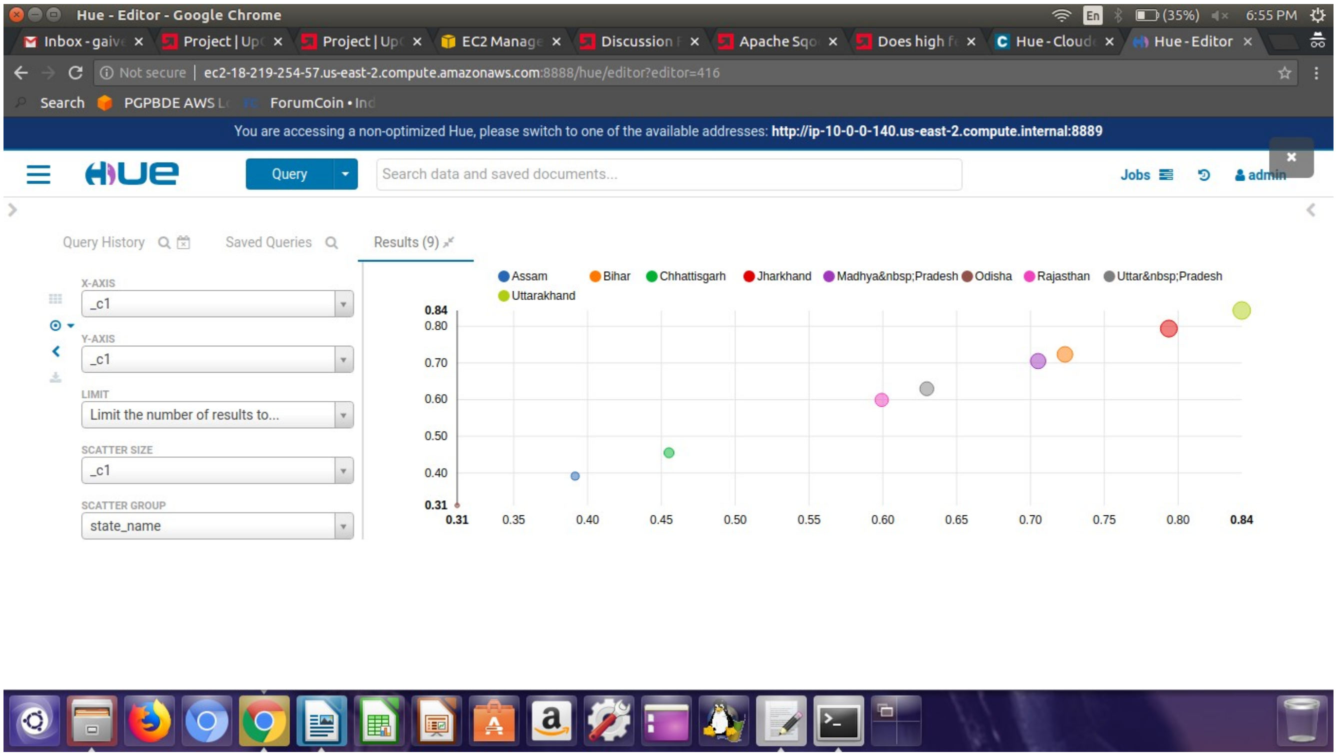


Hue - Editor - Google Chrome
Not secure | ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=407#!id=application_1532257224621_0001
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889
Query
Search data and saved documents...
Jobs
admin
india_annual_health_survey_2012_13_d
Tables (6)
iahs_2012_13
iahs_2012_13_orc
iahs_2012_13_partitioned_orc
iahs_2012_13_partitioned_orc_format
iahs_2012_13_partitioned_text
iahs_2012_13_text
Assistant
Functions
Tables
india_annual_health_survey_2012_13_db.i
id smallint
state_name string
state_district_name string
aa_households_total double
aa_population_total double
cc_sex_ratio_all_ages_total double
ll_total_fertility_rate_total double
yy_under_five_mortality_rate_u5... double
job_1532258643342_0010
Query History
Saved Queries
Results (9)

```
INFO  : Completed executing command(queryId=hive_20180722112727_587edd9f-64d7-4b4e-9618-3ffc56ac0
3d1); Time taken: 195.745 seconds
INFO  : OK
```

	state_name	_c1
1	Assam	0.3915829744764519
2	Bihar	0.7233339695538527
3	Chhattisgarh	0.4551421203097001
4	Jharkhand	0.7936967288511909
5	Madhya Pradesh	0.7051529438563546
6	Odisha	0.31167885766913667
7	Rajasthan	0.5992209550552275
8	Uttar Pradesh	0.629752996928712
9	Uttarakhand	0.8430609600364917

Chart



Hue - Editor - Google Chrome

Not secure | ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=416

You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

X-AXIS: _c1

Y-AXIS: _c1

LIMIT: Limit the number of results to...

SCATTER SIZE: _c1

SCATTER GROUP: state_name

Note

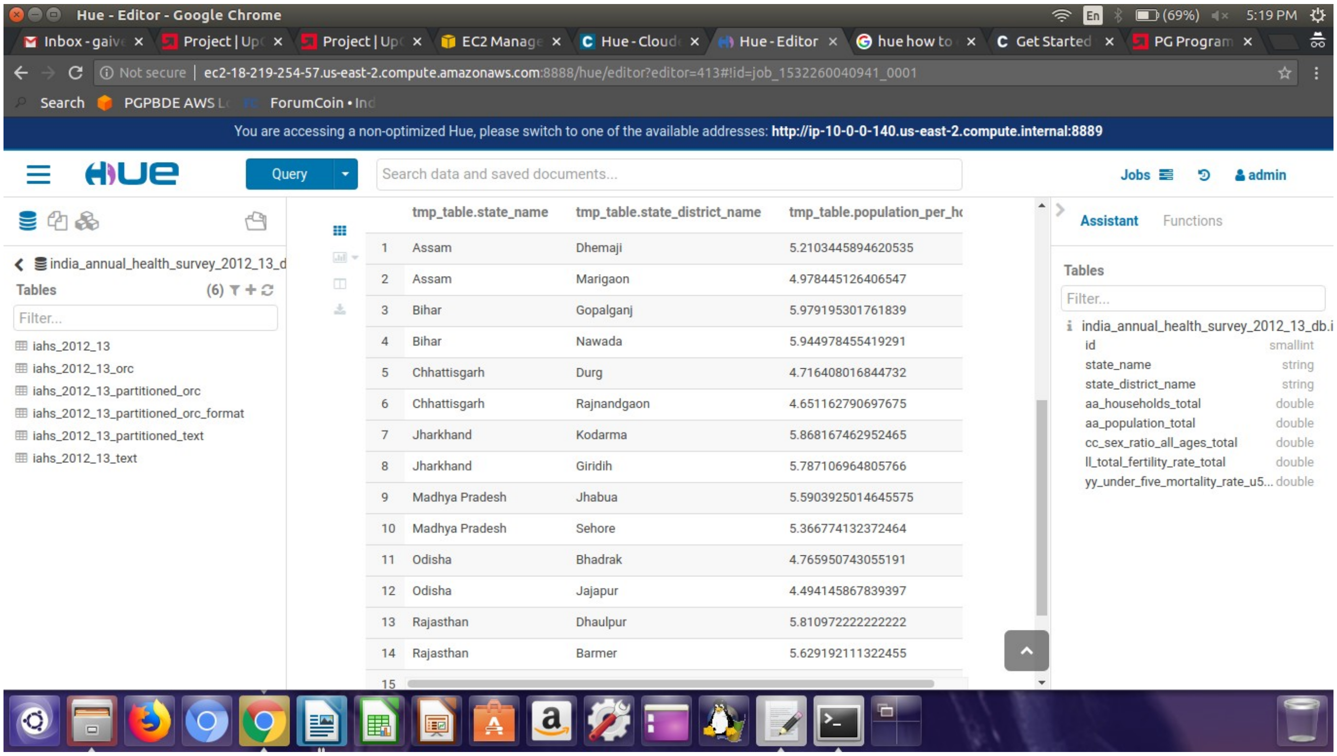
✔ Based on the analysis of the output, we see a positive slope in the scatter plot above as all the correlation co-efficient lie in the range of 0.3 to 0.8

4. Find top 2 districts per state with the highest population per household

Query

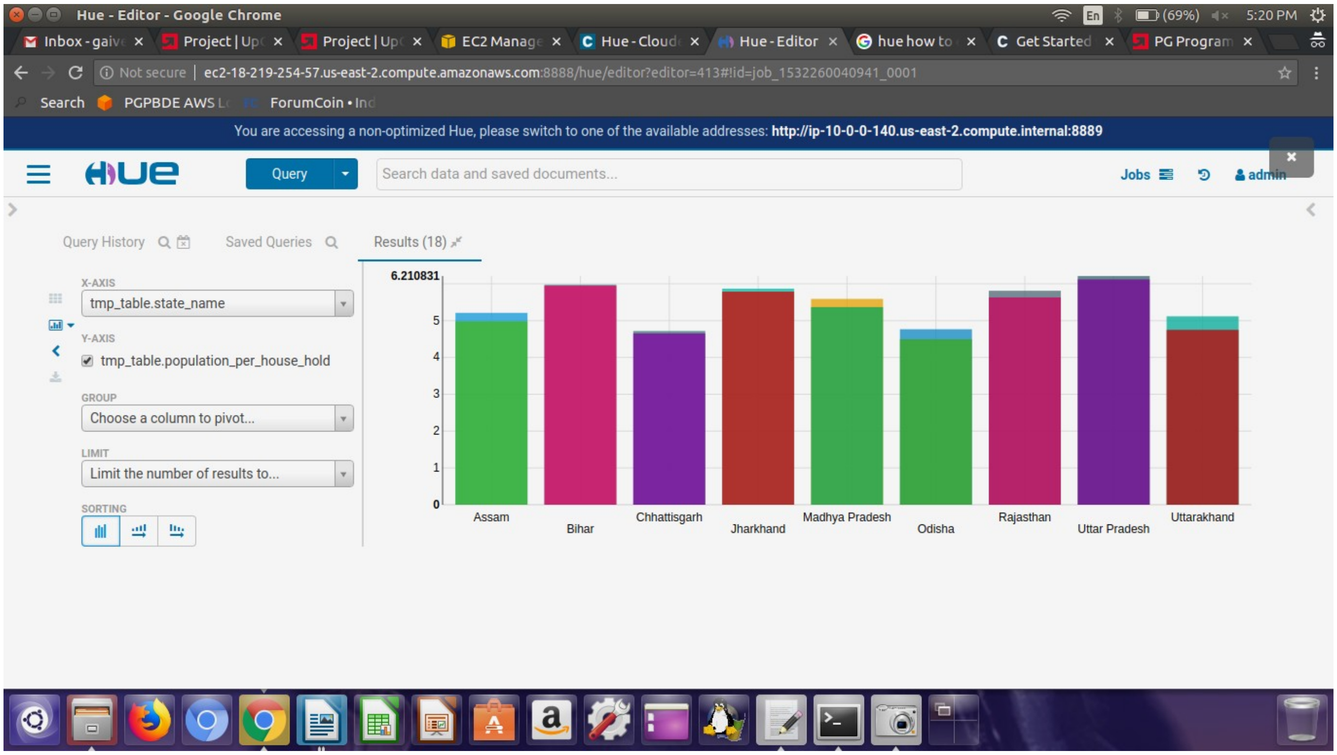
```
SELECT
tmp_table.State_Name,
tmp_table.State_district_name,
tmp_table.Population_Per_House_Hold
FROM (
SELECT
State_Name,
State_district_name,
(AA_Population_Total/AA_Households_Total) AS
Population_Per_House_Hold,
RANK() OVER (PARTITION BY State_Name ORDER BY
(AA_Population_Total/AA_Households_Total) DESC) AS Rank
FROM iahs_2012_13_orc
) tmp_table WHERE Rank < 3;
```

Screenshot of the result


Hue - Editor - Google Chrome
ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=413#!id=job_1532260040941_0001
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-0-0-140.us-east-2.compute.internal:8889
Query
Search data and saved documents...
Jobs
admin
india_annual_health_survey_2012_13_d
Tables (6)
iahs_2012_13
iahs_2012_13_orc
iahs_2012_13_partitioned_orc
iahs_2012_13_partitioned_orc_format
iahs_2012_13_partitioned_text
iahs_2012_13_text
Assistant
Functions
Tables
india_annual_health_survey_2012_13_db.i
id smallint
state_name string
state_district_name string
aa_households_total double
aa_population_total double
cc_sex_ratio_all_ages_total double
ll_total_fertility_rate_total double
yy_under_five_mortality_rate_u5... double

	tmp_table.state_name	tmp_table.state_district_name	tmp_table.population_per_h
1	Assam	Dhemaji	5.2103445894620535
2	Assam	Marigaon	4.978445126406547
3	Bihar	Gopalganj	5.979195301761839
4	Bihar	Nawada	5.944978455419291
5	Chhattisgarh	Durg	4.716408016844732
6	Chhattisgarh	Rajnandgaon	4.651162790697675
7	Jharkhand	Kodarma	5.868167462952465
8	Jharkhand	Giridih	5.787106964805766
9	Madhya Pradesh	Jhabua	5.5903925014645575
10	Madhya Pradesh	Sehore	5.366774132372464
11	Odisha	Bhadrak	4.765950743055191
12	Odisha	Jajapur	4.494145867839397
13	Rajasthan	Dhaulpur	5.810972222222222
14	Rajasthan	Barmer	5.629192111322455
15			

Chart



ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=413#!id=job_1532260040941_0001

You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

X-AXIS
tmp_table.state_name

Y-AXIS
☑ tmp_table.population_per_house_hold

GROUP
Choose a column to pivot...

LIMIT
Limit the number of results to...

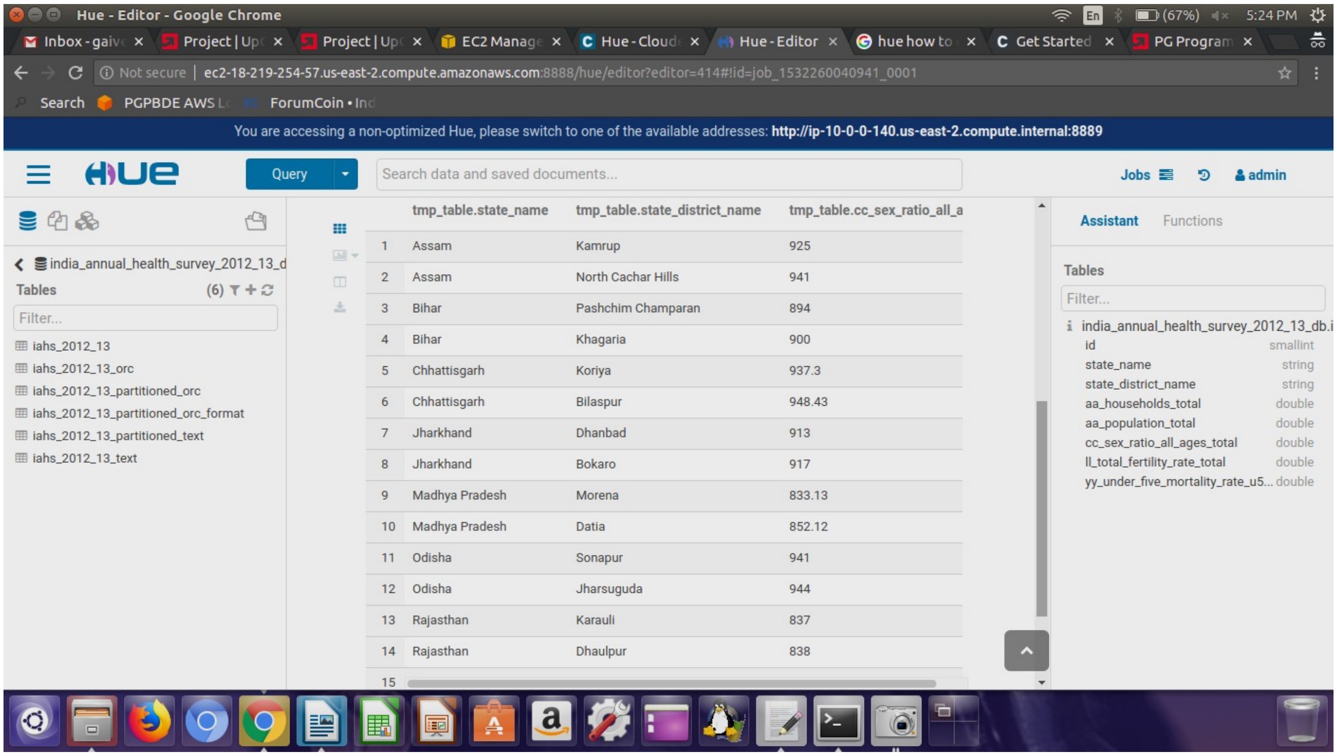
SORTING

5. Find top 2 districts per state with the lowest sex ratios

Query

```
SELECT
tmp_table.State_Name,
tmp_table.State_district_name,
tmp_table.CC_Sex_Ratio_All_Ages_Total
FROM (
SELECT
State_Name,
State_district_name,
CC_Sex_Ratio_All_Ages_Total,
RANK() OVER (PARTITION BY State_Name ORDER BY
CC_Sex_Ratio_All_Ages_Total ASC) AS Rank
FROM iahs_2012_13_orc
) tmp_table WHERE Rank < 3;
```

Screenshot of the result

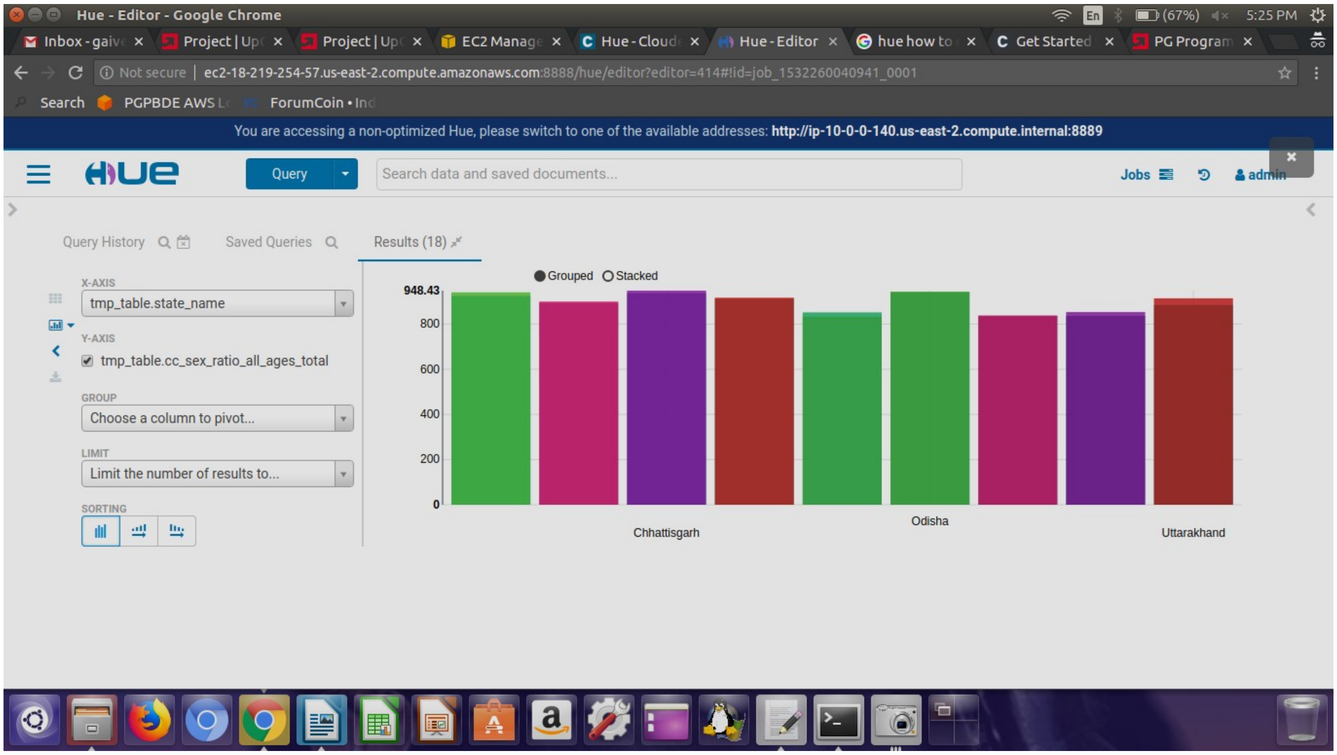


You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

	tmp_table.state_name	tmp_table.state_district_name	tmp_table.cc_sex_ratio_all_a
1	Assam	Kamrup	925
2	Assam	North Cachar Hills	941
3	Bihar	Pashchim Champaran	894
4	Bihar	Khagaria	900
5	Chhattisgarh	Koriya	937.3
6	Chhattisgarh	Bilaspur	948.43
7	Jharkhand	Dhanbad	913
8	Jharkhand	Bokaro	917
9	Madhya Pradesh	Morena	833.13
10	Madhya Pradesh	Datia	852.12
11	Odisha	Sonapur	941
12	Odisha	Jharsuguda	944
13	Rajasthan	Karauli	837
14	Rajasthan	Dhaulpur	838
15			

Hue - Editor - Google Chrome
Not secure | ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=414#!id=job_1532260040941_0001
Query
Search data and saved documents...
Jobs
admin
india_annual_health_survey_2012_13_d
Tables (6)
iahs_2012_13
iahs_2012_13_orc
iahs_2012_13_partitioned_orc
iahs_2012_13_partitioned_orc_format
iahs_2012_13_partitioned_text
iahs_2012_13_text
Assistant
Functions
Tables
india_annual_health_survey_2012_13_db.i
id smallint
state_name string
state_district_name string
aa_households_total double
aa_population_total double
cc_sex_ratio_all_ages_total double
ll_total_fertility_rate_total double
yy_under_five_mortality_rate_u5... double

Chart



Hue - Editor - Google Chrome

ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=414#!id=job_1532260040941_0001

You are accessing a non-optimized Hue, please switch to one of the available addresses: **http://ip-10-0-0-140.us-east-2.compute.internal:8889**

X-AXIS
tmp_table.state_name

Y-AXIS
☑ tmp_table.cc_sex_ratio_all_ages_total

GROUP
Choose a column to pivot...

LIMIT
Limit the number of results to...