

Data Ingestion Using Scoop And Data Analysis Using Hive

DATA INGESTION FROM RDS TO HDFS USING SQQOP

1. Sqoop import command ([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 ([viewData.hdfs](#))

HIVE EXTERNAL TABLE CREATION AND LOADING INGESTED DATA

1. Command to create the external table
([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 ([loadDataInHiveTable.hql](#))

3. Queries to verify that the ingestion is correctly accomplished

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

MySQL Workbench ([verificationQuery1.sql](#))
Hue ([verificationQuery1.hql](#))

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 ([verificationQuery2.sql](#))
Hue ([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.

1. *select count(*) from <Table Name>;*
2. *select State_Name, count(*) from <Table Name> group by State_Name;*
3. *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
([createAndInsertDefaultFormat.hql](#))

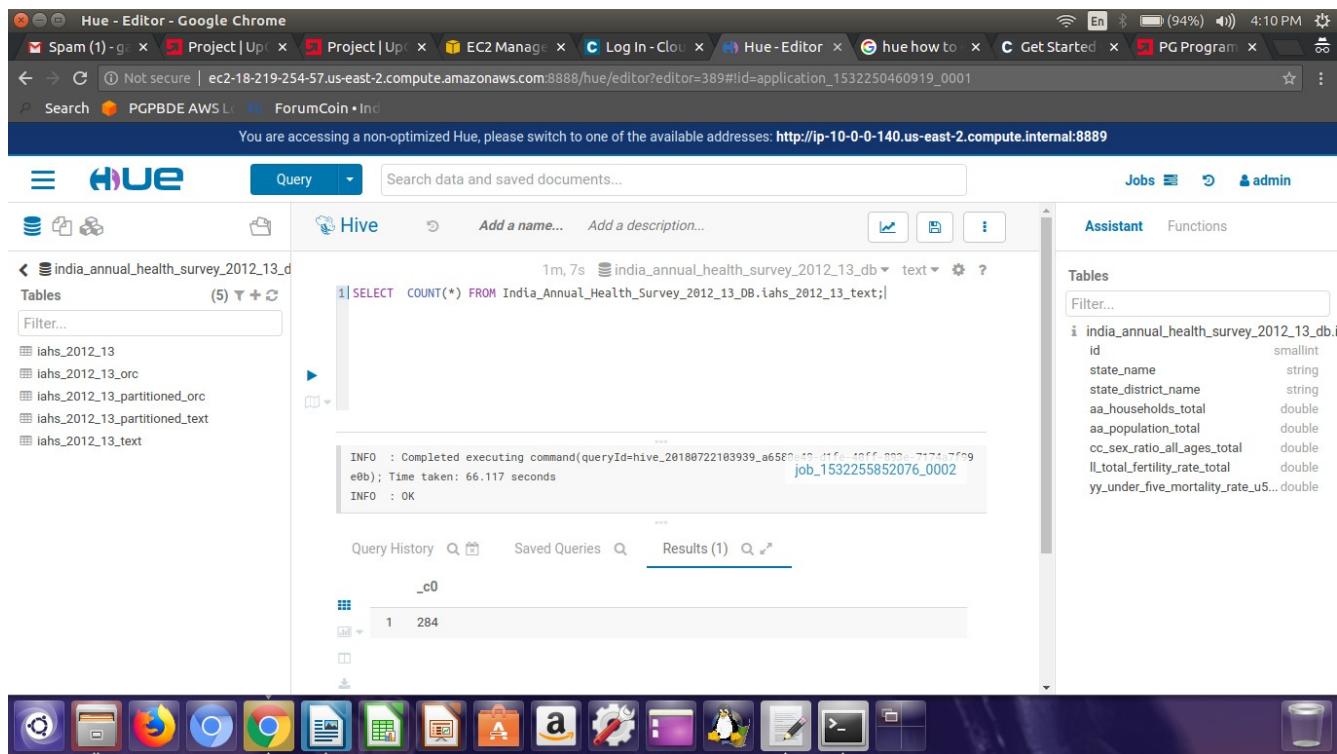
4. Create and insert command for the formats such as ORC
([createAndInsertORCFormat.hql](#))

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



The screenshot shows the Hue Editor interface in Google Chrome. The URL is http://ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=389#id=application_1532250460919_0001. The main area displays a Hive query:

```
1m,7s 1| SELECT COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text;
```

The output shows the execution completed successfully with a time taken of 66.117 seconds:

```
INFO : Completed executing command(queryId=hive_20180722183939_a658010_11f_10000_7171769
e0b); Time taken: 66.117 seconds
job_1532255852076_0002
INFO : OK
```

The sidebar on the left shows tables in the database: iahs_2012_13, iahs_2012_13_orc, iahs_2012_13_partitioned_orc, iahs_2012_13_partitioned_text, and iahs_2012_13_text. The right sidebar shows the table schema:

Table	Column	Type
india_annual_health_survey_2012_13_db	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

The screenshot shows the Hue Editor interface running on Google Chrome. The URL is http://ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=390#id=application_1532250460919_0001. The page title is "Hue - Editor - Google Chrome". The main content area displays a Hive query:

```
1m, 2s india_annual_health_survey_2012_13_db text ?  
1| SELECT COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc;  
INFO : Completed executing command(queryId=hive_20180722104141_5776) in 1m 2s 42ms (job_1532255852076_0005)  
INFO : OK
```

The results section shows a single row:

_c0
1 284

On the right side, there is a sidebar titled "Tables" which lists columns for the table:

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

The bottom of the screen shows a dock with various icons, including a terminal, file explorer, browser, and system tools.

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

The screenshot shows the Hue interface for Apache Hive. At the top, a browser tab for 'Hue - Editor - Google Chrome' is open, displaying the URL http://ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=395#id=job_1532256448123_0004. Below the browser, a message says: "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".

The main area shows a query editor with the following code:

```
3m, 12s  india_annual_health_survey_2012_13_db  text  ?  
1|SELECT State_Name, COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text GROUP
```

Below the editor, a log message indicates the query was completed successfully:

```
INFO : Completed executing command(queryId=hive_20180722105050_4729) in 3m 12s 185.335 seconds  
INFO : OK
```

The results section shows the output of the query:

state_name	_c1
1 Assam	23
2 Bihar	37
3 Chhattisgarh	16

On the left sidebar, under 'Tables', there are five entries: iahs_2012_13, iahs_2012_13_orc, iahs_2012_13_partitioned_orc, iahs_2012_13_partitioned_text, and iahs_2012_13_text.

On the right sidebar, the 'Tables' section lists the schema for the 'india_annual_health_survey_2012_13_db' table:

Table	Column	Type
india_annual_health_survey_2012_13_db	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

The screenshot shows the Hue web interface for Apache Hive. The top navigation bar includes links for Spam, Project, EC2 Manager, Log In, hue how to, Get Started, and PG Program. The main area has tabs for Jobs, Functions, Assistant, and Tables. The Tables tab is active, showing a table named 'india_annual_health_survey_2012_13_db' with columns: id, state_name, state_district_name, aa_households_total, aa_population_total, cc_sex_ratio_all_ages_total, ll_total_fertility_rate_total, and yy_under_five_mortality_rate_u5...'. A query editor window displays the following SQL command:

```
2m, 29s  india_annual_health_survey_2012_13_db  text  ?  
1| SELECT State_Name, COUNT(*) FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc GROU  
INFO : Completed executing command(queryId=hive_20180722104646_1abec17_0007_0000_0c1d_0a7745e113  
aaf); Time taken: 144.895 seconds  
INFO : OK
```

The results section shows a table with three rows:

state_name	_c1
1 Assam	23
2 Bihar	37
3 Chhattisgarh	16

The bottom navigation bar contains icons for various tools and services.

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

The screenshot shows the Hue - Editor interface in Google Chrome. The URL is http://ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=400#id=application_1532257224621_0001. The title bar says "Hue - Editor - Google Chrome". The main area displays a Hive query:

```
1|SELECT * FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_text WHERE State_Name = "Uttar Pradesh";
```

Execution details: 52.29s, 1 row, 1 file(s) processed. INFO : Completed executing command(queryId=hive_20180722110303_67610; time taken: 47.038 seconds). Time taken: 47.038 seconds job_1532257287025_0005 INFO : OK

The sidebar shows tables under "india_annual_health_survey_2012_13_db":

- Tables: iahs_2012_13, iahs_2012_13_orc, iahs_2012_13_partitioned_orc, iahs_2012_13_partitioned_text, iahs_2012_13_text

The results section shows a table with 3 rows:

	iahhs_2012_13 Orc ID	iahhs_2012_13 Orc State Name	iahhs_2012_13 Orc State District
1	202	Uttar Pradesh	Agra
2	203	Uttar Pradesh	Aligarh
3	204	Uttar Pradesh	Basti

The bottom navigation bar includes icons for various applications like HDFS, HIVE, HUE, and others.

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

The screenshot shows the Hue interface in a Google Chrome browser window. The URL is http://ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=397#id=job_1532256448123_0004. The title bar says "Hue - Editor - Google Chrome". The main area shows a Hive query editor with the following code:

```
52.47s 52.47s india_annual_health_survey_2012_13_db text ?  
1| SELECT * FROM India_Annual_Health_Survey_2012_13_DB.iahs_2012_13_orc WHERE State_Name = "Uttar  
INFO : Completed executing command(queryId=hive_20180722105656_18c95a42_a0a1_a5a0_792d42764  
a1f); Time taken: 46.913 seconds  
job_1532256711965_0008  
INFO : OK
```

The results pane shows a table with three columns: `iahs_2012_13_text.id`, `iahs_2012_13_text.state_name`, and `iahs_2012_13_text.state_dis`. The data is:

iahs_2012_13_text.id	iahs_2012_13_text.state_name	iahs_2012_13_text.state_dis
1	202	Agra
2	203	Aligarh
3	204	

The right sidebar shows the table definition for `india_annual_health_survey_2012_13_db`:

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.

([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
([queryForAnalysis1.hql](#))

Screenshot of the result

The screenshot shows the Hue Editor interface in Google Chrome. The URL is http://ec2-18-219-254-57.us-east-2.compute.amazonaws.com:8888/hue/editor?editor=404#lid=application_1532257224621_0001. The page displays the results of a query named "state_wise_average_child_mortality_rate".

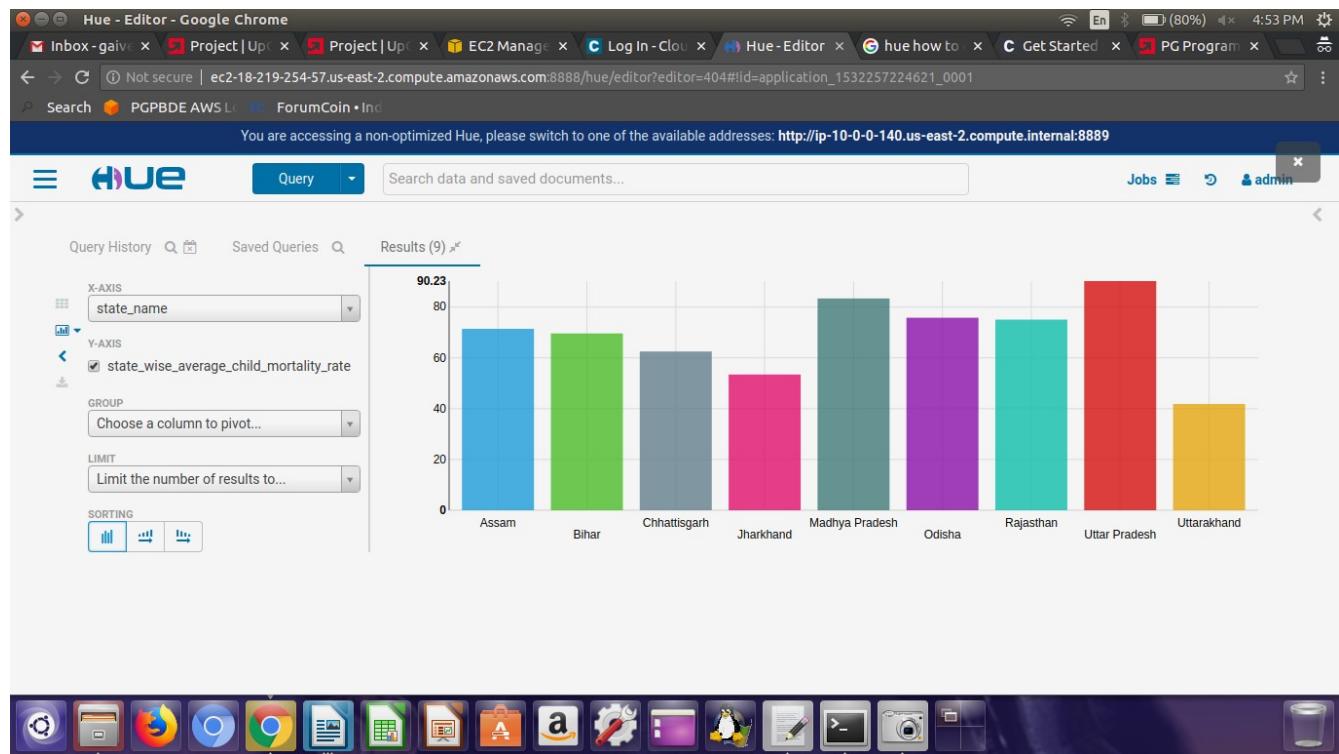
Query Results:

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

Table Definition:

```
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
state_name  string
```

Chart



2. State wise fertility rate ([queryForAnalysis2.hql](#))

Screenshot of the result

The screenshot shows the Hue Editor interface. On the left, there's a sidebar with a 'Tables' section containing several tables related to the 'india_annual_health_survey_2012_13_db' database. The main area displays the results of a query:

```

INFO : Completed executing command(queryId=hive_20180722112424_d51c788a-27-f4-4085-bf5c-1155b7e85
749); Time taken: 69.507 seconds
INFO : OK

```

Results (9)

	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

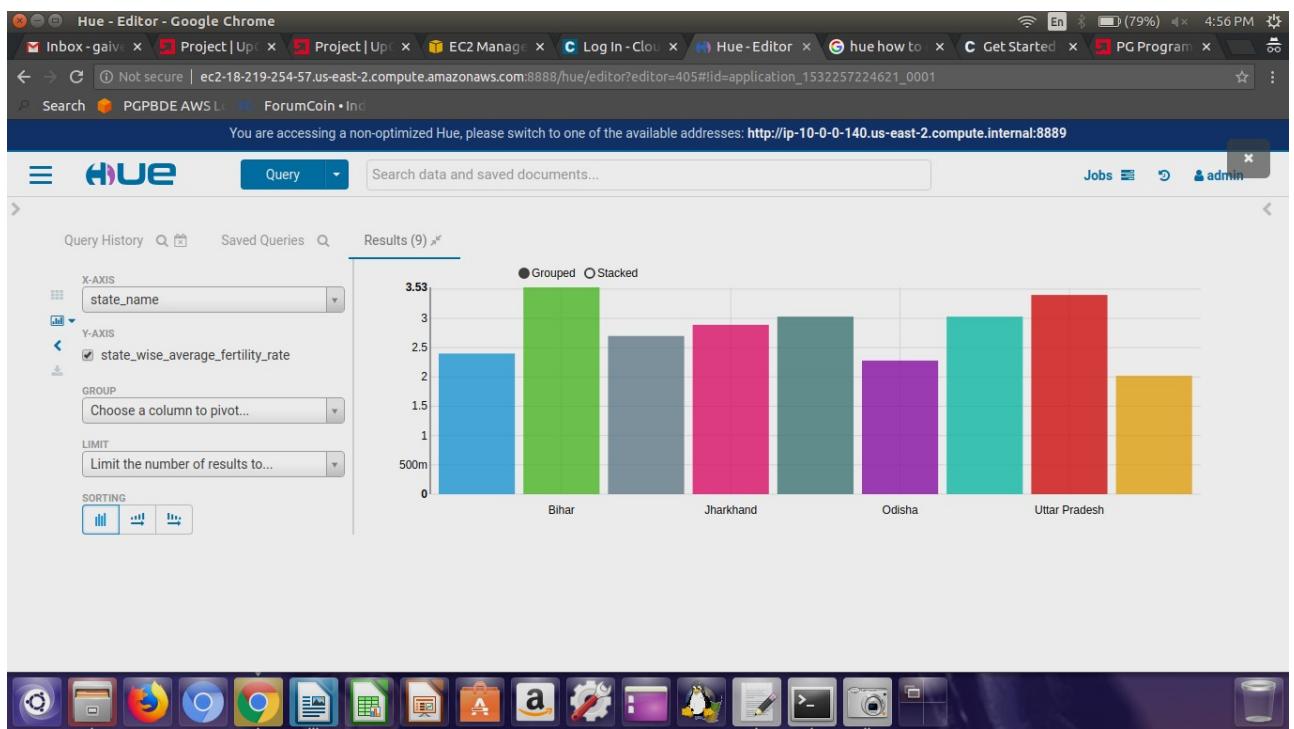
On the right, the 'Tables' section shows the schema for the 'india_annual_health_survey_2012_13_db' table:

```

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_u... double
state_name string

```

Chart



3. Does high fertility correlate with high child mortality? ([queryForAnalysis3.hql](#))

Screenshot of the result

Hue - Editor - Google Chrome
Project | Up | Project | Up | EC2 Manage | Log In - Cloud | 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=407#id=application_1532258643342_0010
Search PGPBDE AWS Log ForumCoin • Inc
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

HUE Query Search data and saved documents...

Tables (6) [+ ↗](#)

Filter...
iahhs_2012_13
iahhs_2012_13 Orc
iahhs_2012_13_partitioned Orc
iahhs_2012_13_partitioned Orc Format
iahhs_2012_13_partitioned Text
iahhs_2012_13 Text

INFO : Completed executing command(queryId=hive_20180722112727_587e10f5-1117-4f4c-8612-366f56a03d1); Time taken: 195.745 seconds
job_1532258643342_0010
INFO : OK

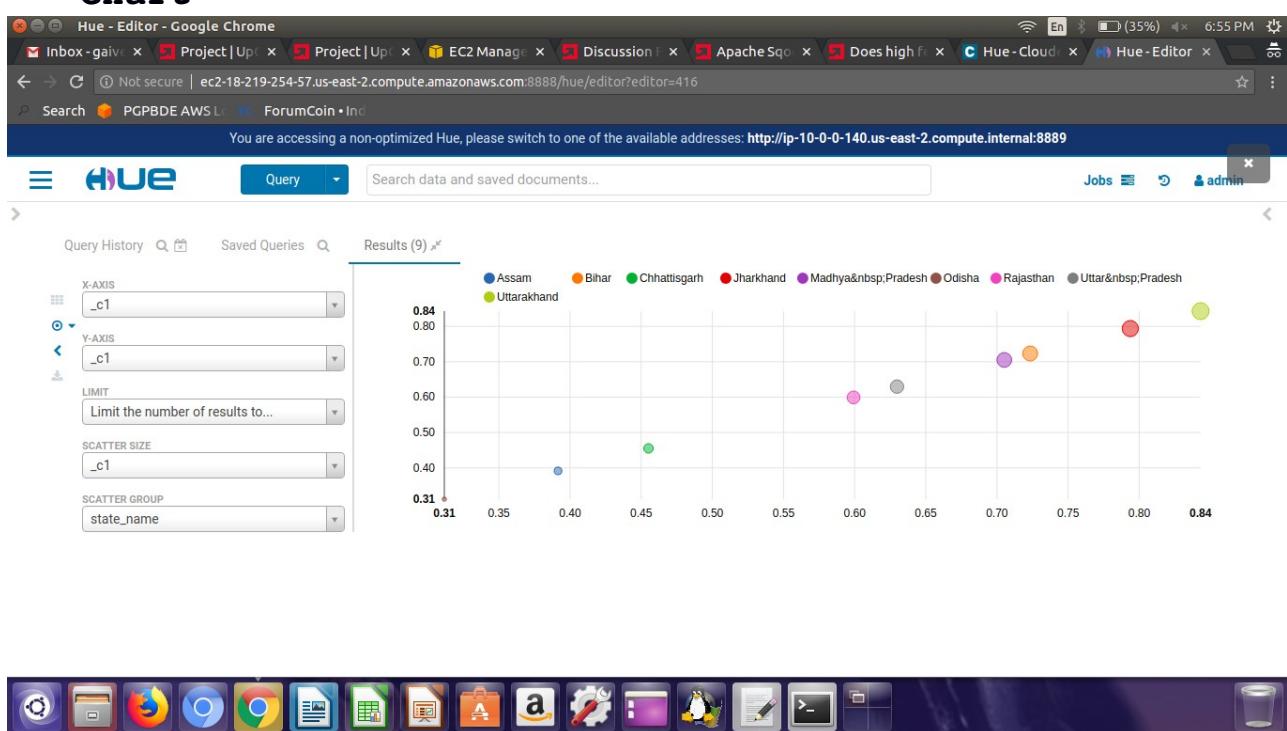
Query History Saved Queries Results (9) [Q ↗](#)

state_name	_c1
1 Assam	0.3915829744764519
2 Bihar	0.723339695538527
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

Assistant Functions

Tables
Filter...
iahhs_2012_13_db
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_us... double

Chart



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

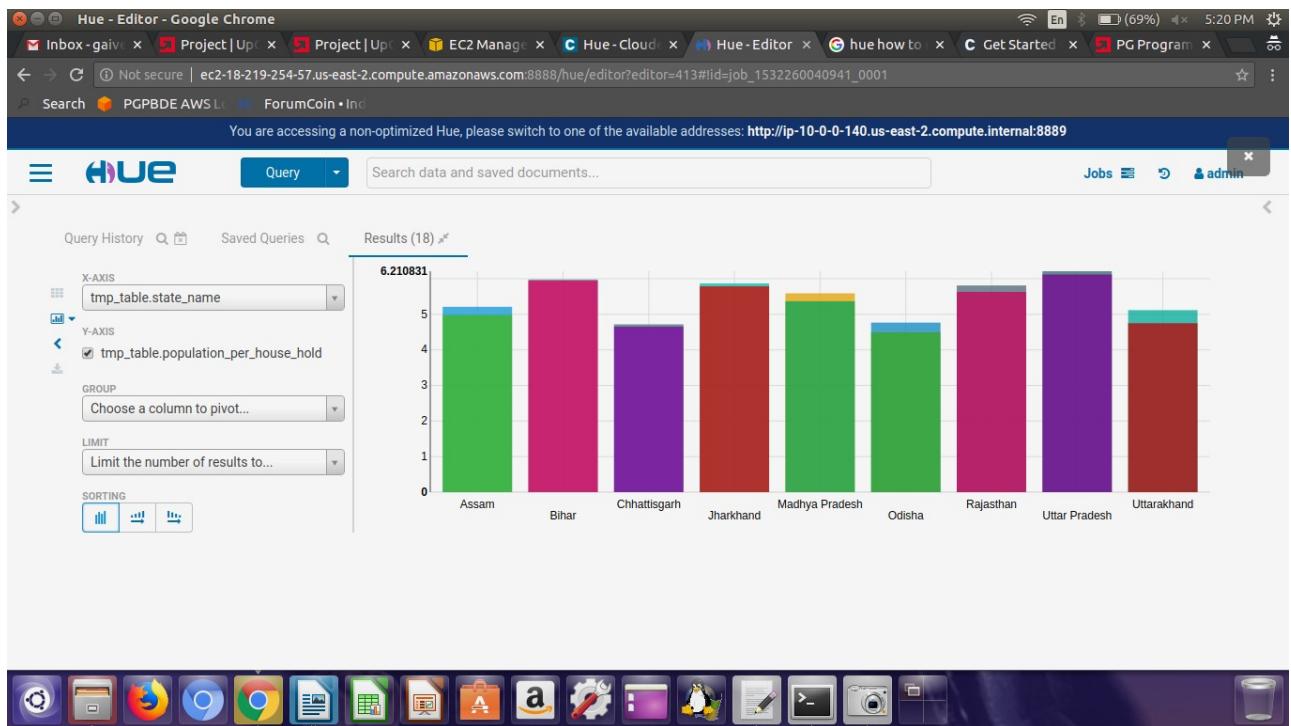
4. Find top 2 districts per state with the highest population per household ([queryForAnalysis4.hql](#))

Screenshot of the result

The screenshot shows the Hue Editor interface with a table of data. The table has three columns: tmp_table.state_name, tmp_table.state_district_name, and tmp_table.population_per_household. The data is as follows:

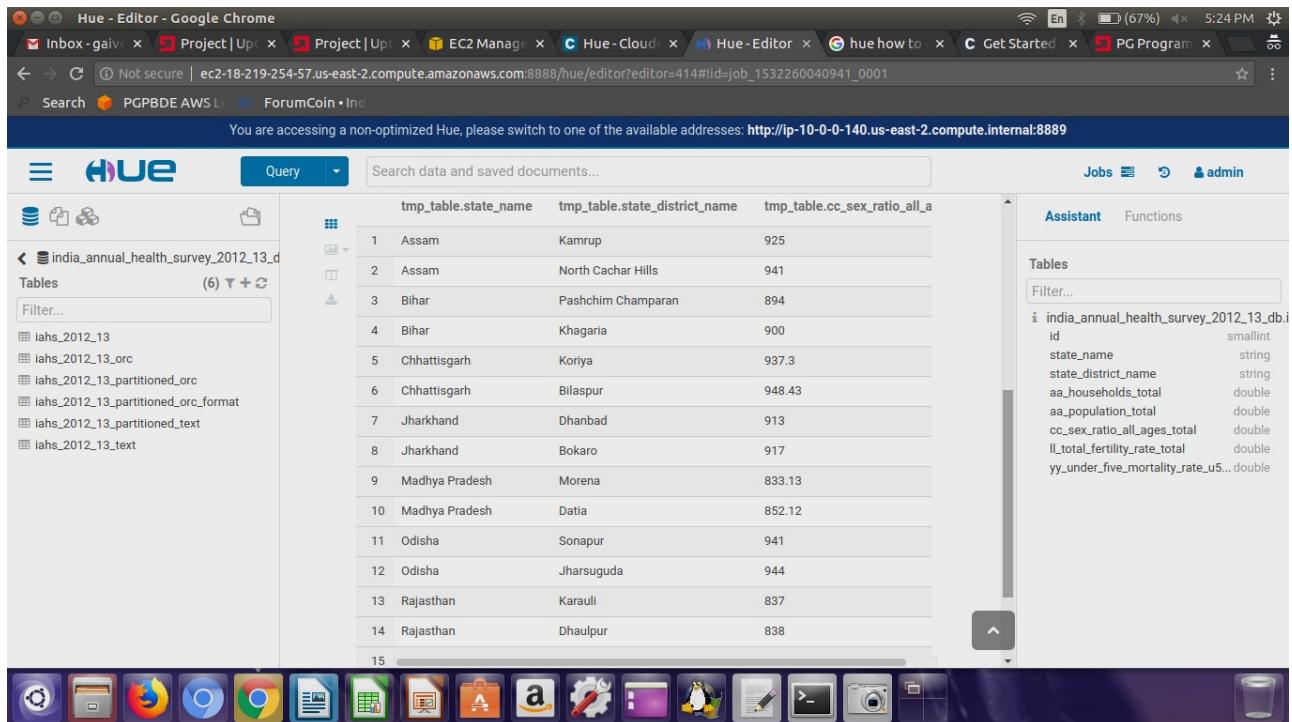
	tmp_table.state_name	tmp_table.state_district_name	tmp_table.population_per_household
1	Assam	Dhemaji	5.2103445894620535
2	Assam	Mariagon	4.978445126406547
3	Bihar	Gopalganj	5.979195301761839
4	Bihar	Nawada	5.944978455419291
5	Chhattisgarh	Durg	4.716408016844732
6	Chhattisgarh	Rajnandgaon	4.651162790697675
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

Chart



5. Find top 2 districts per state with the lowest sex ratios ([queryForAnalysis5.hql](#))

Screenshot of the result



The screenshot shows the Hue Editor interface. On the left, there's a sidebar with a tree view of tables under 'india_annual_health_survey_2012_13_db'. The main area displays a table with three columns: tmp_table.state_name, tmp_table.state_district_name, and tmp_table.cc_sex_ratio_all_a. The table lists 15 rows of data. To the right, there's an 'Assistant' panel showing the schema of the table.

	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			

Chart

