

## Data Ingestion from the RDS to HDFS using Sqoop

### 1. Sqoop Import command

```
sqoop import --connect jdbc:mysql://upgradawsrds.cpclxrkdvwzmz.us-east-1.rds.amazonaws.com:3306/indiaahs2012_13 --username upgraduser --password upgraduser --table Key_indicator_districtwise
```

### 2. Command to see the list of imported data

```
hadoop fs -ls /user/rajasekarssm_gmail/Key_indicator_districtwise
```

## External table creation in Hive and loading the ingested data into it. Data ingestion verification.

### 1. Command to create the external table

```
CREATE EXTERNAL TABLE IF NOT EXISTS HealthSurveyExt( ID STRING,  
State_Name STRING, State_District_Name STRING, etc. )  
COMMENT 'Data about India Annual Health Survey'  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
STORED AS TEXTFILE  
location '/user/rajasekarssm_gmail/Health_Survey_Ext_Folder';
```

### 2. Command to load the ingested data into the external table

```
LOAD DATA INPATH '/user/rajasekarssm_gmail/Key_indicator_districtwise'  
INTO TABLE HealthSurveyExt;
```

### 3. Queries to verify that the ingestion is correctly accomplished

a. Query to count the total number of rows along with the screenshots of the data fetched by the query on MySQL Workbench and Hue

Query:

```
select count(*) from healthsurveyext
```

Hue:

The screenshot displays the Apache Hue web interface for Hive. At the top, there's a header with the 'Hive' logo, a refresh icon, and fields for 'Add a name...' and 'Add a description...'. Below the header, the query editor shows a SQL query: `select count(*) from healthsurveyext`. To the right of the query, it indicates '19.7s' and 'default'. Below the query editor, the execution log is visible, showing a successful execution with the following messages:   
te: 4 SUCCESS  
INFO : Total MapReduce CPU Time Spent: 8 seconds 130 msec  
INFO : Completed executing command(queryId=hive\_20181216192626\_bd589125-437a-4905-8f7a4); Time taken: 17.822 seconds  
INFO : OK  
The job ID 'job\_1540986213' is also displayed. Below the log, there are tabs for 'Query History', 'Saved Queries', and 'Results (1)'. The 'Results (1)' tab is selected, showing a single row of results with the value '284' under the column header '\_c0'.

b. Query to select the top 10 rows and first 8 columns along with the screenshots of the data fetched by the query on MySQL Workbench and Hue

Query:

```
select ID, State_Name, State_District_Name, AA_Sample_Units_Total,  
AA_Sample_Units_Rural, AA_Sample_Units_Urban, AA_Households_Total,  
AA_Households_Rural from healthsurveyext LIMIT 10
```

Hue:

```
2 select ID,State_Name,State_District_Name,AA_Sample_Units_Total,  
3 AA_Sample_Units_Rural,AA_Sample_Units_Urban,AA_Households_Total,AA_Households_Rural  
4 from healthsurveyext LIMIT 10
```

```
AA_Sample_Units_Rural,AA_Sample_Units_Urban,AA_Households_Total,AA_Households_Rural  
from healthsurveyext LIMIT 10  
INFO : Completed executing command(queryId=hive_20181216193333_8301efde-85c2-4f07-8fe5-ab9811eb5  
abc); Time taken: 0.004 seconds  
INFO : OK
```

Query History Saved Queries Results (10)

	id	state_name	state_district_name	aa_sample_units_total	aa_sample_units	
	1	1	Assam	Barpeta	53.0	47.0
	2	2	Assam	Bongaigaon	89.0	73.0
	3	3	Assam	Cachar	105.0	84.0
	4	4	Assam	Darrang	26.0	24.0
	5	5	Assam	Dhemaji	121.0	108.0
	6	6	Assam	Dhubri	42.0	35.0

## Subset schema creation in Hive to support the analyses

1. Columns used in the subset schema

YY\_Infant\_Mortality\_Rate\_Imr\_Total\_Person, LL\_Total\_Fertility\_Rate\_Total,  
AA\_Households\_Total, CC\_Sex\_Ratio\_All\_Ages\_Total

2. Storage format used

TEXTFILE & ORC

3. Create and insert command for the default format

<Create command>

CREATE TABLE IF NOT EXISTS HealthSurveyExtSubDefault  
(State STRING, District STRING, ChildMortalityRate FLOAT,  
FertilityRate FLOAT, HouseHold FLOAT, SexRatio FLOAT) STORED AS TEXTFILE

<Insert command>

```
INSERT INTO TABLE HealthSurveyExtSubDefault select State_Name,  
State_District_Name,  
CAST(YY_Infant_Mortality_Rate_Imr_Total_Person AS FLOAT),  
CAST(LL_Total_Fertility_Rate_Total AS FLOAT) , CAST(AA_Households_Total AS  
FLOAT) , CAST(CC_Sex_Ratio_All_Ages_Total AS FLOAT) from healthsurveyext
```

4. Create and insert command for the formats such as ORC

<Create command>

```
CREATE TABLE IF NOT EXISTS HealthSurveyExtSubORC  
(State STRING, District STRING, ChildMortalityRate FLOAT,  
FertilityRate FLOAT, HouseHold FLOAT, SexRatio FLOAT) STORED AS ORC
```

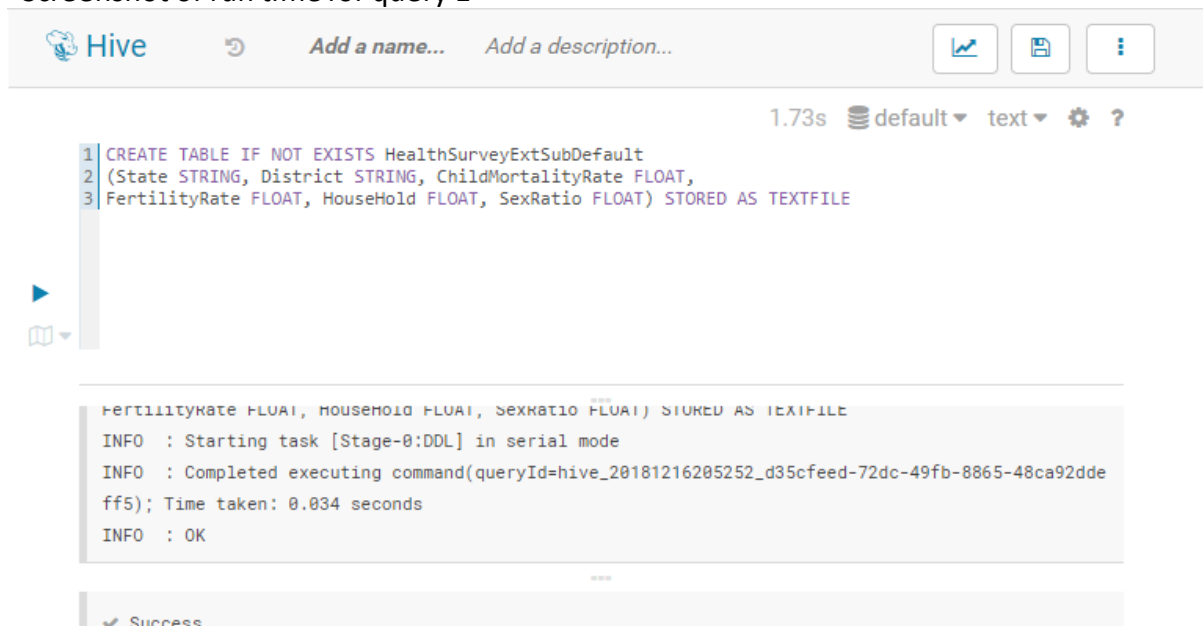
<Insert command>

```
INSERT INTO TABLE HealthSurveyExtSubORC select State_Name,  
State_District_Name,  
CAST(YY_Infant_Mortality_Rate_Imr_Total_Person AS FLOAT),  
CAST(LL_Total_Fertility_Rate_Total AS FLOAT) , CAST(AA_Households_Total AS  
FLOAT) , CAST(CC_Sex_Ratio_All_Ages_Total AS FLOAT) from healthsurveyext
```

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

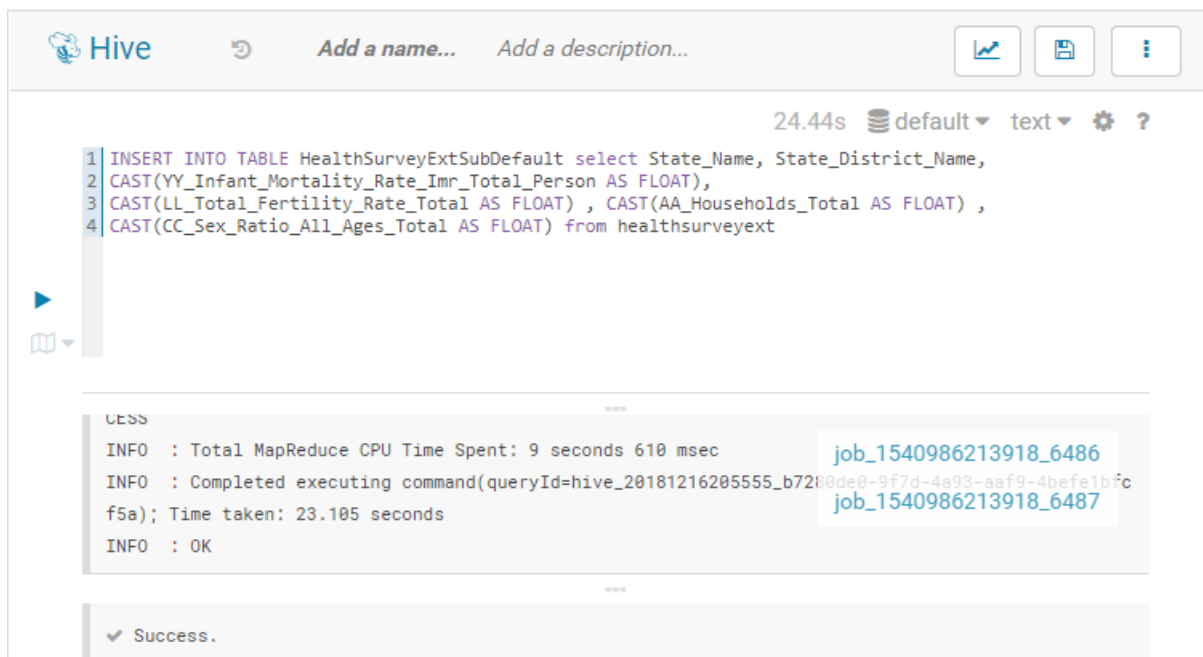
For default format:

<Screenshot of run time for query 1>



The screenshot shows the Hive web interface. At the top, there's a header with the Hive logo, a refresh button, and fields for 'Add a name...' and 'Add a description...'. On the right, there are buttons for 'View', 'Save', and a menu. Below the header, the query execution status is shown as '1.73s' with a default format dropdown, a text format dropdown, and settings icons. The SQL query is displayed in a monospace font: `1 CREATE TABLE IF NOT EXISTS HealthSurveyExtSubDefault`, `2 (State STRING, District STRING, ChildMortalityRate FLOAT,`, `3 FertilityRate FLOAT, Household FLOAT, SexRatio FLOAT) STORED AS TEXTFILE`. To the left of the query is a play button and a book icon. Below the query, the execution log is shown in a light gray box. It starts with 'FERTILITYRATE FLOAT, HOUSEHOLD FLOAT, SEXRATIO FLOAT) STORED AS TEXTFILE', followed by 'INFO : Starting task [Stage-0:DDL] in serial mode', 'INFO : Completed executing command(queryId=hive\_20181216205252\_d35cfeed-72dc-49fb-8865-48ca92ddeff5); Time taken: 0.034 seconds', and 'INFO : OK'. At the bottom, a green checkmark indicates 'Success.'.

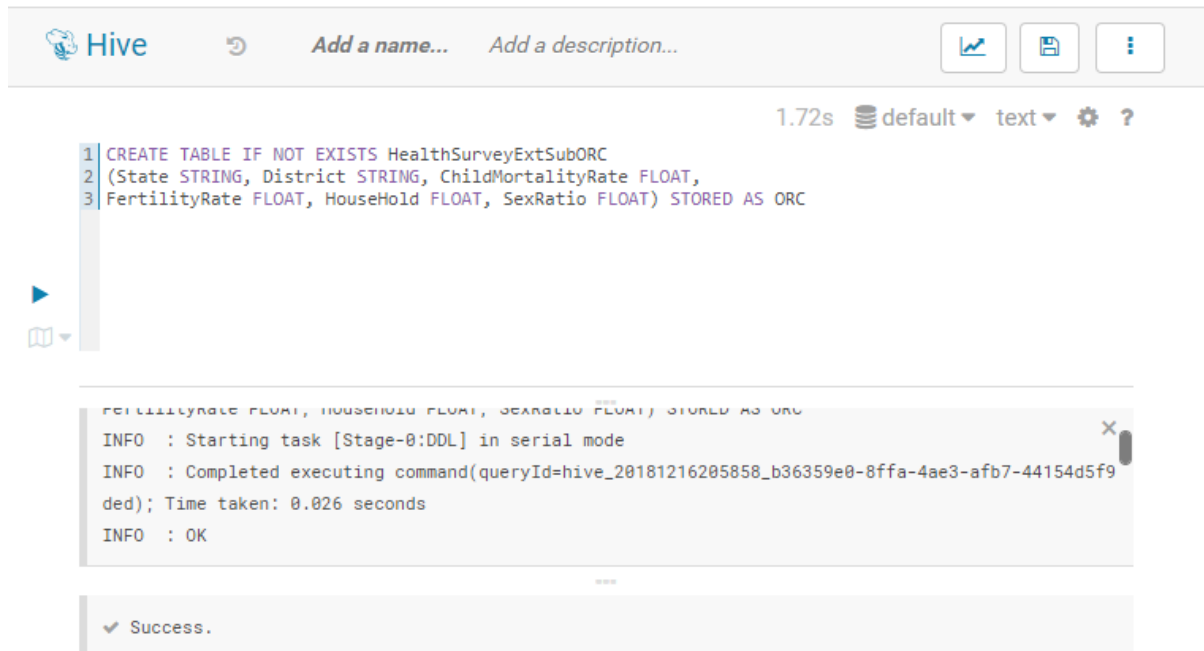
<Screenshot of run time for query 2>



The screenshot shows the Hive web interface. At the top, there's a header with the Hive logo, a refresh button, and fields for 'Add a name...' and 'Add a description...'. On the right, there are buttons for 'View', 'Save', and a menu. Below the header, the query execution status is shown as '24.44s' with a default format dropdown, a text format dropdown, and settings icons. The SQL query is displayed in a monospace font: `1 INSERT INTO TABLE HealthSurveyExtSubDefault select State_Name, State_District_Name,`, `2 CAST(YY_Infant_Mortality_Rate_Imr_Total_Person AS FLOAT),`, `3 CAST(LL_Total_Fertility_Rate_Total AS FLOAT) , CAST(AA_Households_Total AS FLOAT) ,`, `4 CAST(CC_Sex_Ratio_All_Ages_Total AS FLOAT) from healthsurveyext`. To the left of the query is a play button and a book icon. Below the query, the execution log is shown in a light gray box. It starts with 'SUCCESS', followed by 'INFO : Total MapReduce CPU Time Spent: 9 seconds 610 msec', 'INFO : Completed executing command(queryId=hive\_20181216205555\_b7280de0-9f7d-4a93-aaf9-4befe1bfcf5a); Time taken: 23.105 seconds', and 'INFO : OK'. To the right of the log, there are two job IDs: 'job\_1540986213918\_6486' and 'job\_1540986213918\_6487'. At the bottom, a green checkmark indicates 'Success.'.

For formats such as ORC:

<Screenshot of run time for query 1>



The screenshot shows the Hive web interface. At the top, there's a header with the Hive logo, a refresh button, and fields for 'Add a name...' and 'Add a description...'. On the right, there are buttons for 'Run', 'Save', and a menu. Below the header, the query execution time is shown as '1.72s' along with a dropdown menu set to 'default', a 'text' dropdown, and settings icons. The SQL query is displayed in a monospace font: `1 CREATE TABLE IF NOT EXISTS HealthSurveyExtSubORC  
2 (State STRING, District STRING, ChildMortalityRate FLOAT,  
3 FertilityRate FLOAT, Household FLOAT, SexRatio FLOAT) STORED AS ORC`. Below the query, there's a blue play button and a book icon. The execution log is shown in a light gray box with a close button (X). It contains the following text: `INFO : Starting task [Stage-0:DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20181216205858_b36359e0-8ffa-4ae3-afb7-44154d5f9  
ded); Time taken: 0.026 seconds  
INFO : OK`. At the bottom, a green checkmark indicates 'Success.'

<Screenshot of run time for query 2>



The screenshot shows the Hive web interface for a second query. The header is identical to the first screenshot. The query execution time is '23.92s'. The SQL query is: `1 INSERT INTO TABLE HealthSurveyExtSubORC select State_Name, State_District_Name,  
2 CAST(YY_Infant_Mortality_Rate_Imr_Total_Person AS FLOAT),  
3 CAST(LL_Total_Fertility_Rate_Total AS FLOAT) , CAST(AA_Households_Total AS FLOAT) ,  
4 CAST(CC_Sex_Ratio_All_Ages_Total AS FLOAT) from healthsurveyext`. The execution log shows: `INFO : Total MapReduce CPU Time Spent: 9 seconds 910 msec  
INFO : Completed executing command(queryId=hive_20181216205959_837ed037-3bba-4c23-b059-b771648033  
f92); Time taken: 22.688 seconds  
INFO : OK`. A yellow tooltip is visible over the job ID '837ed037-3bba-4c23-b059-b771648033f92', displaying 'job\_1540986213918\_6488' and 'job\_1540986213918\_6489'. The bottom status bar shows 'Success.'

1 select \* from HealthSurveyExtSubDefault

INFO : Executing command(queryId=hive\_20181217104747\_23ae3fe4-af15-4988-a770-8cd8319caaf6): select \* from HealthSurveyExtSubDefault  
INFO : Completed executing command(queryId=hive\_20181217104747\_23ae3fe4-af15-4988-a770-8cd8319caaf6); Time taken: 0.0 seconds  
INFO : OK

Query History Saved Queries Results (100+)

	healthsurveyextsubdefault.state	healthsurveyextsubdefault.district
1	Assam	Barpeta
2	Assam	Bongaigaon

1 select \* from HealthSurveyExtSubORC

INFO : Executing command(queryId=hive\_20181217113535\_49f1f47f-05a3-4b79-87dd-7f4f04f04da5): select \* from HealthSurveyExtSubORC  
INFO : Completed executing command(queryId=hive\_20181217113535\_49f1f47f-05a3-4b79-87dd-7f4f04f04da5); Time taken: 0.001 seconds  
INFO : OK

Query History Saved Queries Results (100+)

	healthsurveyextsuborc.state	healthsurveyextsuborc.district
1	Assam	Barpeta
2	Assam	Bongaigaon
3	Assam	Cachar
4	Assam	Darrang
5	Assam	Dhemaji
6	Assam	Dhubri
7	Assam	Dibrugarh
8	Assam	Goalpara
9	Assam	Golaghat

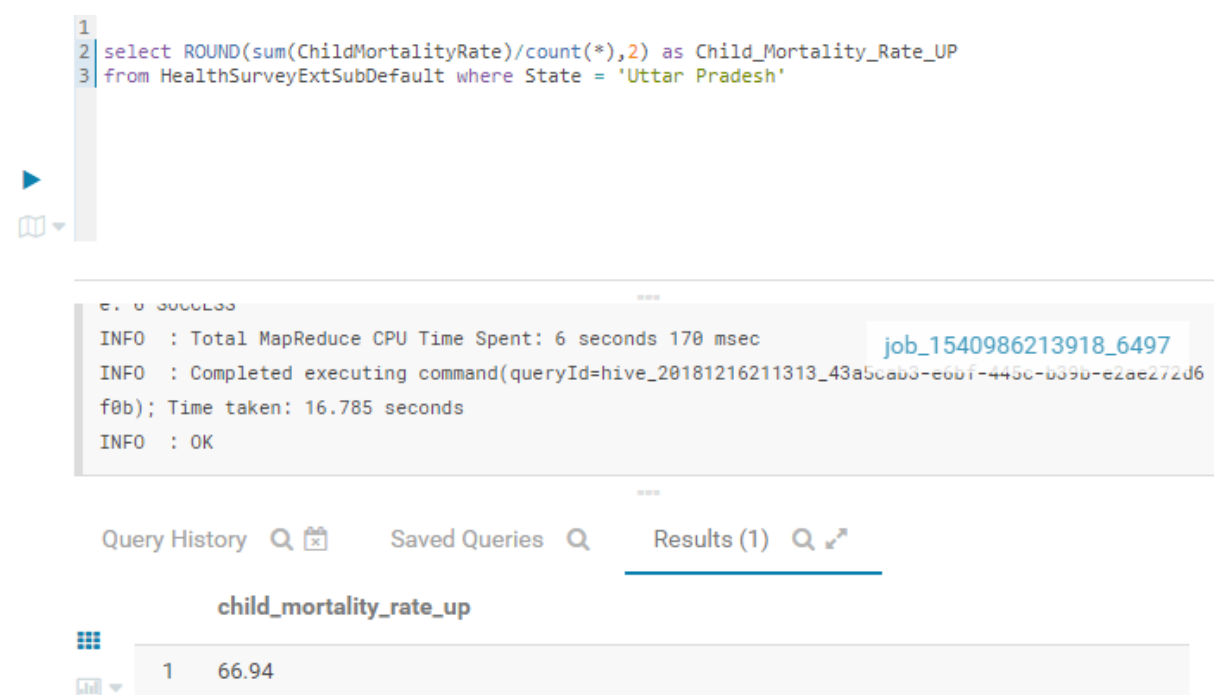
The result of each analysis along with the query and the corresponding chart generated in Hue. Keep optimizations in mind

1. The child mortality rate of Uttar Pradesh

<Query>

```
select ROUND(sum(ChildMortalityRate)/count(*),2) as  
Child_Mortality_Rate_UP  
from HealthSurveyExtSubDefault where State = 'Uttar Pradesh'
```

<Screenshot of the result>



The screenshot displays the Hue interface. At the top, a SQL query is entered in a text area:

```
1  
2 select ROUND(sum(ChildMortalityRate)/count(*),2) as Child_Mortality_Rate_UP  
3 from HealthSurveyExtSubDefault where State = 'Uttar Pradesh'
```

Below the query area, a log window shows the execution status:

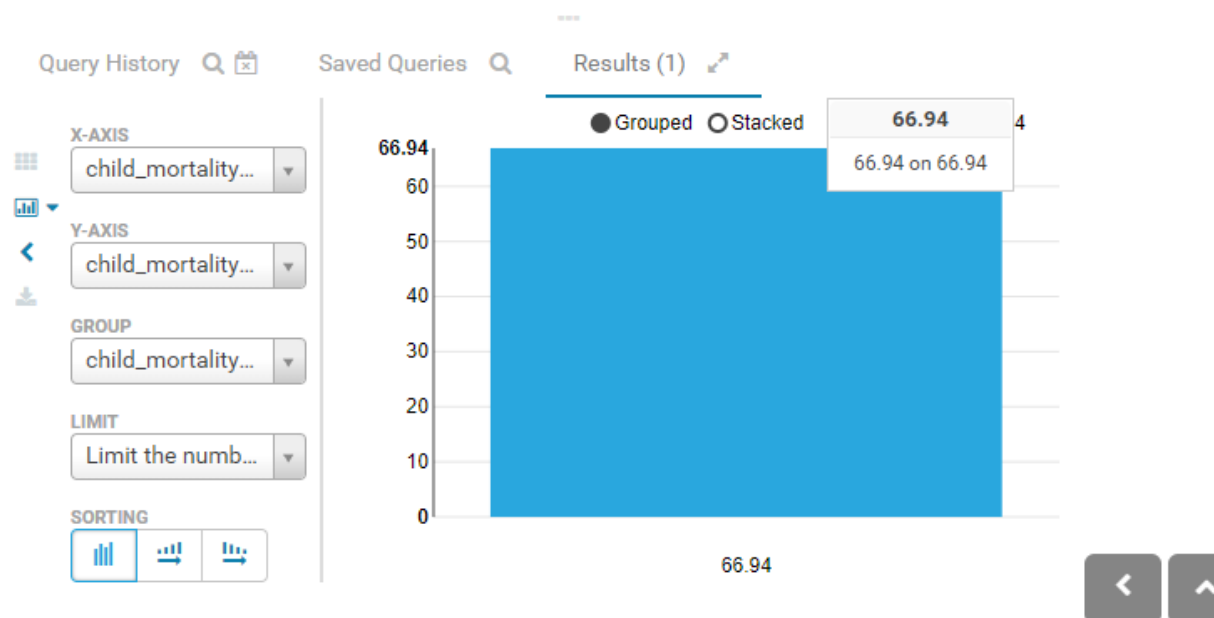
```
INFO : SUCCESS  
INFO : Total MapReduce CPU Time Spent: 6 seconds 170 msec  
INFO : Completed executing command(queryId=hive_20181216211313_43a5cab3-e6bf-445c-b39b-e2ae272d6f0b); Time taken: 16.785 seconds  
INFO : OK
```

The job ID `job_1540986213918_6497` is visible in the log. Below the log, the 'Results (1)' tab is selected, showing a table with one row of data:

child_mortality_rate_up	
1	66.94



<Chart>



2. The fertility rate of Bihar

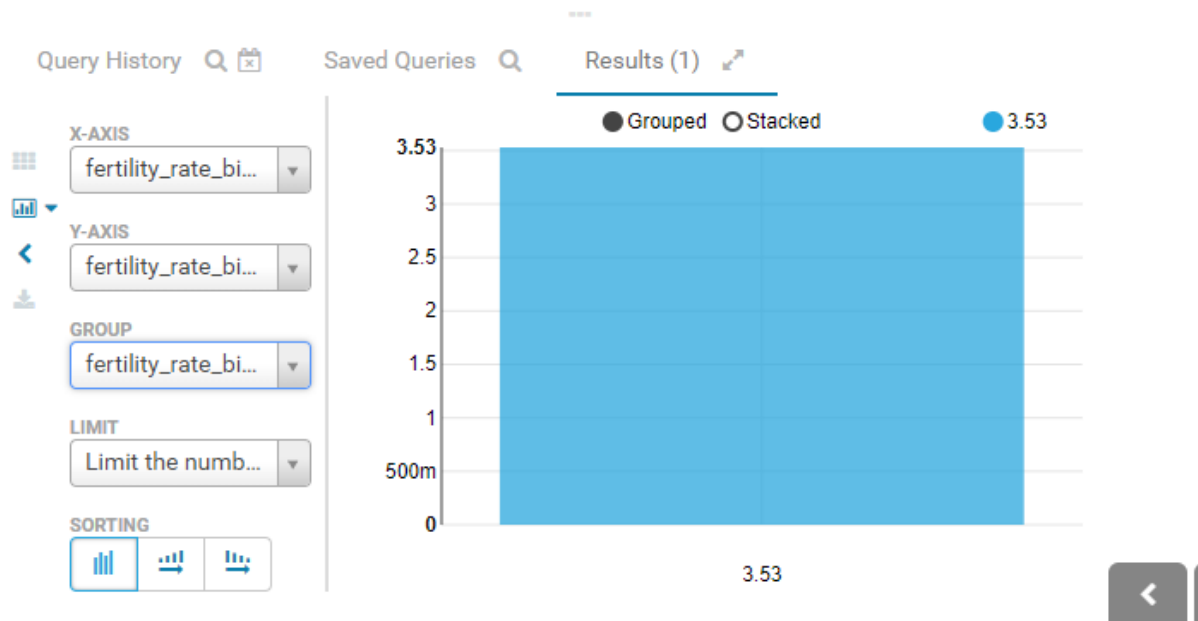
<Query>

```
select ROUND(sum(FertilityRate)/count(*),2) as Fertility_Rate_Bihar
from HealthSurveyExtSubDefault where State = 'Bihar'
```

<Screenshot of the result>



<Chart>



3. State wise child mortality rate and state wise fertility rate and does high fertility correlate with high child mortality?

<Query>

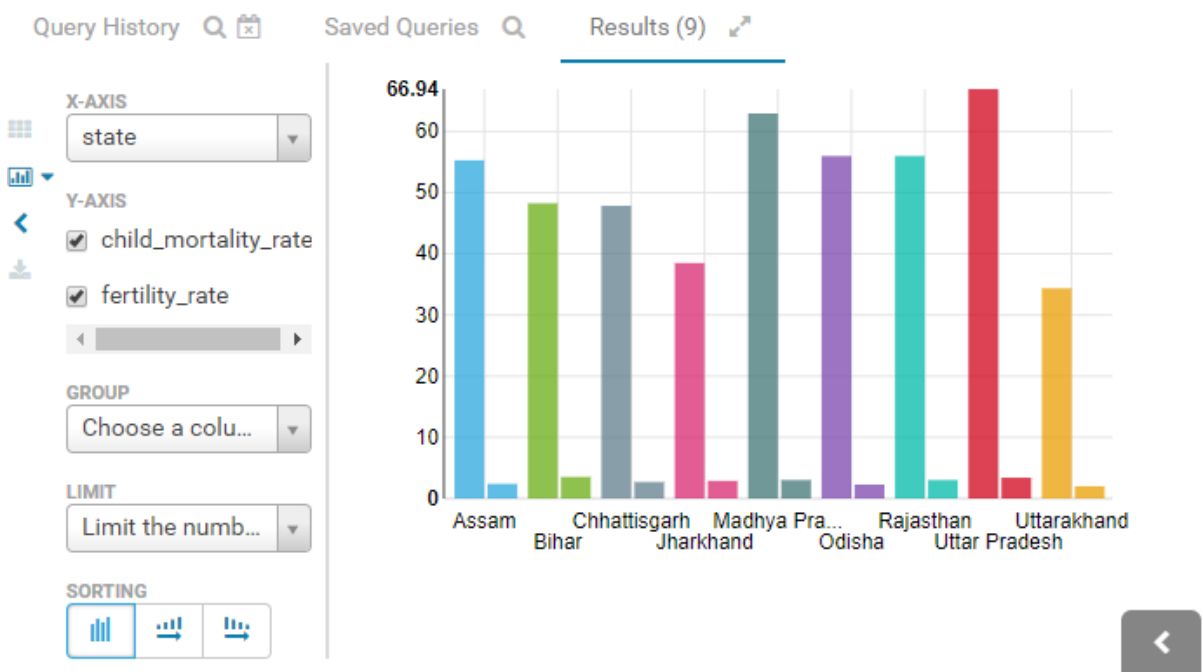
```
select ROUND(sum(ChildMortalityRate)/count(*),2) as Child_Mortality_Rate,  
ROUND(sum(FertilityRate)/count(*),2) as Fertility_Rate  
from HealthSurveyExtSubDefault Group By State
```

<Screenshot of the result>

Query History Saved Queries Results (9)

	state	child_mortality_rate	fertility_rate
1	Assam	55.3	2.4
2	Bihar	48.27	3.53
3	Chhattisgarh	47.86	2.7
4	Jharkhand	38.5	2.89
5	Madhya Pradesh	62.95	3.03
6	Odisha	56	2.28
7	Rajasthan	56	3.03
8	Uttar Pradesh	66.94	3.4
9	Uttarakhand	34.38	2.02

<Chart>

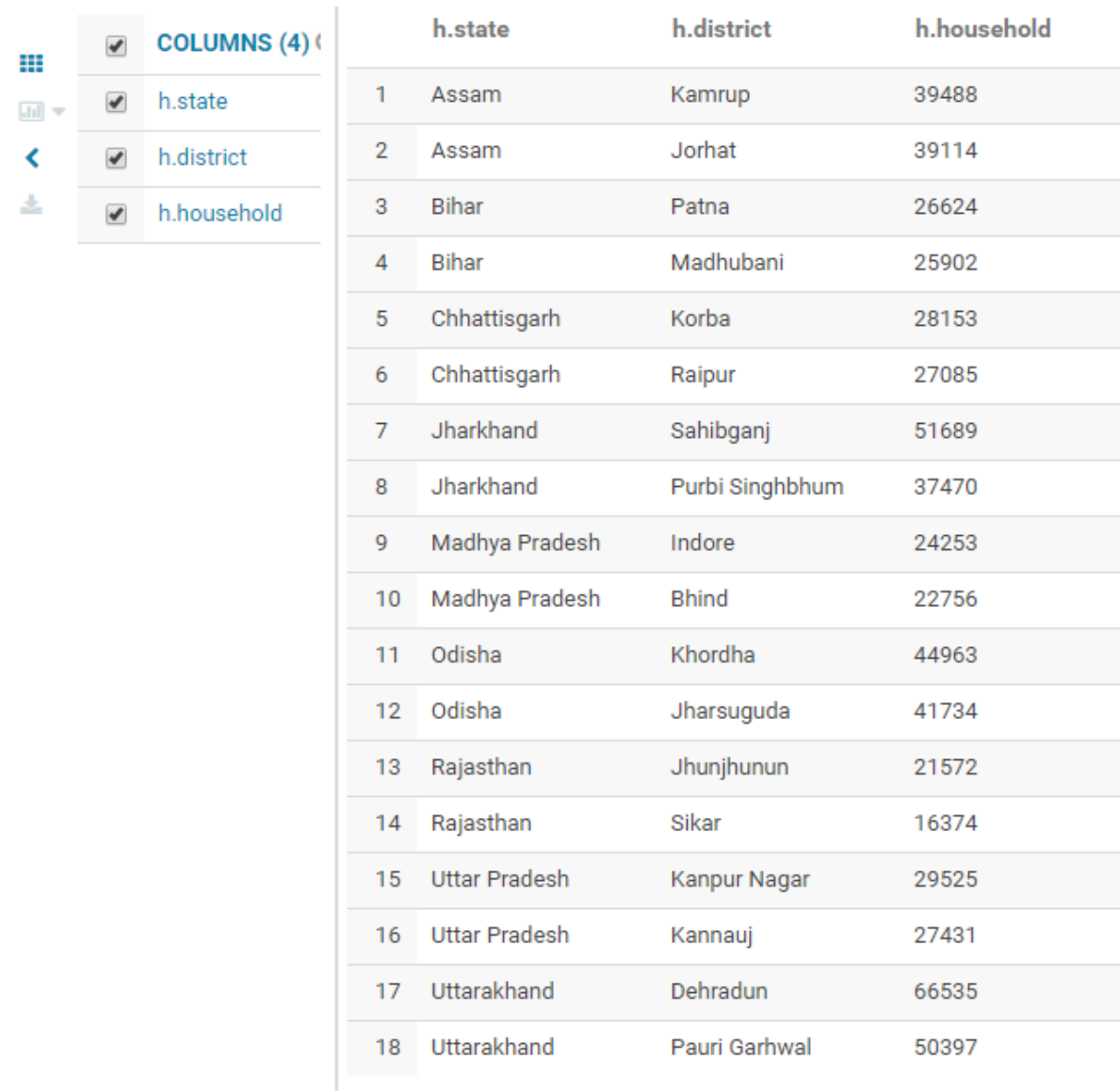


4. Find top 2 districts per state with the highest population per household

<Query>

```
select H.State, H.District, H.HouseHold from
( select State, District, HouseHold, row_number() over (partition by State
order by HouseHold desc) as RN from HealthSurveyExtSubDefault ) as H
where H.RN <= 2 order by H.State asc, H.HouseHold desc
```

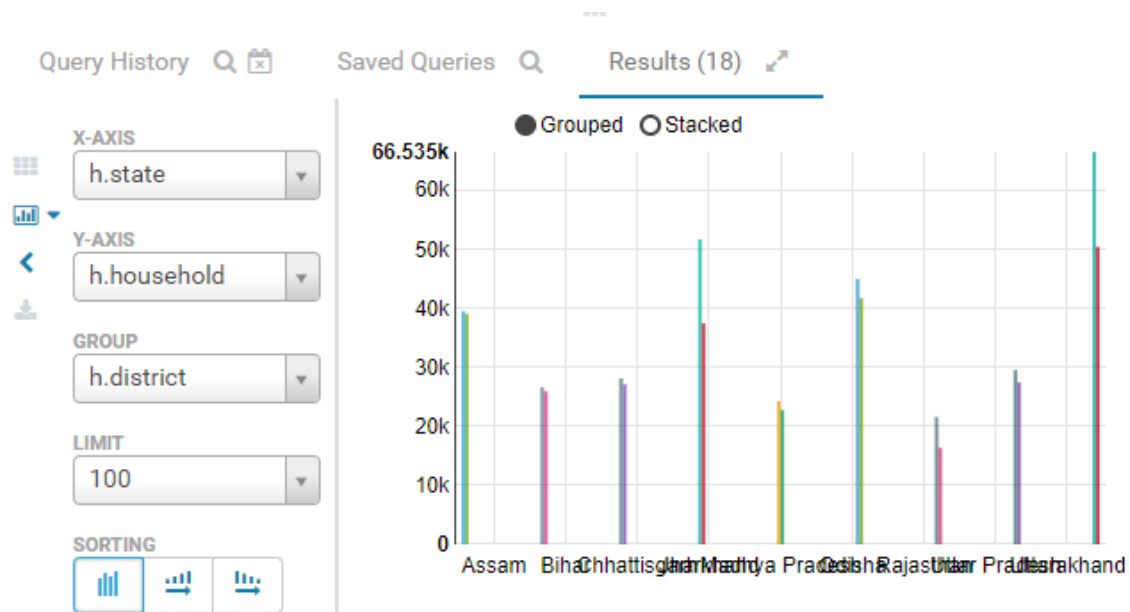
<Screenshot of the result>



The screenshot shows a data query interface. On the left, there is a column list titled 'COLUMNS (4)' with four columns: 'h.state', 'h.district', and 'h.household', each with a checkmark icon. On the right, there is a table with 18 rows and 4 columns. The columns are 'h.state', 'h.district', and 'h.household'. The rows are numbered 1 to 18. The data is sorted by state (ascending) and then by household count (descending).

	h.state	h.district	h.household
1	Assam	Kamrup	39488
2	Assam	Jorhat	39114
3	Bihar	Patna	26624
4	Bihar	Madhubani	25902
5	Chhattisgarh	Korba	28153
6	Chhattisgarh	Raipur	27085
7	Jharkhand	Sahibganj	51689
8	Jharkhand	Purbi Singhbhum	37470
9	Madhya Pradesh	Indore	24253
10	Madhya Pradesh	Bhind	22756
11	Odisha	Khordha	44963
12	Odisha	Jharsuguda	41734
13	Rajasthan	Jhunjhunun	21572
14	Rajasthan	Sikar	16374
15	Uttar Pradesh	Kanpur Nagar	29525
16	Uttar Pradesh	Kannauj	27431
17	Uttarakhand	Dehradun	66535
18	Uttarakhand	Pauri Garhwal	50397

<Chart>



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

<Query>

```
select H.State, H.District, H.SexRatio from
( select State, District, SexRatio, row_number() over (partition by State
order by SexRatio asc) as RN from HealthSurveyExtSubDefault ) as H
where H.RN <= 2 order by H.State asc, H.SexRatio asc
```

<Screenshot of the result>

	h.state	h.district	h.sexratio
1	Assam	Kamrup	925
2	Assam	North Cachar Hills	941
3	Bihar	Pashchim Champaran	894
4	Bihar	Khagaria	900
5	Chhattisgarh	Koriya	937.2999877929688
6	Chhattisgarh	Bilaspur	948.4299926757812
7	Jharkhand	Dhanbad	913
8	Jharkhand	Bokaro	917
9	Madhya Pradesh	Morena	833.1300048828125
10	Madhya Pradesh	Datia	852.1199951171875
11	Odisha	Sonapur	941
12	Odisha	Jharsuguda	944
13	Rajasthan	Karauli	837
14	Rajasthan	Dhaulpur	838
15	Uttar Pradesh	Gautam Buddha Nagar	836.8200073242188
16	Uttar Pradesh	Shahjahanpur	853.6699829101562
17	Uttarakhand	Haridwar	884.9299926757812
18	Uttarakhand	Udham Singh Nagar	914.3099975585938

<Chart>

