Data Ingestion from the RDS to HDFS using Sqoop

1. Sqoop Import command

```
sqoop import --connect jdbc:mysql://upgradawsrds.cpclxrkdvwmz.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;
```

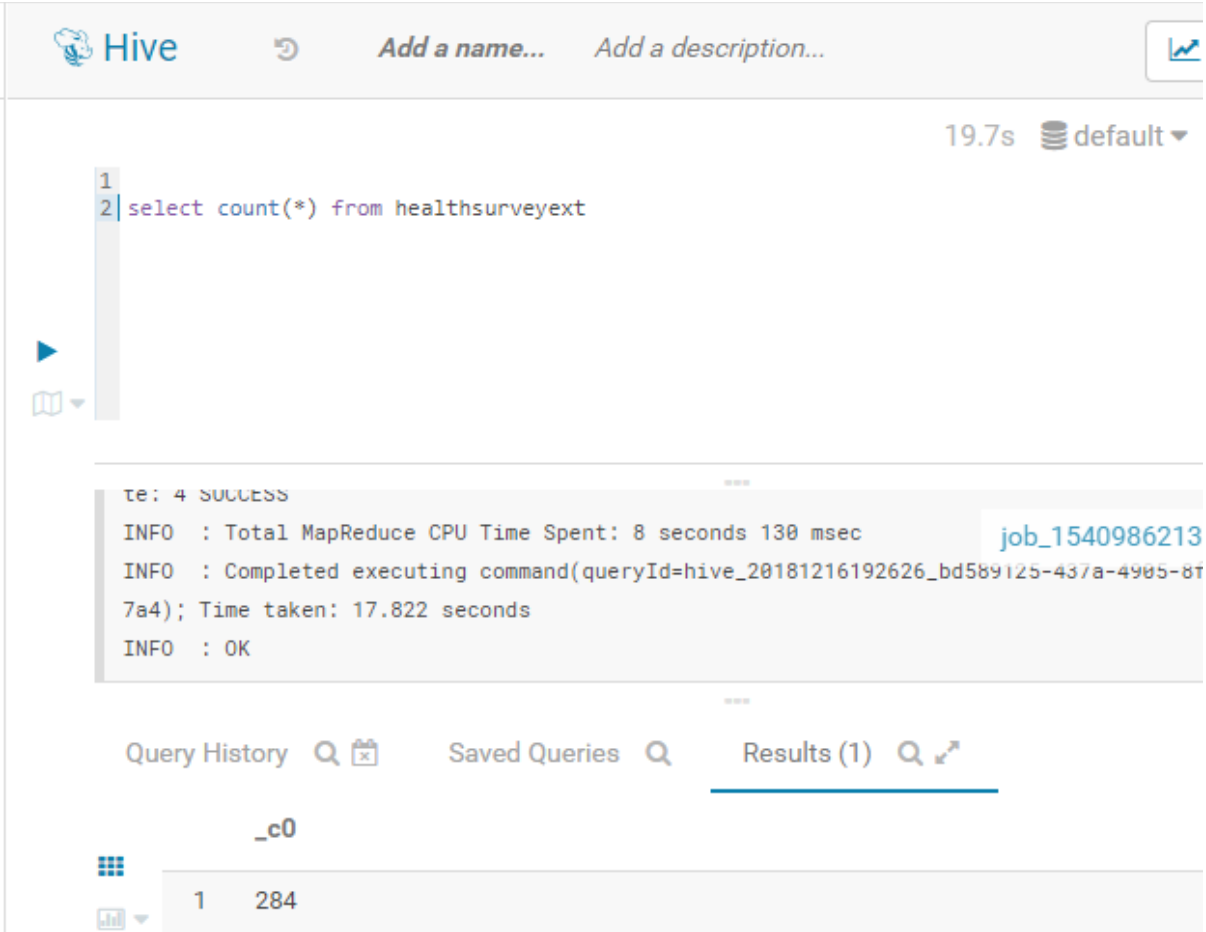
3. Queries to verify that the ingestion is correctly accomplished

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:



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:

```
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
```

```
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>

Hive Add a name... Add a description...

1.73s default text

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

```
FertilityRate FLOAT, HouseHold FLOAT, SexRatio FLOAT) STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20181216205252_d35cfeed-72dc-49fb-8865-48ca92dde
ff5); Time taken: 0.034 seconds
INFO  : OK
```

✔ Success.

<Screenshot of run time for query 2>

Hive Add a name... Add a description...

24.44s default text

```
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
```

```
INFO  : Total MapReduce CPU Time Spent: 9 seconds 610 msec
INFO  : Completed executing command(queryId=hive_20181216205555_b72
f5a); Time taken: 23.105 seconds
INFO  : OK
```

job_1540986213918_6486
job_1540986213918_6487

✔ Success.

For formats such as ORC:

<Screenshot of run time for query 1>

Hive Add a name... Add a description...

1.72s default text

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

```
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
```

✔ Success.

<Screenshot of run time for query 2>

Hive Add a name... Add a description...

23.92s default text

```
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
```

```
INFO  : Total MapReduce CPU Time Spent: 9 seconds 910 msec
INFO  : Completed executing command(queryId=hive_20181216205959_837ad037-3bba-4c23-b059-b77164053
f92); Time taken: 22.688 seconds
INFO  : OK
```

job_1540986213918_6488
job_1540986213918_6489

✔ Success.

2.33s default text ?

```
1 select * from HealthSurveyExtSubDefault
```

```
INFO  : Executing command(queryId=hive_20181217104747_23ae3fe4-af15-4988-a770-8cd
8319caaf6): select * from HealthSurveyExtSubDefault
INFO  : Completed executing command(queryId=hive_20181217104747_23ae3fe4-af15-498
8-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): se
lect * from HealthSurveyExtSubORC
INFO  : Completed executing command(queryId=hive_20181217113535_49f1f47f-05a3-4b79-87dd-7f4f04f
04da5); Time taken: 0.001 seconds
INFO  : OK
```

Query History Saved Queries Results (100+)

COLUMNS (7)
- healthsurveyext
- healthsurveyext
- healthsurveyext
- healthsurveyext
- healthsurveyext
- healthsurveyext

	healthsurveyextsuborc.state	healthsurveyextsuborc.distric
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

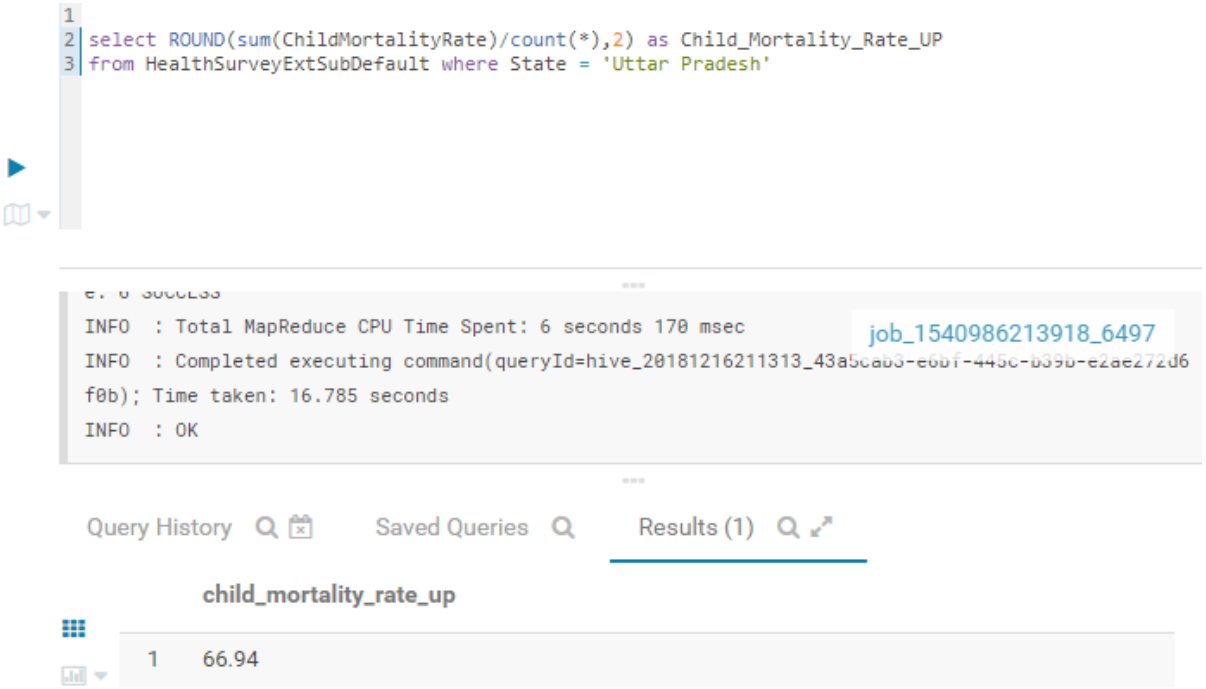
The result of each analysis along with the query and the corresponding chart generated in Hue. Keep optimizations in mind

1. The child mortality rate of Uttar Pradesh

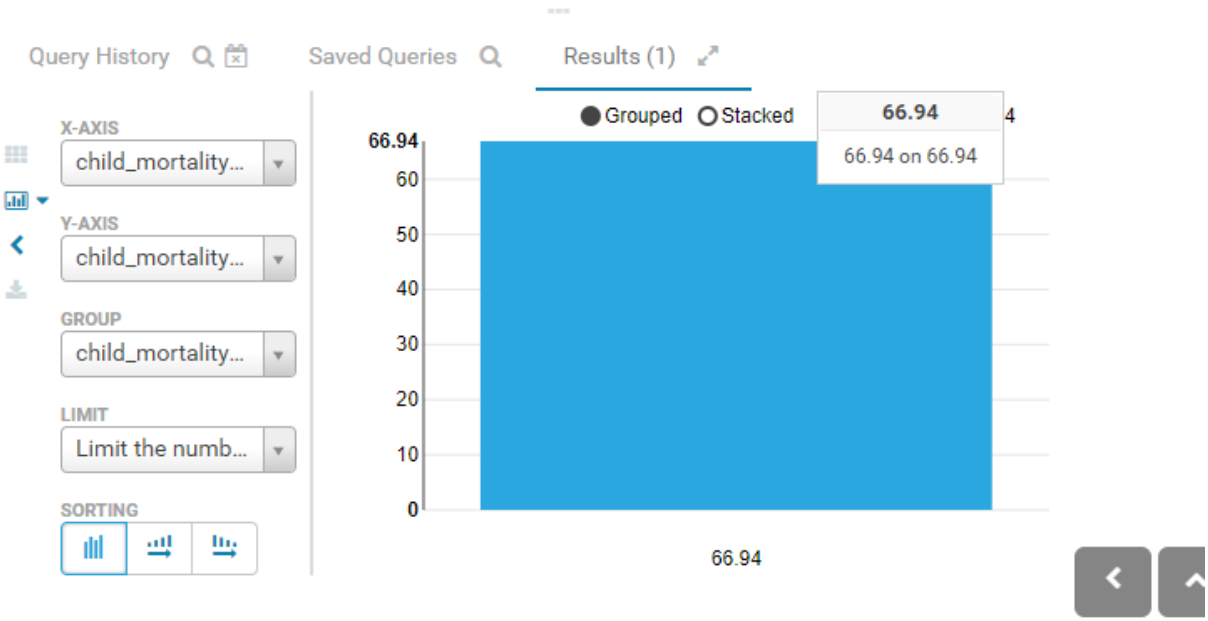
<Query>

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

<Screenshot of the result>



<Chart>



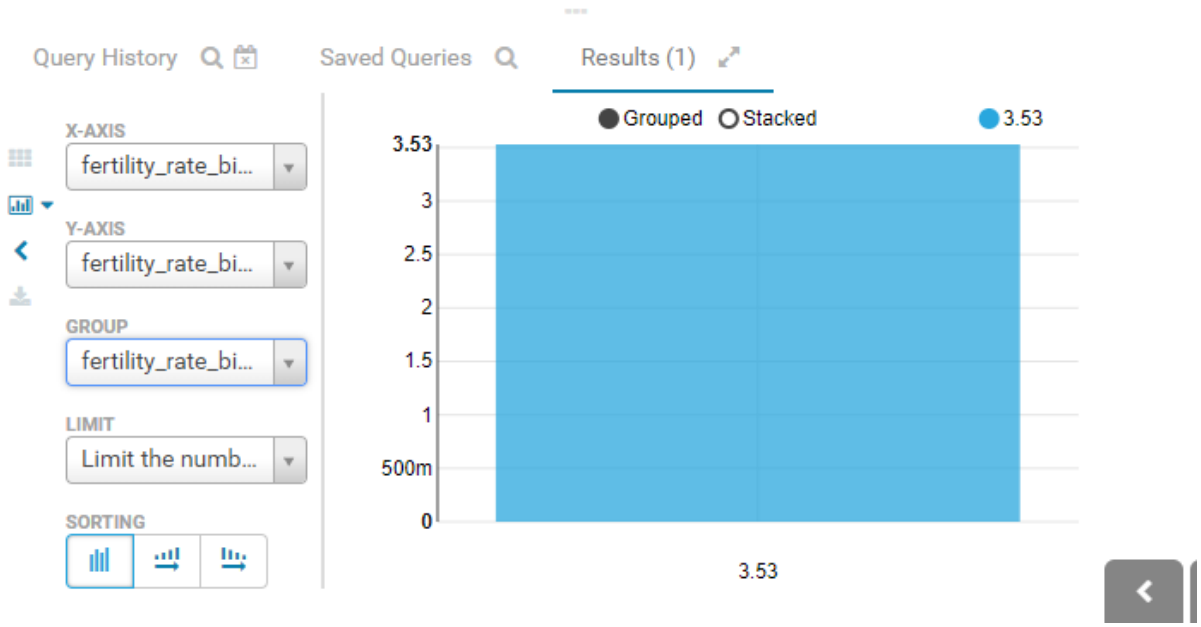
2. The fertility rate of Bihar

<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
```

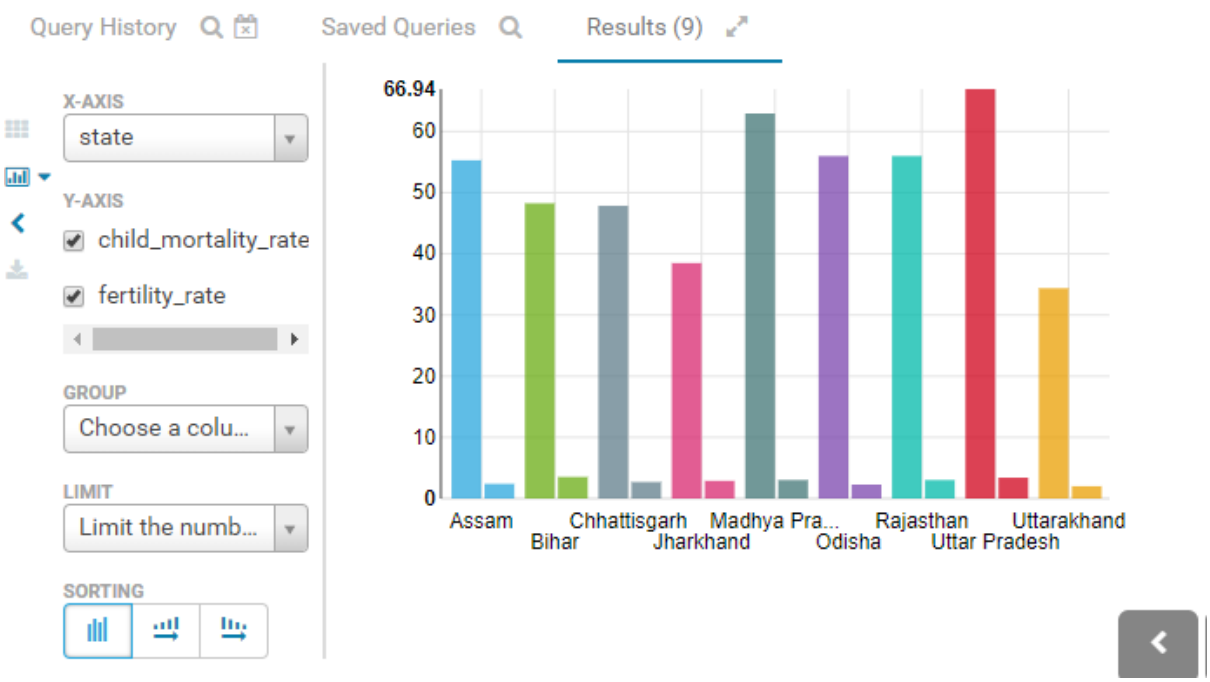
<Screenshot of the result>

Query History | Saved Queries | Results (9)

COLUMNS (4)
- state
- child_mortality_r
- fertility_rate

	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
```

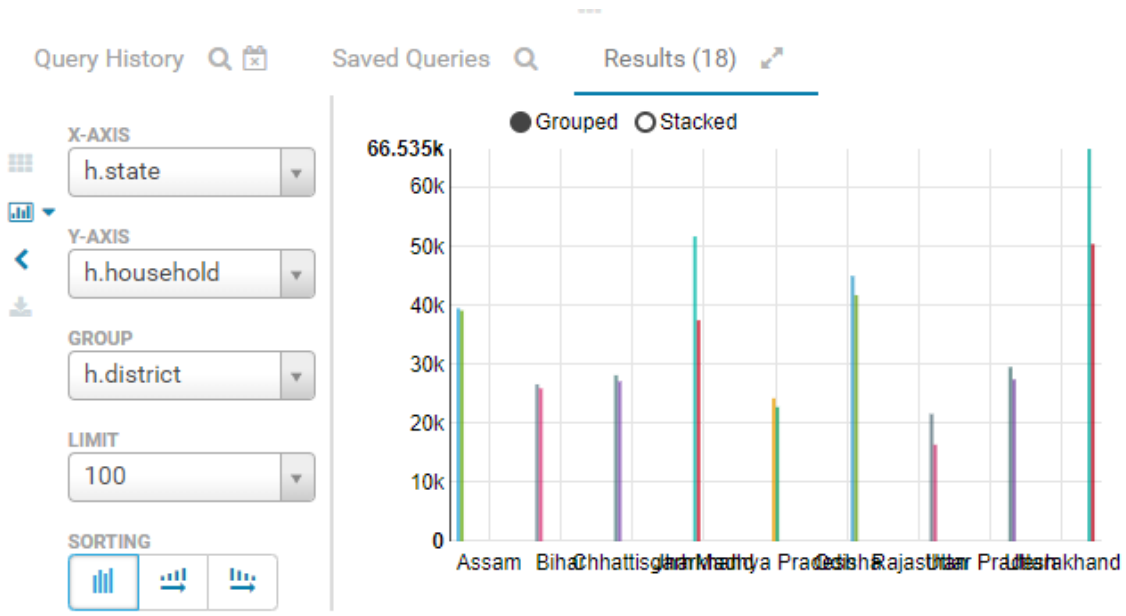
<Screenshot of the result>

COLUMNS (4)

- h.state
- h.district
- h.household

	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>

COLUMNS (4)

- h.state
- h.district
- h.sexratio

	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>

Query History | Saved Queries | Results (18)

X-AXIS: h.state

Y-AXIS: h.sexratio

GROUP: h.district

LIMIT: Limit the num...

SORTING