Data Set

India Annual Health Survey (AHS) 2012-13

The dataset comprises a survey conducted in Empowered Action Group (EAG) states Uttarakhand, Rajasthan, Uttar Pradesh, Bihar, Jharkhand, Odisha, Chhattisgarh & Madhya Pradesh and Assam. These nine states, which account for about 48 percentage of the total population, 59 percentage of Births, 70 percentage of Infant Deaths, 75 percentage of Under 5 Deaths and 62 percentage of Maternal Deaths in the country, are the high focus States in view of their relatively higher fertility and mortality.

A representative sample of about 21 million population and 4.32 million households were covered which is spread across the rural and urban area of these 9 states.

The objective of the AHS is to yield a comprehensive, representative and reliable dataset on core vital indicators including composite ones like Infant Mortality Rate, Maternal Mortality Ratio and Total Fertility Rate along with their covariates (process and outcome indicators) at the district level and map the changes therein on an annual basis. These benchmarks would help in better and holistic understanding and timely monitoring of various determinants on well-being and health of population particularly Reproductive and Child Health. [Source]

Content

This dataset contains the data about the below 26 key indicators.

Problem Statement

Ingest the **India Annual Health Survey (AHS) 2012-13** data hosted on Amazon RDS into Hadoop correctly and process it to generate the following analyses:

Analyses

1. State wise child mortality rate
2. State wise fertility rate
3. Does high fertility correlate with high child mortality?
4. Find top 2 districts per state with the highest population per household
5. Find top 2 districts per state with the lowest sex ratios

Such analyses would help in vivid understanding and timely monitoring of different determinants on well-being and health of population particularly Child and Reproductive Health.

Guidelines

- Ingest data from Amazon RDS to HDFS using Sqoop.
- Create an external table in Hive for the ingested data containing all the columns as given in this document. Ingest the data from HDFS to the Hive table. Verify that the ingestion is successfully accomplished.

file_download**Download**

- Create a subset schema in Hive to store the data for the analyses to be done. The schema should be optimized to support ONLY the analyses to be done. You will be graded on your choice of the chosen columns, storage format (Parquet, RC, ORC, CSV), etc. Benchmark the performance of the storage formats before finalizing the one to be used.

- Write queries against each category of analyses. You will be graded on the relevance of your query to the analytical use case and the optimizations used. Generate the corresponding analyses' charts on Hue.

Note: To access Amazon RDS, refer to the resources section for more details.

Note: The size of the dataset is around 2.5 MB. This is a representative sample and the actual dataset will be of a bigger size. This sample is specifically taken keeping in mind that the engineering process for the data of any size remains the same. Some optimizations might vary as the dataset grows larger. However, while designing the solution, keep optimization in mind and submit a solution that would work even if we increase the size of this dataset.

Upload a PDF document containing the following:

Data Ingestion from the RDS to HDFS using Sqoop

1. Sqoop import command
2. Command to see the list of imported data

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

1. Command to create the external table
2. Command to load the ingested data into the external table
3. Queries to verify that the ingestion is correctly accomplished
    - 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 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

Subset schema creation in Hive to support the analyses

1. Columns used in the subset schema
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';]
3. Create and insert command for the default format
4. Create and insert command for the formats such as ORC
5. Screenshot of runtimes against each query given above for the default format as well as for the formats such as ORC
6. Create and insert command for the partition table for analyses 1 & 2. The partition table should be created using the table created above.

Note: If the default format is giving less time, guess the reasons (in smaller datasets, such anomalies might exist). However, proceed ahead with the format such as ORC as in the practical scenarios formats such as ORC are more efficient.

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

1. State wise child mortality rate
   - Query
   - Screenshot of the result

   - Chart

2. State wise fertility rate
   - Query
   - Screenshot of the result
   - Chart
3. Does high fertility correlate with high child mortality?
   - Query
   - Screenshot of the result
   - Chart
4. Find top 2 districts per state with the highest population per household
   - Query
   - Screenshot of the result
   - Chart
5. Find top 2 districts per state with the lowest sex ratios
   - Query
   - Screenshot of the result
   - Chart