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.

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.

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.