REQUIREMENT ANALYSIS

Problem Statement (What should the application or software do?)

The project for carrying out the following anaysis on a freely available dataset -

- Number of Aadhaar Identities which are generated in each state.
- Number of Aadhaar Identities which are generated by each Enrollment Agency.
- Top 10 districts with maximum Aadhaar identities generated for both Male and Female.

The system must take in data and analyze the data according to the problem statement given above. The result of the analysis helps us understand the way data is distributed over the given parameters.

Description (How should the system behave?)

The system is an implementation of Hadoop. Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

Aadhaar dataset is given as input to the Hadoop system. The system makes use of Hive to run queries on the Dataset. Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives a SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

The output of these Hive queries lets us unsderstand the Dataset better.

Specific Requirements (What are the requirements in term of performance?)

- We need an implantation of Hive over Hadoop along with other essential components like YARN and MapReduce.
- A workable Aadhaar dataset is required.

Hardware Requirements:-

- 4GB RAM (Suggested : 8 GB)
- 20-25 GB of Free Disk Space
- i3 processor or above

Conclusion

- Hive Queries for each problem statement.
- Corresponding MapReduce code (that runs internally whenever the Hive query gets executed) including the Mapper,Reducer and Driver code implemented in Java.
- Analysis Report for each problem statement.
- Project published in the Github respository.