An Analytical Study of Aadhar based Information System via Big Data’s MapReduce Framework

Ashutosh Shankhdhar, Akhilesh Kumar Singh
Department of Computer Science and Engineering, GLA University, Mathura (U.P)
ashutosh.shankhdhar@glau.ac.in, akhilesh.kumar@glau.ac.in

Abstract. Aadhaar is a 12-digit identification number provided to every Indian individual which also includes children and infants. This is used to identify every Indian individual uniquely. This is issued by UIDAI (The unique Identification Authority of India) for free of cost. This unique identification number carries many wonderful advantages. Firstly, it provides a unique identity to every individual whom everybody can trust, no matter in which state they are, in whole country. This unique number can be used to avail different schemes like Pradhan mantri ujjwala yojna and gas subsidy, to open bank accounts and for other government transactions or benefits. This paper shows the use of Aadhaar data set to answer several queries which can be very useful for the government or the concerned authority in various scenarios. Since the Aadhaar data set of all citizens of Indian is a huge data set, big data with its different methods such as Map and Reduce algorithm is very useful in fetching the required query’s answer in a more efficient and faster way.

1. Introduction
Big data Analytics refer to collecting, organizing and analysing the huge amount of data that can be structured, semi-structured or unstructured. It helps in extracting the meaningful and useful information from huge amount of data so that one can decide whether particular data is to keep or not. But it needs powerful analytics like the transfer learning approach when the amount of data is very high and unlabelled [10].

Data Analytics tool perform following activities: Collects and perform the data analytics from various sources, helps to achieve real time analysis of data, Extract meaningful and useful information.

2. Big data Applications
Big data has many significant applications in various fields whether it may be in hospital, in government, in banking etc.

2.1. In Banking Sectors
There is always privacy and security issues in this field. Big data analytics could reveal sensitive personal information [3]. Earlier researches say that data always has privacy issue so bankers are cautions in their use.

2.2. In Government Sector
In this sector, big data analytics is very useful. As government deals at every level, be it gram panchayat or central level, storing this high amount of data requires proper attention. Here big data analytics play very significant role which is further useful for various schemes or events initiated by government.

Like Analysis of Aadhaar data set may be useful in extracting information that can be used while elections, for distributing gas subsidies and various other areas where an estimation of number of people eligible for a scheme is required. Secure transmission of such data needs also be studied effectively [8] [9].

3. Aadhar data analysis
Aadhaar is a valid 12-digit identification number that is unique for every Indian citizen and issued by UIDAI. As Aadhaar data comes in big data category, it uses different methods and technique of big data analytics which can be used to find useful information from Aadhaar data. Aadhaar card contains Demographic and biometric details. Demographic details include name, age, gender, address etc. and biometric details include photograph, fingerprint and iris scan [5]. Aadhaar card can be used to applying for PAN, bank account and driving License etc.

4. Existing System
In traditional system, ration card or voter ID card were used in place of Aadhaar card. But there were several disadvantages of using these cards for unique identification. Families may have more than one ration card. Even people emigrated from neighboring countries like Bangladesh, Myanmar could have ration card. So, this was not alone useful for identification of Indian citizen uniquely. Also, similar case was with voter ID card as it does not contain any biometric verification. Many people of India have more than one voter ID card illegally. This results into poor conduct of elections. Analysis of these data didn’t help in extraction of correct and meaningful data, but after Introduction of Aadhaar card, Ration card and voter ID was linked with Aadhaar. This gave security and unique identity to every citizen of India.

5. Proposed System
Aadhaar gives a unique identification number to every individual of India. So, analyzing the data of Aadhaar may give a better result. By using this method first of all data is going to be preprocessed, so that useful insight is gain. These data can be used for generating reports that can be useful for government agencies. We can get a lot of useful information from the Aadhaar data.

Following information can be analysis from proposed system:
- Counting number of Aadhaar generated in each state
- Counting number of Aadhaar rejected in each state
- Total number of agency and count number of Aadhaar number generated by them
- Counting total number of district and finding top 10 based on maximum number of Aadhaar card produced for Male and Female
- Counting number of voters based on their age mentioned
- Get the male to female ratio based on their gender analysis

6. Implemented algorithm
Map-Reduce is a technique which is used to process huge amount of data in parallel and also a program model for distributed computing based on java [4]. This include mainly two important tasks - map and reduce. Map task is to take the input and convert it into form of tuple which is in the form of key-value pair and reduce task takes mapper output as an input and produce smaller tuple or key-value pair. The main advantage of Map-Reduce is that it helps in scalability by scaling data processing over multiple computing nodes. There are three stages occur during Map- Reduce program execution which are map stage, shuffle stage and reduce stage.
6.1.1. **Map Stage**
At this stage mapper process the input data by the mapper function line by line and then it creates several small chunks of data. Here the dataset from UIDAI is broken into different chunks for parallel processing. For example, to find in each state, total number of Aadhaar generated, the mapper function maps the Aadhaar card generated columns to the state columns and so on.

6.1.2. **Shuffle Stage**
At this stage small chunks of data produced as output by mapper is shuffled and sorted. Shuffling is important to get the data processed efficiently.

6.1.3. **Reduce Stage**
Reducer process the data produced by mapper and after processing it generate new set of outputs. After the mapper function maps the data from the data set, they are shuffled and sorted. The Aadhaar card generated data is most commonly used, so the reducer process it first and then outputs the data based on key value pair.

The Map-Reduce works on key and value pair. Means it takes key-value pair as the input and produce the output in the form of key-value pair. when the job finished then these key-value pair is considered as output of the job.

6.1.4. **Map Reduce working on given dataset**
7. Advantages of Map Reduce

7.1.1. Scalability
As Map-Reduce is capable of processing huge amount of dataset, it helps to run applications from a huge number of servers that could involve the usage of huge amount of data. The UIDAI dataset is a huge data set and therefore it is easier to process the data on map-reduce than any other processing environment.

7.1.2. Fast
Map-Reduce is capable of processing large amount of data. It can process a huge amount of data within some minutes or hour, depending on size of data. It can be in Terabytes or in Petabytes. This helps in faster response to our queries and the model is considered time efficient.
7.1.3. Parallel Processing

Map-Reduce allow their task to execute parallel. It helps in dividing the task to multiple processors. So, this reduces the time to execute the program.

8. Result

The input is taken through a CSV file which contains the data set in the form of comma separated values. There is various information stored like names of enrollment agencies, registrar, states, districts and their pin codes, age of the person, gender, email, mobile numbers etc. After loading a CSV file in the system, hive query generated according to problem statement whether it is to find states who has generated maximum number of Aadhaar or finding various agencies in each state. Then sorting work is done by Map Reduce. Hive query is also useful in querying other problem statements like how many Aadhaar numbers generated by each enrollment agency.

9. Conclusion

This paper describes the Aadhaar based information system, and the analysis of Aadhaar dataset will be carried out to find out the answers to questions related to Aadhaar data analysis and research will be directed towards creating a system that can answer all the queries effectively and efficiently. Map-Reduce Algorithm is used to process the input data and after shuffling by mapper, Reducer generate new set of output. The findings of Aadhaar data analysis are expected to have very significant effect over existing system in Indian context across over different geographies, different political environment. It also suggests the way by which fraud in elections can be reduced. As Aadhaar provide a unique identity to each and every citizen of India, so one person cannot have two identities. It is also useful in providing different schemes initiated by Indian government to needy people. It was reported many cases of fraud in government scheme using earlier system. The analysis of Aadhaar gives many advantages. It has the power to reduce corruption and fraud detection. It is also very useful during election by identifying the total number of fake voters and reduce fake votes. But there are certain issues related to privacy of Indian citizen. It is claimed that government is failed to protect the private data of their citizen. Under Aadhaar, theoretically it is very easy to build the profile of Indian citizen by aggregating their daily life data. It seems bid risk to privacy of Indian citizen.

There is no doubt in saying that Aadhaar system is much effective and useful in many areas but it not confirms the privacy of Indian citizen. So, the further study will be based on security of Aadhaar related data. How one can provide safety to these data. Also, to ensure that no one person has two identities generated in India. We can also extend this work with the technique of Artificial Intelligence. An automatic query generator system can also be developed.

References

[1] Kaushik, M. (2010). Configuring the UID. Business Today, 19(22), 12.
[2] Beyer, M. A., & Laney, D. (2012). The importance of ‘big data’: a definition. Stamford, CT: Gartner, 2014-2018.
[3] Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. interactions, 19(3), 50-59.
[4] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 83-124.
[5] Shukla, R. (2013). Technologies in Aadhaar-A sociotechnical view. the wake of Aadhaar, 372-399.
[6] UIDAI, Unique identification authority of India, Planning Commission, Government of India.
(2010). Aadhaar- communicating to a billion—An awareness and communication Report.

[7] Varma, P. K. (2010). Aadhaar: Scalability and data management challenges.

[8] Sen, B. (2020). Information and the Indian State: A Genealogy. South Asia Multidisciplinary Academic Journal, (23).

[9] Singh, P. (2020). Aadhaar: Platform over Troubled Waters. In Platform Capitalism in India (pp. 201-219). Palgrave Macmillan, Cham.