IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 01-04

Aadhar Based Data Migration, Analysis and Performance using Big Data Analytics and Data Science

Mrs.Lakshmi Piriya.S[1], T. Sri Nithi[2]
[1](Research Scholar / Assistant Professor, KG College of Arts and Science, Bharathiar University.)
(laks.research@gmail.com)
[2](Student, KG College of Arts and Science, Bharathiar University.)
(srinithi.tamilselvan2001@gmail.com)

Abstract: *The current Aadhar based centralized database, face an extensive denunciation because of their intrinsic limitations in securing privacy. Central Identities Data Repository (CIDR), an Indian government agency is responsible for storing and managing the entire country's data related to Aadhar project. Additionally, Unique Identification Authority of India (UIDAI) verifies the authenticity of every Indian citizen. In recent surveys, it is observed that CIDR is facing lot of privacy issues bacause of it's centralized database and plug leakages. Owing to the increasing prevelance of Aadhar and the Indian government's ram towards cashless transactions and digital economy, it is essential to work on data migration, analysis and performance using Big Data Analytics and Data Science.*

Keywords: *Aadhar, Big Data Analytics, Central Identities Data Repository, plug leakages, Unique Identification Authority of India.*

I. INTRODUCTION

All Indian residents are issued a 12-digit unique-identity Aadhar number based on their biometric and demographic data. Fig 1. For this purpose, the Government of India, established a statuory authority on 12 July 2016 known as Unique Identification Authority of India (UIDAI), under the Ministry of Electronics and Information Technology, under the provisions of the Aadhaar Act 2016. Fig 2. The duty of UIDAI is to collect the biometric and demographic data for all Indian citizens. All these details are stored an an Aadhaar database on a central server known as Central Identity Data Repository(CIDR). Enrolment of the residents is computerised, and information is exchanged between Registrars and the CIDR over a network. [2]

In a recent survey, it was realized that the centralised databases have suffered a widespread criticism because of their inherent limitations, in keeping the data secure. In the current situation of Aadhar, CIDR is facing problems not only because it is a centralized database, but also many plug leakages were reported. This is an era of Big Data with the growing pervasiveness of Aadhaar. Also, the government's objective towards a cashless and digital economy has led to a recurrence of interest in data migration, analysis and performance using Big Data Analytics and Data Science in India. [3]

Big Data Analytics assents a valuable insight for cost cutting, efficient operation and discover innovative ways to increase profits in a business or public sector. The idea of Big data can prevent costly problems by predicting the customer behaviors and requirements. The usual approach of other methodologies will be reacting to the problems which incurr huge cost, but Big Data Analytics predicts the problem and therefore eliminates the cost incurred in solving the problem after it occurrs. Applying Big Data analytics can thus increase revenue both in business and public sectors.

The knowledge can be extracted from any data by working out mathematical modeling and statistical modeling. This methodology when integrated with Decision-making, it is termed as Data Science. With an objective of increasing the value of data assets, the Data Science yields a comprehensive overview of extracting, exploring, analyzing, modeling and interpreting data.

The main objective of this paper is to perform Data Analytics and Data Migration on Aadhar data. Also, data security will be enhanced to limit the plug-leakages. Finally, data storage system is caliberated to increase the computing efficiency.

II. DATA MIGRATION

The concept of transferring all Aadhar related biometric and demographic data from one system to another is Data Migration. The Data Migration of Aadhar data can be best done using the tools of ETL (Extract-Transform-Load) process by changing the storage. It is the process of transferring data from source database to the destination data warehouse. There are three different sub-processes like E for Extract, T for Transform and L for Load. In the extraction process, the data is extracted from the source database. In the transformation process, the extracted data is transformed into the required format and then in the loading process, the transformed data is loaded to the destination data warehouse. All these functions are performed using certain tools which are called the ETL tools. [1] Fig 3.

For Data Warehousing tasks, the following steps describes the ways to select ETL Tools and their significance. [8] The first step is to identify the most compatible tool for an enterprise. This can be done using a data integration survey using Google. [6] The proper selection of tools leads to a successful ETL functioning. Also, the selection of proper tools enables a better data transfer between databases. If the functionality is performed with less appropriate tools, then there will be an issue in the functioning of the complete data transfer process. Even for every sub process, the proper selection of tools is significant. Hence, carefulness must be maintained by a proper selection of the ETL tools. [4]

III. DATA ANALYTICS

The Data Analytics is the concept of examining a massive amount of data to research hidden patterns, correlations and other insights. The traditional Data Analysis system is slower and less efficient. With Big Data Analytics, it's comparatively easy to analyze the data and get results from it within very short span of time. Since the Aadhar data is completely unstructured and semi-structured, such data types cannot go good with our structural relational database. In the current context of stock trading, mobile applications and web tracking, our traditional data warehouses lack to handle the data that needs continuous and repeated updations. This issue is better solved using Big Data Analytics – Hadoop and NoSQL databases with the support of their companion tools like YARN, MapReduce, Spark, Hbase, Hive. Kafka and Pig. [10]

The Aadhar data can be examined straight away in a Hadoop cluster. As an alternative they can run through a processing engine like Spark. As a prerequisite, the data before stored in the Hadoop Distributed File System ought to be organized, configured and partitioned correctly to get better results on performance on both extract, transform and load (ETL) integration jobs and analytical queries. [11]

IV. ENHANCED DATA SECURITY AND CALIBERATION OF DATA STORAGE SYSTEM

To enhance the Data Security of Aadhar data, an Unified Interface is proposed to overcome the existing plug leakages. A new metadata-oriented client interface, to consistently provide better results in enhancing security is known as an Unified Interface. [7] This interface is especially designed in such a way that it is consistent, regular user interface, global access and optimum visibility. The form experiences for updating and viewing the records are the same. All devices in this approach have interactive dashboards to view the information and to analyze them. Also, Unified Interface supports reference panel and RTL (right-to-left) languages. The improved accessibility option still enhances the security. Hence, this Unified Interface can be proposed for Aadhar system to enhance their security and to minimize the plug-in-leakages. [5]

Caliberation of Data Storage System is an infrastructure that spreads storage and compute power over many nodes, to deliver near-instantaneous results to complex queries. Also, to maintain accuracy, Real-time data analytics needs a right sensor. Poor calibration leads to the damage of hardware and general infrastructure. Intelligent Sensor Management (ISM) is a digital technology platform for processing data. This technology uses the power of on-board microprocessors to provide worry-free measurement points and maximum process confidence. [9]

V. FIGURES

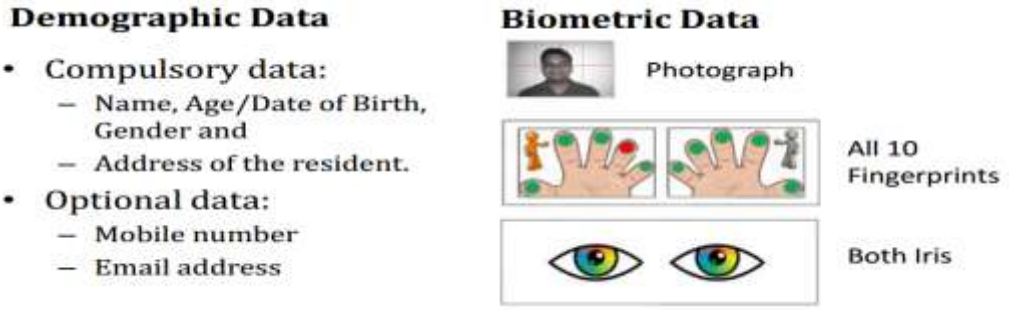

Fig 1. 12-digit Aadhar Number, Unique, lifetime, biometric based identity

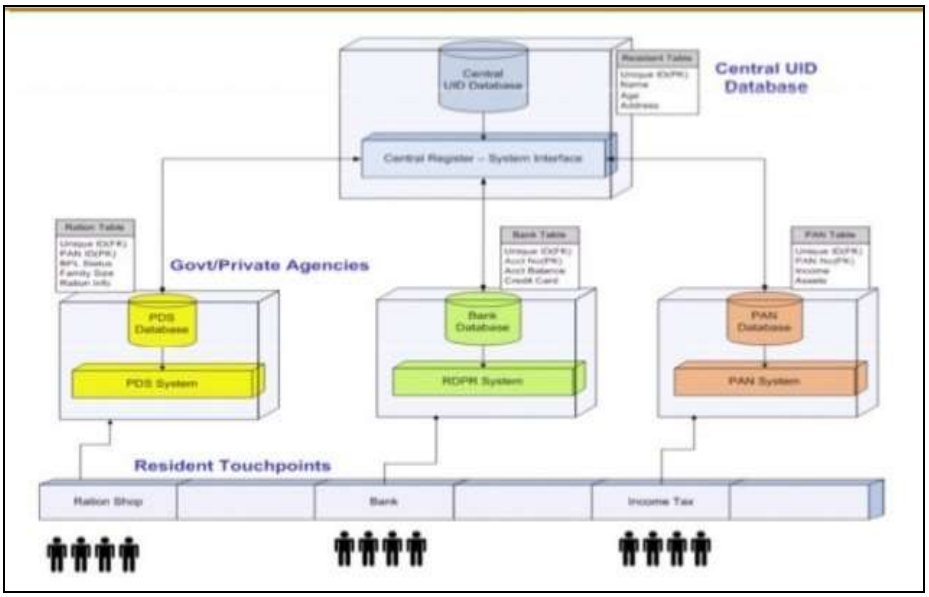

Fig 2. UIDAI Architecture

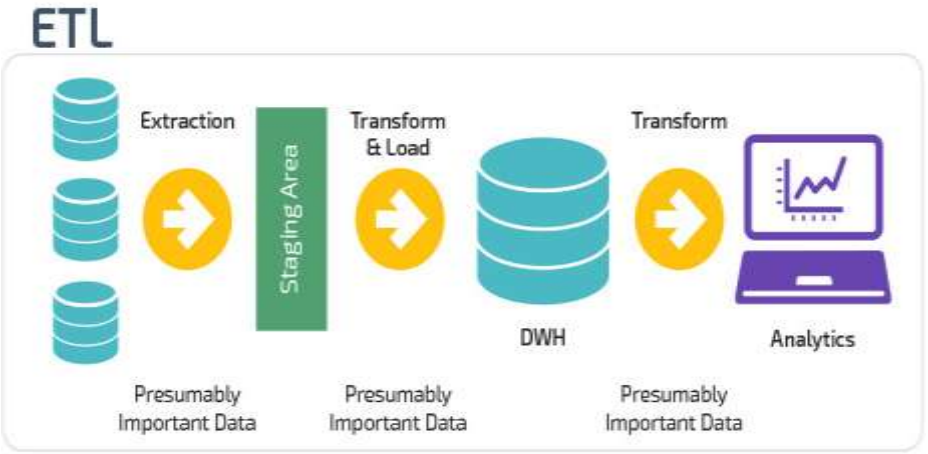

Fig 3. Extract, Transform, Load – an automatic process in Data Migration

VI. CONCLUSION

The Aadhar based Data Migration, Analysis and Performance using Big Data Analysis and Data Science is expected to give more benefits than the current system. Retrieval of almost all the details from various sectors like RTO, Income Tax, Election Commission, Civil Supply and Mediclaims are possible. Fraud can be detected the moment it happens, and proper measures can be taken to limit the damage. This architecture

also paves way im minimizing the plug leakages. Early warning systems for medical emergencies, natural disasters, occupation-based alerts through SMS can also be activated. The main advantage of implementing Big Data and Data Science is cost savings. The Data Caliberation can also be done using HDFS, Data Migration by using Hadoop tool, Data Cleansing through Data Wrangler, Data Analytics by R Programming. The Data Security can also be enhanced by using encryption scheme with a Unified Interface. Finally, the architecture testing is performed in the Hadoop Environment and Performance of the system is caliberated using Map Reduce tools. Thus the advent of Big Data and their associated tools makes the Aadhar system more efficient and error free.

REFERENCES

[1] B. Dufrasne, A. Warmuth, J. Appel, et al. *Introducing disk data migration. DS8870 Data Migration Techniques. IBM Redbooks. pp. 1–16.* ISBN 9780738440606. (2017).

[2] B. Goes, Paulo, Design science research in top information systems journals. *MIS Quarterly: Management Information Systems.* 38 (1) (2014).

[3] Boyd, dana; Crawford, Kate, Six Provocations for Big Data. *Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.* doi:10.2139/ssrn.1926431. (21 September 2011).

[4] C. Seiwert, P. Klee, L. Marinez, L, et al. Migration techniques and processes. *Data Migration to IBM Disk Storage Systems. IBM Redbooks. pp. 7–30.* ISBN 9780738436289. (2012).

[5] Hellerstein, Joe, *Parallel Programming in the Age of Big Data.* Gigaom Blog. (9 November 2008).

[6] J. Morris, Data Migration: What's All the Fuss? *Practical Data Migration (2nd ed.). BCS Learning & Development Ltd. pp. 7–15.* ISBN 9781906124847. (2012).

[7] Laney, Doug, 3D data management: Controlling data volume, velocity and variety. *META Group Research Note.* 6 (70). (2001).

[8] MJ. Denney, Validating the extract, transform, load process used to populate a large clinical research database. *International Journal of Medical Informatics. 94: 271–4.* doi: 10.1016/j.ijmedinf.2016.07.009. PMC 5556907. PMID 27506144. (2016).

[9] O.J Reichman, Jones, M.B. Schildhauer, *M.P Challenges and Opportunities of Open Data in Ecology Science. 331* (6018): 703–5. Bibcode:2011Sci... 331..703R. doi:10.1126/science.1197962. PMID 21311007. (2011).

[10] Ralph., Kimball, *the data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data.* Caserta, Joe, 1965-. Indianapolis, IN: Wiley. ISBN 978-0764579233. OCLC 57301227. (2004).

[11] Segaran, Toby, Hammerbacher, Jeff Beautiful Data: *The Stories Behind Elegant Data Solutions. O'Reilly Media.* p. 257. ISBN 978-0-596-15711-1. (2009).