A PROJECT REPORT

on

"Pehchaan: Analysis of the 'Aadhar Dataset' to Facilitate a Smooth and Efficient Conduct of the Upcoming NPR"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE AND ENGINEERING

BY

Harshit Anand	1705238
Nishan Acharya	1705247
Soumyadev Mukherjee	1705274
Subham Char	1705277
Pritam Ghosh	1705427

UNDER THE GUIDANCE OF
Prof. Dr. Minakhi Rout

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

May 2020

A PROJECT REPORT

on

"Pehchaan: Analysis of the 'Aadhar Dataset' to Facilitate a Smooth and Efficient Conduct of the Upcoming NPR"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE AND ENGINEERING

BY

Harshit Anand	1705238
Nishan Acharya	1705247
Soumyadev Mukherjee	1705274
Subham Char	1705277
Pritam Ghosh	1705427

UNDER THE GUIDANCE OF
Prof. Dr. Minakhi Rout

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

May 2020

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"Pehchaan: Analysis of the 'Aadhar Dataset' to Facilitate a Smooth and Efficient Conduct of the Upcoming NPR"

submitted by

Harshit Anand	1705238
Nishan Acharya	1705247
Soumyadev Mukherjee	1705274
Subham Char	1705277
Pritam Ghosh	1705427

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2019-2020, under our guidance.

Date: / /

(Prof. Dr. Minakhi Rout)
Project Guide

Acknowledgement

We are profoundly grateful to Prof. Dr. Minakhi Rout for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. It has been a cherishing journey working under her, the constant motivation and enthusiasm with which she has ignited our minds has inspired us to bring the best out of us.

Harshit Anand
Nishan Acharya
Soumyadev Mukherjee
Subham Char
Pritam Ghosh

ABSTRACT

The Government of India has sanctioned Rs. 3,941.35 crore for maintaining the National Population Register. The 'usual residents' of a nation are reflected in the NPR. Any individual who has stayed in an area for the past six months or plans to stay in an area for the next six months is referred to as a 'usual resident'. 'Aadhar' is an authentic identity number comprising of 12 digits that can be issued at will by people who reside in the nation or individuals who hold passports of India, subjective to their demographic and biometric information.

Analyzing the 'Aadhar Dataset' and drawing meaningful insights out of the same will surely ensure a fruitful result and facilitate a smoother conduct of the upcoming NPR. The sole objective of 'Hadoop' in this research is storing and processing huge amount of semi structured data. Hence, our proposed work uses 'Hadoop' for processing the data gathered. The input data is processed using MapReduce and finally the result is loaded into the Hadoop Distributed File System (HDFS).

Keywords: National Population Register, Aadhar, Identity Crisis, UIDAI, Big Data, Hadoop, MapReduce, HDFS, Tableau

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 A brief history and motivation behind the work

The Indian Government will shortly be coming up with a plan to maintain and update the 'National Population Register' nationwide. A sudden outbreak of COVID-19 has brought the entire procedure to a halt. Let us have a sight at what this National Population Register is all about.A list of usual residents of the nation is what the NPR comprises of. The same is being maintained right from the grassroots level including *panchayats* and suburbs to the state and national levels subject to the Citizenship Act of 1955 as well as the 2003 issued Citizenship Rules. It is the compulsion of every individual residing in India to be a part of the NPR. Any individual who has stayed in an area for the past six months or plans to stay in an area for the next six months is referred to as a 'usual resident'

India witnessed a lot of protest against the 'Citizenship Amendment Act'. 'Identity Crisis' is expected to be an obvious consequence of the same when blended with the 'National Register of Citizens', people believe. The protests involving the CAA has compelled the state governments to to pause the process of gathering data to facilitate the NPR for the time being. A particular section of the society might fall prey to this is what the authorities fear. NPR is not merely a census exercise, there are several other reasons solid enough to worry about the same, it might result in the government putting its residents under custody and monitor their activities strictly, this is turn is a major threat to the constitutional rights and the secular image of the nation. The 'Aadhar' getting linked up with NPR plays a crucial role here.

We do all fear that someday our identity will be at stake. It is the desire of every individual being a part of this nation to call himself/herself a citizen of this country. 'Aadhar' has been issued to most of the folks with certain exceptions and anomalies which if not taken care of will put forth a huge problem in front of the Government. Soon, The Registrar General of the country is expected to request the Unique Identification Authority of India (UIDAI) to check credentials of people for making them a part of the recently planned National Population Register (NPR) process, fresh collection of biometrics would otherwise be a tedious task. However, there are many who haven't yet verified their 'Aadhar' credentials despite of being a part of this nation for a very long period of time. Don't you think they are citizens too? Don't they deserve the right to be referred to as Indians? They do.

1.1.2 Problems that may be encountered

Lost Aadhar: Many a times individuals lose their Aadhar cards. For instance, Raghu, a fisherman residing near Puri, lost his Aadhar card during 'Fani' and did not reissue the same.

Irrelevant and fake Aadhar Numbers: Possessing a fake identity is an offence, however, the story of duplicate Aadhar numbers being circulated is no more an unsaid tale. It is well known that many individuals possess more than one cards and pose a major threat to the proper governance.

Unrecognized immigrants: Aadhar cards have not been issued in bordering states like Assam as many immigrants have forcefully entered the state from Bangladesh. 'Rohingyas' coming from Myanmar have come across similar problems too.

Linking Issues: Many innocent people have missed the deadline. Ahalya, a cancer patient from Cuttack could not be a part of the process as she had no one to support her and take her for the formalities.

Updation of biometrics: The biometric information of an individual changes gradually over a span of time and has to be updated to avoid anomalies. However, a huge number of incorrect credentials pose an immense problem to the progress of the nation.

1.1.3 The following steps will resolve

1. Spot out the age group more prone to the anomalies associated with the Aadhar Database and resolve the same.

2. Figure out the count of Aadhar cards accepted and directly facilitate the NPR.
3. Figure out the count of Aadhar cards not accepted across states, districts and blocks and the reason behind the same.
4. Verify biometric and other personal credentials and update the same if not,
5. Check if the mobile number is linked with the Aadhar or not.
6. Resolve all peculiarities for valid citizens by passing on the results to UIDAI.
7. Spread a social message across the nation that there is no need to panic, proper analysis of data resulting in meaningful insights will protect your identity and help the Government in resolving conflicts.

Chapter 2

Literature Survey

In the past few years, a lot of research works as well as projects have been carried out in--UIDAI domain of updation of the National Population Register using various Big Data Analytics and algorithms. A thorough study and observation on the works of few of such papers is done under this section.

2.1 An Anatomization of Aadhaar Card Data Set-A Big Data Challenge.

Mohit Dayal et al. (2016) proposed an analysis of UIDAI Data Set against various queries using Hadoop Cluster and Pig Latin. They have tried to find the solutions of few of the research queries based on the total number of cards accepted or removed. But the work is accompanied with a few limitations like using fully distributed Hadoop cluster mode. Moreover, Apache Pig doesn't have an explicit Data Schema.

2.2 Big Data Analytics on Aadhaar Card Dataset in Hadoop Ecosystem.

D. Durga Bhavani et al. (2019) has suggested a model to infer fluctuations in enrolment due to the effect of demonetization and PAN linking in UIDAI dataset using Hadoop and Hive. The respected work is acknowledged with few flaws like the row level updates and real time queries aren't much likely to be handled. Moreover, the author has ignored the fact of high Hive latency along with the issue that the model isn't likely to support sub-queries.

2.3 Analysis of Aadhaar Card Dataset Using Big Data Analytics.

R. Jayashree (2020) exclaimed a prototype for the retrieval of blood donor details and crime investigation using Sqoop (SQL + Hadoop) and Hive. In our point of thought, we observed an inefficient data transfer from RDBMS to Hadoop along with few issues in carrying out the sub-queries and inline queries. Moreover, the work is highly market based.

2.4 Big Data Applications in Aadhar Card Fraud Detection.

K. Ramya et al. (2019) has implemented data mining techniques in classification of Naive Bayesian (NB), c4.5 and Back Propagation (BP) to identify patterns leading to fraud. But, using Naive Bayes may lead to the overfitting of the model, in addition to that there is also a probability of Network Paralysis while using BP.

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

Mrs.Lakshmi Piriya.S et al. (2018) has put forward a proposal to use HDFS in data calibration, data wrangler in data cleansing, R in data analytics and MapReduce tools in architecture testing. The ideas presented seems pleasing but there is a lack of practical implementation or an elucidated approach to the solution of the problem. It is based on statistical summaries and data quality visuals rather than a practical approach.

Chapter 3

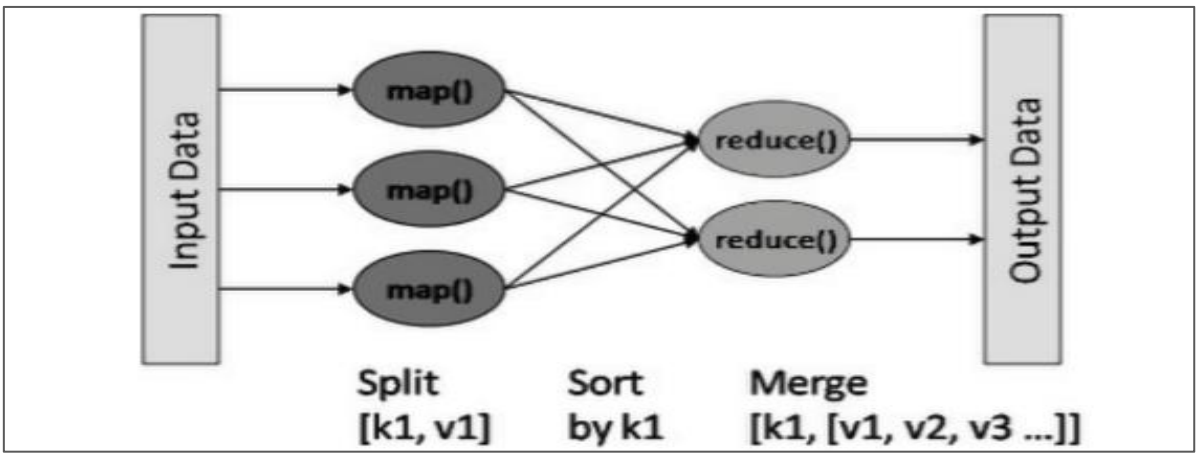
Software Requirements Specification

Hadoop 2.7.7

Hadoop is an open-source platform based on Apache which facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It is written in Java and a collection of other open-source software utilities. It provides a software framework for distributed storage and processing of big data and uses Google's MapReduce and Google File System as its foundation. Inherent data protection, storage flexibility, low cost, scalability and high computing power are few of the key elements which made us use Hadoop as a tool to solve our problem statement.

MapReduce (MRv2)

MapReduce is a programming model or a software framework used in Apache Hadoop for writing applications which process and analyze large data sets in parallel on large multi-node clusters of commodity hardware in a scalable, reliable and fault tolerant manner. Data analysis and processing uses two different steps namely, Map Phase and Reduce Phase.



Hadoop Distributed File System (HDFS)

HDFS is a distributed file system that provides reliable, scalable and fault tolerant data storage on commodity hardware. It works closely with MapReduce by distributing storage and computation across large clusters by combining storage resources that can scale depending upon requests and queries while remaining inexpensive and in budget. HDFS accepts data in any format like text, images, videos etc. regardless of architecture and automatically optimizes for high bandwidth streaming. HDFS exploits master/slave architecture with NameNode daemon and secondary NameNode running on master node and DataNode daemon running on every single salve node.

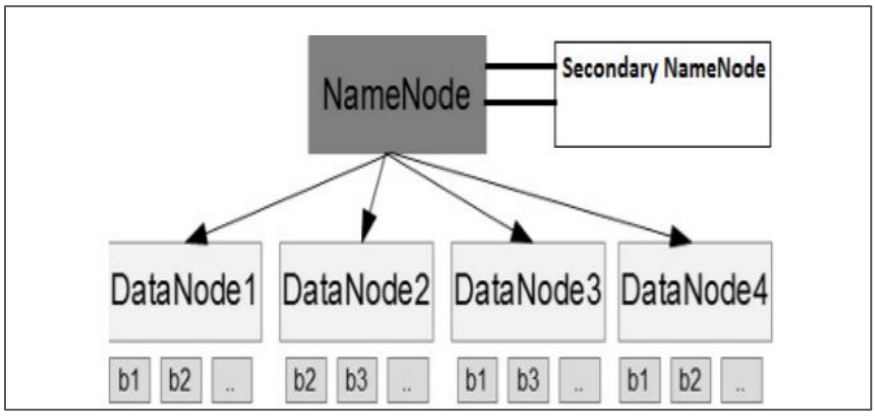


Tableau 2020.1

Tableau software is a platform that helps make Big Data small, and small information quick and noteworthy along with the extraction of insights and catching of hidden trends. The primary utilization of Tableau is to assist individuals with observing and comprehending data. It brings multiple data sources into one single point of truth and performs basic ETL operations quickly. Exploratory Data Analysis is easier in Tableau as compared to other tools. Moreover, automate reporting, interactive user interface and dashboards, freely switching between local computer and the cloud are few of the key features because of which it is being used in the project.

Chapter 4

Requirement Analysis

Along with the software requirements as specified in the Software Requirement Specifications (SRS) section, below are the hardware prerequisites we needed for the completion of our project. The major two factors we were concerned about were-

- Portability
- Processing Power

The higher the processing power, the heavier and complex the system gets, making it less portable.

Virtual Machines (VM Ware Workstation)

A virtual machine is a software simulation of own or another machine hardware configuration that can run an operating system. It acts as a layer between the virtual OS and system OS which then communicates with the lower level hardware to provide support to the virtual OS. Few of the features of VM are:-

- *Isolation:* When we come across a scenario when we are running many applications simultaneously, it is required to be extra cautious which gives the need of isolation of different working codes or applications from one another so as to avoid undesirable interactions or outright conflicts.
- *Standardization:* A standardized platform helps in cutting costs and ensure a proper distribution of resources that can be utilized in getting an upper hand in finding the solution to the problem statement.
- *Ease of testing:* VMs provides us with snapshots and rollback capabilities. Along with that, they have options for saving each sessions and versions. This is very handy in scenarios of debugging and system as well as unit testing.
- Mobility: VMs can be moved to different physical machines with least issues. They store a disk in the guest environment in the host environment. Transferring a VM to another physical machine is same as moving a virtual disk file and some configuration files to another system.

RAM

The minimum RAM that was required for the Big Data Analytics was 8 GB. Although, 16 GB is recommended for faster processing, but we lacked a few of the hardware specifications and had to shift some of the tasks to ‘Future Work’.

GPU & Processor

Although for GPU, an NVIDIA or any other graphics card was required but we tried to keep an 8-core AMD for the completion of the task. It is tried to achieve an optimum and efficient approach for the solution of the problem statement with whatsoever available resources we had. Also, the 64-bit Intel i5 7th Gen processor is being used.

Operating System

CentOS which is a Linux distribution is chosen over Windows for the completion of the project as Hadoop Services are running at the top of Linux Operating System like IBM Infosphere Biginsights (IBM Hadoop) is built at the top of SUSE Linux OS and Cloudera Hadoop Distribution is running at the top of CentOS.

Chapter 5

System Design

5.1 Use Case, Entity Relationship and Class Diagrams

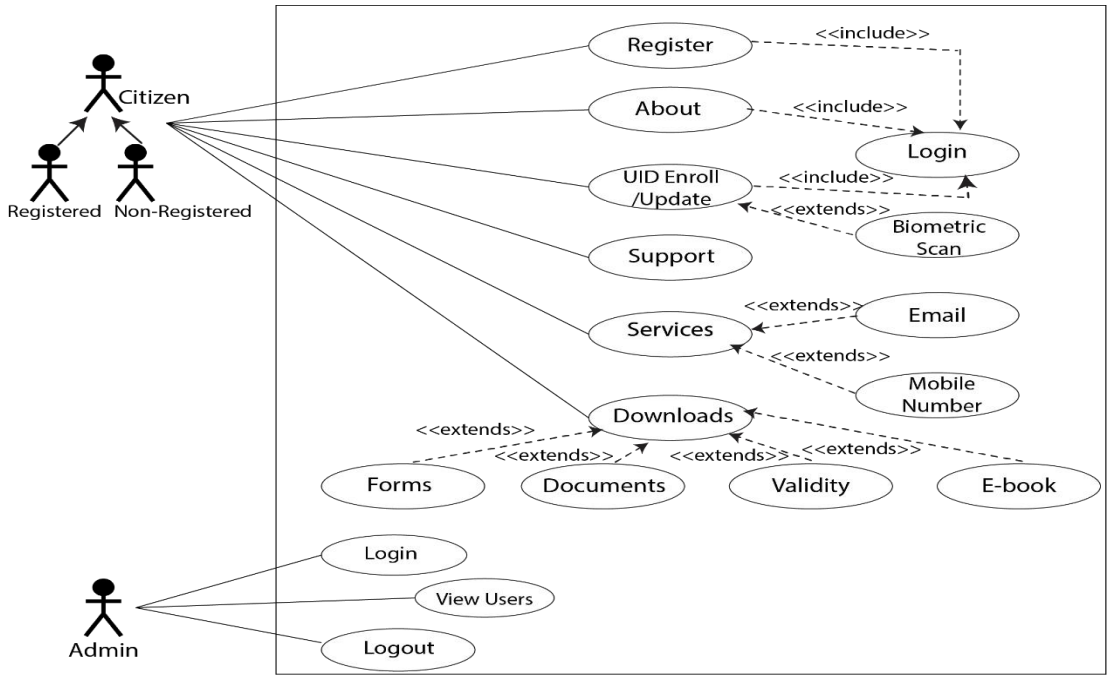


Fig. 9. Use Case Diagram

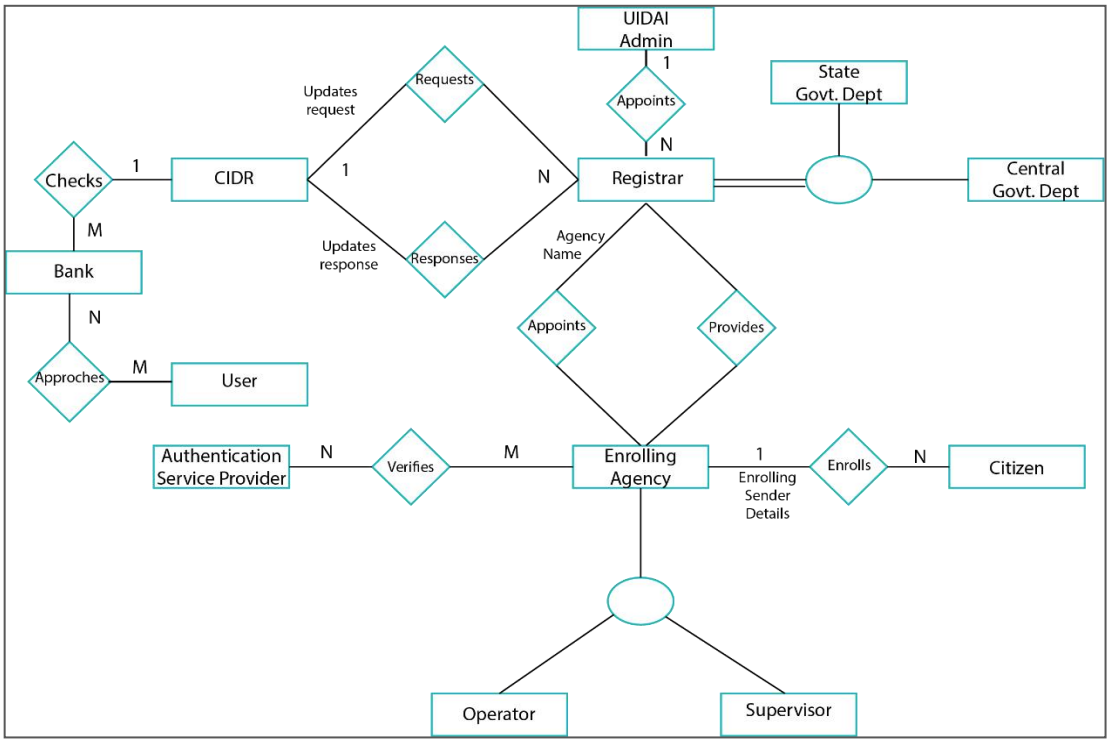


Fig. 10. Entity Relationship Diagram

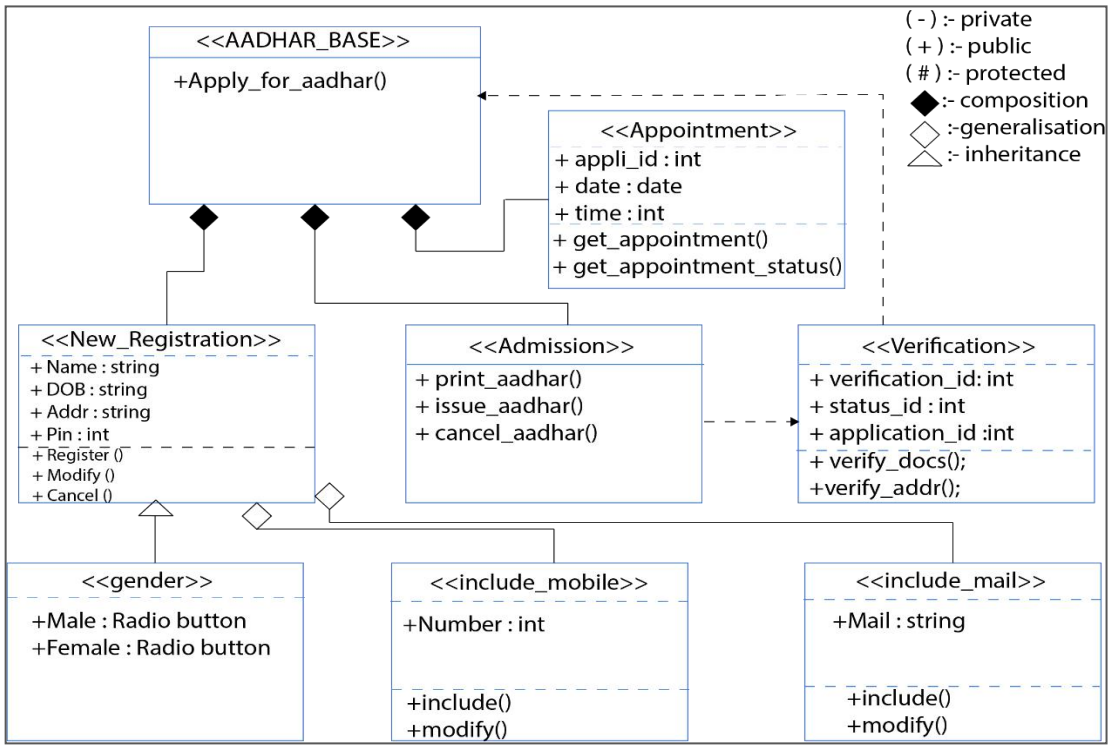


Fig. 11. Class Diagram

Chapter 6

System Testing

Testing is the process of exercising a program with specific intent of finding errors prior to delivery to the end user. It is the process of analyzing a software item to detect the differences between existing and required conditions and to evaluate the features of the software.

6.1 Test Cases and Test Results

Sl. No.	Test Case Title	Test Condition	System Behaviour	Expected Result	Test Case Type
01	Entry-correct-Aadhar-number	Number of digits of Aadhar number must be equals to 12 .	Accepts the Aadhar number and generates steps to proceed further .	Accepts the Aadhar number and generates steps to proceed further .	Positive test case
02	Entry-correct-Aadhar-number	Number of digits of Aadhar number must be equals to 12 .	Displays error as incorrect Aadhar number	Accepts the Aadhar number and generates steps to proceed further .	Negative test case
03	Entry-incorrect-Aadhar-number	Number of digits of Aadhar number must be equals to 12 .	Displays error as incorrect Aadhar number	Displays error as incorrect Aadhar number	Positive test case
04	Entry-incorrect-Aadhar-number	Number of digits of Aadhar number must be equals to 12 .	Accepts the Aadhar number and generates steps to proceed further .	Displays error as incorrect Aadhar number	Negative test case
05	Update-correct-mobile-number	Mobile number must be of 10 digits	Mobile number gets updated	Mobile number gets updated	Psitive test case
06	Update-correct-email-address	Entered email address must match email pattern	User email gets updated	User email gets updated	Positive test case
07	Update-correct-email-address	Entered email address must match email pattern	Displays error as Enter correct email	User email gets updated	Negative test case

Chapter 7

Project Planning

Development Step	Duration	Start	Finish
1 Business Case	**10 days**	**26/12/2019**	**4/01/2020**
1.1 High Level Requirements	5 days	26/12/2019	30/12/2019
1.2 Estimation	2 days	31/12/2019	1/01/2020
1.3 Draft Business Case	3 days	2/01/2020	4/01/2020
2 Analysis	**15 days**	**5/01/2020**	**19/01/2020**
2.1 Requirement Gathering	5 days	5/01/2020	9/01/2020
2.2 Draft Requirements Document	8 days	10/01/2020	17/01/2020
2.3 Requirements Review	1 day	18/01/2020	18/01/2020
2.4 Requirements Sign-off	1 day	19/01/2020	19/01/2020
3 Design	**13 days**	**20/01/2020**	**3/02/2020**
3.1 High Level Design Document	10 days	20/01/2020	31/01/2020
3.2 Design Review	3 days	1/02/2020	3/02/2020
4 Build	**46 days**	**4/02/2020**	**22/03/2020**
4.1 Component 1	**18 days**	**4/02/2020**	**21/02/2020**
4.1.1 Sub Component	12 days	4/02/2020	15/02/2020
4.1.2 Unit Testing	6 days	16/02/2020	21/02/2020
4.1 Component 2	**23 days**	**22/02/2020**	**16/03/2020**
4.1.1 Sub Component	16 days	22/02/2020	9/03/2020
4.1.2 Unit Testing	7 days	10/03/2020	16/03/2020
4.3 Integration Testing	4 days	18/03/2020	21/03/2020
4.4 Configure QA Environment	1 day	22/03/2020	22/03/2020
5 Quality Assurance	**11 days**	**25/03/2020**	**5/04/2020**
5.1 Draft Test Case	3 days	25/03/2020	27/03/2020
5.2 Review Test Case	1 day	28/03/2020	28/03/2020
5.3 Sign-off Test Case	2 days	29/03/2020	30/03/2020
5.4 Test Execution	5 days	1/04/2020	5/04/2020

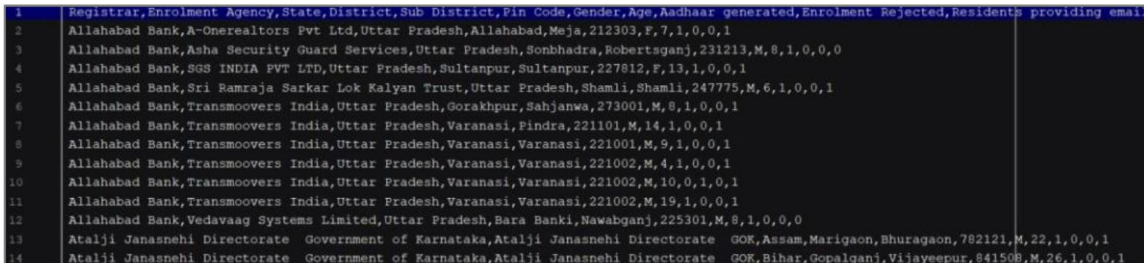
Chapter 8

Implementation

We used MapReduce jobs to break and divide the input data into chunks and work on it.
In our project we used 5 MapReduce key functions to get the desired output. The functionalities of each functions are as follows:-

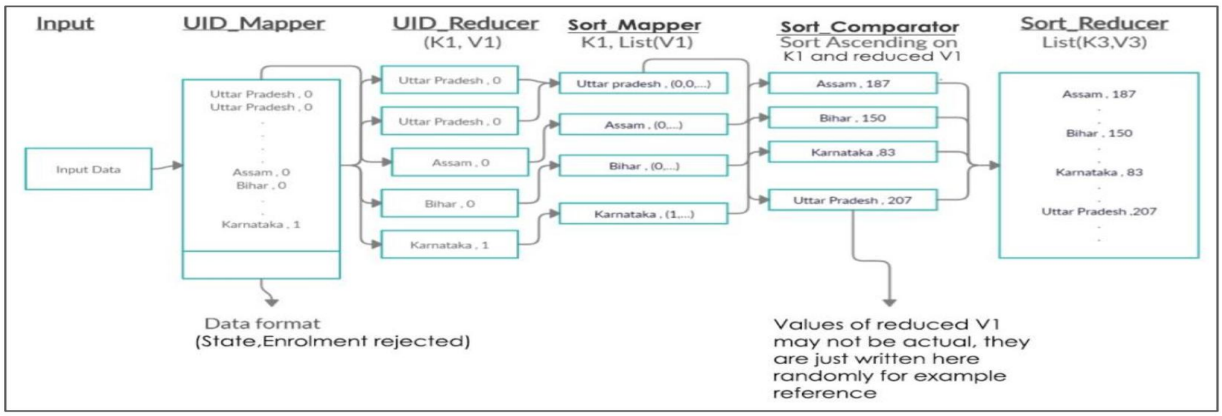
- UID_Mapper.java: Filters the header and writes to mapper output.
- UID_Reducer.java: Aggregate values for each state and outputs state wise identity.
- Sort_Mapper.java: Receive output from previous UID_Mapper – Reducer phase
- Sort_Comparator.java: Sorts and reduces the output in descending order.
- Sort_Reducer.java: Swap (x,y) pair into (state,count) format and produce output.
- Driver.java: It's the main driver program for the MapReducer job.

Snap of the data set we worked on :



```
Registrar,Enrolment Agency,State,District,Sub District,Pin Code,Gender,Age,Aadhaar generated,Enrolment Rejected,Residents providing emai
Allahabad Bank,A-Onerealtors Pvt Ltd,Uttar Pradesh,Allahabad,Meja,212303,F,7,1,0,0,1
Allahabad Bank,Asha Security Guard Services,Uttar Pradesh,Sonbhadra,Robertsganj,231213,M,8,1,0,0,0
Allahabad Bank,SGS INDIA PVT LTD,Uttar Pradesh,Sultanpur,Sultanpur,227812,F,13,1,0,0,1
Allahabad Bank,Sri Ramraja Sarkar Lok Kalyan Trust,Uttar Pradesh,Shamli,Shamli,247775,M,6,1,0,0,1
Allahabad Bank,Transmoovers India,Uttar Pradesh,Gorakhpur,Sahjanwa,273001,M,8,1,0,0,1
Allahabad Bank,Transmoovers India,Uttar Pradesh,Varanasi,Pindra,221101,M,14,1,0,0,1
Allahabad Bank,Transmoovers India,Uttar Pradesh,Varanasi,Varanasi,221001,M,9,1,0,0,1
Allahabad Bank,Transmoovers India,Uttar Pradesh,Varanasi,Varanasi,221002,M,4,1,0,0,1
Allahabad Bank,Transmoovers India,Uttar Pradesh,Varanasi,Varanasi,221002,M,10,0,1,0,1
Allahabad Bank,Transmoovers India,Uttar Pradesh,Varanasi,Varanasi,221002,M,19,1,0,0,1
Allahabad Bank,Vedavaag Systems Limited,Uttar Pradesh,Bara Banki,Nawabganj,225301,M,8,1,0,0,0
Atalji Janasnehi Directorate  Government of Karnataka,Atalji Janasnehi Directorate  GOK,Assam,Marigaon,Bhuragaon,782121,M,22,1,0,0,1
Atalji Janasnehi Directorate  Government of Karnataka,Atalji Janasnehi Directorate  GOK,Bihar,Gopalganj,Vijayeepur,841508,M,26,1,0,0,1
```

Snap of the MapReduce Job :



It's about implementation, only planning is not sufficient !!

Chapter 9

Screen shots of Project

9.1 Visualizations: Observations and Inferences

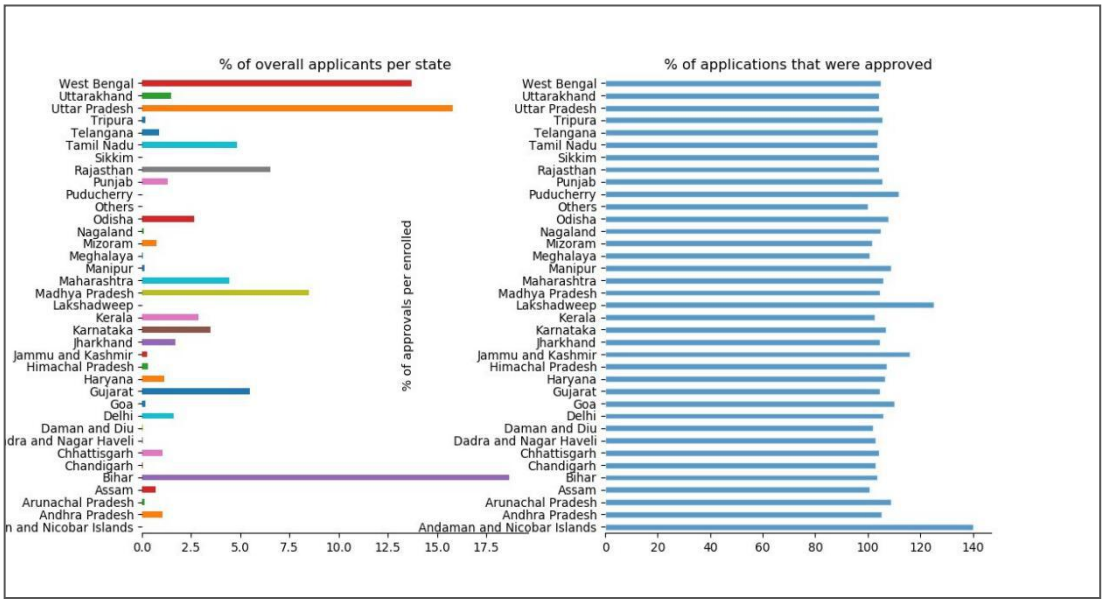


Figure 9.1: Percentage of Overall Applications per State

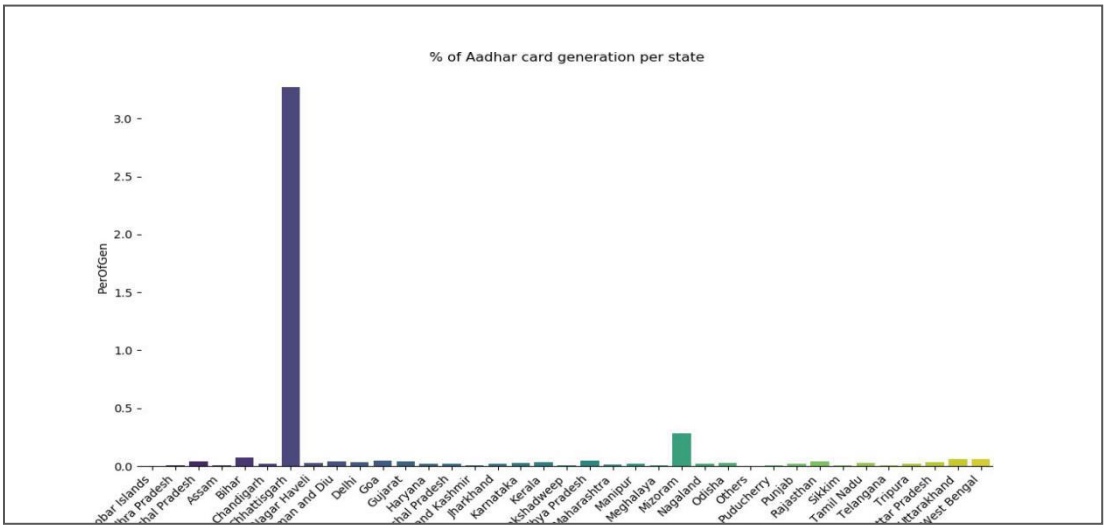


Fig. 13. Percentage of Aadhar Cards Generated per State

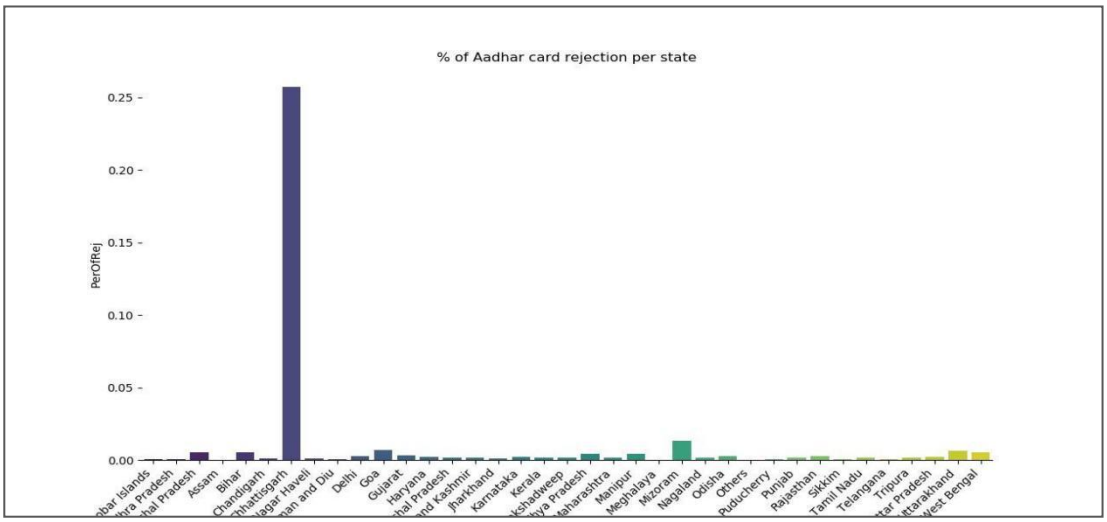
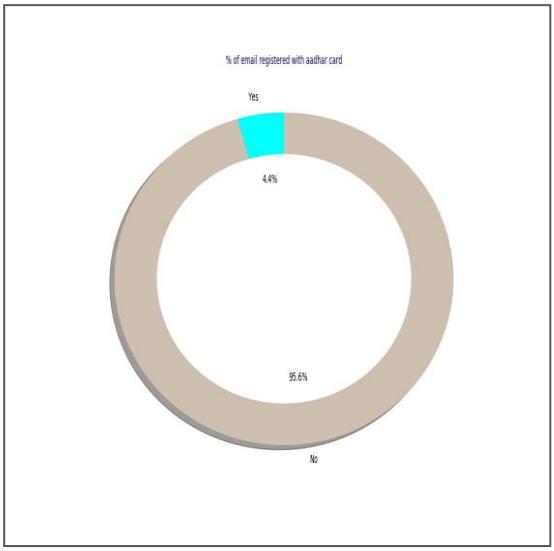


Fig. 14. Percentage of Aadhar Card Rejected per State



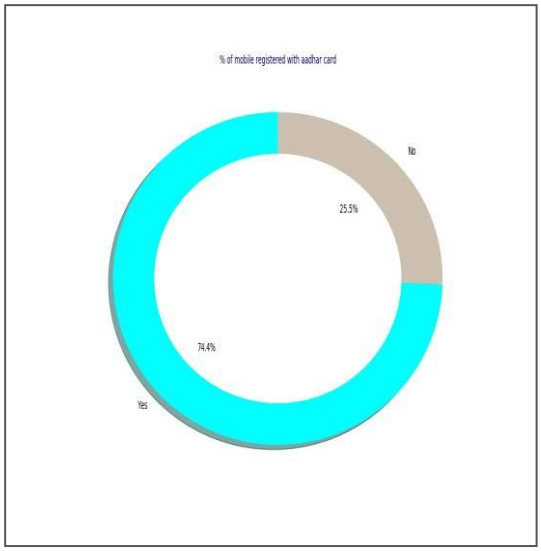


Fig. 15. Percentage of Emails Registered with Aadhar Card **Fig. 16.** Percentage of mobiles Registered with Aadhar Card

Chapter 10

Conclusion and Future Scope

10.1 Conclusion

After having concluded the analysis, authorities can undertake prompt measures to counter the anomalies observed in the 'Aadhar' dataset, the citizens will in turn breathe a sigh of relief. Ambiguous results will be filtered out helping the State and Central Governments resolve conflicts. The proposed idea is an attempt to sort out the long standing dilemma post the Citizenship Amendment Act. It has been observed that the number of female applicants for Aadhar is less than the number of male applicants. It has also been noticed that a lot of Aadhar applications are getting rejected, many approved applications contain ambiguous results, showcasing the inefficiency of the registration process. The need of the hour is to resolve all anomalies and facilitate a smoother conduct of the upcoming NPR Programme . Let each and every individual being a part of this beautiful nation be proud of his/her identity and let us all cheer together, "Mera Aadhar, Meri Pehchaan".

10.2 Future Scope

We have figured out the anomalies in the existing Aadhar database in this project work of ours, this can be further enhanced, the subsequent approach could be to analyse the biometric credentials and verify whether they're in accordance to the requirements or possibly to create a database to monitor the entire procedure subject to the upcoming National Population Register programme.

References

All references should be numbered and cited in the content.

1. Mohit Dayala, Nanhay Singha, An Anatomization of Aadhaar Card Data Set-A Big Data Challenge, International Conference on Computational Modeling and Security (CMS 2016).
2. Durga Bhavani D., Rajeswari K., Srinivas Naik N. (2019) Big Data Analytics on Aadhaar Card Dataset in Hadoop Ecosystem. In: Bapi R., Rao K., Prasad M. (eds) First International Conference on Artificial Intelligence and Cognitive Computing. Advances in Intelligent Systems and Computing, vol 815. Springer, Singapore.
3. Jayashree R. (2020) Analysis of Aadhaar Card Dataset Using Big Data Analytics. In: Hemanth D., Kumar V., Malathi S., Castillo O., Patrut B. (eds) Emerging Trends in Computing and Expert Technology. COMET 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 35. Springer, Cham.
4. K. Ramya, A. Sumathi, Big Data Applications in Aadhar Card Fraud Detection, International Journal of Computer Sciences and Engineering (2019).
5. Mrs.Lakshmi Piriya.S, T. Sri Nithi, Aadhar Based Data Migration, Analysis and Performance using Big Data Analytics and Data Science, International Conference on Computing Intelligence and Data Science (ICCIDS 2018), Department of Computer Studies Sankara College of Science and Commerce Saravanampatty, Coimbatore.
6. Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White and Seung Yon Rhee, "Big data: The future of biocuration", Nature, international weekly journal of science 455, 47- 50,4 September 2008.
7. Clifford Lynch," Big data: How do your data grow? ",Nature , international weekly journal of science ,455, 28-29.
8. Adam Jacobs,"The pathologies of big data", Communications of the ACM - A Blind Person's Interaction with Technology,Volume 52 Issue 8, August 2009.
9. 'Aadhaar' most sophisticated ID programme in the world: World Bank, Daiji World (2017).

10. Linda, B.:Analysis of the Social Security Number Validation Component of the Social Security Number, Privacy Attitudes, and Notification Experiment. In: Report of Census 2000 Testing, Experimentation, and Evaluation Program (2003).
11. Madhavi, V.: Parallel Processing of cluster by Map Reduce. In: International Journal of Distributed and Parallel Systems (IJDPS), vol. 3, no.1, (2012).
12. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. 26(1), 97–107 (2014).
13. Xu, L., Jiang, C., Wang, J., Yuan, J., Ren, Y.: Information security in big data: privacy and data mining. China Commun. (Suppl. 2) (2014).
14. Matturdi, B., Zhou, X., Li, S., Lin, F.: Big data security and privacy: a review. IEEE Trans. Content Min. (2014).
15. Mark A. Beyer and Douglas Laney. "The Importance of 'Big Data': A Definition". Gartner, 2012. For book.
16. D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics," interactions, vol. 19, no. 3, pp. 50–59, May 2012. For journal.
17. Pang -Ning Tan, Vipin Kumar,Micheal Steinbach, "Introduction to data mining", First Edition, 2012.
18. B. Dufrasne, A. Warmuth, J. Appel, et al. Introducing disk data migration. DS8870 Data Migration Techniques. IBM Redbooks. pp. 1–16. ISBN 9780738440606 (2017).
19. B. Goes, Paulo, Design science research in top information systems journals. MIS Quarterly: Management Information Systems (2014).
20. 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).

INDIVIDUAL CONTRIBUTION REPORT:

Pehchaan: Analysis of the 'Aadhar Dataset' to Facilitate a Smooth and Efficient Conduct of the Upcoming NPR

Nishan Acharya
1705247

Abstract: Analyzing the 'Aadhar Dataset' and drawing meaningful insights out of the same will surely ensure a fruitful result and facilitate a smoother conduct of the upcoming NPR. The sole objective of 'Hadoop' in this research is storing and processing huge amount of semi structured data. Hence, our proposed work uses 'Hadoop' for processing the data gathered. The input data is processed using MapReduce and finally the result is loaded into the Hadoop Distributed File System (HDFS).

Individual contribution and findings:

- ✓ Incorporating Software Engineering in my project.
- ✓ Enhancing the credibility of the project.
- ✓ Implementation of Use-Case, Entity-Relationship and Class Diagrams.
- ✓ Throwing light on the future scope of the project and giving a brief on how the Deployment process should take place by integrating with Emerging Technologies and making it fully automated.

Individual contribution to project report preparation:

- ✓ Following a structured approach and prepare the timeline schedule of the activities for a smooth implementation of the project.
- ✓ Following a daily schedule and host meetings at the end of the day.
- ✓ Enhance the look of the Presentation and Highlight the key points.

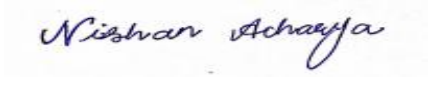
Individual contribution for project presentation and demonstration:

- ✓ Finding out exceptional entry into the database.
- ✓ Resolve anomalies in the Aadhar database.
- ✓ Showcasing the technical aspects involved.
- ✓ Showcasing the utility of the project.

Full Signature of Supervisor:

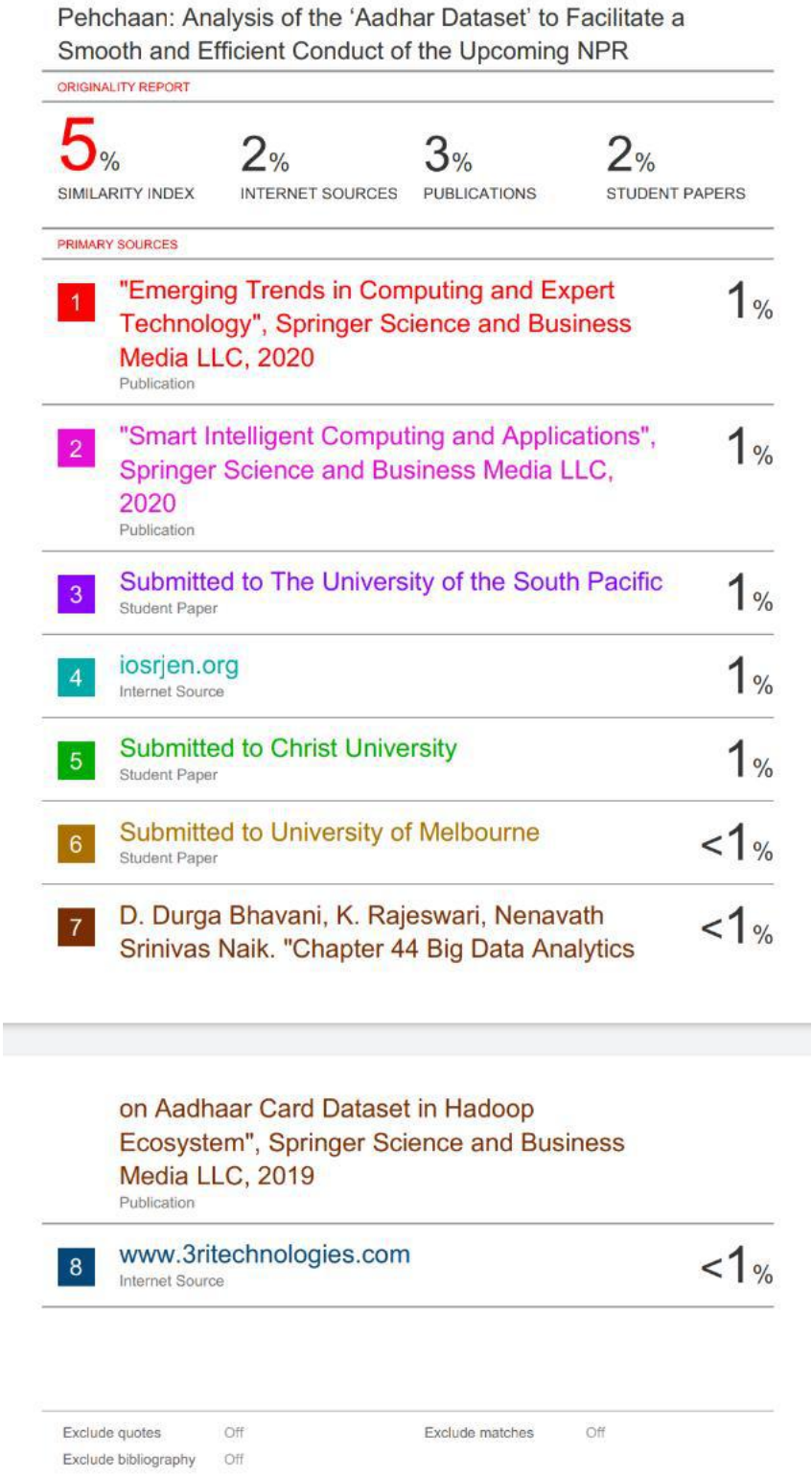
……

Full signature of the student:



TURNITIN PLAGIARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)



Pehchaan: Analysis of the 'Aadhar Dataset' to Facilitate a Smooth and Efficient Conduct of the Upcoming NPR

ORIGINALITY REPORT

5%	2%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Emerging Trends in Computing and Expert Technology", Springer Science and Business Media LLC, 2020 Publication	1%
2	"Smart Intelligent Computing and Applications", Springer Science and Business Media LLC, 2020 Publication	1%
3	Submitted to The University of the South Pacific Student Paper	1%
4	iosrjen.org Internet Source	1%
5	Submitted to Christ University Student Paper	1%
6	Submitted to University of Melbourne Student Paper	<1%
7	D. Durga Bhavani, K. Rajeswari, Nenavath Srinivas Naik. "Chapter 44 Big Data Analytics on Aadhaar Card Dataset in Hadoop Ecosystem", Springer Science and Business Media LLC, 2019 Publication	<1%
8	www.3ritechnologies.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		