The Aadhaar Project: India's Unique Identification Card

Rebecca Rosser Tomás

HADOOP INDIVIDUAL ASSIGNMENT

MBD 2020-A

Table of Contents

INTRODUCTION

Who are you if you do not have an identity card? How do you prove your identity without an identity card and what are you entitled to without an identity card?

With a population size of over 1.2 billion people, only 3% of Indian citizens pay taxes and 75% struggle to survive with less than 2$ per day[1]. The Indian government has a number of poverty schemes in place, amounting to more than $50 billion, to alleviate social suffering and enable individuals to access the benefits that they are entitled to[2]. However, without proof of identity, people in need cannot guarantee that they are Indian residents and are, consequently, out of the spectrum. In addition, whilst millions of individuals live without food and shelter, there are many fraudulent individuals who take advantage of the flaws of this system and are able to create multiple identities – commonly known as 'ghost identities' - to receive numerous benefits to which they are not entitled.

A year after the World Bank published that only 50% of births in India are registered, the Aadhaar Project began – an ambitious Big Data initiative that aims to create Unique Identity Cards based on biometric data for the whole Indian population. In less than ten years, the project has successfully created approximately 1 billion identities.

This brief paper examines the top-notch technology that sustains the Aadhaar Project, which is being successfully deployed to both urban and countryside areas and has received the approval and confidence of private companies, now also requiring this Unique Identification Number to access their systems.

OBJECTIVE

The Aadhaar Project is primarily a mission-driven project with two specific objectives:

- Ensuring good governance and effective and efficient poverty-relief programmes.
- Empowering every citizen with a unique identity and a digital platform to authenticate themselves anytime and anywhere.[3]

[1] https://cdn.oreillystatic.com/en/assets/1/event/119/Architecting%20World_s%20Largest%20Biometric%20Identity%20System%20-%20Aadhaar%20Experience%20Presentation.pdf

[2] https://cdn.oreillystatic.com/en/assets/1/event/119/Architecting%20World_s%20Largest%20Biometric%20Identity%20System%20-%20Aadhaar%20Experience%20Presentation.pdf

[3] https://uidai.gov.in/about-uidai/unique-identification-authority-of-india/vision-mission.html

To accomplish these objectives, the Aadhaar Project had to enable enrolment and authentication and be highly scalable, since the objective was to ultimately register the identities of the whole population of India. Furthermore, data architects had to design a data structure that could be effectively deployable in every corner of the country, including both urban areas and the countryside.

To accomplish these objectives, Dr Pramod Varma, Chief Data Architect, stated that each decision was made bearing in mind the intended positive impact – empowering individuals with a unique identity – and the fact that the technology would definitely fail at some point.[4] Hence, the team proactively sought multiple open sources, horizontal data scaling and distributed and partitioned data storage systems where little computing and processing occurred. In this way, Dr Varma envisioned that, should a technology crash occur, the entire system would not be compromised, and they would be able to resolve it independently.

DATA SOURCES

Aadhaar collects each individual's biometrics (digital fingerprints and iris scan) together with a digital photograph and other relevant text-based data (name, date of birth and place of residence) through multiple third-party devices using a standard Application Programming Interface (API) layer supported VDM, Automatic Biometric Identification Systems (ABIS), Language Support, etc.

[4] https://www.youtube.com/watch?v=08sq0y8V1sE&t=2255s

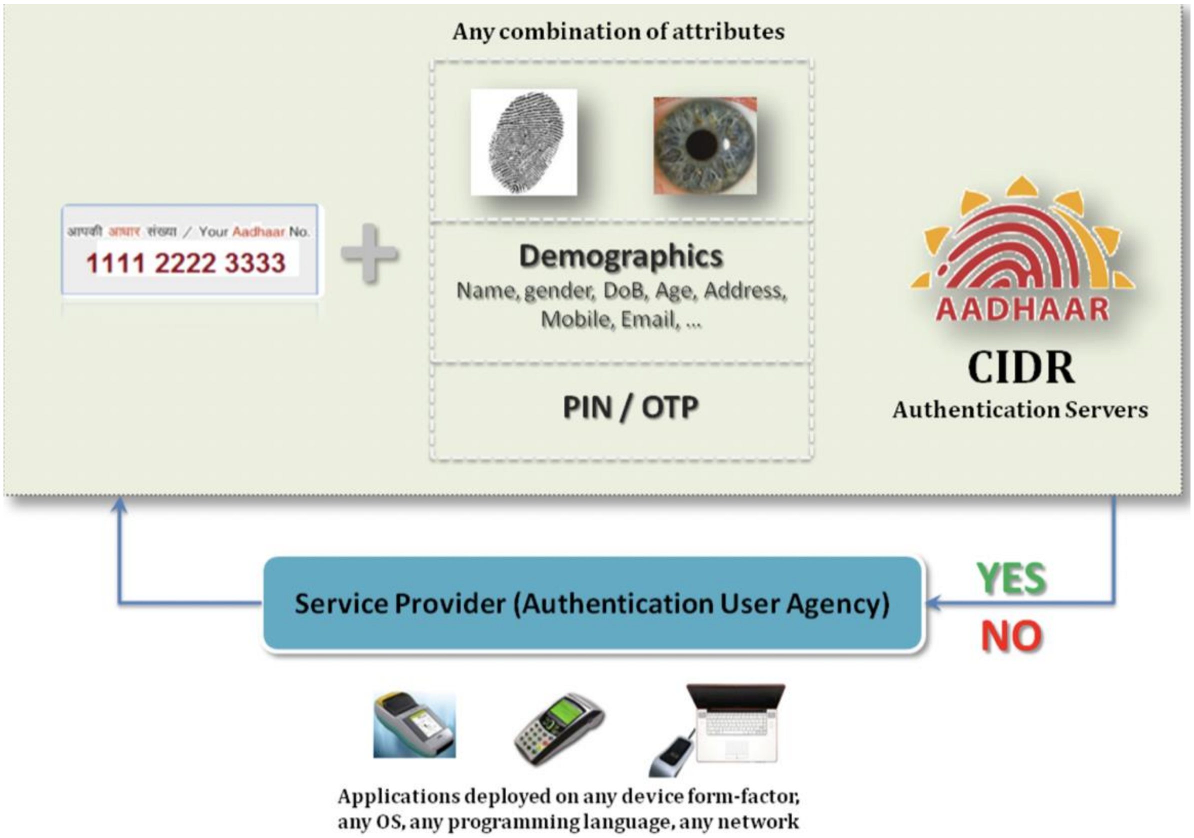

Any combination of attributes
आपकी आधार संख्या / Your Aadhaar No.
1111 2222 3333
Demographics
Name, gender, DoB, Age, Address, Mobile, Email, ...
PIN / OTP
AADHAAR
CIDR
Authentication Servers
Service Provider (Authentication User Agency)
YES
NO
Applications deployed on any device form-factor, any OS, any programming language, any network

Figure 1: Input and output data and technological devices involved.

Aadhaar indicates that the user interface is made possible thanks to HTML 5, Liferay, Java Swing and Android. The rationale behind using mainstream technologies here is that the model must be deployable to common electronic devices found in urban and rural communities, since the ultimate purpose of the project is to provide each citizen with an individual and unique identity that enables him/her to access social care.

DATA INGESTION

The Aadhaar Project has a multi-vendor approach, thus exploiting the benefits of each vendor and system for the desired purposes. Having collected the data, Aadhaar now has to transferred in order to begin the data architecture. Safety in this process is essential in this phase, since Aadhaar has to transfer large amounts of sensitive data, and the team entrusts Apache Sqoop, as part of the Hadoop environment, for this purpose.

DATA STORAGE

Through MapR Converged Database, Aadhaar creates and maintains the largest biometric database ever created. Using Hadoop HDFS, MapR Database builds a NoSQL database management system within the MapR Data Platform, offering Aadhaar just what it needed – high scalability and a multi-modal database.

Scientists driving the Aadhaar project highlighted the importance of horizontal upscaling. Aadhaar has to produce unique identities for 1.2 billion people – the population size of India – which means that there are approximately 1 million entries on a daily basis and approximately 200 matches each day. Therefore, to avoid bottlenecks and data losses in the event of tech failures, the Aadhaar team preferred horizontal storage, as the diagram below illustrates.

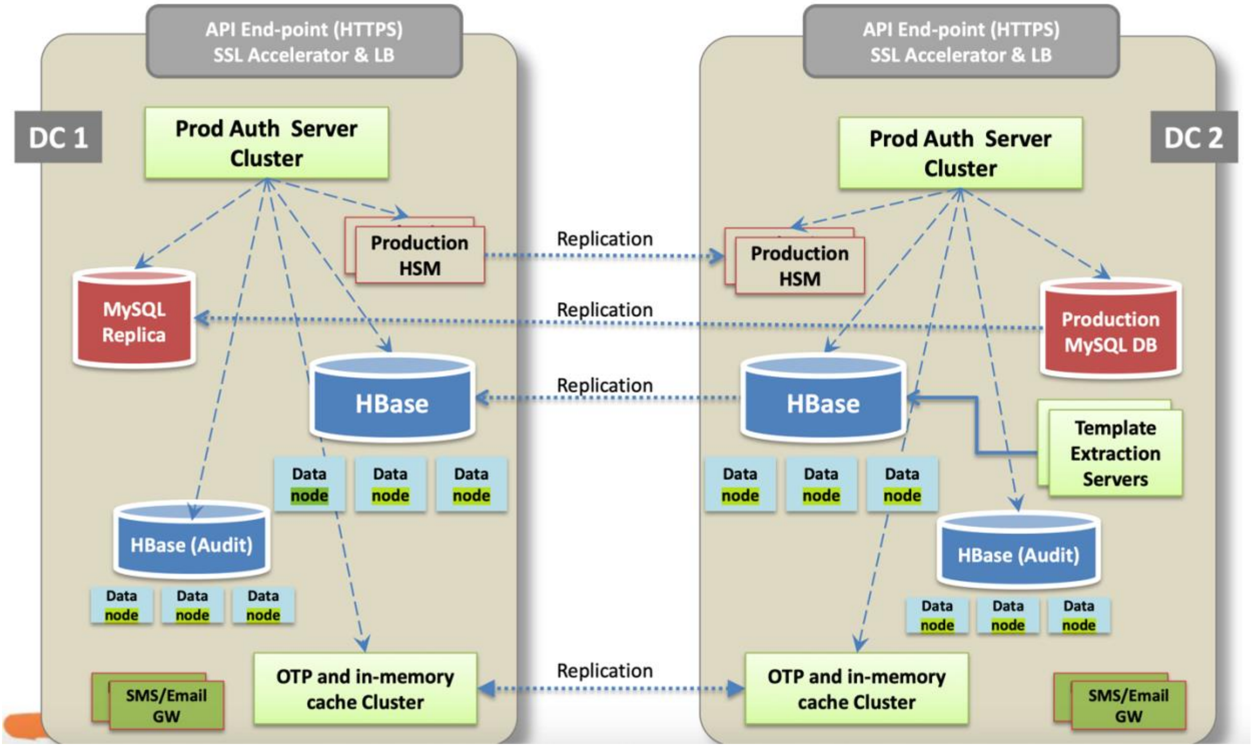


Figure 2: Insights into data management showcasing horizontal data storage.

Yet, Aadhaar's Chief Data Architect emphasised at The Fifth Elephant conference that it was important that the data was not stored on the same platform where the computing would later occur since he believed that it was highly probable that either process failed at some point.[5] Due to the vast data they were going to create and process on a daily

[5] https://www.youtube.com/watch?v=08sq0y8V1sE&t=2255s

basis, they did not conceive the possibility of MySQL automatically resolving technology crashes nor did they trust other agencies to resolve tech crashes. Hence, India's Unique Identification (Aadhaar's Project) is developed across different data storage and processing platforms, as the diagram below showcases.

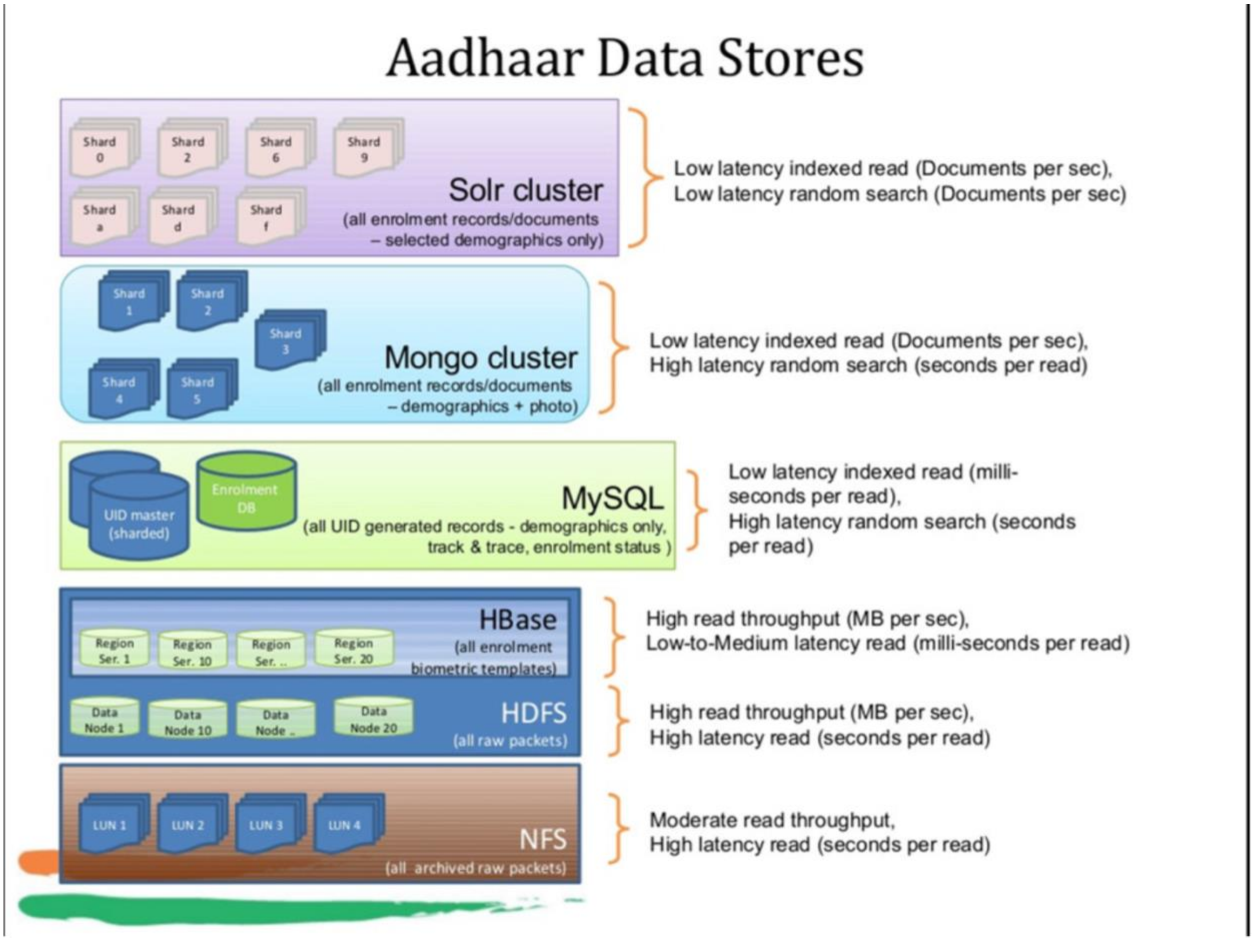

Aadhaar Data Stores
Shard 0
Shard 2
Shard 6
Shard 9
Shard a
Shard d
Shard f
Solr cluster
(all enrolment records/documents – selected demographics only)
Low latency indexed read (Documents per sec),
Low latency random search (Documents per sec)
Shard 1
Shard 2
Shard 3
Shard 4
Shard 5
Mongo cluster
(all enrolment records/documents – demographics + photo)
Low latency indexed read (Documents per sec),
High latency random search (seconds per read)
UID master (sharded)
Enrolment DB
MySQL
(all UID generated records - demographics only, track & trace, enrolment status)
Low latency indexed read (milli-seconds per read),
High latency random search (seconds per read)
HBase
(all enrolment biometric templates)
Region Ser. 1
Region Ser. 10
Region Ser. ..
Region Ser. 20
High read throughput (MB per sec),
Low-to-Medium latency read (milli-seconds per read)
Data Node 1
Data Node 10
Data Node ..
Data Node 20
HDFS
(all raw packets)
High read throughput (MB per sec),
High latency read (seconds per read)
LUN 1
LUN 2
LUN 3
LUN 4
NFS
(all archived raw packets)
Moderate read throughput,
High latency read (seconds per read)

Figure 3: Illustration of various data stores used by Aadhaar Project.

DATA PROCESSING

As the core purpose of the project is to provide each individual with a unique identity, an intrinsic part of the process is detecting and eliminating duplicates. Duplicates refer to multiple identity entries in the system referring to the same person. Thanks to biometric data, an individual may accidentally or fraudulently attempt to enter the database multiple times, but s/he will only have one unique identity – once the Automatic Biometric Identification System (ABIS) correctly detects this. This is done by contrasting each entry’s data (approximately 4MB in size) against all others using a variety of open-source technologies that assess both text-based and biometric data through a multi-step procedure known as Staged Event Driven Architecture (SEDA).

Hence, the objective of this phase is to compute each individual’s data in order to verify an identity within 200 milliseconds throughout an asynchronous process that allows for coupling various components as well as for independent component level scaling.

This is done through MapR’s Converged Data Platform, using a variety of platforms such as Hadoop Map Reduce (intrinsic to MapR) as well as Pentaho and Hive, which facilitate reading, writing and managing such large amount of data. Hive, in particular, can summarise, query and analyse data in a comparatively easier manner.

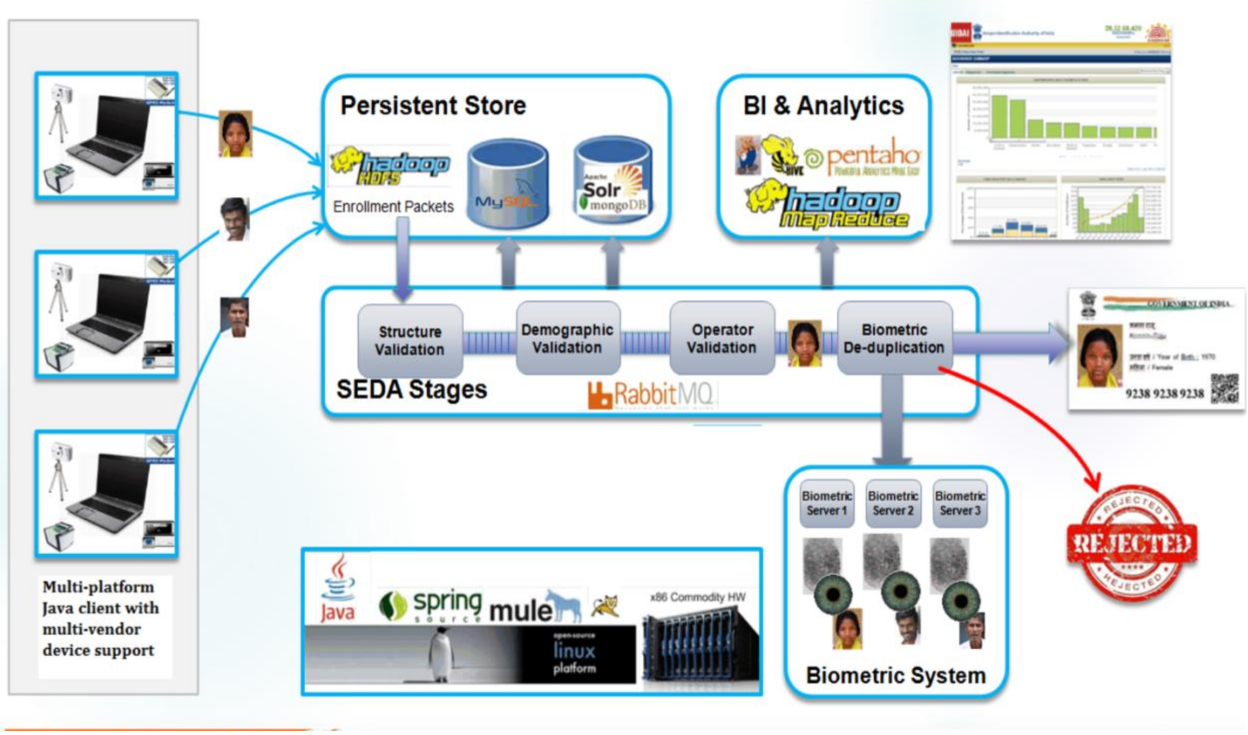


Figure 4: Image indicates the numerous and varied tools used throughout the different stages in the Staged Event Driven Architecture (SEDA).

DATA SERVING

To guarantee that any citizen can authenticate himself or herself anywhere at any moment in time, Aadhaar had to guarantee that when a query was made to the dataset, it always returned a consistent answer.

Thanks to the project's use of NoSQL (Not Only SQL), Aadhaar can view the processed data in a simple, clear and structured manner. NoSQL also ensures a horizontal scalability of the dataset, uniting various data stores and protecting the system from completely collapsing in the event of random failure.

However, as the diagrams we have seen thus far show, the Aadhaar Project strives to implement the best of multiple stores and services, hence leveraging MapR's usage of Hadoop and complimenting this with Mongo, SolR and other tools too.

DATA ARCHITECTURE

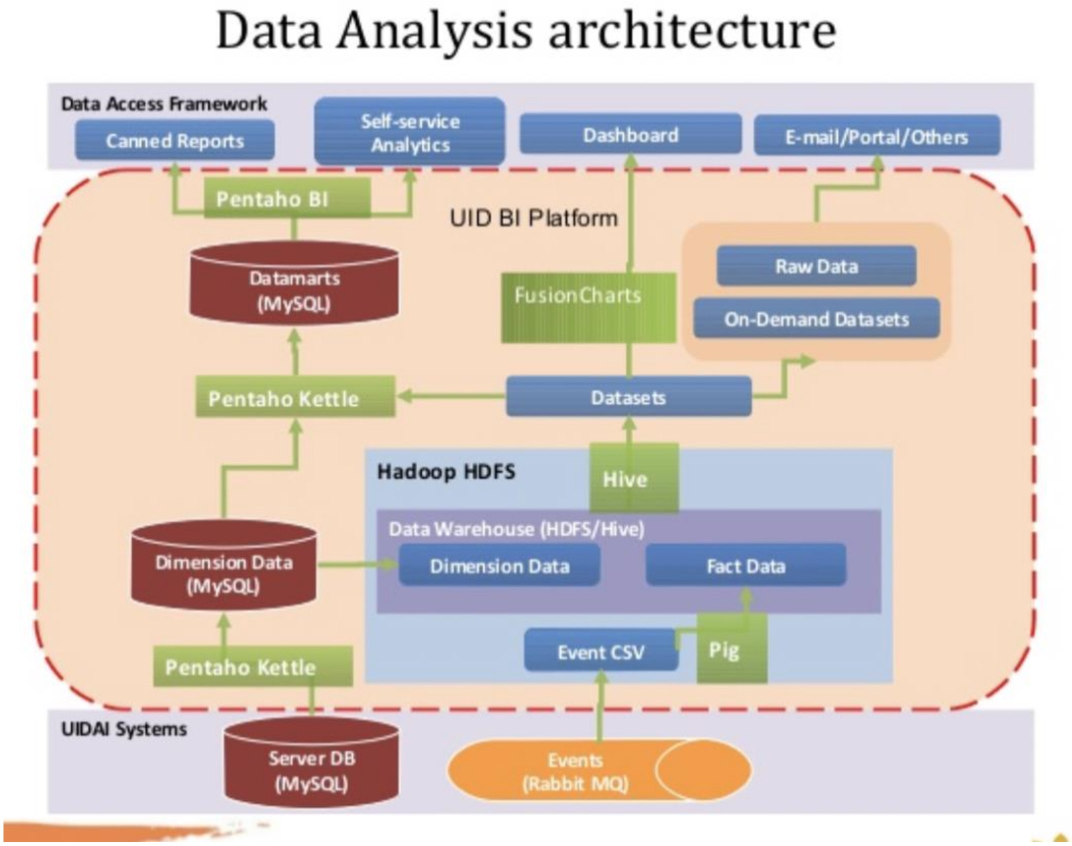


Figure 5: Illustration of Data analysis architecture.

The Aadhaar Project follows a hybrid architecture. Thanks to MapR's Converged Data Platform, Aadhaar can take full advantage of open-source innovations, access a cloud-scale data store and, also integrate certain systems on-premise. As the diagram above showcases, the Aadhaar Project uses open-source systems within the Unique Identity Database Business Intelligence Platform (UID BI Platform).

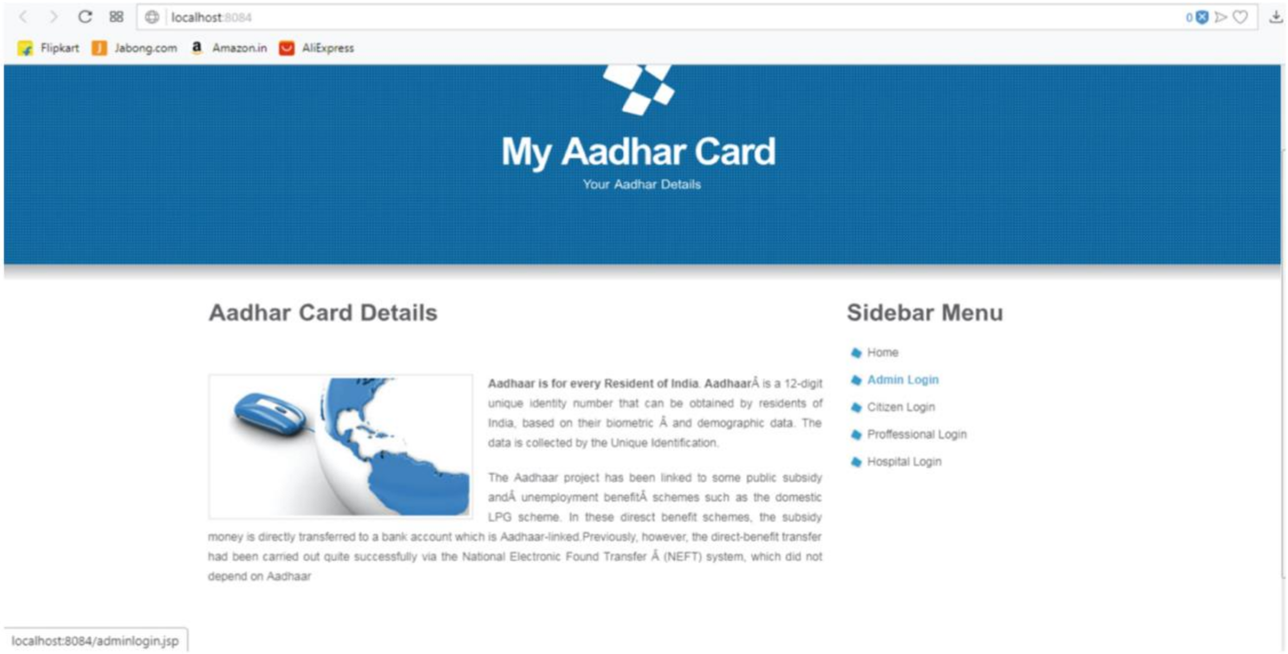
The reasoning behind this type of data architecture is easily understandable: Each day, there are approximately 1 million entries and more than 200 matches (multiple identities) are processed. In addition, this extremely sensitive large database requiring a significant level of safety is used by multiple governmental departments – e.g. social services, health, taxation, employment, etc – and also by private companies wanting to verify the identity of their customers and employees. Therefore, the government wants to have full control of the database – even if this requires having to solve technical failures in-house, as Dr Varma stated at the launch of the platform[6].

RESULTS

Once biometric and text-based data is processed and assessed against other possible matching entries, an identification number is provided to the individual. Whilst the data processing step is completed quickly, the end-user is informed within a few days from the first data entry.

Contrarily to what happened before, this number cannot be usurped since it is associated to a dataset containing biometric information, which is unique to every individual. By reducing the risk of fraudulent duplicate identities, the government can better allocate social aid to the people who need it most in an effective and efficient manner and private organisations can validate the authenticity of their customers' and employee's identity.

[6] https://www.youtube.com/watch?v=08sq0y8V1sE&t=2255s



Aadhar Card Details

Aadhaar is for every Resident of India. AadhaarÂ is a 12-digit unique identity number that can be obtained by residents of India, based on their biometric Â and demographic data. The data is collected by the Unique Identification.

The Aadhaar project has been linked to some public subsidy andÂ unemployment benefitÂ schemes such as the domestic LPG scheme. In these diresct benefit schemes, the subsidy money is directly transferred to a bank account which is Aadhaar-linked.Previously, however, the direct-benefit transfer had been carried out quite successfully via the National Electronic Found Transfer Â (NEFT) system, which did not depend on Aadhaar

Figure 6: Image of the final website where the end-user can find his/her personal data.

CONCLUSION

This paper has briefly indicated a few of the variety of tools and systems with which the Aadhaar Project decided to build the Unique Identification Database to provide each Indian citizen with a biometric-based identity that could be authenticated at any moment and place. Decisions were mostly based on the required safety and horizontal scalability of the database as well as the exhaustive use of the database on a daily basis.

One could say that this ambitious project, founded in 2009, has been highly successful on a technological level. In less than ten years, Aadhaar has registered approximately 1 billion identities. Furthermore, conceived as a mission-driven project that aimed to guarantee good governance and effective deployment of social aid, the Aadhaar Project has created a snowball effect within private companies operating in India, such as Uber, Airbnb and financial institutions[7]. These entities have found in the Unique Identification Authority of India (UIDAI) an opportunity to implement good governance and collect accurate knowledge regarding the identities of their customers and employees.

However, the Aadhaar Project has also been highly criticised. On the one hand, its impact – the rationale and primary intention of the project - is dubious. Many hold that the project has failed to take into consideration the customer journey – in other words, whether the poorest of the poor can successfully register in the dataset, receive their Unique Identification number or use it.

On the other hand, the extensive use that private companies are making of the Unique Identification number provides the government with vast amounts of information regarding citizens' activities, leading to issues of privacy in a country that lacks data protection regulations[8]. To illustrate the concern built around data protection matters, the Aadhaar Project has been labelled in the Indian Times as the largest "Big Brother" conceived.[9]

In conclusion, the UIDAI demonstrates the power of Big Data - and the technology that sustain it - in delivering impactful solutions to societal problems, but also sheds light on issues raised by Big Data, such as privacy controls. How do we find a harmonic marriage between them? This is a question which the Aadhaar Project is currently dealing with – and one over which many of us must ponder.

[7] https://www.buzzfeednews.com/article/pranavdixit/airbnb-uber-and-ola-may-start-using-aadhaar-indias
[8] https://www.buzzfeednews.com/article/pranavdixit/airbnb-uber-and-ola-may-start-using-aadhaar-indias
[9] https://economictimes.indiatimes.com/news/politics-and-nation/government-objects-to-use-of-orwellian-for-aadhaar-says-its-must-for-plugging-leaks/articleshow/59496601.cms?from=mdr

REFERENCES

1. https://uidai.gov.in/about-uidai/unique-identification-authority-of-india/vision-mission.html
2. White Paper on “Mobile as digital identity”. Available at: https://www.mygov.in/frontendgeneral/pdf/white-paper-mobile-as-digital-identity-v0-2.pdf
3. https://www.cse.iitb.ac.in/~comad/2010/pdf/Industry%20Sessions/UID_Pramod_Varma.pdf
4. https://mapr.com
5. Ursula Rao & Vijayanka Nair (2019) Aadhaar: Governing with Biometrics, South Asia: Journal of South Asian Studies, 42:3, 469-481. Available at: https://www.tandfonline.com/doi/pdf/10.1080/00856401.2019.1595343?needAccess=true
6. https://www.slideshare.net/regunathbalasubramanian/hadoop-at-aadhaar-24084015
7. Nupur Aggarwal, Understanding the Technology Empowering India's Aadhar Card. Available at: https://www.linkedin.com/pulse/understanding-technology-empowering-indias-aadhar-card-nupur-aggarwal/
8. Thulasiram Gunipati, Know all about the backbone of Aadhaar – Big Data! Available at: https://www.upgrad.com/blog/the-backbone-of-aadhaar-big-data/#Commodity_Hardware
9. Anthony Kimery, Aadhaar’s architect discusses what went into world’s biggest biometric repository. Available at: https://www.biometricupdate.com/202003/aadhaars-architect-discusses-what-went-into-worlds-biggest-biometric-repository
10. Pramod Varma, Architecting World's Largest Biometric Identity System - Aadhaar Experience. Available at: https://conferences.oreilly.com/strata/stratany2014/public/schedule/detail/36305
11. MapR Industry guide for Big Data in federal agencies & the public sector. Available at: https://mapr.com/whitepapers/mapr-government-federal-industry-guide/assets/mapr-government-federal-industry-guide.pdf
12. https://www.youtube.com/watch?v=08sq0y8V1sE&t=2255s
13. Airbnb, Uber and Ola Are Considering Using India's Creepy National ID Database. Available at: https://www.buzzfeednews.com/article/pranavdixit/airbnb-uber-and-ola-may-start-using-aadhaar-indias

14. Samanwaya Rautray, Government objects to use of ‘Orwellian’ for Aadhaar, says it’s must for plugging leaks, *The Economic Times.* Available at: https://economictimes.indiatimes.com/news/politics-and-nation/government-objects-to-use-of-orwellian-for-aadhaar-says-its-must-for-plugging-leaks/articleshow/59496601.cms?from=mdr
15. Bidisha Chaudhuri & Lion König (2018) The Aadhaar scheme: a cornerstone of a new citizenship regime in India?, *Contemporary South Asia*, 26:2, 127-142.
16. R. Jayashree, ‘Analysis of Aadhaar Card Dataset Using Big Data Analytics’ in *Emerging Trends in Computing and Expert Technology*, p. 11208-1225.