Twitter Thread by Arun Krishnan ■■



Arun Krishnan ■■
@ArunKrishnan_

Thread!

1/

OK Folks, I have gone through the entire document and so here is (hopefully) a simplified account of what Aadhaar is, what it is meant to and NOT meant to do, how it works, the privacy aspects etc. This thread is probably going to be long, so bear with me.

> Since I have the time, I am going through a White Paper on Aadhar (UADAI) and its architecture. It is something stupendous.
>
> The sheer scale and the way it has been architected are brilliant.
>
> I will see if I can put together a thread once I finish going through the document.
>
> — Arun Krishnan \U0001f1ee\U0001f1f3 (@ArunKrishnan_) June 7, 2022

2/ For those who would like to read through the document, here is the link:

https://t.co/YqltVlgJay

The document can be heavy reading for non technical folks, so let me try and simplify it here in this thread.

3/

Given that one of the biggest problems in India was to prove one's identity, Aadhaar was conceived to do exactly that. The Aadhar strategy can be listed as follows:

1) UIDAI only provides identity linked to a person's demographic and biometric information.

4/

2) Aadhaar number provides identity NOT citizenship. All legal residents can get an Aadhaar card.

3) A pro-poor approach. Aadhaar hopes to enable India's poor and underprivileged by providing token-less, online, anytime anywhere authentication system.

5/

4) Enrolment of residents with proper verifications. This also includes ways and means to enrol those without proper documentation by means of "the introducer" model

5) Partnership model - partnerships with Registrars and Enrolling agencies through central/state govts.

6/

6) Enrolment is not mandated but since services now flow through it, it incentivizes people to join in.

7) UIDAI issues a number and NOT an identitiy card.

8) The Aadhaar number does NOT contain any intelligence or embedded personal information.

7/

Info collected by Aadhaar:

1. Name
2. DOB (or Age)
3. Gender
4. Address
5. Mobile # and Email (Optional)
6. Ten fingerprints, 2 iris scans, and a photograph
7. For children < 5 years, Aadhaar name and number of parent/guardian

8/

Aadhaar has a process to ensure no duplicate and I shall talk about it later.

Aadhaar also has a process to keep data up to date. Individuals are incentivized to provide accurate data.

9/

Aadhaar is ONLY an identity provider and many services can use Aadhaar to establish identity along with their own means. Three main usage types

1) Confirming Beneficiary of various social sector programs/schemes. Eg: Subsidized food/kerosene, health service delivery/MNREGA

10/

2) Attendance and Muster rolls - for eg: for students/teachers; MNREGA beneficiaries, pension system.

3) Financial Transactions - eg: banks that authenticate customers using both Aadhaar & bank-related identity information.

11/
People often ask, why Aadhaar when we have DL/Pan/Passport etc.

Aadhaar is designed as the "ROOT" identity. All others can use Aadhaar to establish identity but they aren't the root.
Also, Aadhaar has been designed so that its databases do NOT contain linkages to other DBs

12/

Since this is the case, &databases are federated (fancy word meaning separated), it is not possible 2 really get ALL personal information by linking Aadhar number 2 Pan number to DL to Passport to Health/Insurance card etc.

This is good bcos it minimizes data privacy issues

13/

This also means that Aadhaar has no knowledge of other transactions etc. For eg: If your bank uses your Aadhaar number to establish your identity, all they get back from Aadhar is a YES/NO response.
Any transactions you do post that, are captured by the bank & not Aadhaar

14/

One of the key principles while designing the system was scale.

* 1.3 BILLION storage records
* 600-800 million UIDs in 4 years
* 1-4 million enrolments a day
* 5-8MB data / enrollment
* Nearly 600 TRILLION biometric searches a day in the initial years!

15/

Nearly 25 PETA bytes of data is going to be stored by the system is meant to scale to about 80 billion UIDs eventually.

In order to achieve this scale, some of the design considerations were:

16/

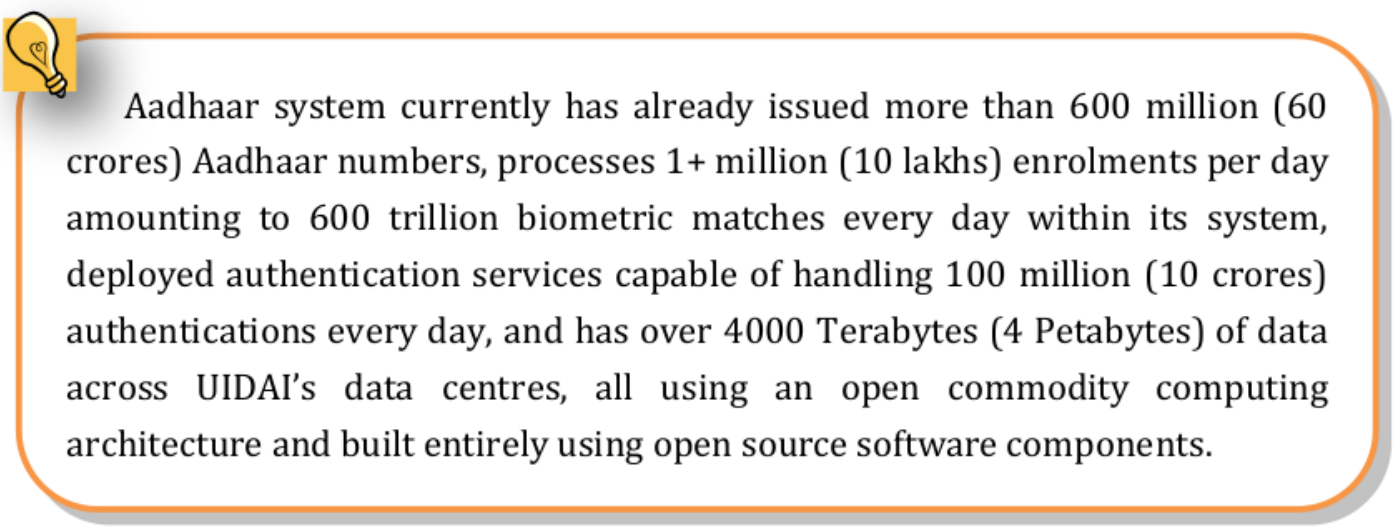
1) Horizontal Scale for compute and storage
2) Using Commonly Off the Shelf products (servers/hard disks)
3) Use of open source software - Java, Spring, Spring-boot, Hibernate, HDFS, Hive, Pig
4) Use of open source protocols - JSON, POJOs, XML

17/
5) Data partitioning and parallel processing
6) Sharding and replication of data across multiple nodes for redundancy
7) Open API and messaging ensuring loose coupling.

All of the above also ensured that there was no dependence on a small subset of vendors.

18/

Just to leave you with the scale. Remember this document is from 2014.



19/
OK. so now let us get into how it works.
I told you that UIDAI works through Registrars and Enrollment Agencies. So they needed to build client applications that these EAs would give out to their people to enroll others.
All Enrollment officers need to be registerd w Aadhaar

20/

When we give our details, you would have noticed that he would be filling them into the client application. That application also has data for auto-filling address elements based on Pincode as well as a transliteration program from English to other Indian languages

21/

One question that is often on people's minds is, how am I to know that my data is secure on that guy's laptop. Well, here it is.
There is an in-memory database that is used to store all your information as you give it, including photo, data, fingerprints and iris scans

22/
So basically none of that data is store on the file system - YET.
Once your enrollment is complete, that in-memory data is encrypted with a 2048-bit key and then signed by the enrolling officer (so his id, station, company he works for) is also encoded.

23/

Often, his supervisor also digitally signs the enrollment. This encrypted data is then sent over SSL (secure layer) to the UIDAI servers.
Note that even if there is a Man in the middle attack, or someone hacks into the machine, all they have is the encrypted data.

24/
This CANNOT be decrypted without the corresponding private key which is only available within the secure UIDAI servers (and with its own safeguards).

Once this data is received at the server, it is kept in a zone called DMZ (fancy name for 'safe place').

25/

The data is checked to see if it is the same as the data that was sent. If the enrolling officer and supervisor are in the system. Only THEN, is it moved to the production server and a copy also made for Archival and another redundant backup (all in the encrypted form)

26/

Once that is done, ONLY then is the copy of the data in the DMZ, deleted.

Now the fun starts. The data goes through a series of processes including cleanup etc. There is a first level of removal of duplication based on the personal details provided. This will get rid of

27/
Obvious duplications. Once this is done, is the piece I found most fascinating.

The BIOMETRIC De-deduplication. Basically what they have done is given this de-duplication to multiple vendors so that there is redundancy built in.

28/
The three Automatic Biometric Identification System (ABIS) compete for work based on their accuracy and throughput with the accuracy being constantly monitored. So more accurate ABIS systems get more of the work.

29/
I need to take a slight detour to help you understand why 3 ABIS systems and why fingerprints + iris scans are used.

The committee found that globally, 99% de-duplication accuracy was achieved by fingerprints alone in a database size of 50 million records.

30/
If you increased the DB size to 1.2 billion records, then the accuracy of identifying duplicates goes down to 95%.

In order to get 99% accuracy at the size of the DB, it required both fingerprints and iris scans. And THAT's WHY they ask for

both.

31/
Anyway, back to the ABIS. If any ABIS identifies a potential duplicate, it is sent to the other ABIS for verification and then the results from the 3 are combined to determine if an enrollment is a duplicate or not.
In the rare event where there is some doubt ...

32/
Manual intervention is required for the same.
The utilization of 3 different de-duplication engines also helps to detect various kinds of software or data collection errors. Additionally, this also permits for continuous monitoring.

33/

Once the de-duplication step is done, if the record is found NOT to have duplicates, an Aadhaar number is generated and master record is created.
Also, all the biometric data is encapsulated in a few metrics so that it becomes easy to access when doing authentication.

34/
What happens to the original data submitted you ask?
You remember I said that the encrypted data is stored for archival and backup. They are also stored under a randomly generated HASH filename so that from the filename itself you can't identify whose record it is.

35/

When say a bank wants to use your Aadhaar number for eKYC, the data cannot be retrieved unless authorized by the person. So you have to either use your iris scan or finger print or an OTP based authentication before any of that data can be accessed by the bank.

36/

And as I said before, the response that is received is just YES/NO. Yes, this person is who he says he is or NO he isn't. None of your personal information is ever provided to anyone else.

37/

For technical folks, I would suggest you go through that document because the thought put into the architecture fascinated me.

There are checkpoints at every stage of the process with the state saved at every step.

38/
The pipeline itself is stateless so it can be easily parallelized and taskfarmed. They use a SEDA (Staged Event-driven Architecture) wherein each task stands on its own and is based on POJOs. There are a series of event queues through which the tasks move.

39/
There are checkpoints at every stage with the state being saved. Cron jobs run frequently to identify zombie or failed processes and these are resubmitted.

40/

They have a lot of logs being written for the BI/Analytics engines to work with. Information through the enrollment or authentication process is given in great detail as well as the health of the physical infrastructure.

41/

They even have ratings in terms of errors for the different enrollment officers, stations, locations, companies etc and these are monitored and corrections made/eduaction provided.

42/42

There is a lot more interesting information in there but hopefully this is enough for the lay person to appreciate the scale at which Aadhaar works.