Aadhaar

Scalability & Data Management Challenges

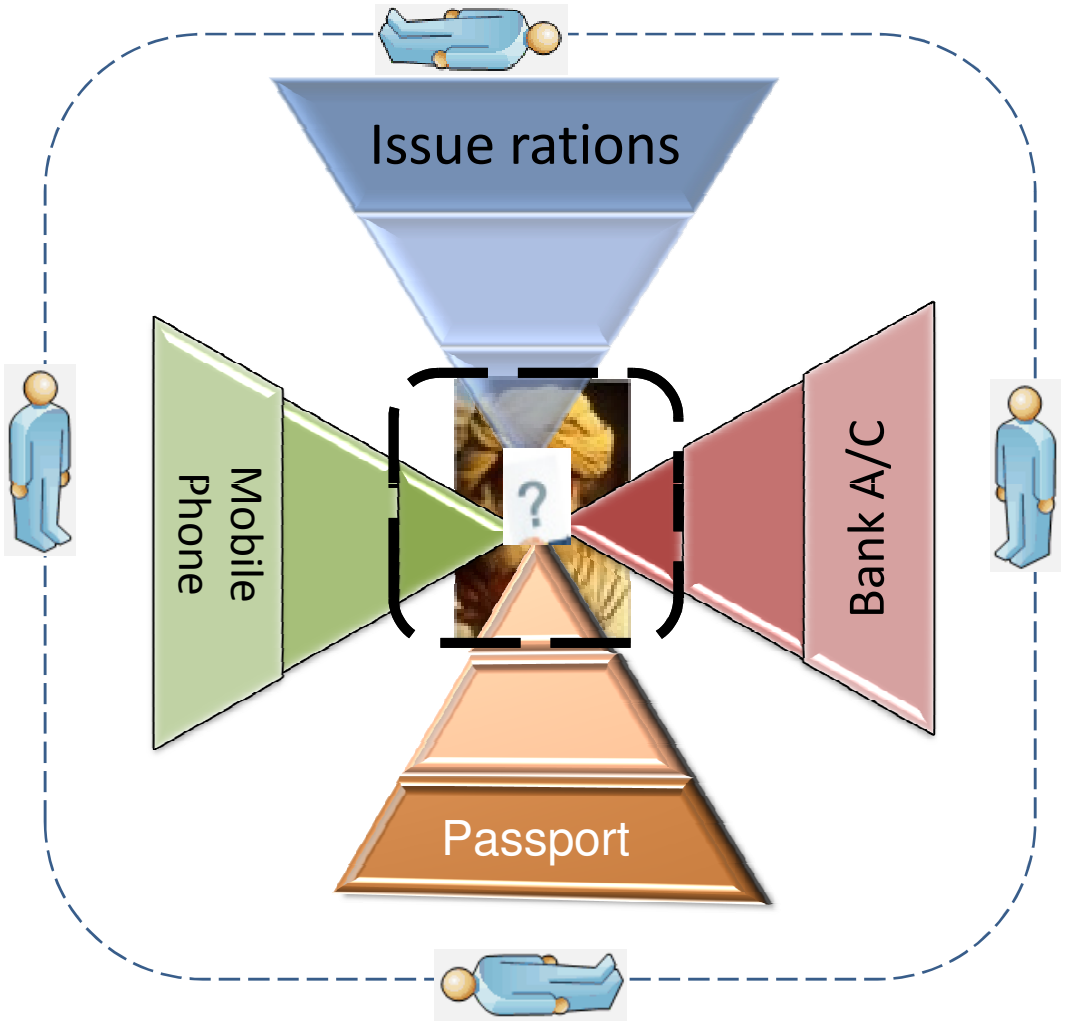
Dr. Pramod K. Varma

Chief Architect, UIDAI

twitter.com/pramodkvarma **pramodkvarma.com**

Understanding Aadhaar System

Establishing ID is a Challenge



A resident typically accesses multiple service providers, at different times

Needs to repeatedly re-establish ID = problem for the poor

Birth records ✗
Address proof ✗
Money to 'beat' the system ✗

= No or limited access to entitlements and opportunities

AADHAAR

Why Aadhaar?

Difficulty in establishing ID ⇒ exclusion

Weak authentication ⇒ inefficient delivery

Entitlements

Food, fuel, fertilizer subsidy = ~Rs. 1 lac crore

Financial

60% unbanked (~700mn)

Social Security Net

45% BPL do not have a ration card

→ Ghost Entries

→ Duplication

→ Multiple layers

"…biometric-based unique identity has the potential to address both these dimensions simultaneously."

- Thirteenth Finance Commission

AADHAAR

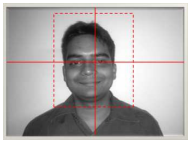
Enroll Once …

- Aadhaar Number - Unique, lifetime, biometric based identity

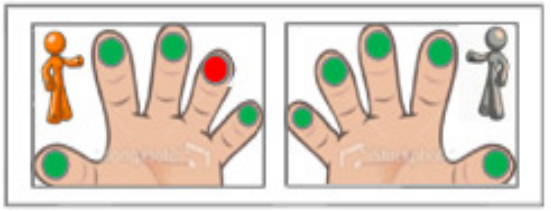
Demographic Data

- Compulsory data:
  - Name, Age/Date of Birth, Gender and
  - Address of the resident.
- Conditional data:
  - Parents/Guardian details
- Optional data:
  - Phone no., email address

Biometric Data



Resident's Photograph



Resident's Finger Prints

Resident's Iris

AADHAAR

… authenticate many times

- Online service to verify the claim – “are you who you claim to be?”

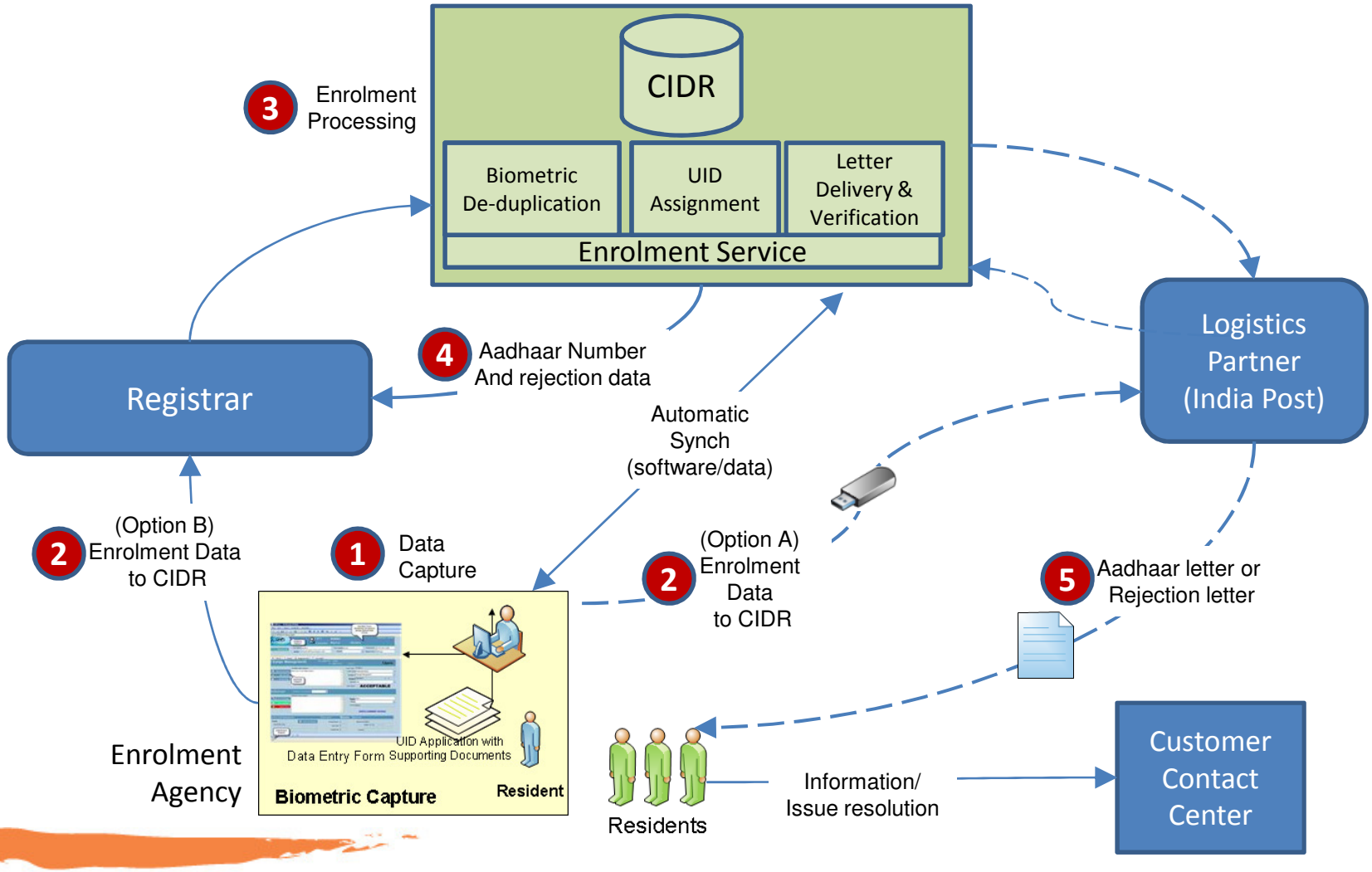
- 1:1 check – only a “yes/no” answer
- Authenticate online
  - Anytime, anywhere, multi-factor
  - Always responds with “yes” or “no”
- Open identity platform
  - Can be used in any service, any domain
  - using any protocol, any device, any network

AADHAAR

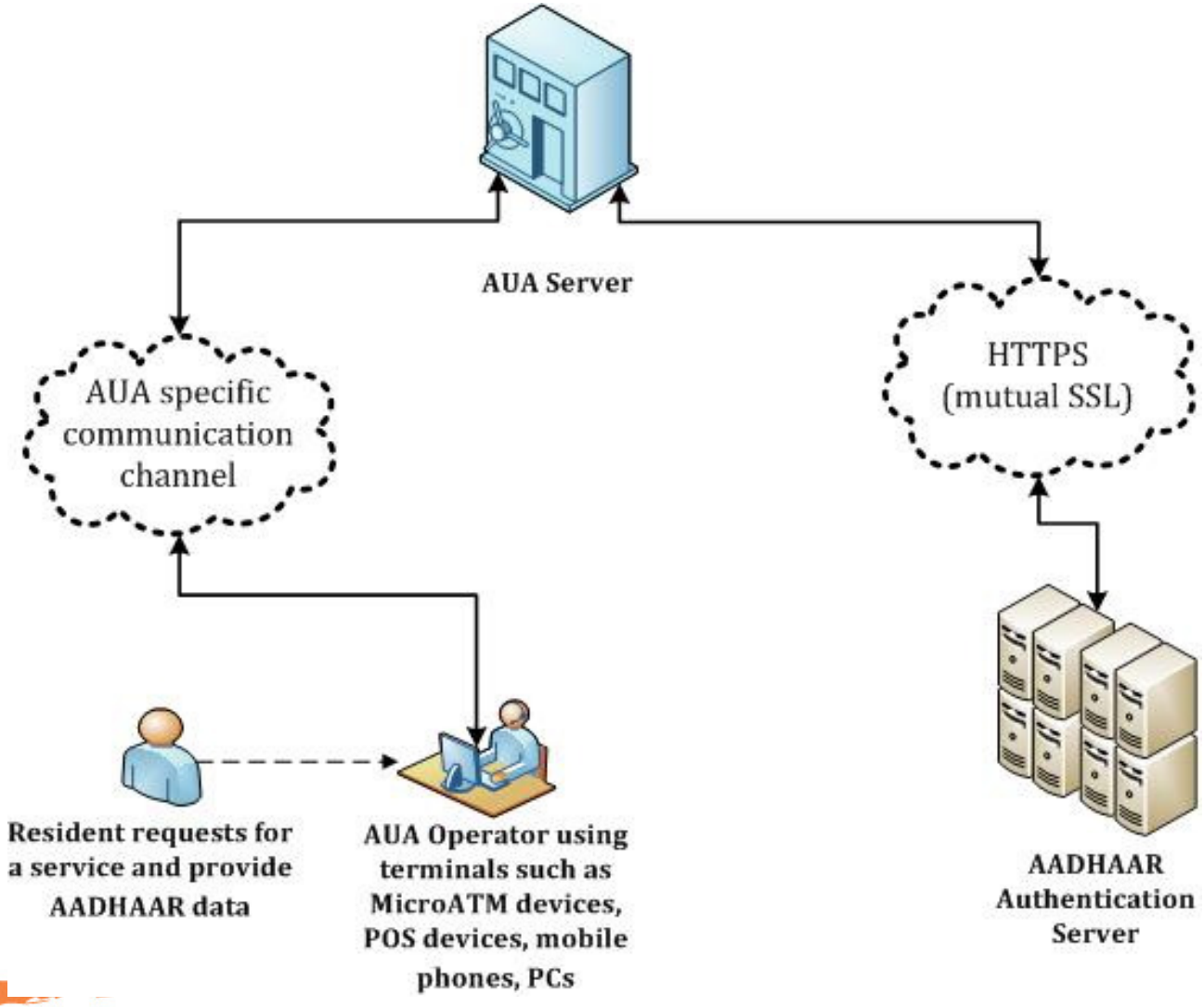
Application Modules

- Enrolment
  - Geographically Distributed Client (mostly offline)
  - Enrolment Server with Multi modal, Multi-vendor ABIS
- Authentication
  - Geographically Distributed Servers
  - Geographically Distributed Devices (several millions)
  - Multi-factor support
- Supporting Systems
  - Business Intelligence
  - Fraud Detection

Enrolment Process



Authentication Process



Enrolment Server

- Manages complete Aadhaar enrolment and lifecycle process
- Features
  - Data validation
  - Operator, supervisor verification
  - Biometric de-duplication (1:N matching)
  - Manual inspection
  - Aadhaar number allocation / rejection
  - Letter generation and delivery tracking
  - Registrar integration

Biometric De-duplication

- Multi-modal matching
  - 1:N matching (Every resident is matched using his/her biometrics against every entry in the ABIS system)
- Multi-vendor interface through ABIS API
  - Dynamic allocation to ABIS vendor based on their accuracy and performance
  - Multi-DC architecture adds complexity
- Exception handling
  - Mostly automated and manual
  - Volumes require highly automated and learning systems to handle exceptions in an effective manner

AADHAAR

Authentication

- Supports answering the question “is a resident the person he/she claims to be”
  - Verifies resident information (demographics, biometrics) for a given Aadhaar number against the stored data
  - Online service that is lightweight, ubiquitous, and secure
  - Only responds with a “yes/no” and no personal identity information is returned as part of the response
- Supports multi-factor authentication using biometrics, PIN, OTP and combinations thereof
- Supports multiple protocols and devices
  - Personal computer, mobile, PoS terminals, etc.
  - Many protocols (USSD, SMS, HTTPS) over data and mobile connections
  - Works with assisted and self-service applications

AADHAAR

Scalability and Data Management Challenges

Architecture Highlights

- Support large scaling of enrolments and authentications
- No vendor lock-in across the system
- Use of open-source technologies wherever available and prudent
- Use of open standards to ensure interoperability
- Ensure wide device driver support for biometric devices through standardization
- Use of widely adopted technology platforms and tools
- Make all performance metrics (no PII) public through business intelligence portal for transparency
- Build strong end-to-end security upfront

AADHAAR

Enrolment Server Architecture

- Throughput is the key
- Fully distributed compute platform
- Data sharded across multiple RDBMS instances and DFS
- Highly asynchronous using a high speed messaging layer
- SEDA (Staged Even Driven Architecture) allows smarter failure handling
- Multi-DC architecture for near-zero RTO and zero RPO (adds complexity in biometric d-deuplication)

AADHAAR

Enrolment Volume

- 600 to 800 million UIDs in 4 years
- 1 to 4 million enrolments a day
- When we cover half the country, we will end up doing
  - 4 m * 12 * 500 m * 12 biometric matches a day!!!
- Data updates and new enrolments will continue for ever
- Enrolment data moves from very hot to cold needing multi-layered storage architecture

AADHAAR

Enrolment Data Management

- Enrolment require handling of large binary data for all residents
  - ~5 MB per resident biometrics
  - ~3 MB for supporting docs
  - Maps to about 8 PB of raw data!
  - With replication, it means managing about 25 PB of source data
  - Replication and backup across DCs of 4+ TB of incremental data every day for near-zero RTO
- Additional workflow/process/event data
  - 15+ million events on an average moving through async channels
  - Needing complete update and insert guarantees across data stores
- Lifetime updates adds several more petabytes

AADHAAR

Authentication Server

- Authentication poses response time issue
  - Match demographics (partial, fuzzy, Indian language matching)
  - Match biometrics (balancing FPIR)
- Needs to scale to handle 100's of million requests every day with sub-sec response
- Edge cached, in-memory operation
- Async data updates to the cache
- Stateless service
- Audits maintained asynchronously on HDFS

AADHAAR

Authentication Volume

- Few 100 million authentications per day
  - mostly during 10 hr period
  - High variance on peak and average
  - Requires async request handling on HTTP server
  - Sub second response with support for OTP, guaranteed audits
- Multi-DC architecture
  - Fully load balanced
  - Mostly reads with some updates (OTP, Audit)
- All changes needs to be propagated from enrolment data stores to all authentication sites
  - PIN updates, OTP requests, and less occasional demographic data updates

AADHAAR

Authentication Data Management

- Minutiae based authentication request is about 1 K
  - Image based ones are about 10 K on an average
- 100 million authentications / day means
  - 1 billion audit records in 10 days
  - 1 TB encrypted audit logs in 10 days
  - Need to keep recent audits online accessible any time and older ones in achieve until deleted
  - Audit write must be guaranteed

Analytics/Mining Architecture

- Analyzing terabytes of data generated out of billion+ events every day
  - Constantly aggregating data across billions of records on a distributed compute grid to analyze and create patterns for operational and strategic decision making
- Fraud detection
  - Detecting fraud during enrolment
  - Detecting identity fraud scenarios near real-time during authentication
  - Building mining, clustering, learning tools to work on top of billions of events

Technology Stack

- Java application deployed on Linux stack with virtualization
- Multiple MySQL instances as RDBMS
- Apache Hadoop (HDFS, Hive, HBase, Pig) stack for large scale compute and distributed storage
- RabbitMQ (AMQP standard) as messaging framework
- Drools for rules engine
- Several other open source libraries
- All 3rd party interfaces abstracted through standard API layer (VDM, ABIS, Language Support, etc)

Final Thoughts

- Largest biometric identity system is about 120 million. Scaling needs are unprecedented.
- Completely built on open standards and open source platforms
- Scalability, Security, interoperability, and vendor neutrality a must
- Next generation e-governance applications require cloud based, large data-driven, open platforms
- Research community support required

AADHAAR

Thank You!

Dr. Pramod K. Varma

Chief Architect, UIDAI

twitter.com/pramodkvarma **pramodkvarma.com**