AADHAAR

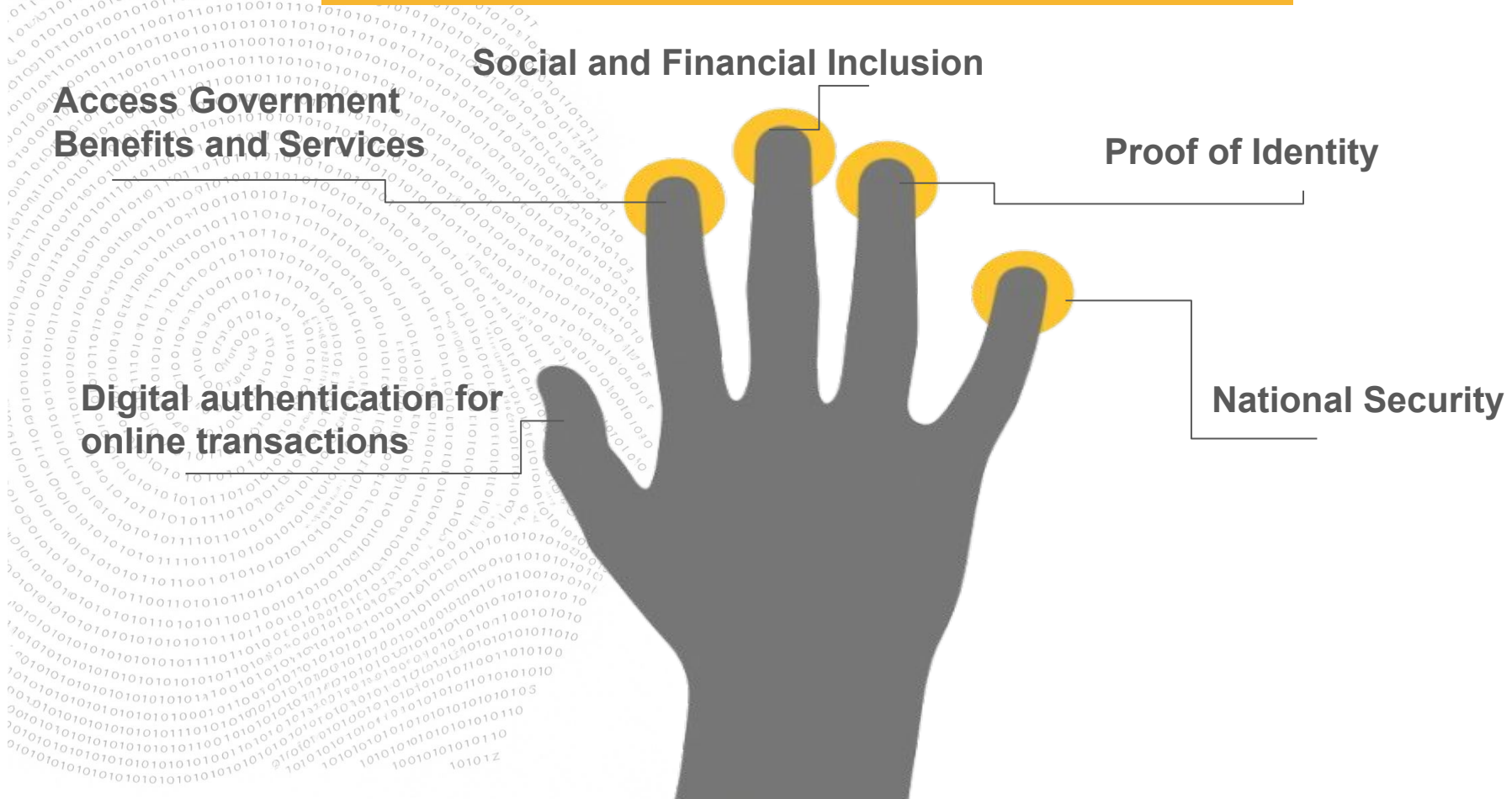
Understanding India's National Identification System.

- Hari Haran
- Hua-Hsing Huang
- Rohith CR

What is Aadhaar ?

- Aadhar is India's National Identification program, assigning a 12- digit random number.
- World's largest Biometric Identification system, issued to ~1.4 Billion residents. (~15 Petabytes!)
- Assigned to residents after satisfying the verification process
- Managed by the UIDAI (Unique IDentification Authority of India)

Benefits and Implications



Data Description

Demographic Data:

- Name
- Date of Birth
- Gender
- Address
- Phone number

Biometric Data:

- Ten fingerprints
- Two iris scans
- Facial photograph

Unique Identifier:

- A randomly generated 12-digit Aadhaar number assigned to each resident.

Other potential tables

- Verification logs
- Enrollment logs
- Benefits register
- … and more

Data Governance

The Aadhaar database contains highly sensitive personal data and requires strong data governance policies

Legal Frameworks :

- Aadhar Act 2013 and Digital Personal Data Protection Act, 2023
- Provides rules on consent, purpose of the PII

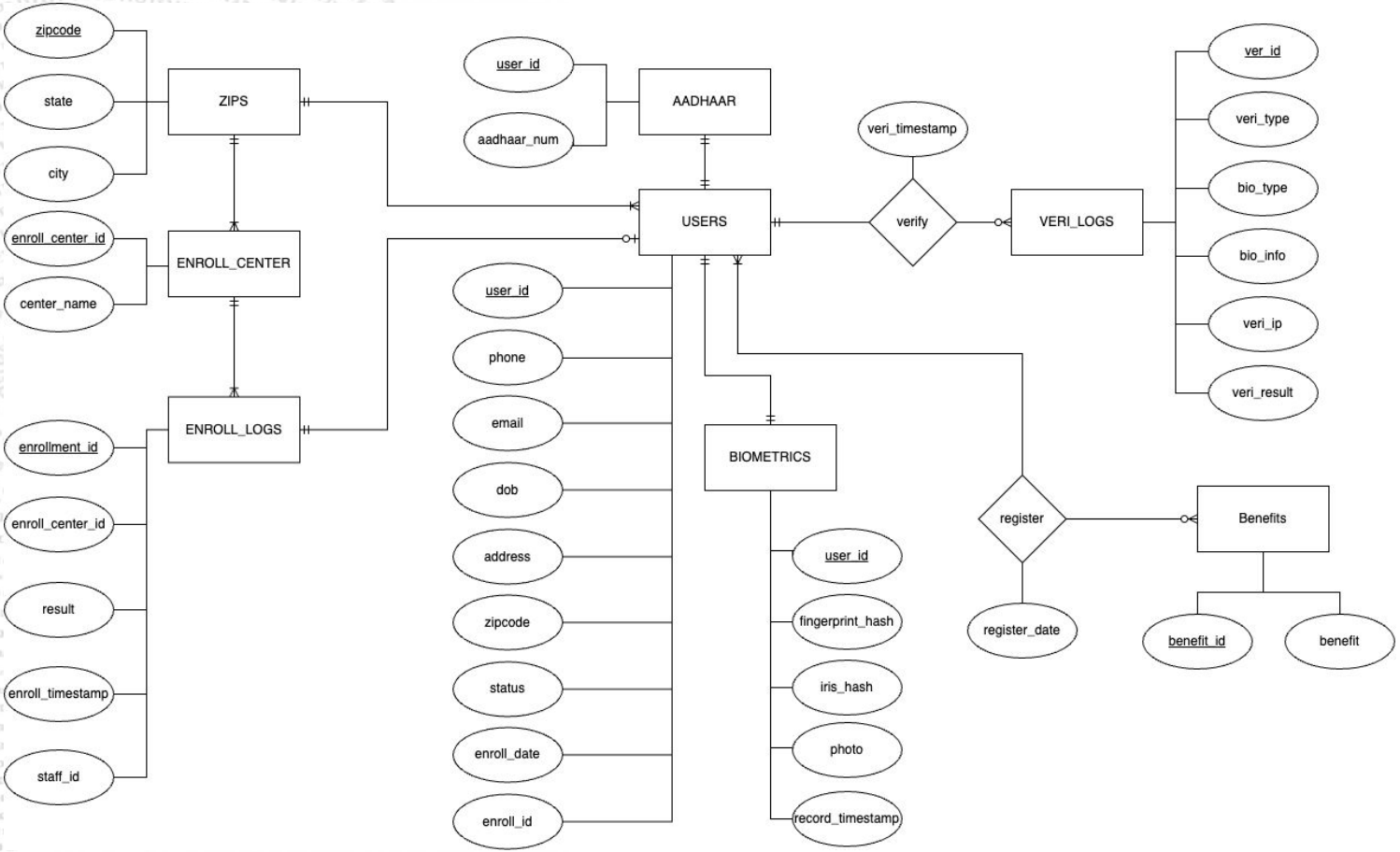
Data Minimization & Anonymization:

- Authentication only returns a "Yes/No" response
- UIDAI promotes virtual IDs (VIDs) to avoid using the actual Aadhaar number
- RBAC implemented at appropriate levels

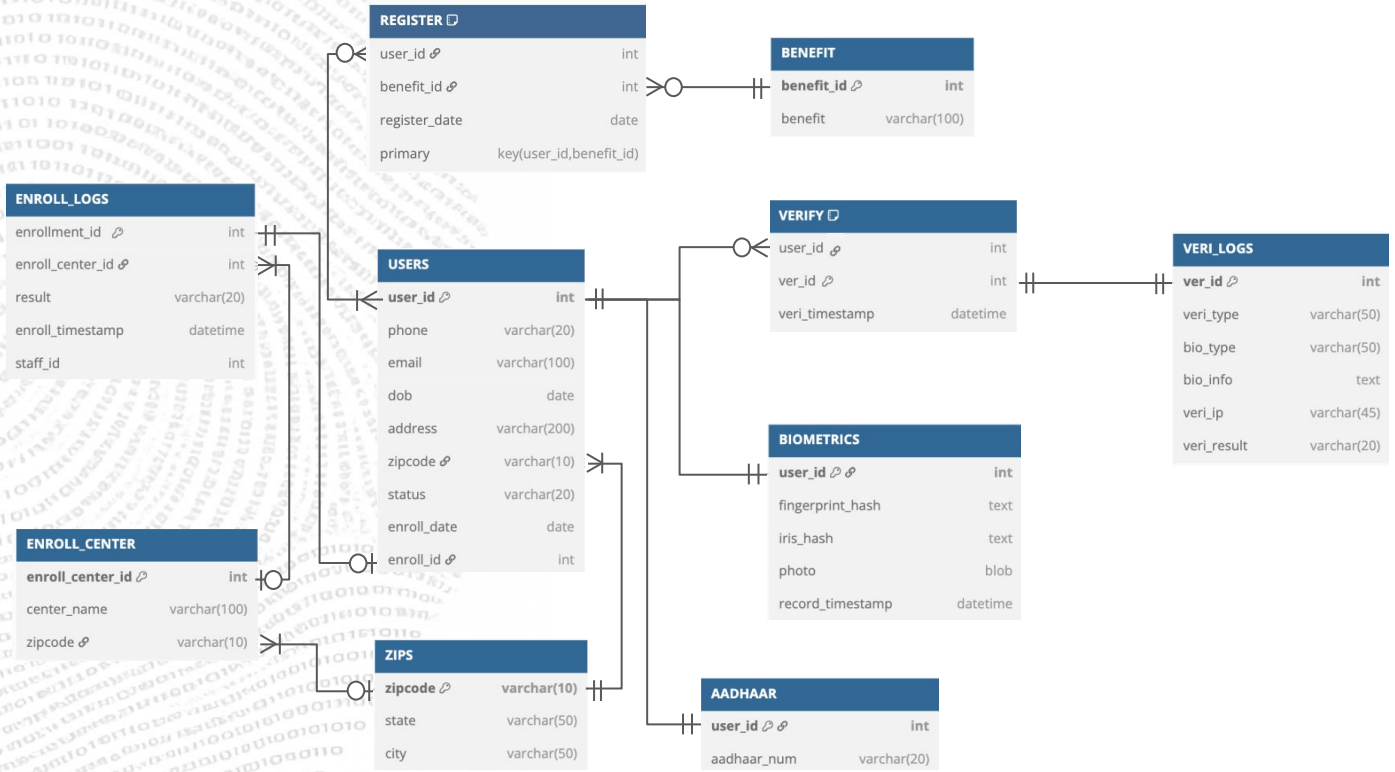
Storage and Compliance:

- Biometric data is never shared and only stored securely in the Central Identities Data Repository (CIDR).
- Periodic security audits and external Compliance Audits may be performed
- Access Control; Only authorized personnel can access to the sensitive data

ER Diagram


zipcode
state
city
ZIPS
enroll_center_id
center_name
ENROLL_CENTER
enrollment_id
enroll_center_id
result
enroll_timestamp
staff_id
ENROLL_LOGS
user_id
aadhaar_num
AADHAAR
USERS
user_id
phone
email
dob
address
zipcode
status
enroll_date
enroll_id
BIOMETRICS
user_id
fingerprint_hash
iris_hash
photo
record_timestamp
veri_timestamp
verify
VERI_LOGS
ver_id
veri_type
bio_type
bio_info
veri_ip
veri_result
register
register_date
Benefits
benefit_id
benefit

Schema


REGISTER
user_id int
benefit_id int
register_date date
primary key(user_id,benefit_id)
BENEFIT
benefit_id int
benefit varchar(100)
ENROLL_LOGS
enrollment_id int
enroll_center_id int
result varchar(20)
enroll_timestamp datetime
staff_id int
USERS
user_id int
phone varchar(20)
email varchar(100)
dob date
address varchar(200)
zipcode varchar(10)
status varchar(20)
enroll_date date
enroll_id int
VERIFY
user_id int
ver_id int
veri_timestamp datetime
VERI_LOGS
ver_id int
veri_type varchar(50)
bio_type varchar(50)
bio_info text
veri_ip varchar(45)
veri_result varchar(20)
BIOMETRICS
user_id int
fingerprint_hash text
iris_hash text
photo blob
record_timestamp datetime
ENROLL_CENTER
enroll_center_id int
center_name varchar(100)
zipcode varchar(10)
ZIPS
zipcode varchar(10)
state varchar(50)
city varchar(50)
AADHAAR
user_id int
aadhaar_num varchar(20)

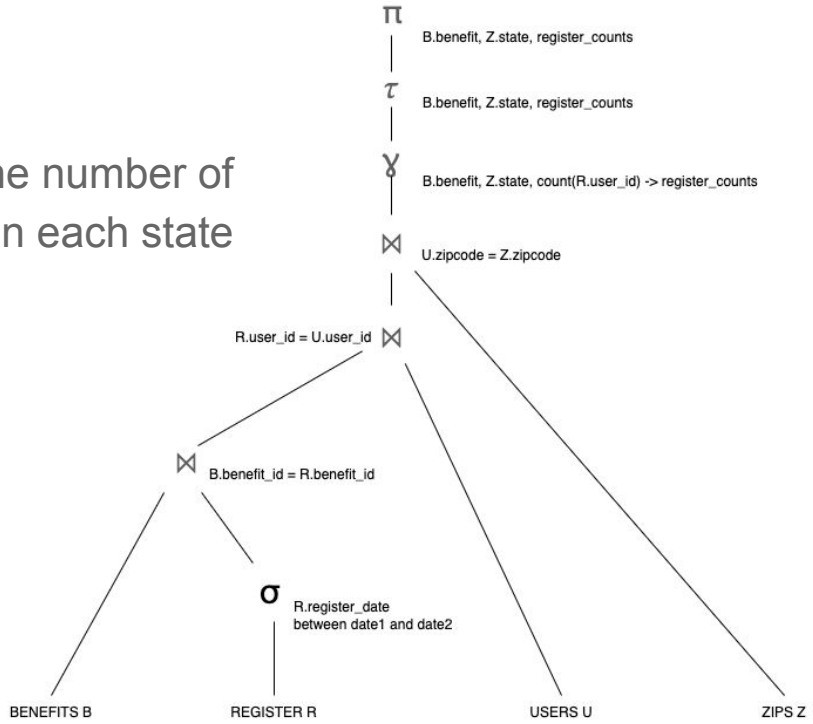
Use Case 1: Social Benefit Register Report

Objective: Understand the monthly status of benefit registrations in each state

Frequency: Monthly

Query: Group data by benefit and state to count the number of registrants per benefit within a certain time period in each state

```sql
SELECT
  B.benefit
  , Z.state
  , count(user_id) as register_counts
FROM BENEFITS B
JOIN REGISTER R
  on B.benefit_id = R.benefit_id
JOIN USERS U
  on R.user_id = U.user_id
JOIN ZIPS Z
  on U.zipcode = Z.zipcode
WHERE R.register_date between date1 and date2
GROUP BY B.benefit, Z.state
ORDER BY B.benefit, Z.state, count(user_id)
```



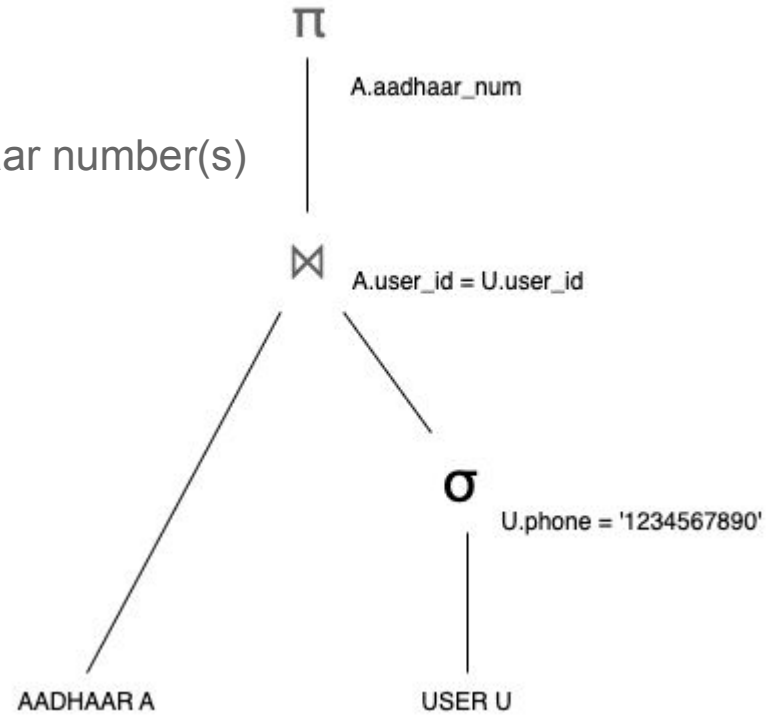
Use Case 2: Authentication

Objective: Identify the Aadhaar number(s) linked to individuals who own a phone number reported for criminal activity

Frequency: Ad hoc (a few times per month)

Query: Given a phone number, retrieve the Aadhaar number(s) of its current registered owner(s)

```
SELECT A.aadhaar_num
FROM AADHAAR A
JOIN USERS U
  ON A.user_id = U.user_id
WHERE U.phone = '1234567890'
```



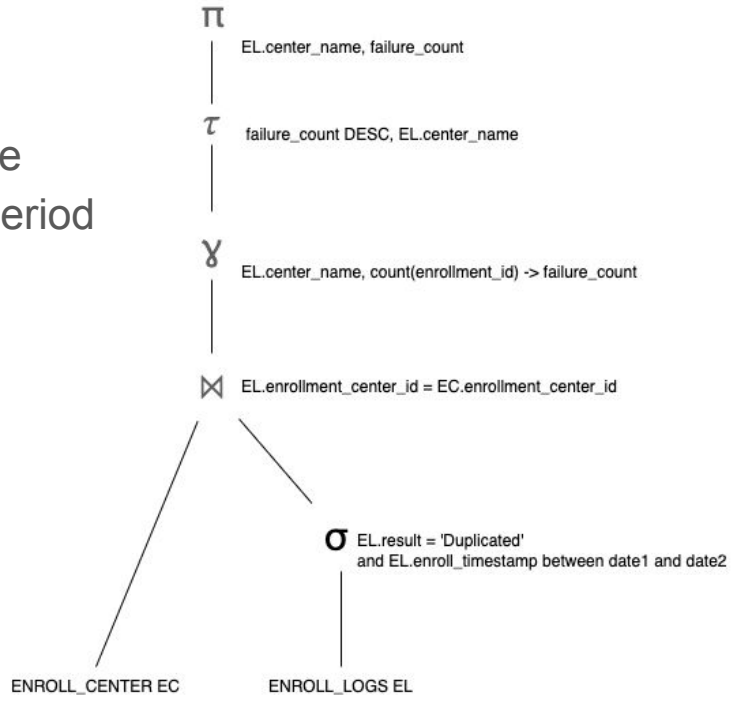
Use Case 3: Enrollment Duplication

Objective: Summarize duplicate enrollments across enrollment centers

Frequency: Monthly

Query: Group data by enrollment center and count the number of duplicate applications within a given time period

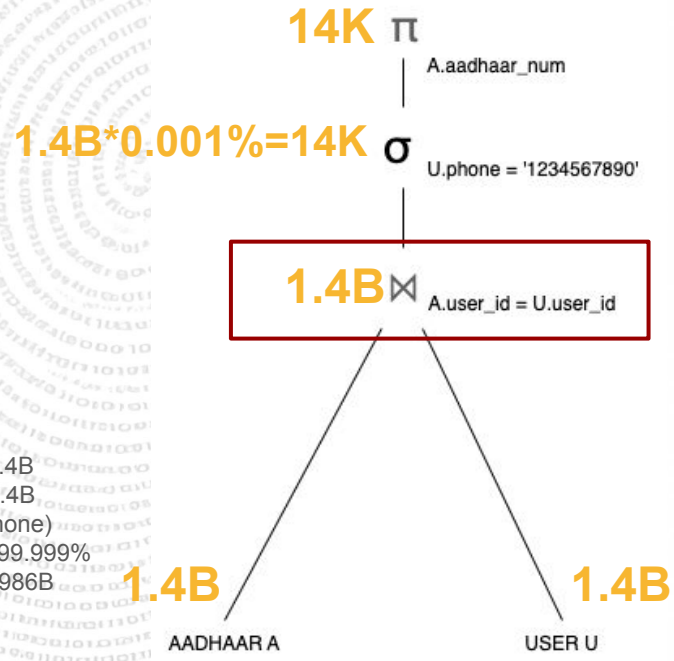
```sql
SELECT
  EL.center_name
  , count(EL.enrollment_id) as failure_count
FROM ENROLL_LOGS EL
JOIN ENROLL_CENTER EC
  ON EL.enrollment_center_id = EC.enrollemnt_center_id
WHERE EL.result = 'Duplicated'
  and EL.enroll_timestamp between date1 and date2
GROUP BY EL.center_name
ORDER BY count(EL.enrollment_id) DESC, EL.center_name
```



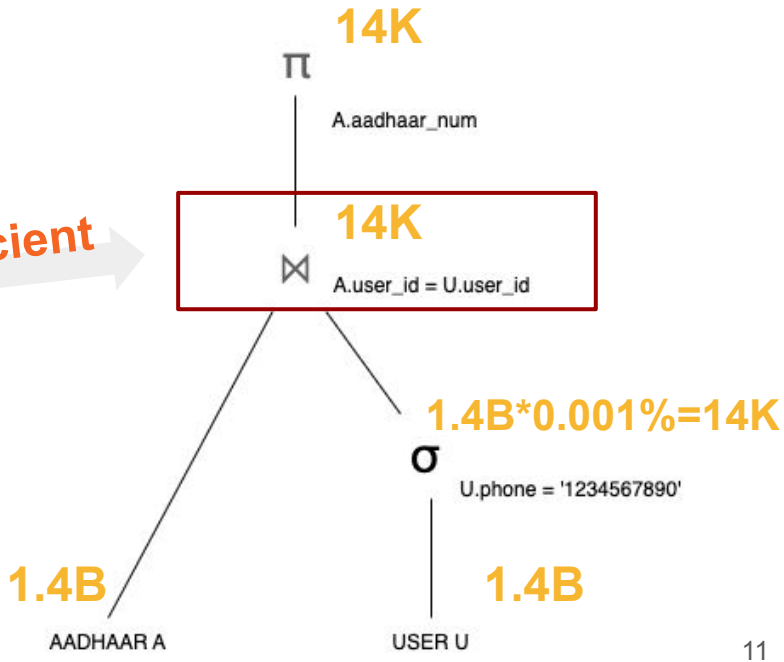
Logical Plan - Case 2

Assumptions

- 1.4 billion individuals in India have an Aadhaar number
- 0.001% of phone numbers are associated with duplicate records



100000x efficient



T(A)=1.4B
T(U)=1.4B
T(U, phone)
=1.4B*99.999%
=1.399986B

Normalization

Table	Primary Index (clustered)	Secondary Index	Partition	Note
USERS	user_id	zipcode phone enroll_id enroll_date	user_id enroll_date	• Frequently join user_id • [Case 2] point search phone • [Case 3] join zipcode • Partition by user_id due to the large number of users.
AADHAAR	user_id	aadhaar_num	user_id	• Frequently join user_id and search aadhaar_num
ENROLL_CENTER	enroll_center_id	center_name		• [Case 3] join and group by centers
ENROLL_LOGS	enrollment_id	enroll_center_id result (result, enroll_timestamp)	enroll_date enroll_center_id	• [Case 3] join and group by centers • [Case 3] point search result • [Case 3] range search enrollment date
REGISTER	user_id	benefit_id register_date	register_date	• [Case 1] group by benefit_id
VERI_LOGS	veri_id	veri_timestamp user_id veri_status	veri_id veri_date	• Not in our use cases, but may need to query verification results by time
BIOMETRICS	user_id	record_timestamp	user_id	• Not in our use cases, but may need to search bio_info record by time

Data Architecture and System Recommendation

- Oracle Database Enterprise Edition - For storing relational data
- Hadoop Distributed File System (HDFS) - For Biometrics Storage
- Apache Solr – Index for full text search
- Openstack- Private Cloud for Infrastructure & deployment


THANK
YOU
GRAZIE
DANKE
SPASIBO
ARIGATO
TACK
DAW-DYEH
KOP KHUN
TODA
KIITOS
GRACIAS
TAKK
DO JEH
MERCI
XIEXIE
MULTUMESC
TERIMA KASIH
OBRIGADO
HVALA
NANDRI
SHUKRAN
GRAZZI
DANK U
TAK
GRACIES