• PII and Regulatory Compliance

⇒ We will now see some approaches to store sensitive information securely, from the perspective of PII.
These approaches can be applied to any data that can impact Business compliance.

⇒ Most companies need to implement policies, that comply with both GDPR and CCPA, to have business in Europe and California

⇒ Lakehouse Architecture reduces copies of PII, and finds personal info quickly.

⇒ While Databricks support table ACLs, physically seperating personal information files and private data into Storage containers, must be done.
Then, they can also be guarded by Cloud IAM settings & configs.

Cached data would not persists after a job completes.

⇒ Also, from demographic views, identifiable details such as keys, must be removed.

⇒ Applying PSEUDONYMISATION and ANONYMIZATION techniques to datasets reduces the risk of data exfiltration and reduce visibility to most users of the datasets.

Pseudonymization

- Switches original data point with pseudonym for later re-identification
- Only authorized users will have access to keys/hash/table for re-identification
- Protects datasets on record level for machine learning
- A pseudonym is still considered to be personal data according to the GDPR

approaches

(not the Hashing that we mean in cryptography)

Hashing

⇒ Data (like passwords) are hashed and then stored.

⇒ Salt values are added to hash.

⇒ Salt values used are stored in secured places and are accessible to authorised users only.

⇒ Increases size of data.

Tokenization

⇒ Stores PII as keys. Values are stored securely

⇒ Slow to write, fast to read.

⇒ optionally, some data are segregated before hashing. For ex:- Street no. & name can be first separated, and then hashed.

⇒ GDPR & CCPA are better implemented with values.

In Psedonymisation:-
Data is reversibly altered, and original data can be reidentified.

In Anonymisation:-
Data is irreversibly altered, such that data subjects can no longer be identified directly or indirectly.

Anonymisation Approaches :-

Data Suppression

⇒ Use of dynamic access controls

⇒ Exclude column with PIIs from views.

Generalisation

⇒ Anonymising data by removing specificity.

⇒ Ex:- Aggregation of data, by states or regions, can be done without individual person names & contact n.o.s.

⇒ Remove rows, where demog. rows are small, & indiv. can be identified.

⇒ Removing precision of data (categorical gen.)

a) Group small cities into larger regions, or nations.

⇒ Binning

a) Age column containing particular aged records, can be converted to age range (10-20 yrs, etc.)

b) Grouping of salary in ranges.

⇒ Truncation of IP addresses, phone no.s, emails (seen on UPI apps, banking apps & sms by banks.)

⇒ Rounding off.

These steps done on data, becomes irreversible to identify original.