Data Warehousing
Privacy and Security

Spring Term 2018
Dr. Andreas Geppert
geppert@acm.org
Outline of the Course

- Introduction
- DWH Architecture
- DWH-Design and multi-dimensional data models
- Extract, Transform, Load (ETL)
- Metadata
- Data Quality
- Analytic Applications and Business Intelligence
- Implementation and Performance
- Privacy and Security
Privacy

- **(US-)American Tradition**
  - no real constitutional right to privacy
  - 4th amendment: “…right of the people to be secure in their persons, houses, papers, and effects…”
  - «right to be left alone»

- **European/German Tradition**
  - privacy is a fundamental human right
  - understood as a person’s right to determine when to publish which personal information ("informationelles Selbstbestimmungsrecht")
  - based on the German Constitutional Court ruling about the German Census (1983)

- **Privacy as Contextual Integrity**
Privacy in Practice

- The current situation is characterized by organizations collecting huge amount of (personal) consumer data and exploiting these data in the course of their business (e.g., targeted marketing)
- data brokers are a new kind of company in the data landscape which collect, integrate, process, and distribute consumer data

Source: https://netzpolitik.org/2015/netzrueckblick-kommerzielle-ueberwachung-2015-digitales-tracking-echtzeit-datenhandel-und-scoring-
Privacy in Practice (2)

- 「Notice-and-consent」 has turned out to be a failed approach to establish privacy
- Anonymization is considered to be broken
- Consumers must actively strive for privacy (「digital aikido」)
- see, e.g., https://siprivacy.wordpress.com/
GDPR (General Data Protection Regulation)

- Datenschutzgrundverordnung (DSGVO)
- Approved by the European Union in 2016, in effect since May 25
- European law, directly binding in all member states
- binding not only for European companies, but for non-European companies as well, if they offer products or services within the EU or if they analyze people in the EU
- significant fines, up to 4% of global turnover
GDPR (General Data Protection Regulation)

- personal data: data that refer directly or indirectly to an identifiable person
- processing of personal data is generally forbidden
  - user provided consent with a clear and understandable purpose description
  - processing is required to fulfill the service
  - processing is in the legitimate interest of the processor
- Sensitive personal data (such as religion, sexual orientation or biometric data) can be processed only in exceptional cases
GDPR (2)

- permission to process data in return for an un-related service is no longer allowed
- consumers have the right to know which data companies have about them
- right to get forgotten
- data portability
Security

- Security Requirements
- Threat Model
- Authentication, Authorization, Auditing
- Data Protection
Security Requirements

- In general, there can be a goal conflict between consumer privacy and the company’s information needs.
- DWH must obey the respective laws and regulations:
  - data protection laws
  - industry-specific laws and regulation (e.g., Swiss banking secrecy, DFS, FINMA)
- Security makes sure that privacy of customers and consumers are guaranteed.
- Particularly, DWH security protects personally identifiable information (PII), respectively potentially PII.
- Security also protects the company’s data assets.
Threat Model

- Security design always refers to a threat model, i.e., the threats that the security design is supposed to protect against.
- External threats to be addressed by the enterprise-wide security design and implementation (firewall etc).
- Here we focus on internal threats emanating from internal users, e.g.
  - data leakage
  - fraud
Authentication

- **users:**
  - database users vs. external users (e.g., enterprise users)
  - personal vs. non-personal/technical users
  - ideally use personal/enterprise users to ensure (better) auditability/traceability

- **authentication:**
  - ensures that the user has indeed the claimed identity
  - to authenticate, the user has to provide two or more credentials (e.g., username and password)

- **example user creation:**
  - create user scott identified by tiger;
Authorization

- authorization ensures that only users with the appropriate permissions and privileges perform actions on the data
- principle of least-privileged access: grant users only those privileges they really need to perform their function
- it is generally bad practice to assign permissions to users directly
- better create roles (or groups) according to required application functionality and assign the required privileges to the role
  - create role customer_analysis_role;
  - grant select on customer_table to customer_analysis_role; …
  - grant customer_analysis_role to scott;
Column-level Security

- some roles may (not) be allowed to see or modify certain attributes
- These columns can be excluded in grant statements
  - grant ALL on patients to doctors;
  - grant ALL (id, name, doctor) on patients to accountants;

<table>
<thead>
<tr>
<th>Id</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>HIV</td>
<td>Mabuse</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>Cancer</td>
<td>Frankenstein</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>Asthma</td>
<td>Eisenbart</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>Flu</td>
<td>Eisenbart</td>
</tr>
</tbody>
</table>
Row-level Security

- aka fine-grained access control
- some roles may (not) be allowed to see or modify certain rows

principle
- implement restriction to enforce as a function
- create function/policy and attach to table
- at runtime, queries against tables will be written against

create policy patients_pol on patients to doctors
using (lower(doctor) = current_user);

#eisenbart: select * from patients;

<table>
<thead>
<tr>
<th>id</th>
<th>age</th>
<th>diagnosis</th>
<th>doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>25</td>
<td>Asthma</td>
<td>Eisenbart</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>Flu</td>
<td>Eisenbart</td>
</tr>
</tbody>
</table>
Auditing

- collect information about “sensitive” actions in audit log
- sensitive actions can e.g. be reading or modifying personal data
- balance audit requirements vs. performance overhead
- analyze audit log to detect suspicious behavior
- log analysis as a typical Big Data use case
Anonymization

- Anonymization removes/replaces personally identifiable attribute values from a record
- The intention is that private information is no longer visible from the record, while the record can still be used for analysis
- Data sets often can be de-anonymized by combining the remaining attributes with other data sets
- Aggregation alone is also not always sufficient
- In DWHs, production data is typically not anonymized (why?). Copying to test/development typically requires an anonymization step
Encryption

- Encryption replaces the plaintext of sensitive data with a cypher
- Tablespace encryption
  - similar to disk encryption and protects against unauthorized copies of data files etc.
- Column encryption
  - encrypts individual sensitive columns. Encryption keys are with the application
  - protects against nosy/malicious DBAs
  - specific features may not be supported anymore, respectively may be restricted (e.g., indexing)
Data Separation

- Data separation separates sensitive from non-sensitive information, into distinct schemas or databases.
- The set of DBAs with access to sensitive data can thus be kept smaller or be restricted to DBAs with specific properties (e.g., location).
- Auditing can then be used to identify attempted access to sensitive data by untrusted DBAs, respectively such attempts can be blocked.
Summary

- Privacy as a fundamental right
- GDPR as a potentially game-changing regulation
- Adequate data management as a prerequisite for GDPR compliance
- Security is a crucial task in a DWH, as DWHs contain integrated, enterprise-wide, sensitive information, in particular personal information
- Use appropriate authentication, authorization, and auditing techniques to control access
- Understand tradeoffs of data protection techniques