Security and Privacy

Kristo Raun

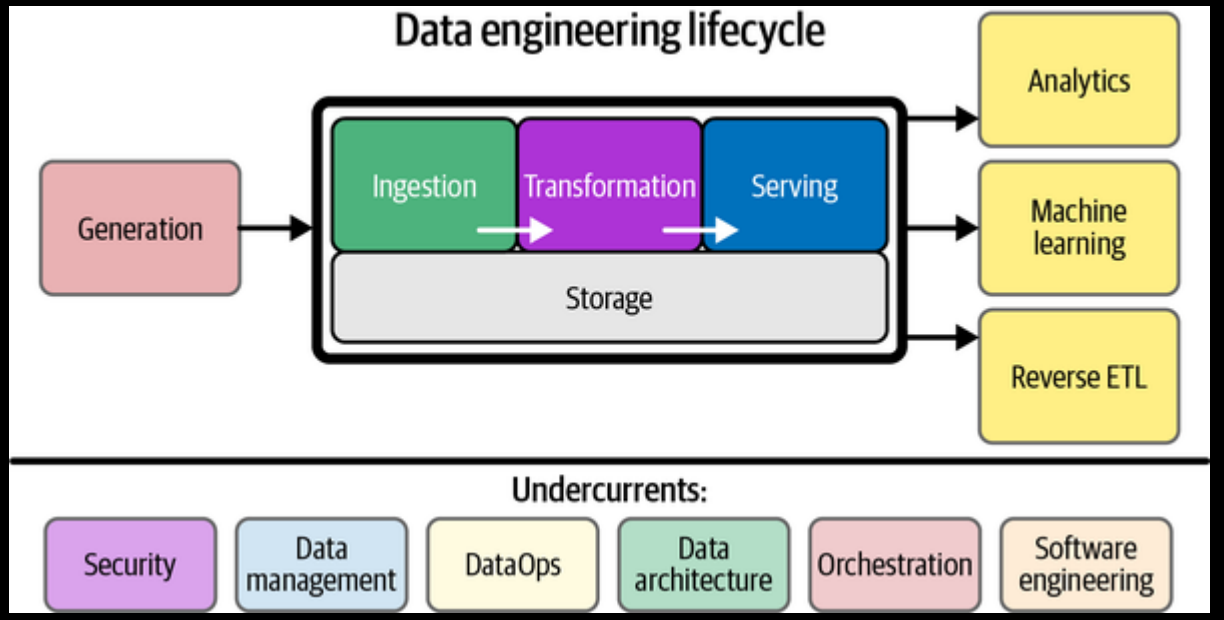
Data Engineering 2024 Fall

UNIVERSITY OF TARTU

Week	Lecture Date	Lecture Topic	Practice Date	Practice Topic
1	2024-09-02	No class	2024-09-05	No class
2	2024-09-09	Intro	2024-09-12	Docker+Postgres
3	2024-09-16	Data processing and orchestration	2024-09-19	Airflow
4	2024-09-23	Data modelling	2024-09-26	Kimball
5	2024-09-30	Data transformation	2024-10-03	dbt
6	2024-10-07	Data storage	2024-10-10	DuckDB
7	2024-10-14	NoSQL (*pre-recorded*)	2024-10-17	MongoDB (*online*)
8	2024-10-21	Data Lakes	2024-10-24	Delta, Iceberg
9	2024-10-28	Graph Databases (*online*)	2024-10-31	Neo4j (*online*)
10	2024-11-04	Security and privacy	2024-11-07	Security and privacy
11	2024-11-11	Data governance	2024-11-14	Open Metadata (*pre-recorded*)
12	2024-11-18	Key-Value stores (*online*)	2024-11-21	Redis (*online*)
13	2024-11-25	Data visualization	2024-11-28	Streamlit
14	2024-12-02	Exam	2024-12-05	Working in class (tutoring for project)
15	2024-12-09	Working in class (tutoring for project)	2024-12-12	Working in class (tutoring for project)
16	2024-12-16	Project presentation (poster session)	2024-12-19	Redo exam

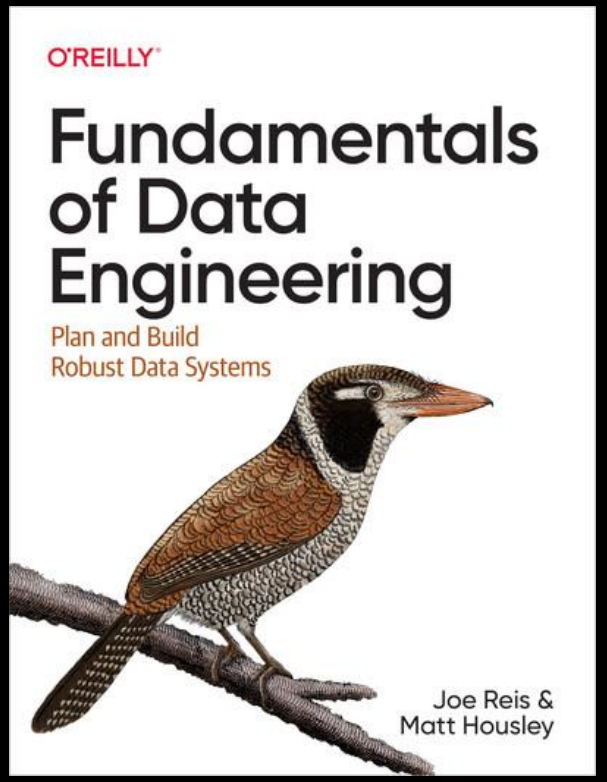
Week	Lecture Date	Lecture Topic	Practice Date	Practice Topic
1	2024-09-02	No class	2024-09-05	No class
2	2024-09-09	Intro	2024-09-12	Docker+Postgres
3	2024-09-16	Data processing and orchestration	2024-09-19	Airflow
4	2024-09-23	Data modelling	2024-09-26	Kimball
5	2024-09-30	Data transformation	2024-10-03	dbt
6	2024-10-07	Data storage	2024-10-10	DuckDB
7	2024-10-14	NoSQL (*pre-recorded*)	2024-10-17	MongoDB (*online*)
8	2024-10-21	Data Lakes	2024-10-24	Delta, Iceberg
9	2024-10-28	Graph Databases (*online*)	2024-10-31	Neo4j (*online*)
10	2024-11-04	Security and privacy	2024-11-07	Security and privacy
11	2024-11-11	Data governance	2024-11-14	Open Metadata (*pre-recorded*)
12	2024-11-18	Key-Value stores (*online*)	2024-11-21	Redis (*online*)
13	2024-11-25	Data visualization	2024-11-28	Streamlit
14	2024-12-02	Exam	2024-12-05	Working in class (tutoring for project)
15	2024-12-09	Working in class (tutoring for project)	2024-12-12	Working in class (tutoring for project)
16	2024-12-16	Project presentation (poster session)	2024-12-19	Redo exam

Week	Lecture Date	Lecture Topic	Practice Date	Practice Topic
1	2024-09-02	No class	2024-09-05	No class
2	2024-09-09	Intro	2024-09-12	Docker+Postgres
3	2024-09-16	Data processing and orchestration	2024-09-19	Airflow
4	2024-09-23	Data modelling	2024-09-26	Kimball
5	2024-09-30	Data transformation	2024-10-03	dbt
6	2024-10-07	Data storage	2024-10-10	DuckDB
7	2024-10-14	NoSQL (*pre-recorded*)	2024-10-17	MongoDB (*online*)
8	2024-10-21	Data Lakes	2024-10-24	Delta, Iceberg
9	2024-10-28	Graph Databases (*online*)	2024-10-31	Neo4j (*online*)
10	2024-11-04	Security and privacy	2024-11-07	Security and privacy
11	2024-11-11	Data governance	2024-11-14	Open Metadata (*pre-recorded*)
12	2024-11-18	Key-Value stores (*online*)	2024-11-21	Redis (*online*)
13	2024-11-25	Data visualization	2024-11-28	Streamlit
14	2024-12-02	Exam	2024-12-05	Working in class (tutoring for project)
15	2024-12-09	Working in class (tutoring for project)	2024-12-12	Working in class (tutoring for project)
16	2024-12-16	Project presentation (poster session)	2024-12-19	Redo exam



Reading

- Chapter X



Agenda

- Security
- Privacy
- Trivia
- Quiz session

Motivation?

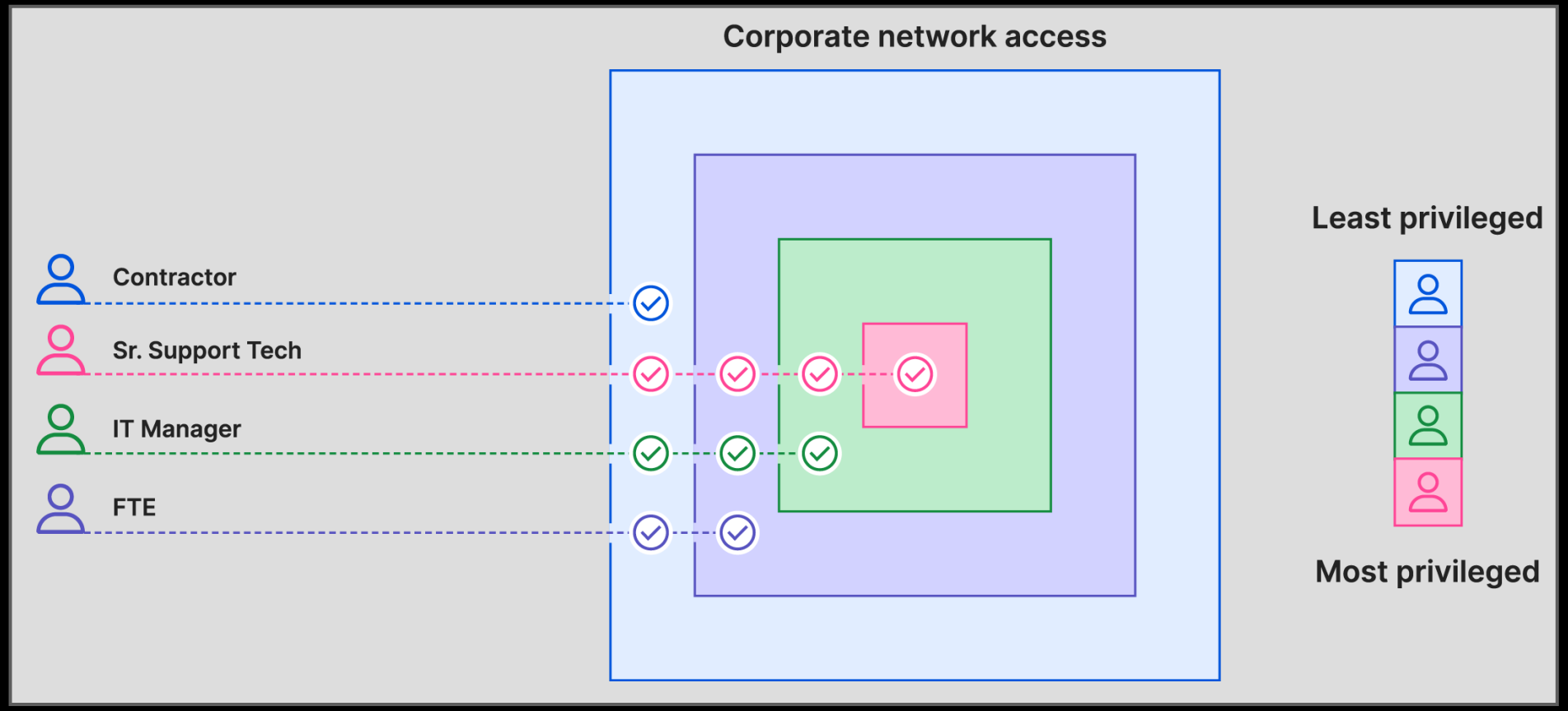
https://en.wikipedia.org/wiki/List_of_data_breaches

Security

- Access control
  - IAM – Identity and Access Management
- Data encryption
  - Data at rest and in transit should be unreadable
- Monitoring
  - Ensure that activities are logged (IP, user, activity type and timestamp, ...)

Security – Access control (IAM)

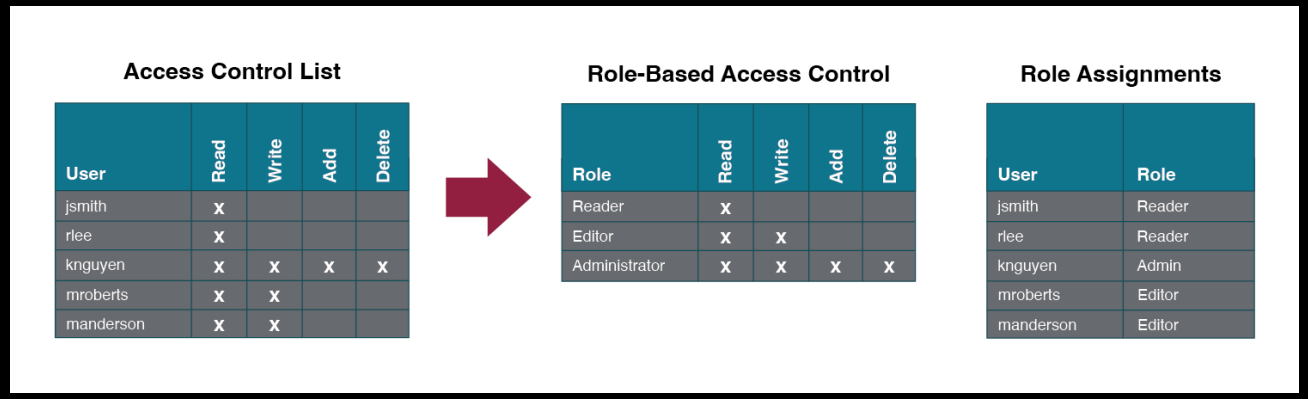
- Principle of Least Privilege



https://www.cloudflare.com/en-gb/learning/access-management/principle-of-least-privilege/

Security – Access control (IAM)

- Access control models
  - Access Control List (ACL)
  - Role-Based Access Control (RBAC)
  - Attribute-Based Access Control (ABAC)



Access Control List

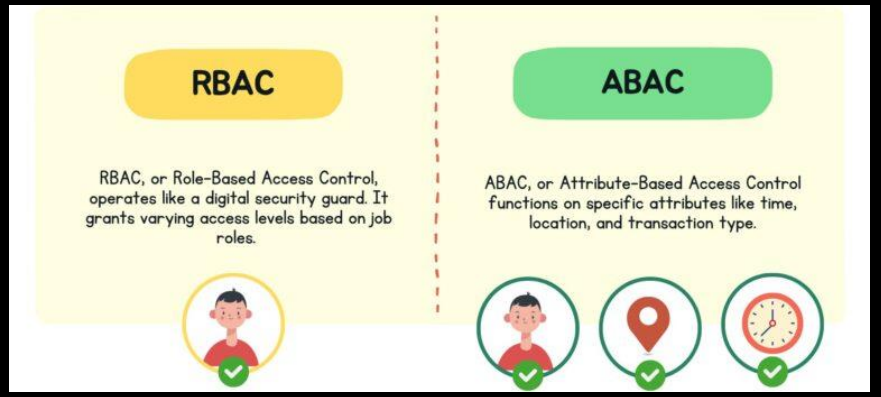
User	Read	Write	Add	Delete
jsmith	x			
rlee	x			
knguyen	x	x	x	x
mroberts	x	x		
manderson	x	x		

Role-Based Access Control

Role	Read	Write	Add	Delete
Reader	x			
Editor	x	x		
Administrator	x	x	x	x

Role Assignments

User	Role
jsmith	Reader
rlee	Reader
knguyen	Admin
mroberts	Editor
manderson	Editor



https://eng.libretexts.org/Courses/Delta_College/Master_CS11_POCR/08:_Information_Systems_-_Security/8.05:_IDS

https://www.airdroid.com/mdm/role-based-access-control-rbac/

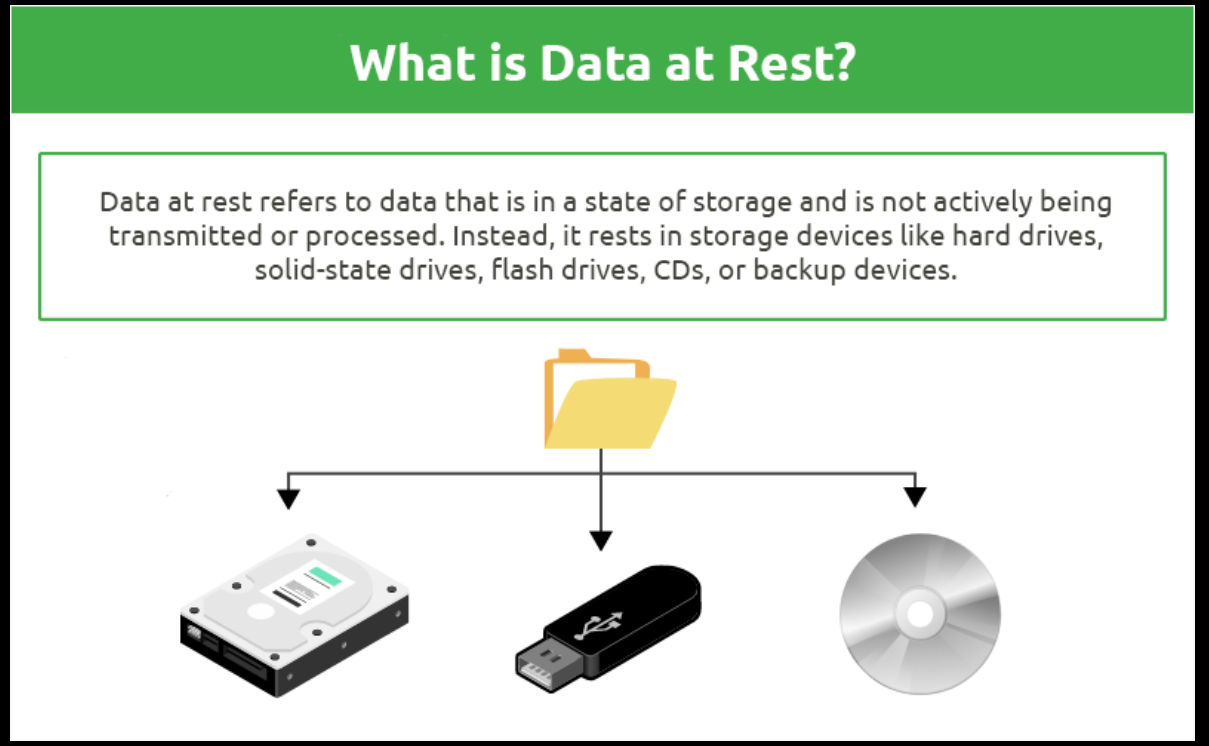
Security – Access control (IAM)

- Authentication
  - Multi-Factor Authentication (MFA)
    - If possible – on everything. Including server access, etc
  - Single Sign-On (SSO)
    - Ideally, on everything
  - Password requirements
    - Types of symbols required? ✖
    - Length? Mandatory min 8, recommended min 15
    - How often to change password? ✖
  - Use password manager

https://cybersecuritynews.com/nist-rules-password-security/

Security – Encryption

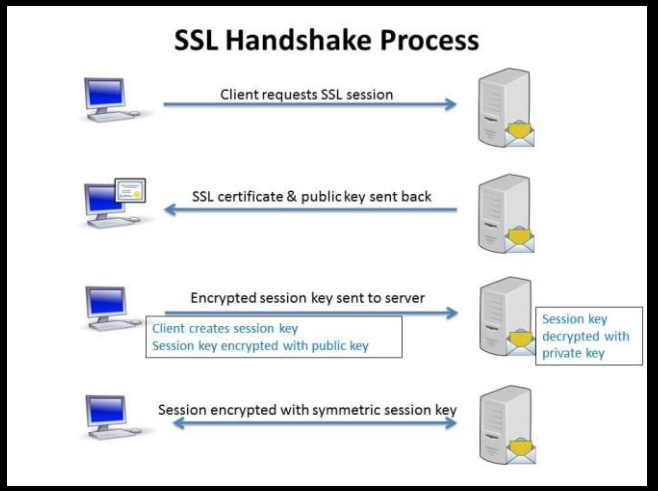
- Data at rest
  - Filesystems
    - Bitlocker, FileVault, etc
  - Cloud (object storage)
    - Usually enabled by default, but verify
  - Databases
    - Transparent Data Encryption (TDE)

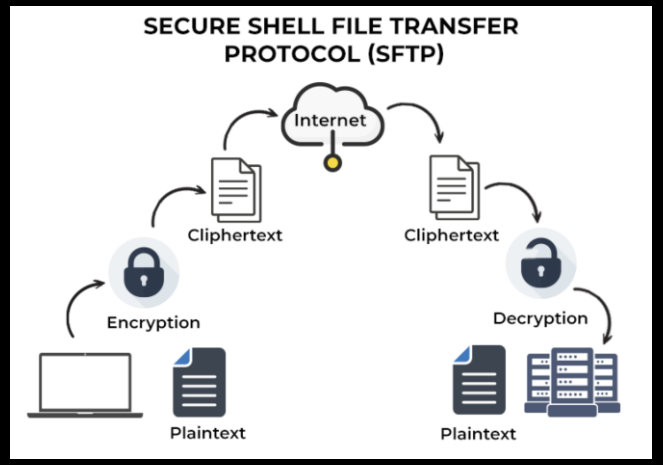


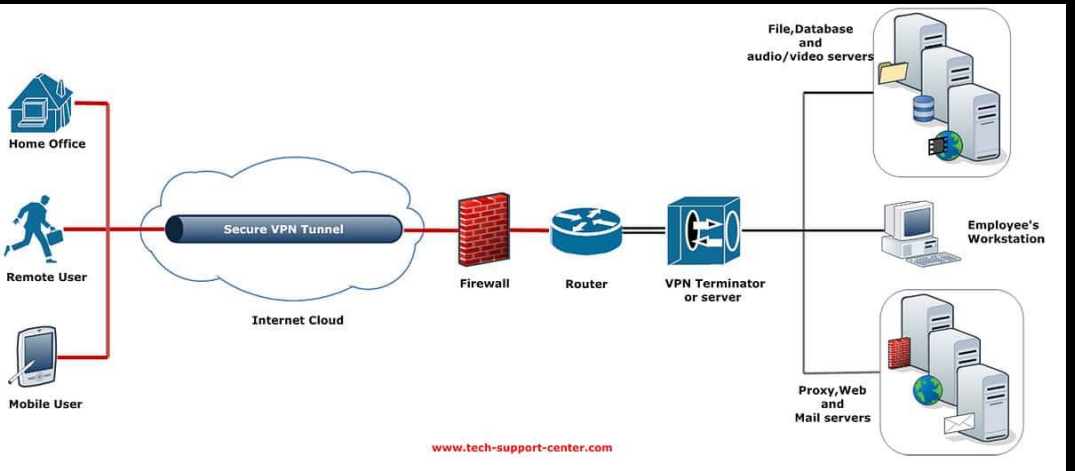
https://www.titanfile.com/blog/data-at-rest-encryption/

Security – Encryption

- Data in transit
  - TLS/SSL (https)
    - Transport Layer Security / Secure Sockets Layer
  - Use SFTP
    - Secure File Transfer Protocol, builds on SSH (Secure Shell)
  - Use VPN
    - Virtual Private Network







https://blog.mdaemon.com/ssl-tls-best-practices

https://intellihr.zendesk.com/hc/en-us/articles/6191307494543-SFTP-and-intelliHR-Explained

https://www.macobserver.com/tips/deep-dive/vpn-can-help/

Security – Monitoring

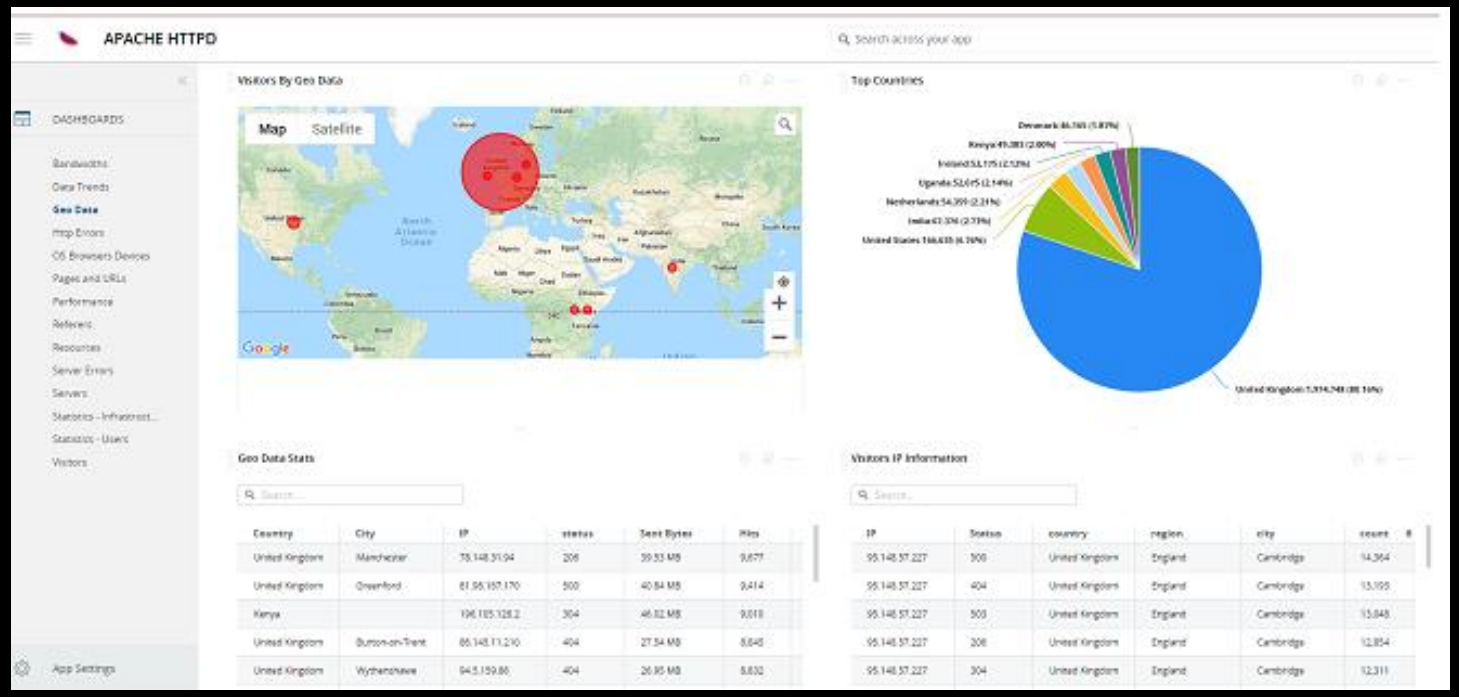
- Access logs
  - Enable logging of activities

```
192.168.1.1 - - [11/Oct/2018:11:36:24 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:25 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:25 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:27 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:28 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:29 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:31 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:32 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:34 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:35 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:35 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:37 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:39 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:39 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:40 +0200] “GET /index.php/apps
192.168.1.1 - - [11/Oct/2018:11:36:42 +0200] “GET /index.php/apps
```

https://10web.io/glossary/access-logs/

Security – Monitoring

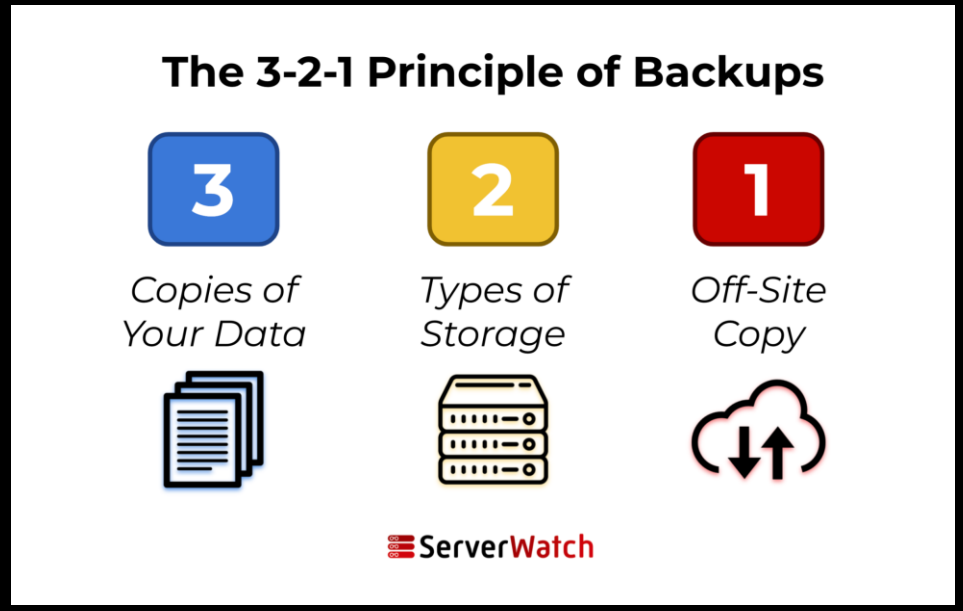
- Tracking
  - Monitor the logs!
  - Create alerts on suspicious behavior



https://www.xplg.com/access-log-viewer-analyzer/

Security – Monitoring

- Incident response
  - Regularly create backups of critical data
  - Make sure this data is stored separately BUT encrypted!



https://www.serverwatch.com/guides/server-backup/

Agenda

- ~~Security~~
- Privacy
- Trivia
- Quiz session

Privacy

- Data minimization
  - Collect only what's necessary
    - No data = no problem
  - Data retention
    - How long is it necessary to keep this data?
    - Do customers/employees have a right to delete this data?

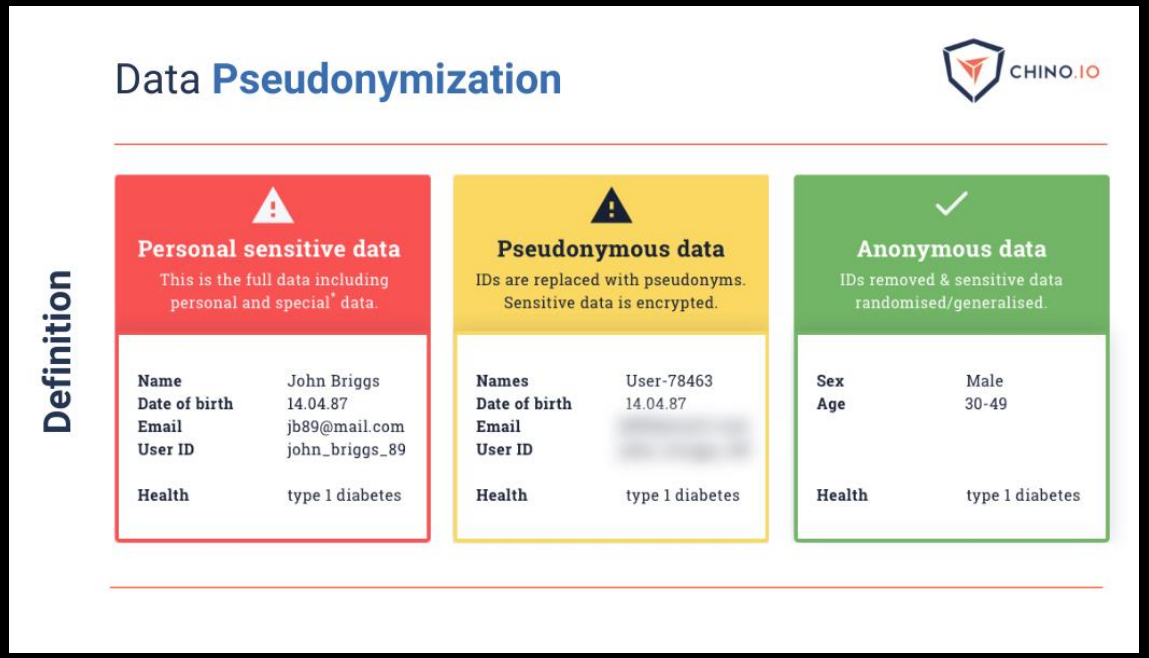
Human Resources Retention Schedule

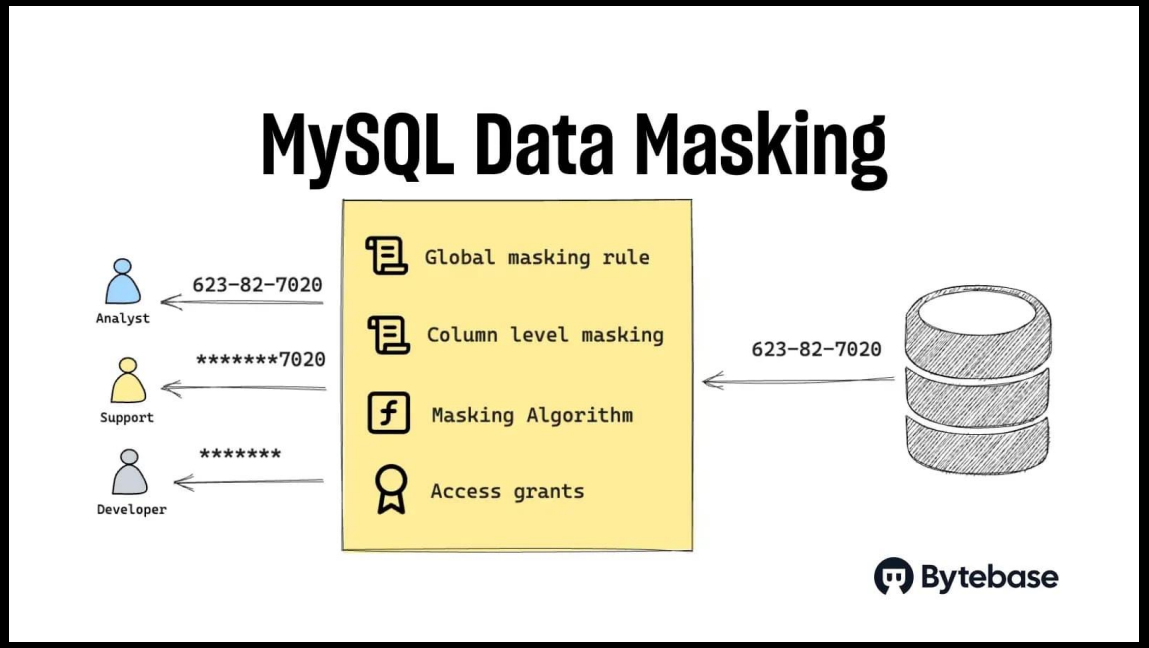
HR.1	Administering Employees		As the Employee file			
HR.1.1	Absence Monitoring		Records documenting an employees absence due to sickness			
Ref.	Series	Doc type Examples	Trigger	Retention	Action	Rationale / Comments
HR.1.1.1	Employee File	Absence record due to sickness: Statutory Sick Pay records, calculations, certificates, self-certificates, fit notes	End of the tax/financial year to which they relate	4 years	Destroy	The Statutory Sick Pay (General) Regulations 1982 (SI 1982/894) as amended plus 1 year for local requirements. All documents to be sent to HR Service Centre. **N.B.** Record of actual sick pay part of payroll record and not included here.
HR.1.1.2	Health File	Complete Record of Sickness	Termination of Employment	Overall minimum retention period applies: 6 or 10 or 25 years	Destroy	Contains Sensitive data under Data Protection Act to be kept as part of Health record not as part of employee record. Information held to be relevant and up-to-date.
HR.1.1.3	Case File	Ill Health case work (For Asbestos/Hazardous substances/Ionising Radiations/lead/Major Injury at work please see longer retention at HR.5.3)	Termination of Employment	Overall minimum retention period applies: 6 or 10 or 25 years	Destroy	Contains Sensitive data under Data Protection Act. Files held by both Occupational Health and HR Advisory team not as part of employee record. Information to be relevant and up-to-date.
See also HR.1.9 Leave and HR.5 Occupational Health.						

https://templatesumo.com/business/data-retention-policy-template-the-essential-guide-to-gdpr/

Privacy

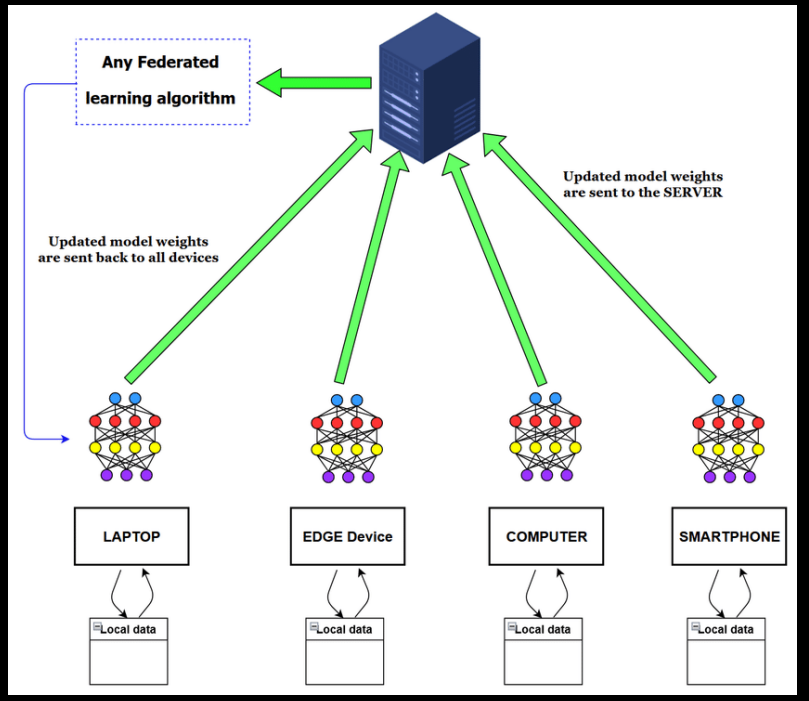
- Data anonymization
  - Removing or modifying data so it can't be tracked to a person
- Pseudonymization
  - Same as anonymization, but having a mapping back to the original person
- Masking
  - Partially obscured, but authorized users may be able to access the original data

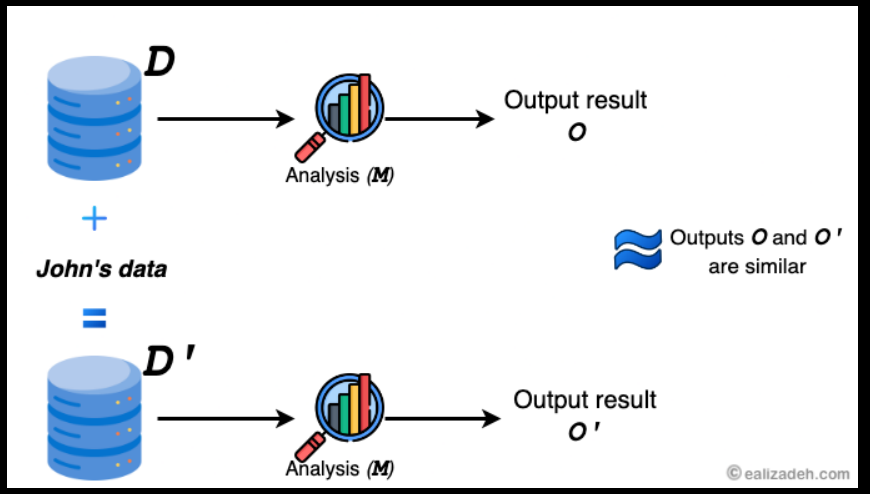




Privacy

- State-of-the-art concepts
  - Federated learning
    - Only aggregates/parameters shared with a central server
  - Differential privacy
    - Adding noise to data
- Mostly relevant for ML
- Less applicable for analytics or BI





Application of federated learning in manufacturing. Hegiste, Legler, Ruskowski

https://ealizadeh.com/blog/abc-of-differential-privacy/

Agenda

- ~~Security~~
- ~~Privacy~~
- Trivia
- Quiz session

Trivia

- What's 127.0.0.1 ?

Trivia

- What's 0.0.0.0/0 ?

Trivia

- What's pandas==2.2.2 ?

- https://osv.dev/

Trivia

- What’s chmod 777 ?

Trivia

- What’s sudo ?

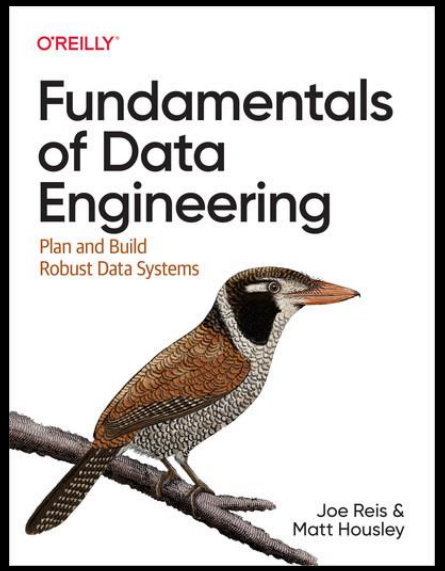
Trivia

- What’s .gitignore ?

Agenda

- ~~Security~~
- ~~Privacy~~
- ~~Trivia~~
- Quiz session

Further reading



- Chapter X
  - People
  - Processes
  - Technologies
- Article
  - Data Security In Data Engineering on Google Cloud https://medium.com/google-cloud/data-security-in-data-engineering-on-google-cloud-4938074ee005