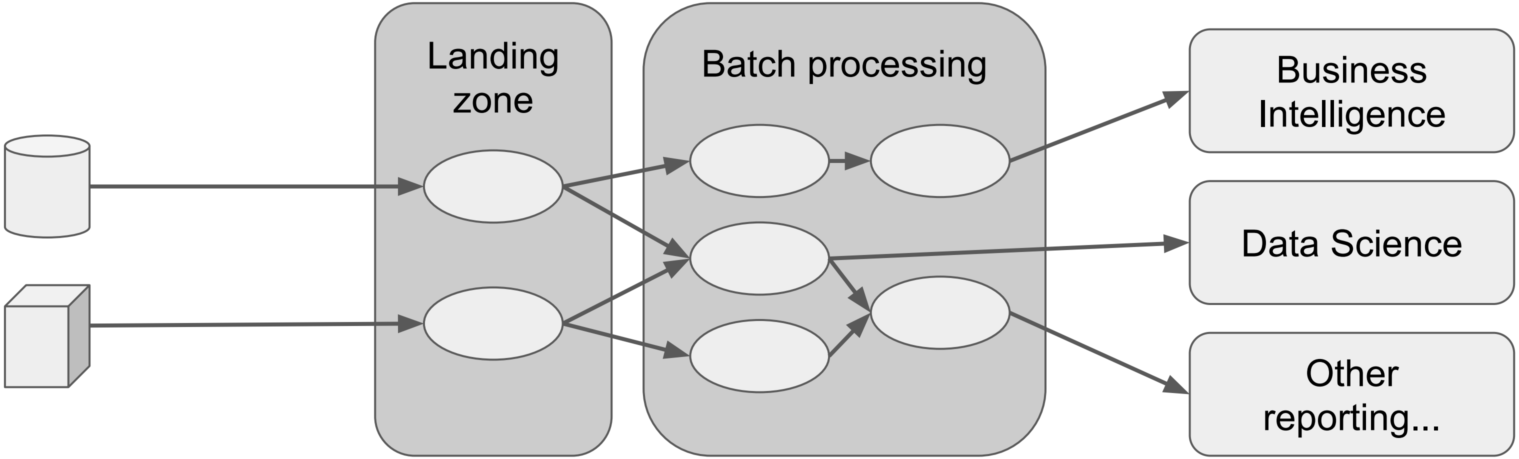
Building a data lake at a bank

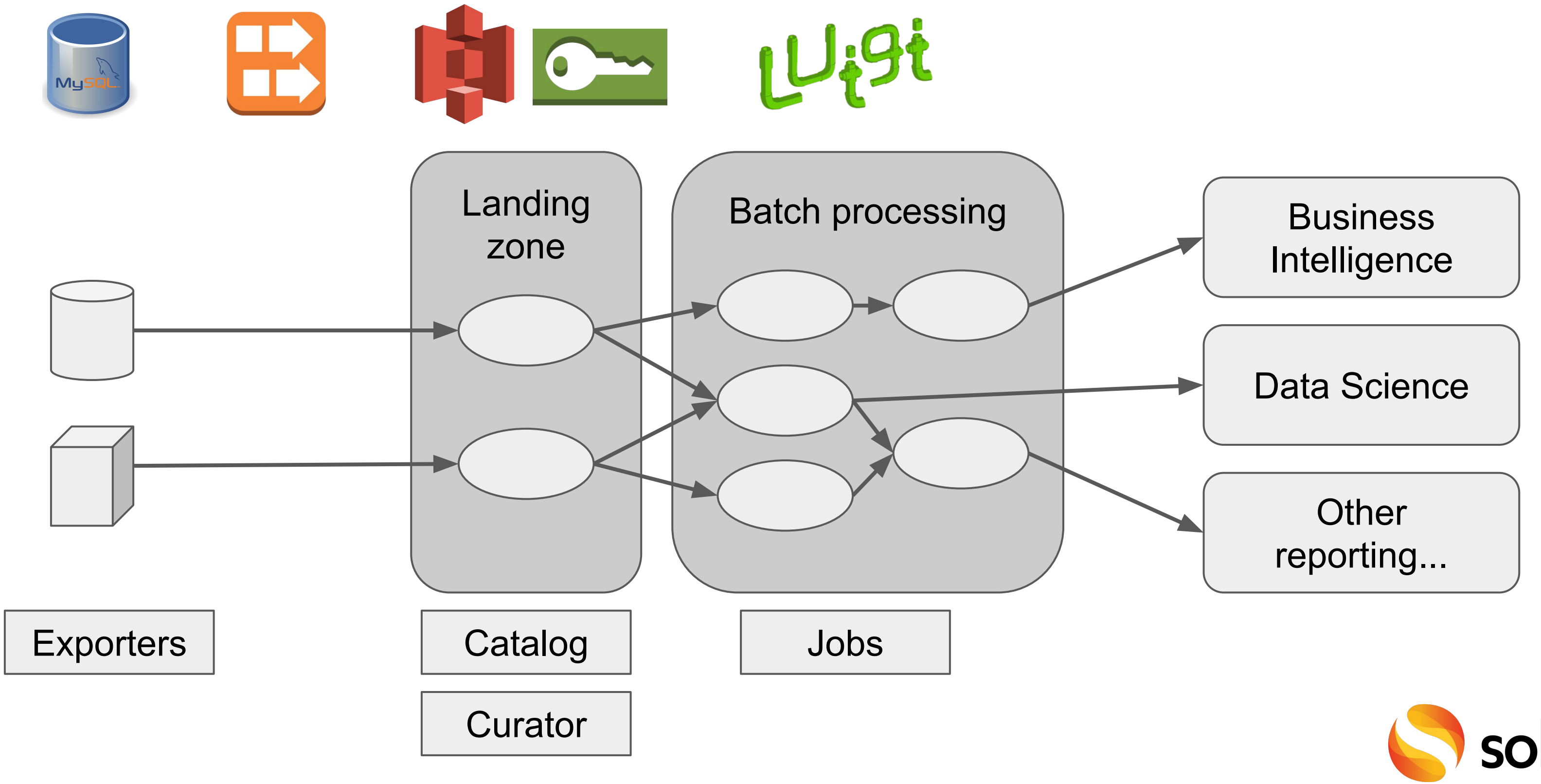
solarisBank

Overview



solarisBank

Overview



Landing zone

Batch processing

Business Intelligence

Data Science

Other reporting...

Exporters

Catalog

Curator

Jobs

Regulation / GDPR

Disclaimer
Danger Zone: tech guy talking about the law!
Ask your local lawyer for dependable advice!

Typical technical requirements

Regulation / GDPR

Typical technical requirements

- Encryption
  - At rest
  - In transit

Regulation / GDPR

Typical technical requirements

- Encryption
  - At rest
  - In transit
- Strong authentication (MFA)

Regulation / GDPR

Typical technical requirements

- Encryption
  - At rest
  - In transit
- Strong authentication (MFA)
- Auditability, automation and version control

HashiCorp Terraform

Concourse

amazon web services

Regulation / GDPR

Typical technical requirements

- Encryption
  - At rest
  - In transit
- Strong authentication (MFA)
- Auditability, automation and version control
- Roles, separation of concerns

Regulation / GDPR

Typical technical requirements

- Encryption
  - At rest
  - In transit
- Strong authentication (MFA)
- Auditability, automation and version control
- Roles, separation of concerns
- Pseudonymization/masking of data

Regulation / GDPR

Typical technical requirements

- Encryption
  - At rest
  - In transit
- Strong authentication (MFA)
- Auditability, automation and version control
- Roles, separation of concerns
- Pseudonymization/masking of data
- Other IT security measures

Regulation / GDPR

Typical technical requirements

- Encryption
  - At rest
  - In transit
- Strong authentication (MFA)
- Auditability, automation and version control
- Roles, separation of concerns
- Pseudonymization/masking of data
- Other IT security measures

“Easy” part

solarisBank

Data protection by design and by default (Art. 25)

Data protection by design and by default (Art. 25)

Bla bla bla …

- implement appropriate technical and organisational measures …
- which are designed to implement data-protection principles …
- and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation

… bla bla bla

CODE

GDPR Principles

High level GDPR (Art. 5)

- Lawful processing
  - Purpose: you must have a good reason for processing
  - Legal basis: you also need a reason why your purpose is legal
- Data minimization
  - Collection and processing: only use what you really need
  - Deletion: once you are done, you must delete
- Accountability
  - onus of proof is on the company

Lawful processing

Art. 6

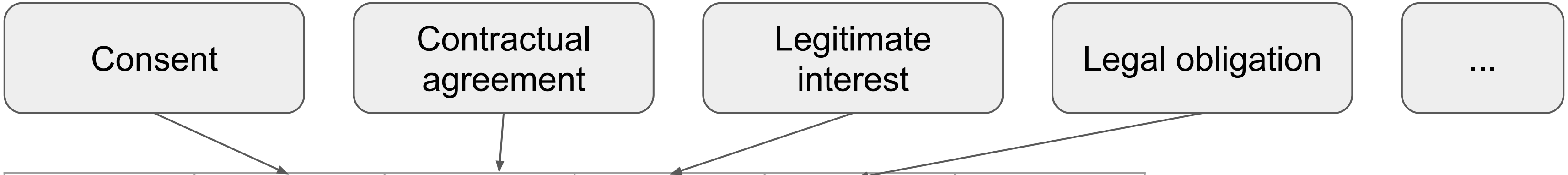
solarisBank

Lawful processing

Art. 6

- Consent
- Contractual agreement
- Legitimate interest
- Legal obligation
- ...

Lawful processing



	Legal Basis 1	**Legal Basis 2**	**Legal Basis 3**	**Legal Basis 4**	**...**
Cust. 1		✔	✔	✔	...
Cust. 2	✔		✔	✔	...
Cust. 3	✔	✔		✔	...
Cust. 4	✔	✔	✔		...
...	...	...	...	...	...

Withdraws consent

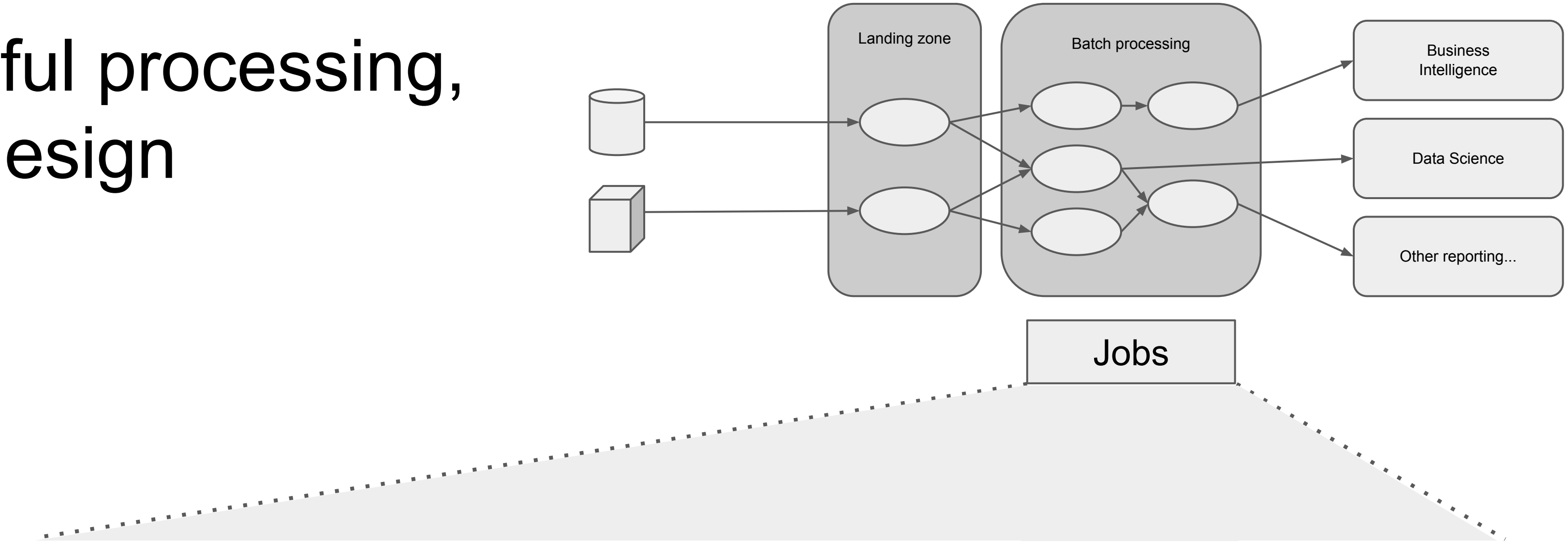
Doesn’t want all your products

Objects to processing

Lives in different country

Art. 12-23

Lawful processing, by design



Landing zone

Batch processing

Business Intelligence

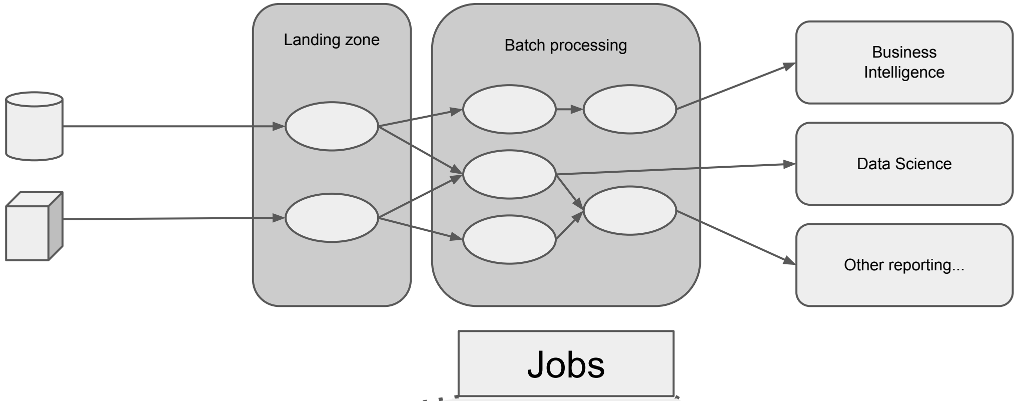
Data Science

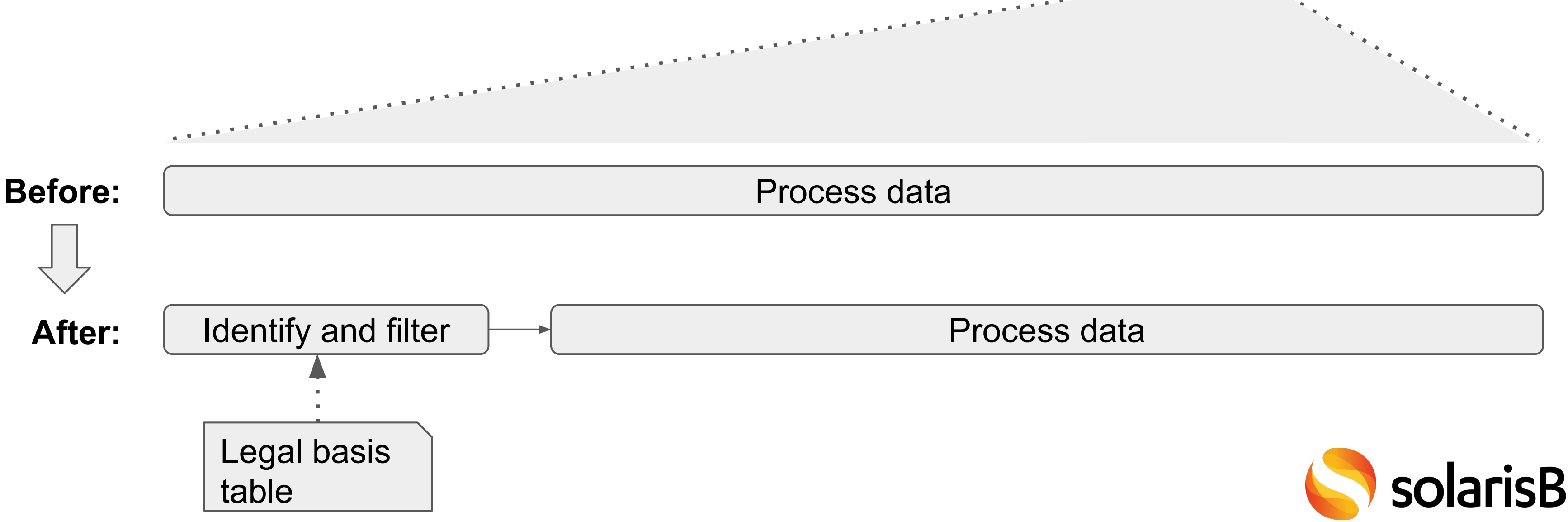
Other reporting...

Jobs

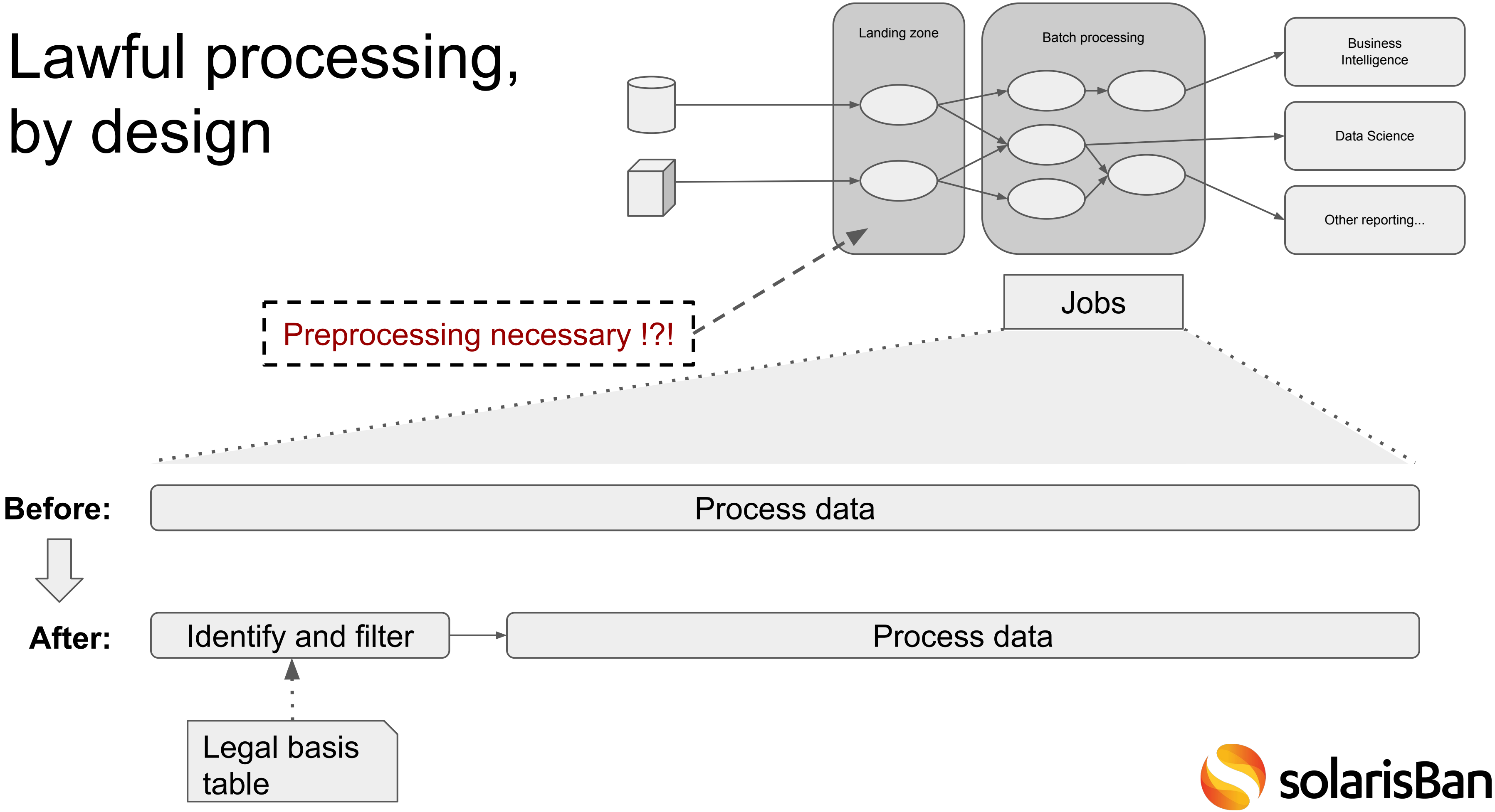
Before: Process data

Lawful processing, by design





Lawful processing, by design



Landing zone

Batch processing

Business Intelligence

Data Science

Other reporting...

Preprocessing necessary !?!

Jobs

Before:

Process data

After:

Identify and filter

Process data

Legal basis table

Data minimization

Only store/use personal data you need, and store it only as long as you need it.

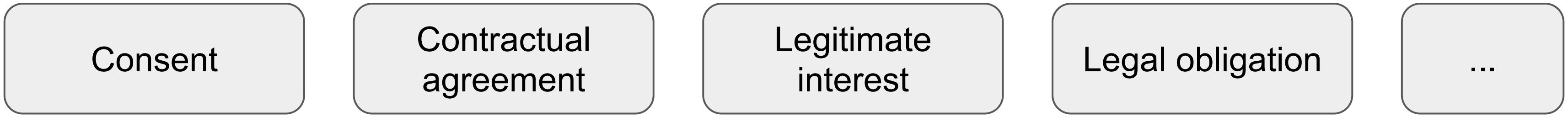
→ Use only the data you really need: very difficult topic… (e.g, machine learning)

→ Implement data retention policies everywhere: deletion!

Data minimization

Only store/use personal data you need, and store it only as long as you need it.

→ Implement data retention policies everywhere



Consent

Contractual agreement

Legitimate interest

Legal obligation

...

Triggers: every legal basis may an expiry date

- Consent can be withdrawn
- Contracts may end
- Legitimate interests may be subject to objection (Art. 21)
- Legal obligations usually have a time-frame
- ...

→ Definitely stop processing

→ Maybe also delete

Data minimization - deletion

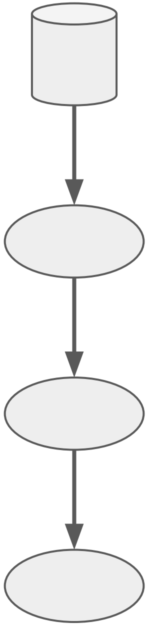
Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

- Could be problematic
  - Ideal: “Functional style”
  - Immutable intermediate results, often materialized
  - Deletion by “marking as deleted” does not count

...



Data minimization - deletion, by design

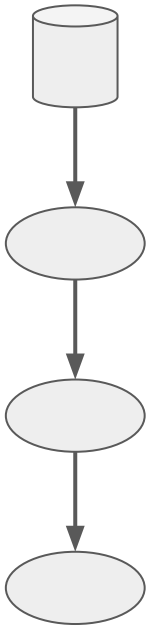
Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

- Could be problematic
    - Ideal: “Functional style”
    - Immutable intermediate results, often materialized
    - Deletion by “marking as deleted” does not count

    ...
- Only 3 ways out:



Data minimization - deletion, by design

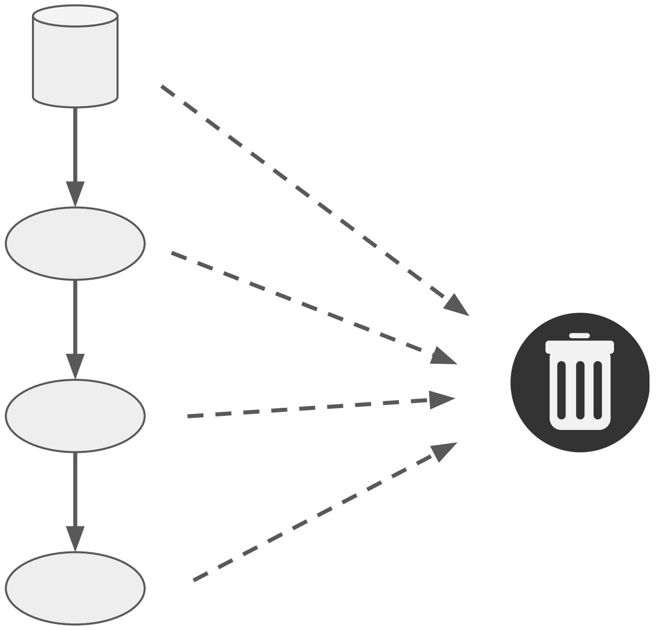
Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

- Could be problematic
    - Ideal: “Functional style”
    - Immutable intermediate results, often materialized
    - Deletion by “marking as deleted” does not count

      ...
- Only 3 ways out:
    - Low enough retention by default



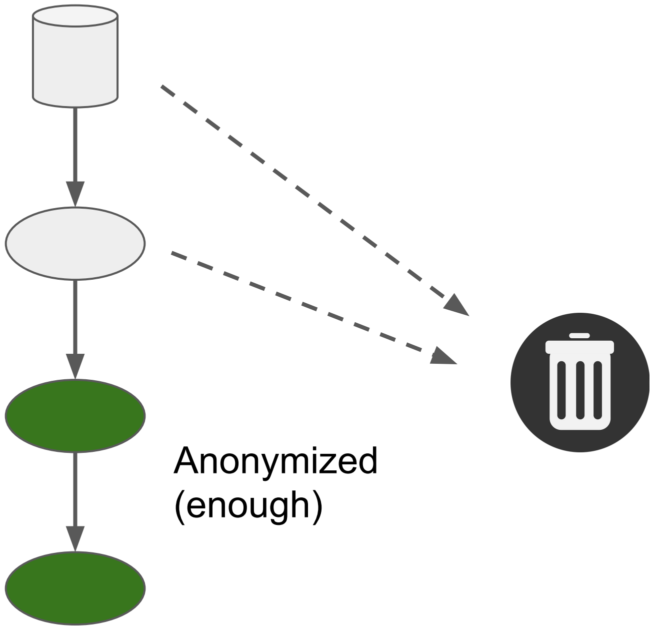
Data minimization - deletion, by design

Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

- Could be problematic
  - Ideal: “Functional style”
  - Immutable intermediate results, often materialized
  - Deletion by “marking as deleted” does not count
  - ...
- Only 3 ways out:
  - Low enough retention by default
  - Make data anonymized



https://2017.berlinbuzzwords.de/17/session/protecting-privacy-practice

Data minimization - deletion, by design

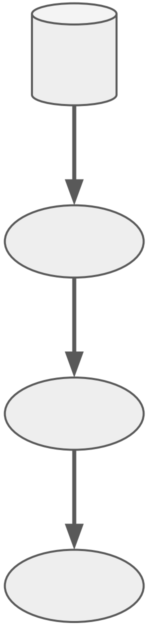
Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

- Could be problematic
  - Ideal: “Functional style”
  - Immutable intermediate results, often materialized
  - Deletion by “marking as deleted” does not count

  ...
- Only 3 ways out:
  - Low enough retention by default
  - Make data anonymized
  - Clean up after you



solarisBank

Data minimization - deletion, by design

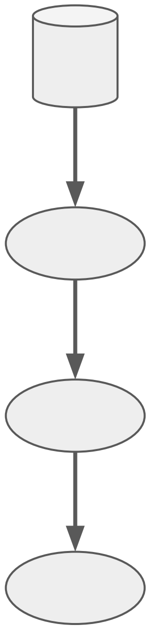
Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

- Could be problematic
  - Ideal: “Functional style”
  - Immutable intermediate results, often materialized
  - Deletion by “marking as deleted” does not count

  ...
- Only 3 ways out:
  - Low enough retention by default
  - Make data anonymized
  - Clean up after you



Clean up after you

Recomputation

- Clean up source
- Re-run downstream jobs

https://2017.berlinbuzzwords.de/17/session/protecting-privacy-practice

Data minimization - deletion, by design

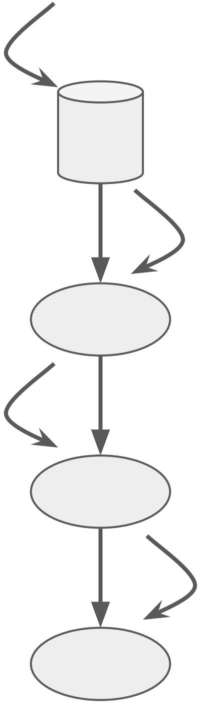
Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

- Could be problematic
  - Ideal: “Functional style”
  - Immutable intermediate results, often materialized
  - Deletion by “marking as deleted” does not count

  ...
- Only 3 ways out:
  - Low enough retention by default
  - Make data anonymized
  - Clean up after you



Clean up after you

Recomputation

- Clean up source
- Re-run downstream jobs

Cleaning pipelines

- Additional job cleans up source data and all intermediate/end results

solarisBank

https://2017.berlinbuzzwords.de/17/session/protecting-privacy-practice

Data minimization - deletion, by design

Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

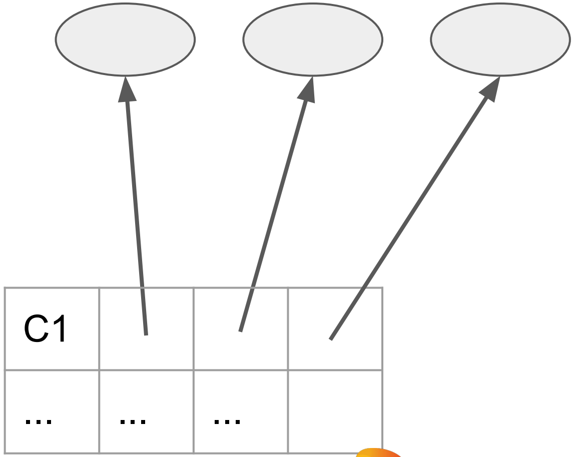
Data Pipelines

- Could be problematic
  - Ideal: “Functional style”
  - Immutable intermediate results, often materialized
  - Deletion by “marking as deleted” does not count
  - ...
- Only 3 ways out:
  - Low enough retention by default
  - Make data anonymized
  - Clean up after you

Factor out

Join table

- Assign a pseudonym per dataset and person
- Maintain one table to join all pseudonyms



C1			
...	...	...	

Data minimization - deletion, by design

Transactional systems:

- Hopefully easy, if backed by something like RDBMS
- Look at DB schema, get and/or delete all data about a person

Data Pipelines

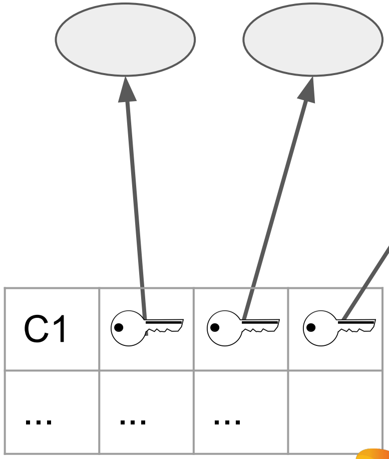
- Could be problematic
  - Ideal: “Functional style”
  - Immutable intermediate results, often materialized
  - Deletion by “marking as deleted” does not count

  ...
- Only 3 ways out:
  - Low enough retention by default
  - Make data anonymized
  - Clean up after you

Factor out

Encryption key join table

- “Lost key pattern”
- Assign an encryption key per dataset and person
- Maintain one table to join all keys



C1	🔑	🔑	🔑
...	...	...	

Compliance as code

Smaller Word documents = world is a better place