codecentric

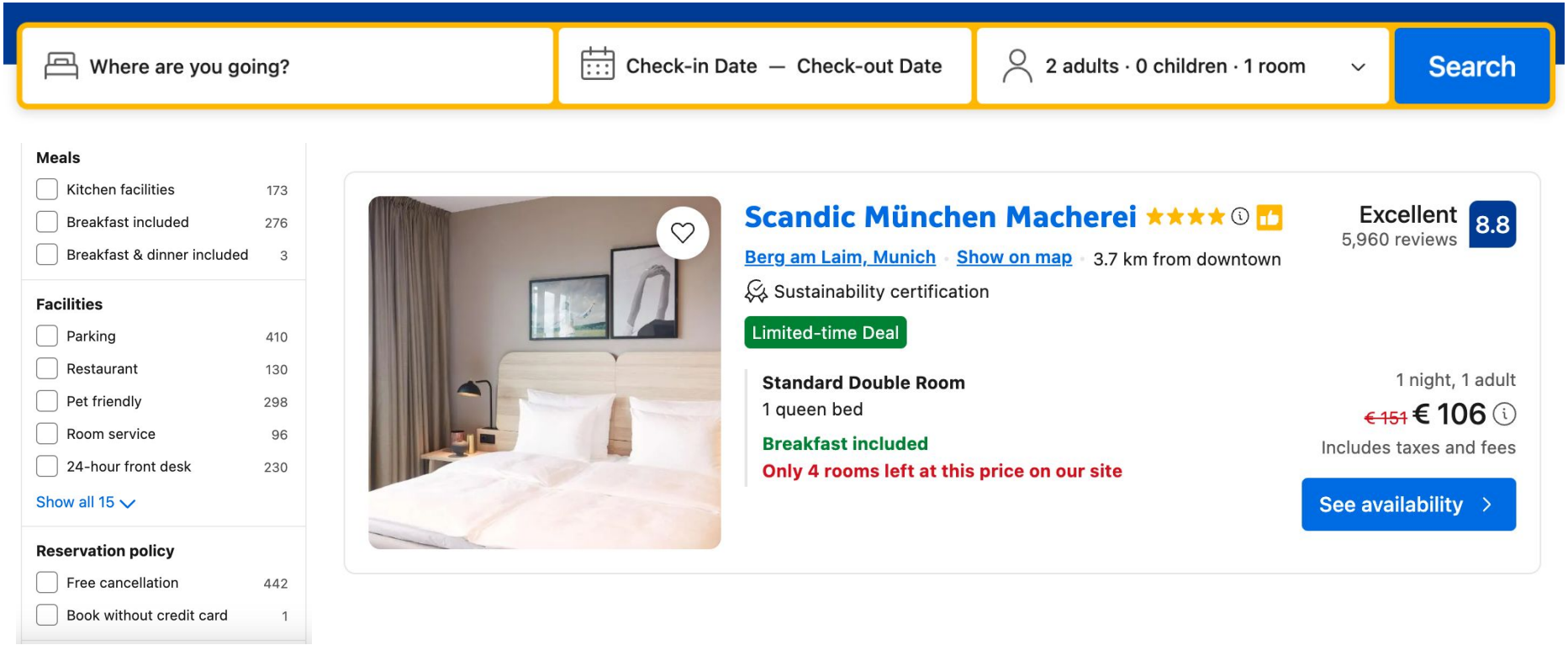
Data-Aware Architectures





Data? That's what the others are doing

Which data and user input is stored?



Changing data

Store the order

order_id	quantity
1001	1

Get order_id and quantity

Mark unfulfilled orders

order_id	quantity	unfulfilled
1002	3	true

Why are number of orders larger than stock?

Measure unfulfilled items

order_id	quantity	unfulfilled
1003	3	1

Breaking pipelines

Only the technical schema of the data is exposed

The data team ends up reverse engineering the business process & data model and isn’t sure if t can be trusted

And the producers don’t even know who uses the data for what purpose

From a business perspective, they belong together

Trustworthy, high-quality data is needed from operational systems towards analytics

- enables faster, better, data-based decisions
- provide new functionality towards users
- improve revenue/costs

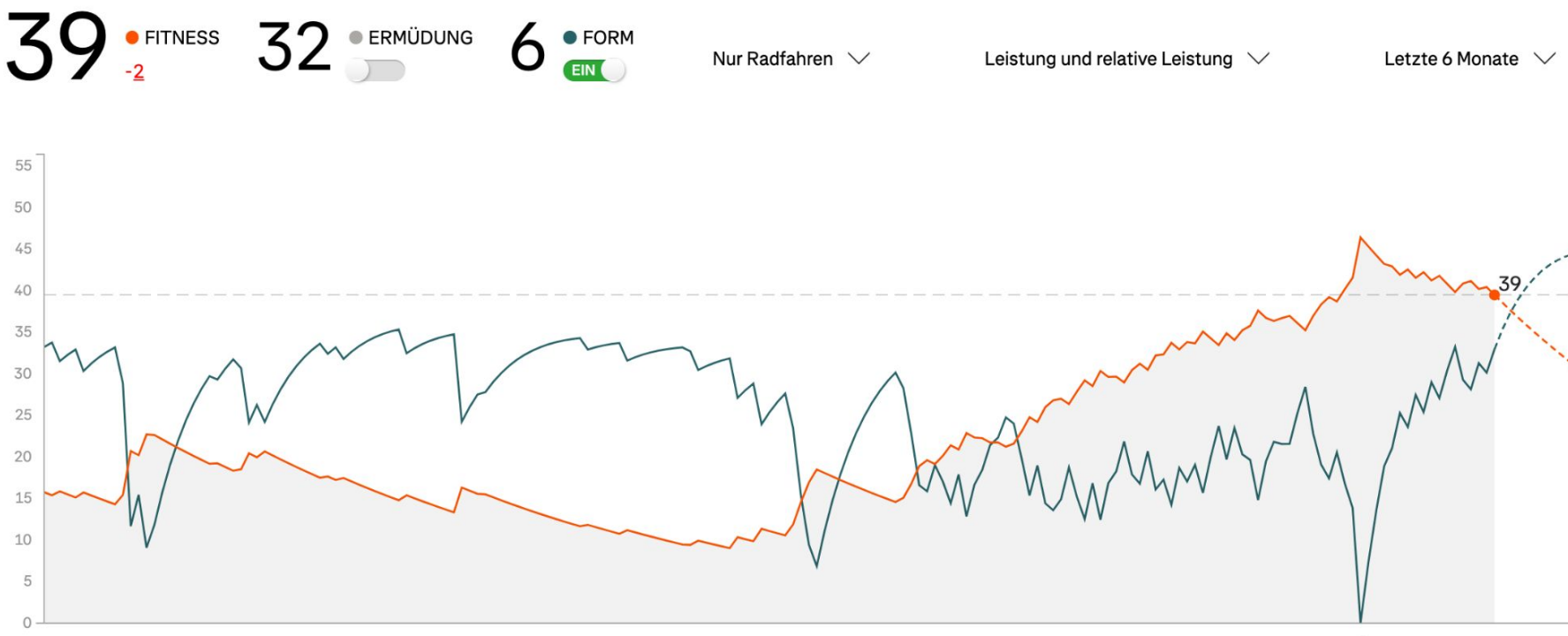
The gap between application and data affects trust in data



Analyse collected data to improve the application



Reporting & analytical capabilities for application users



Integrate ML models into applications



Not only: Application → Analytics

But also: Analytics → Application

Still: Data as second-class citizen

- Product Owner/Engineers doesn’t know or understand the case for data
- Competing with further requirements

⇒ providing data often ends in quick one-off solutions

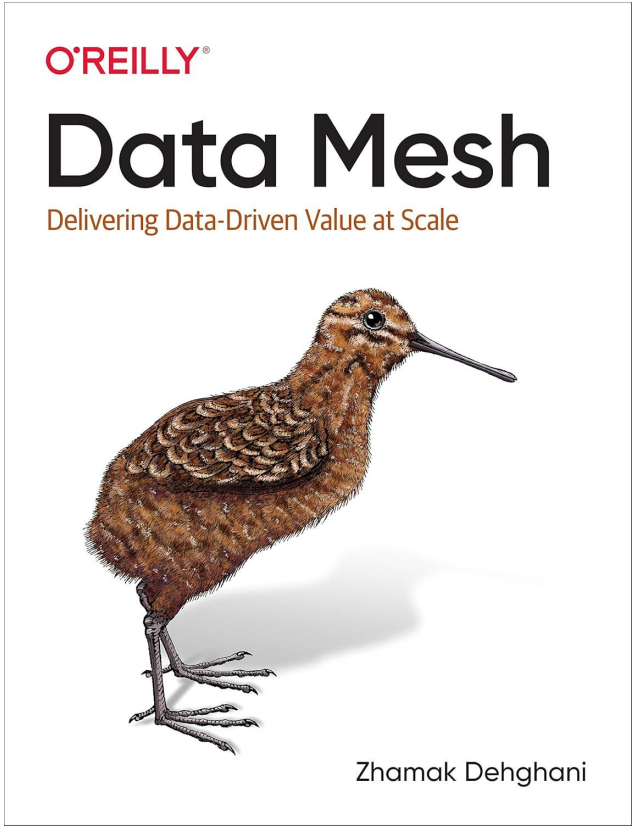
even if data is used for analytical purposes in team:

it mostly stays within the team and isn’t provided to the outside





It’s time for a Data Aware Architecture, covering technology and organisation


O'REILLY
Data Mesh
Delivering Data-Driven Value at Scale
Zhamak Dehghani

What is “Data Aware” in context of Architecture?

Consciously consider the storage, management and accessibility of data when making architectural decisions - both for internal & external usage as well as for operational & analytical purposes

Documentation

Storage Implementation

Schema changes

Quality & SLAs

Analytical usage

Privacy

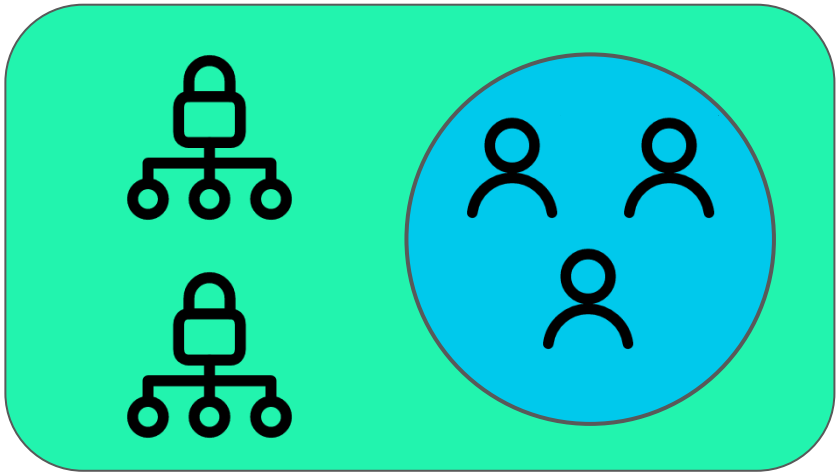
Interface for mass data

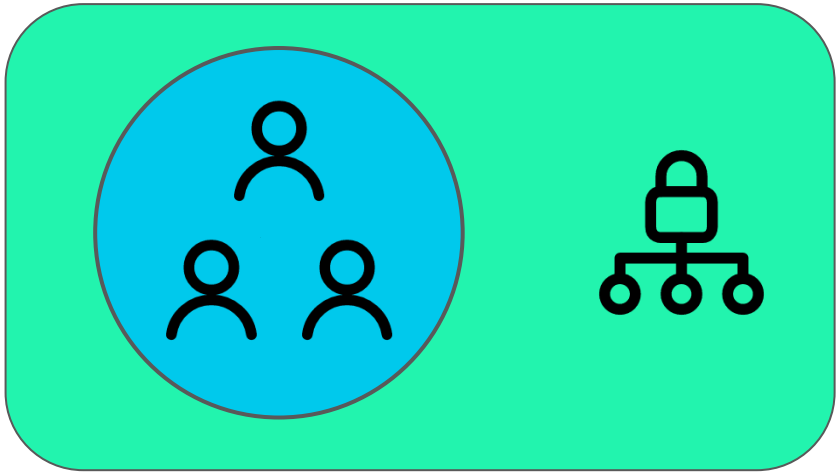
Handle new data

It needs a foundation



Domain / Value Stream Teams





Products with data

- Data as a part of the product
  - Forecasting, Recommender, Self Service Reporting: product features
  - Requirements from other areas, other stakeholders; different goals, culture and technology
- PO/PMs: accept the new requirements
  - include them equally in backlog
  - demand support

Data competency

Technical

- How to do & build stuff: tooling, frameworks, languages, methods
- Another skillset next to frontend, backend, …

Methodical

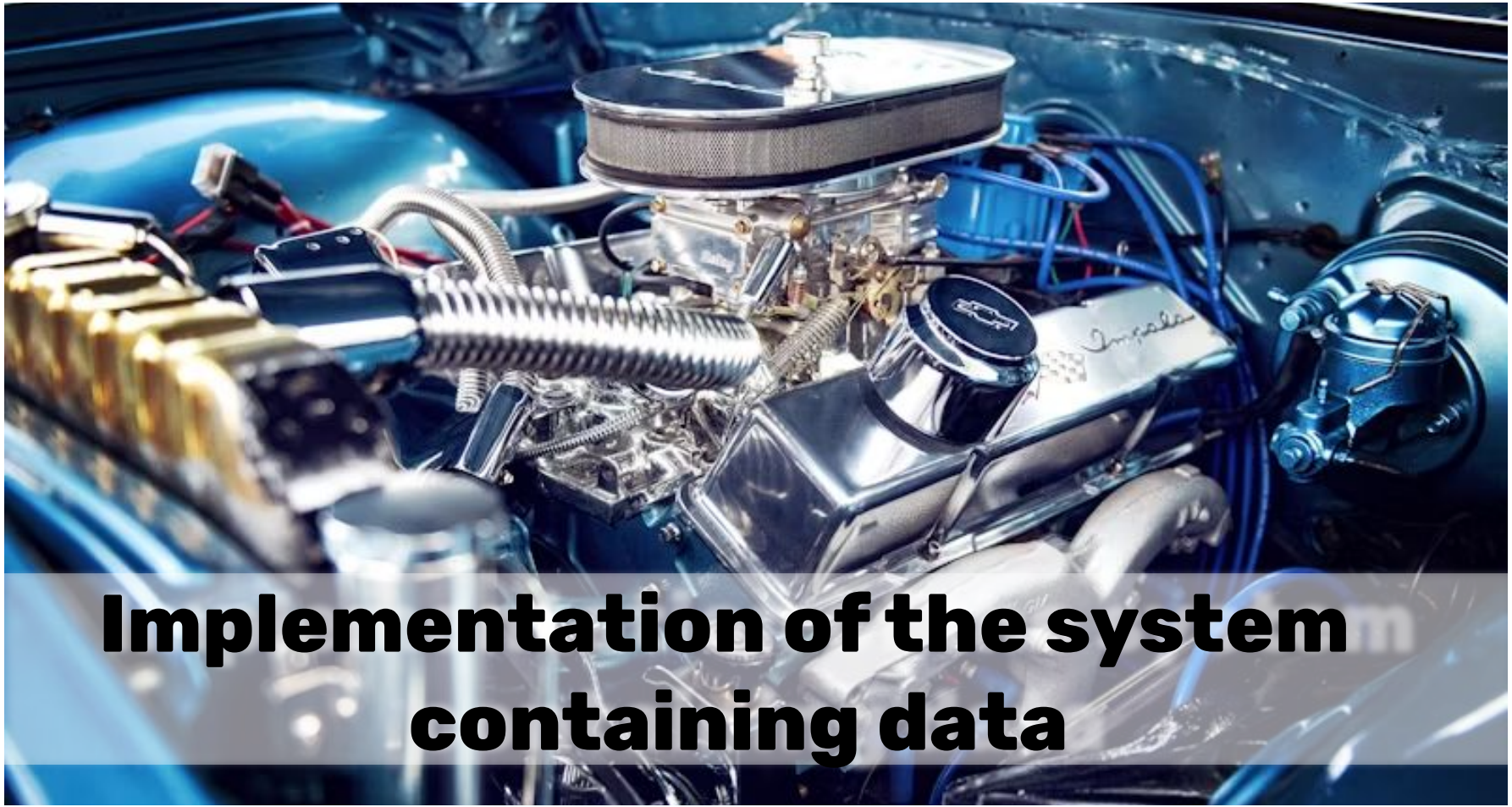
- What can be done with data? What additional data would be helpful?
- e.g.: Asking questions about the own product.
  How is it used?
  Which filters are used?
  How does this reflect in conversion rates?

Different areas to tackle

Implementation of the systems itself

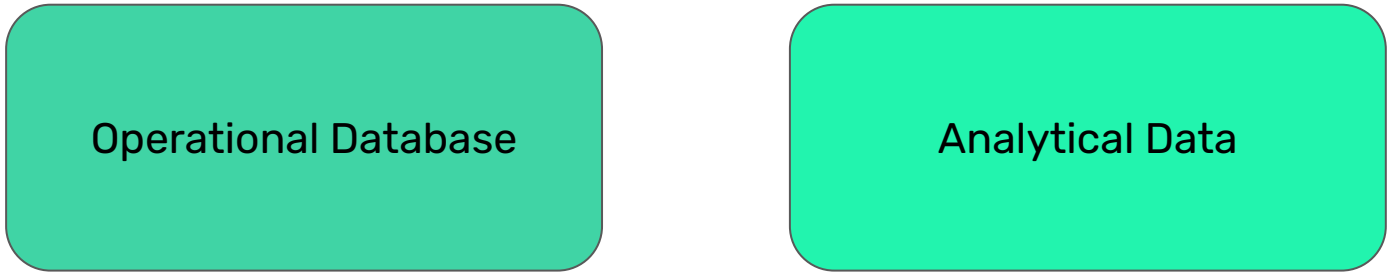
Connecting the systems & transfer data

Process, Platform & Organisation



Implementation of the system containing data

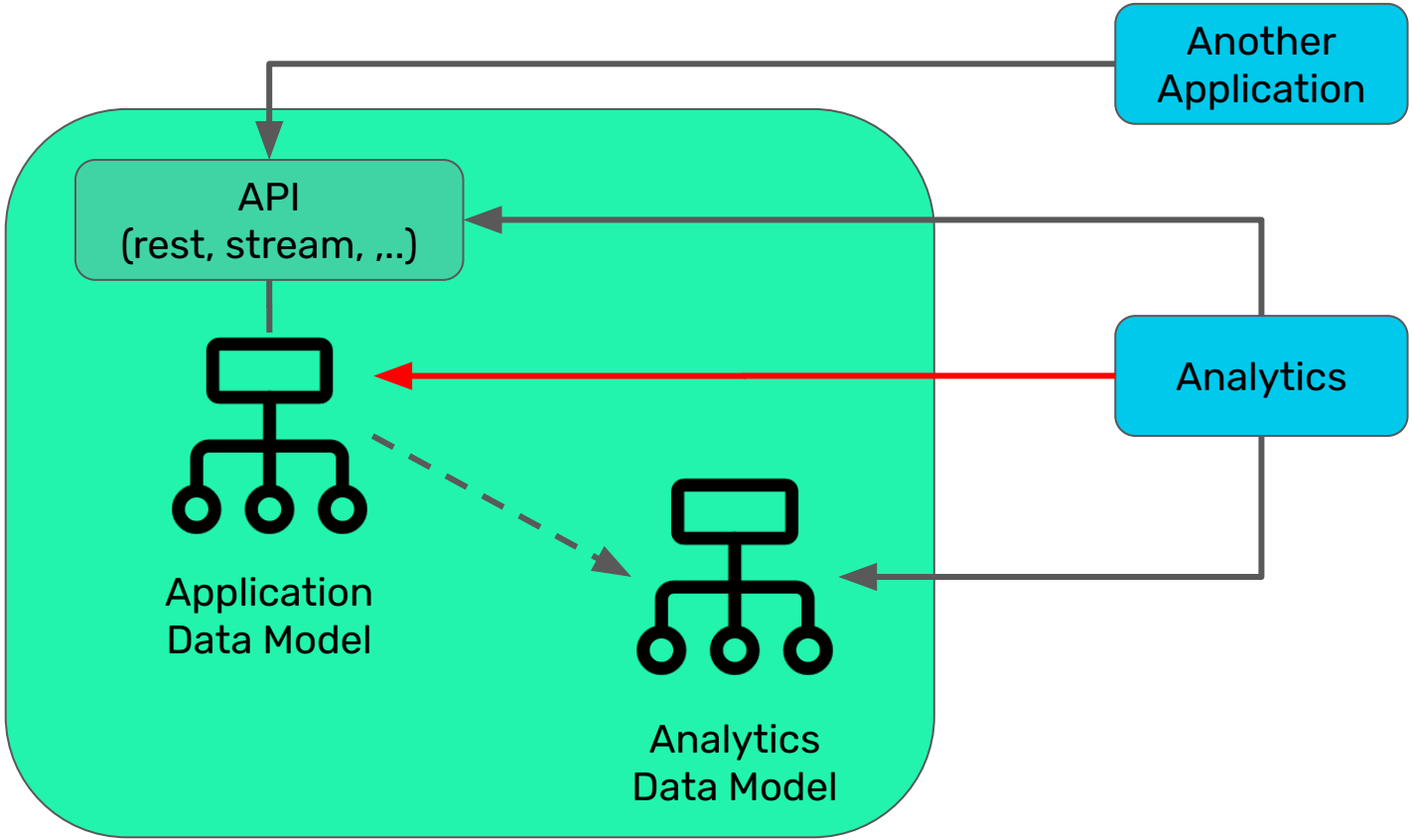
First of all: storage



- The same for both
- Specialized for analytical data
- Analytical on demand only

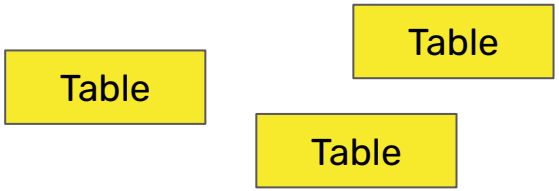
Volume, Variety, Velocity
Access Patterns, Latency, Read/Write Heavy, Availability & Consistency

Encapsulation / Information Hiding



Modeling of an analytics data model

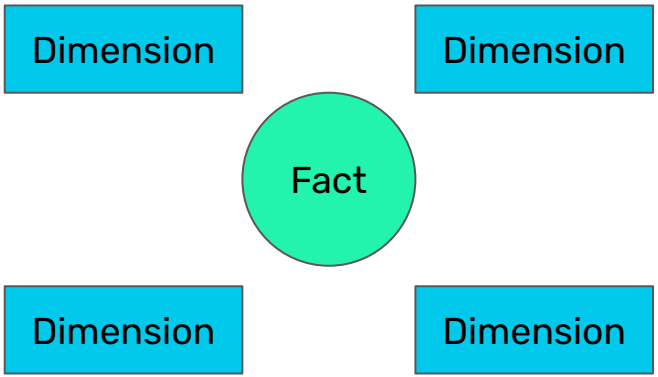
Raw



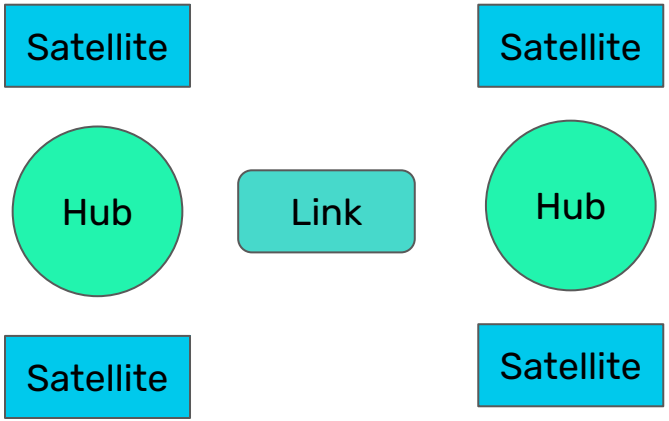
Denormalized / Wide



Star Schema / Dimensional modeling



Data Vault



Trade Offs

	Flexibility in modelling	Ease of use in querying	Understanding business rules	Query performance
Raw	low	low	low	low
Wide	medium	very good	medium	very good
Dimensional	high	very good	good	good
Vault	very high	medium	medium	low - medium

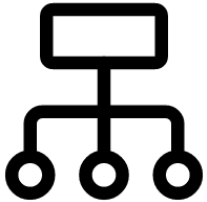
Modeling depends on purpose of the data

provide source data	provide aggregated data	data for consumers / analysts
data as is optional: filtered or selected • operational systems • connectors to off-the-shelf products	transformed and joined data, including business logic. might be cross domain. • data registry • central data store / provider	ready to be used for analytic employees, e.g. in self service, dashboards • DWH • Feature Store / ML • Operational Systems
• raw • data vault	• raw • wide • data vault	• dimensional • wide

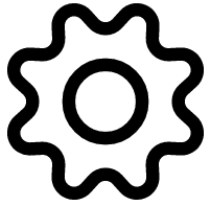
Further artifacts within the system



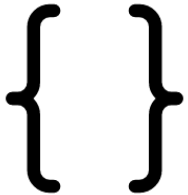
Metadata



Schema



CI/CD



Code



Quality Checks


Connect and transfer data
between systems

Data (Product) Interface Specification

- **Format**

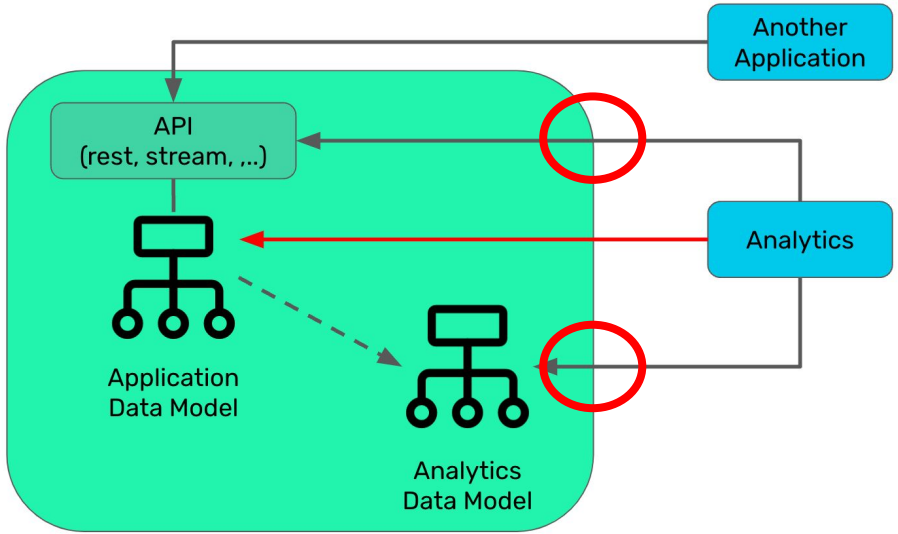
  parquet, iceberg, json, a postgres table
- **Schema**

  dbt schema, json schema, graphql schema
- **Definitions**

  description and definition of business object
- **Quality**

  dbt/great-expectations, soda or other
- **Service Level**

  latency, frequency, freshness



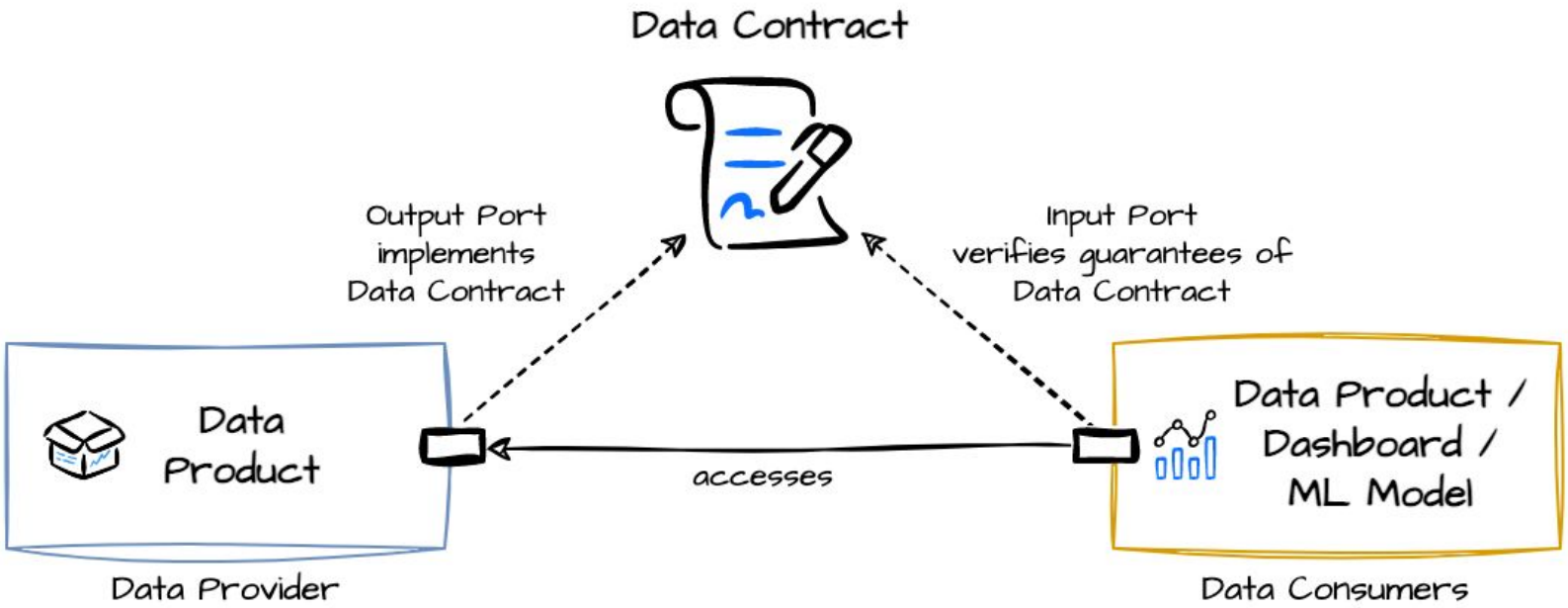
Server fka Ports

Where to get the data

```yaml
servers:
  - type: <server-type>
    description: <server-description>
    environment: <server-environment>
    <server-type-specific-fields> # according to the server type
    roles:
      - <role-details>
    customProperties:
      - <custom-properties>
```

Open Data Contract Specifcation
https://bitol-io.github.io/open-data-contract-standard/latest

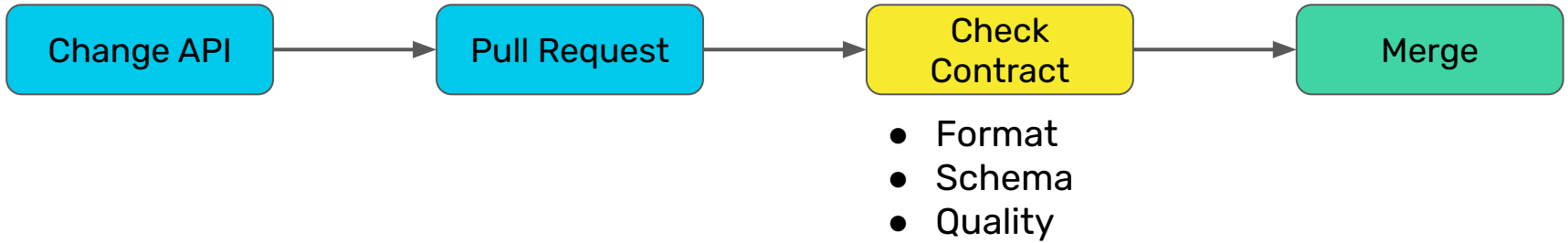
Contract



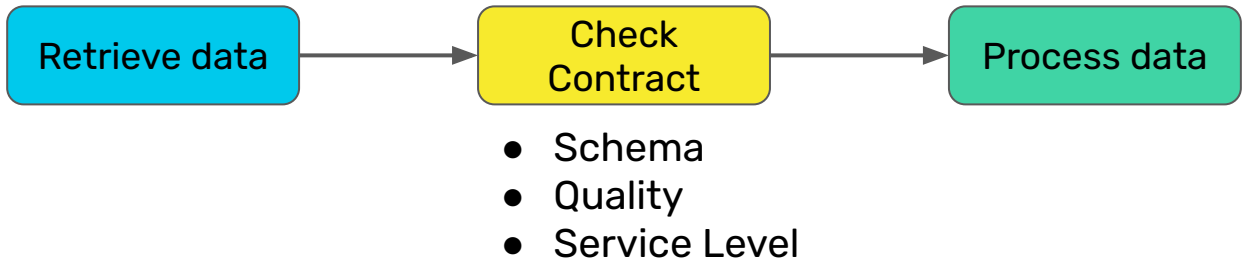
Data Contract Specification
https://datacontract.com/

Enforcing Contracts

On the publisher in CI/CD



On the consumer in the data pipeline



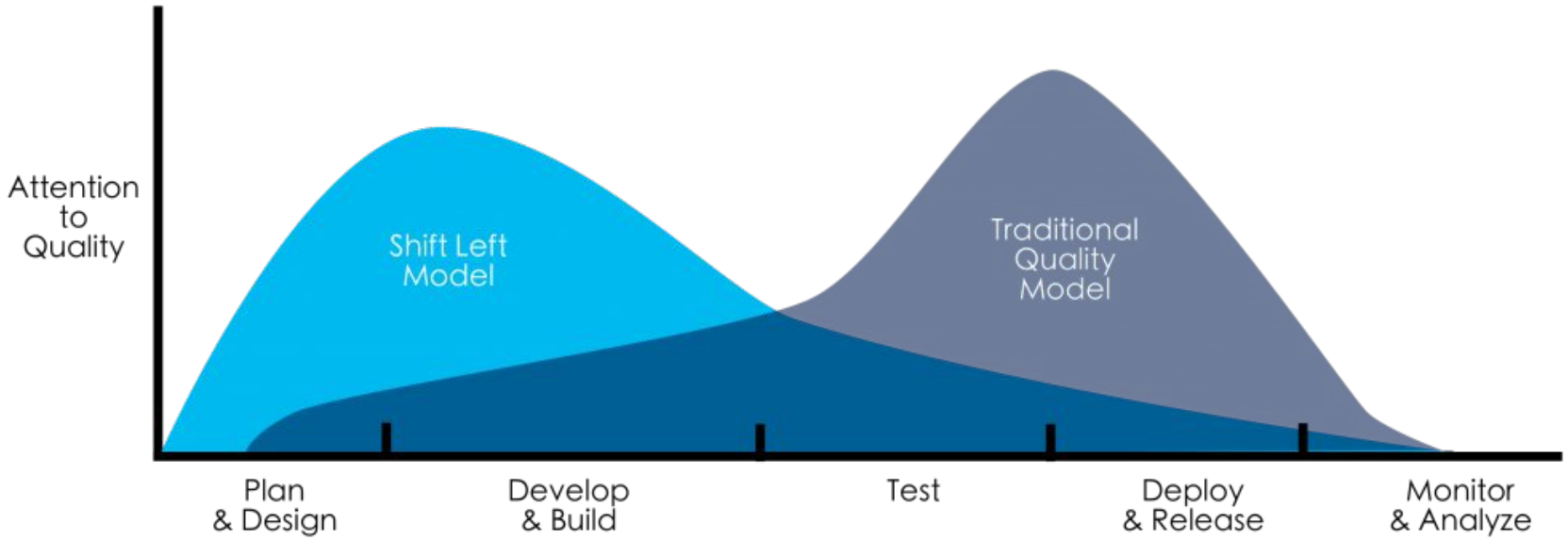
An good approach to start

Do not publish data to a data platform without a data contract

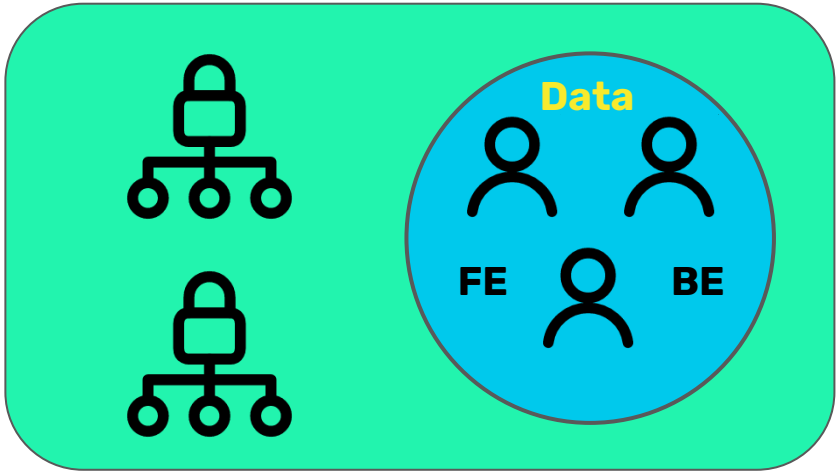
This applies to product teams and data teams!

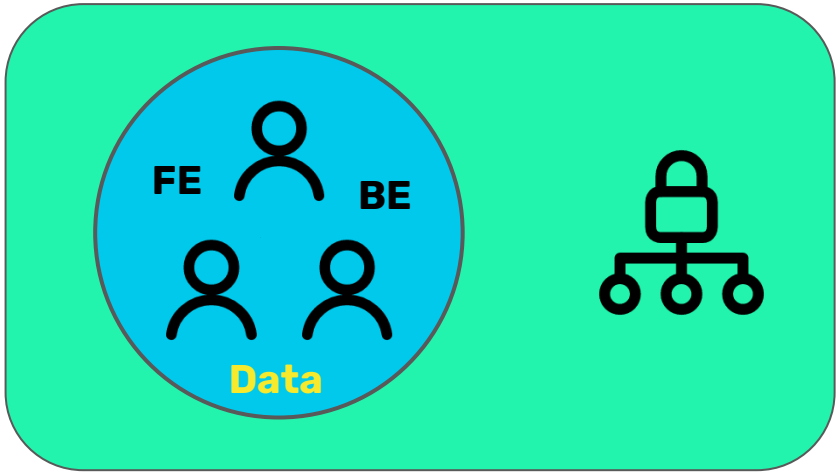

Process, Platform & Organisation

Shift Left everywhere - also for data



Build cross functional teams





dev
sec data
ops



Platform Teams as enabler

- Reduce cognitive load of product team
- Provide necessary and helpful infrastructure
- Data specific help
  - Blue Prints
  - How-Tos
  - “Golden Path”
- Part-Time work in team

Next to enable

- monitor
- govern

the data management



Build Data & Application Platforms!

When the goal is to combine data and application - why build different platforms?

- build data & application platforms
- one platform with all necessary capabilities
- might be different technology stacks (eg. k8s and databricks)

  but integrate them well

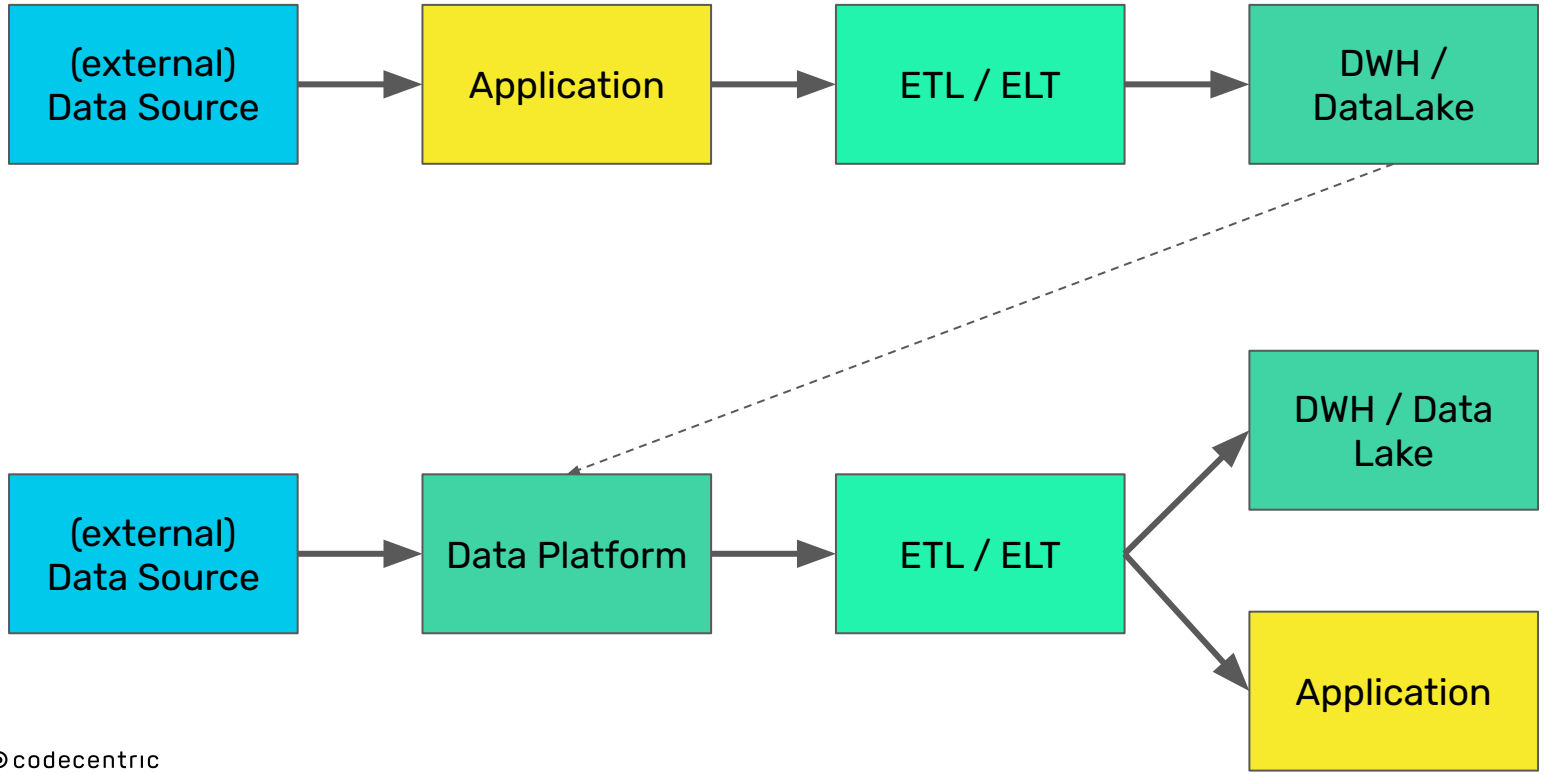
TODO:

sketch building blocks of a data & app platform

- app runtime
- storage
- data processing
- security
- monitoring
- …
-

When shifting left in process, shift left in data flow as well?

At least when building a data application with a WebUI

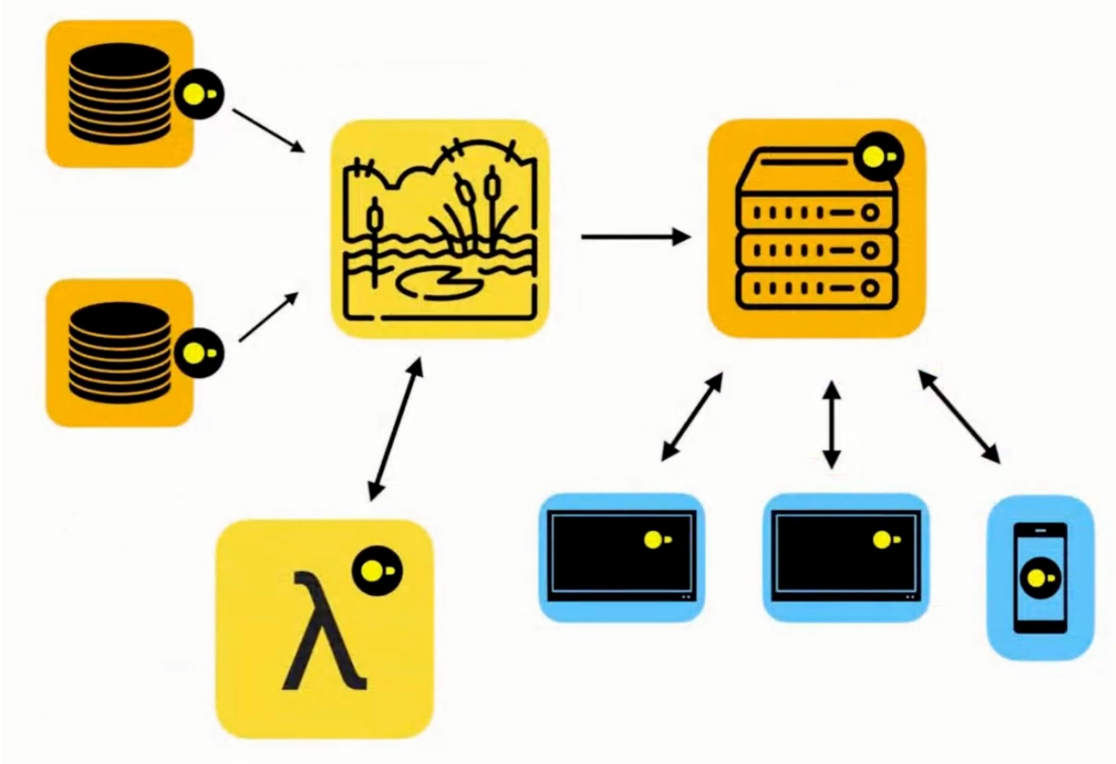


When data analytics & application become one

- Less friction between operational and analytics
- Requirements are more likely to be aligned
- Define and encapsulate responsibilities
  - Aggregates in app or in “data platform”?
  - Pattern detection using ml?
- Optimized data representation for different use case, e.g. visualization vs processing
- Know and use the tooling available on both sides
- Engineering best practices also apply to data



"Going Beyond Two Tier Data Architectures" ?

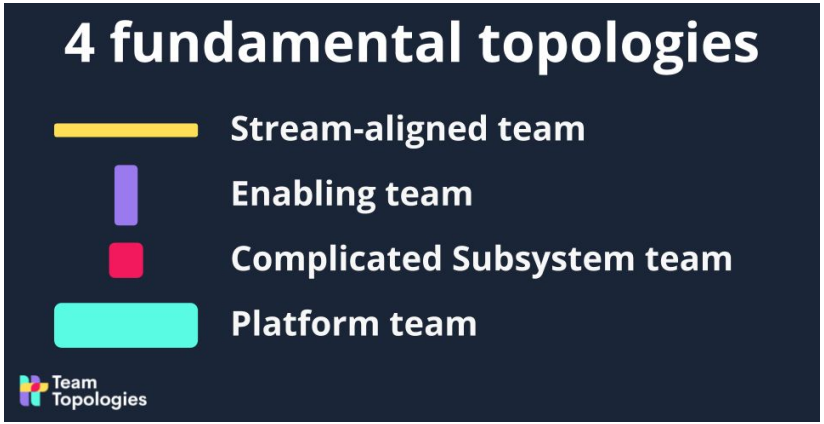


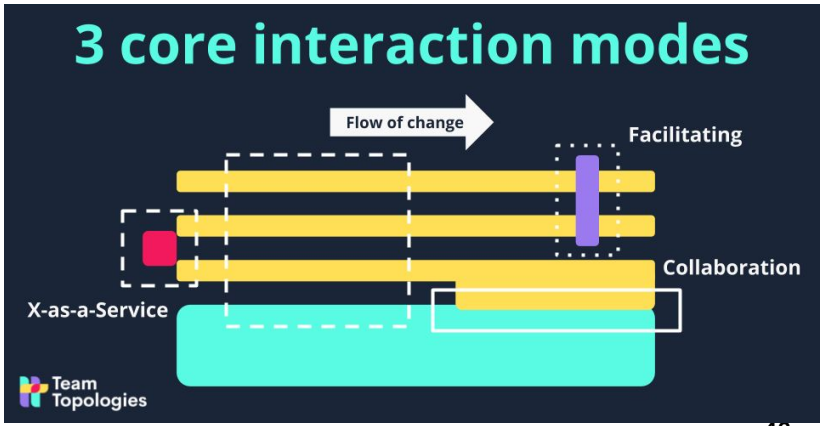
https://youtu.be/xX6qnP2H5wk?t=429

Team Topologies as idea for organisation

- Platorm Team builds the platform - does *not* process or provide data
- Enablement is a must
- Stream aligned teams for source, provider and user teams
- Complicated Subsystem for ML? Or just stream-alinged?

Let me know what you think and what you've seen.






Takeaways

Conclusion

- Data still often a second thought in software architecture
- Time to include data in architectural designs and decisions
- Foundations like data literacy needed also among developers
- Several areas to tackle within and in between applications, responsibilities might change
- Organisational support needed



We support you with your data projects.

Feel free to contact us!

codecentric AG
Hochstraße 11
42697 Solingen

Matthias Niehoff
Head of Data
matthias.niehoff@codecentric.de
www.codecentric.de

codecentric



Feedback for the talk



Feedback for the event





Creating the digital future together.

codecentric AG | Hochstraße 11 | 42697 Solingen