

Data Pipelines in a Business Context

Considerations and Tools for using Data Pipelines effectively and in accordance with the organisations goals

Successful businesses and products need to provide value for their customers and users. In order to achieve that organizations need knowledge about the users intentions and desires. Technology enables commercial digital offerings to collect, process and analyze user-generated data in order to improve their user experience (UX) and strategic decisions-making. Ideally, this goes in hand with achieving the (economical) goals of an organization. What considerations to take into account when implementing a data pipeline and what tools could be used are displayed in the following. A special focus lies on the accordance of the technological possibilities to collect data and the economic implications that certain data can unfold.

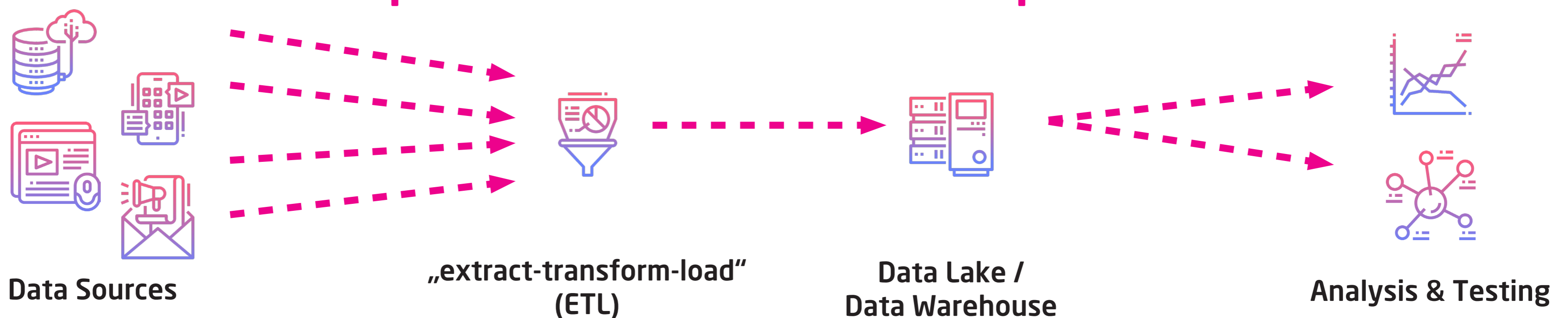
Considerations before implementing a Data Pipeline

- | | |
|---|---|
| 1. What do you want to achieve? | Sell more items? Sell more expensive items? Increase conversions? Increase customer satisfaction? Increase average basket size? |
| 2. Where to collect the data? | Application vs. Browser? At Log-in? At checkout? Post-purchase? Newsletter? Search Engine Ranking? |
| 3. Metrics for expected outcome | Conversion Rate, Revenue Per Visitor, Average Order Value, Repeat Rate, Cost per Click, Net Promoter Score |
| 4. How to extract information? | } Need for Data Pipelines and „extract-transform-load“ (ETL) Tools |
| 5. How to test and prove findings? | |



Implementation of a Data Pipeline

Steps



Tools (examples)



Considerations

Define data sources (E-Mails, website or App, Session Logs) and events (i.e. clicks, scroll, search). Even data that is not yet needed should potentially be collected to use it in the future.

Data cleansing, joining, validation and formatting to database schemas can be part of this step. The usage of Data Warehousing vs. Lakes depends on the company resources, the update frequency of data and its cleanliness. Update scheduling of batch processing should be in accordance with the need of up to date informations and often can be more economical feasible than real-time data processing.

Data Dashboards can be used for simple data displaying, Machine Learning Models for detecting dependencies and deducing optimal solutions for the regarded problem such as finding the best recommendation for customers or the most cost-efficient promotion channel. A/B Testing can validate findings.