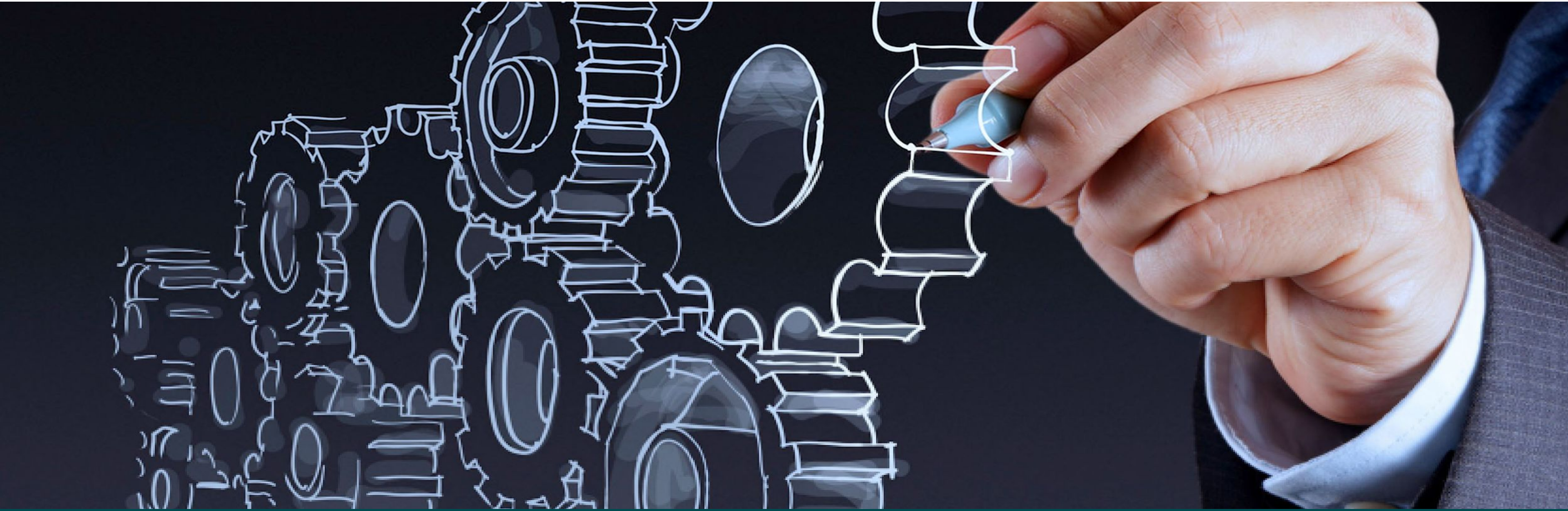
Alligator Company
Data Works!



Data Vault and Machine Learning - Does it fit together?

Alligator Company - Data Works!

Torsten Glunde / CEO

Data Warehouse Automation und Modernisierung

Data Warehousing und Business Intelligence seit 2002.

Kernkompetenzen:
Moderne Datenplattformen, Data Vault Automation, Analytics Engineering, Cloud DBMS

Methoden:
ELT/ETL, SQL, CI/CD, Datavault, Information Modeling, ELM, BEAM

Generative AI

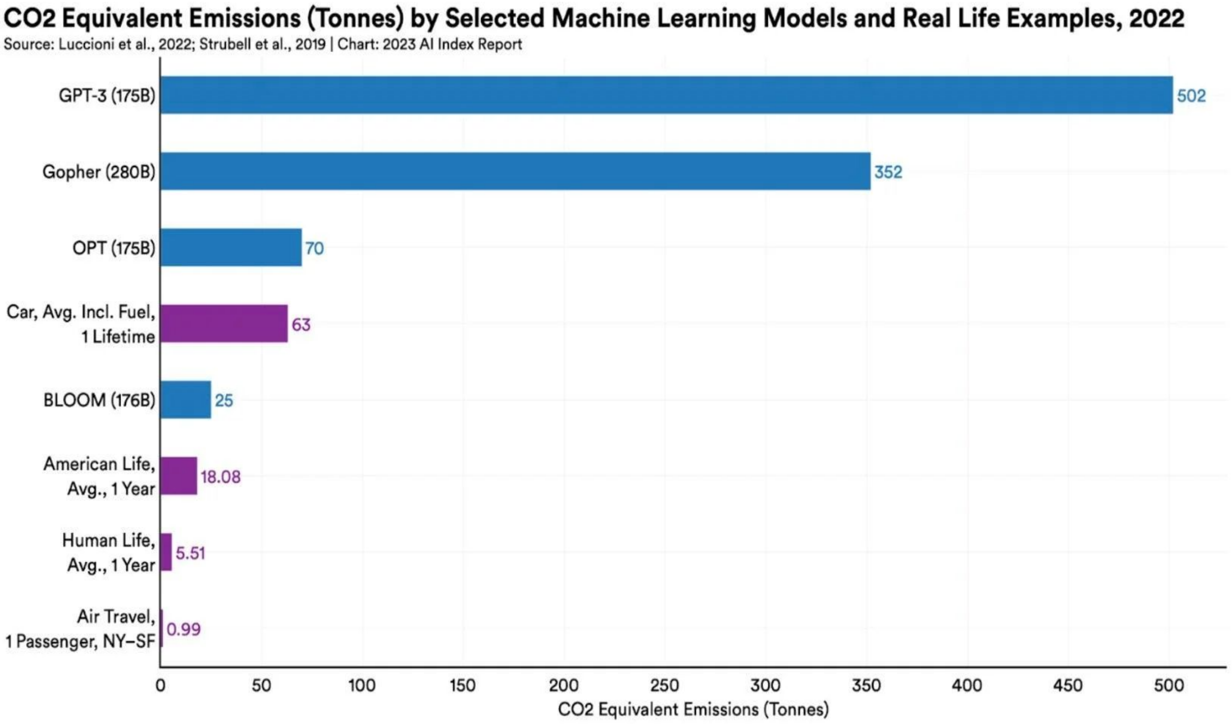
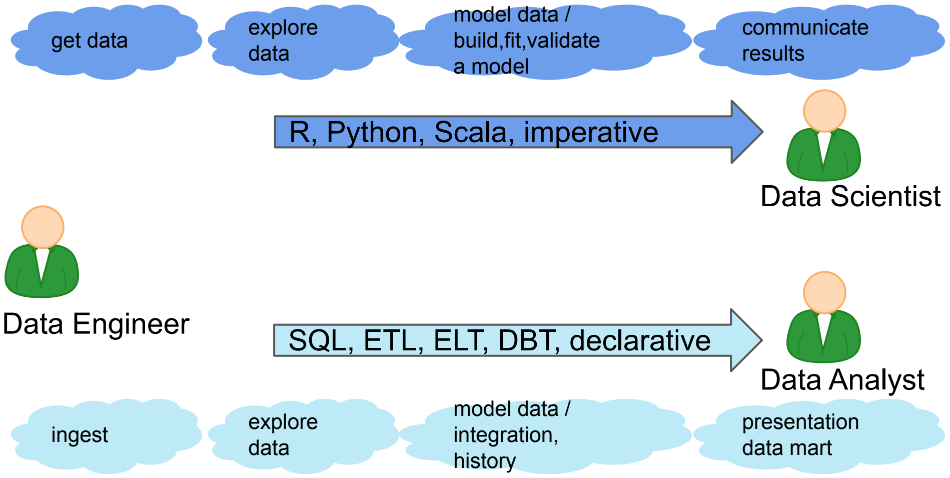


Figure 2.8.2

LLM - Generative AI - Using programming languages is getting easier

Current State of Data Science & Business Intelligence



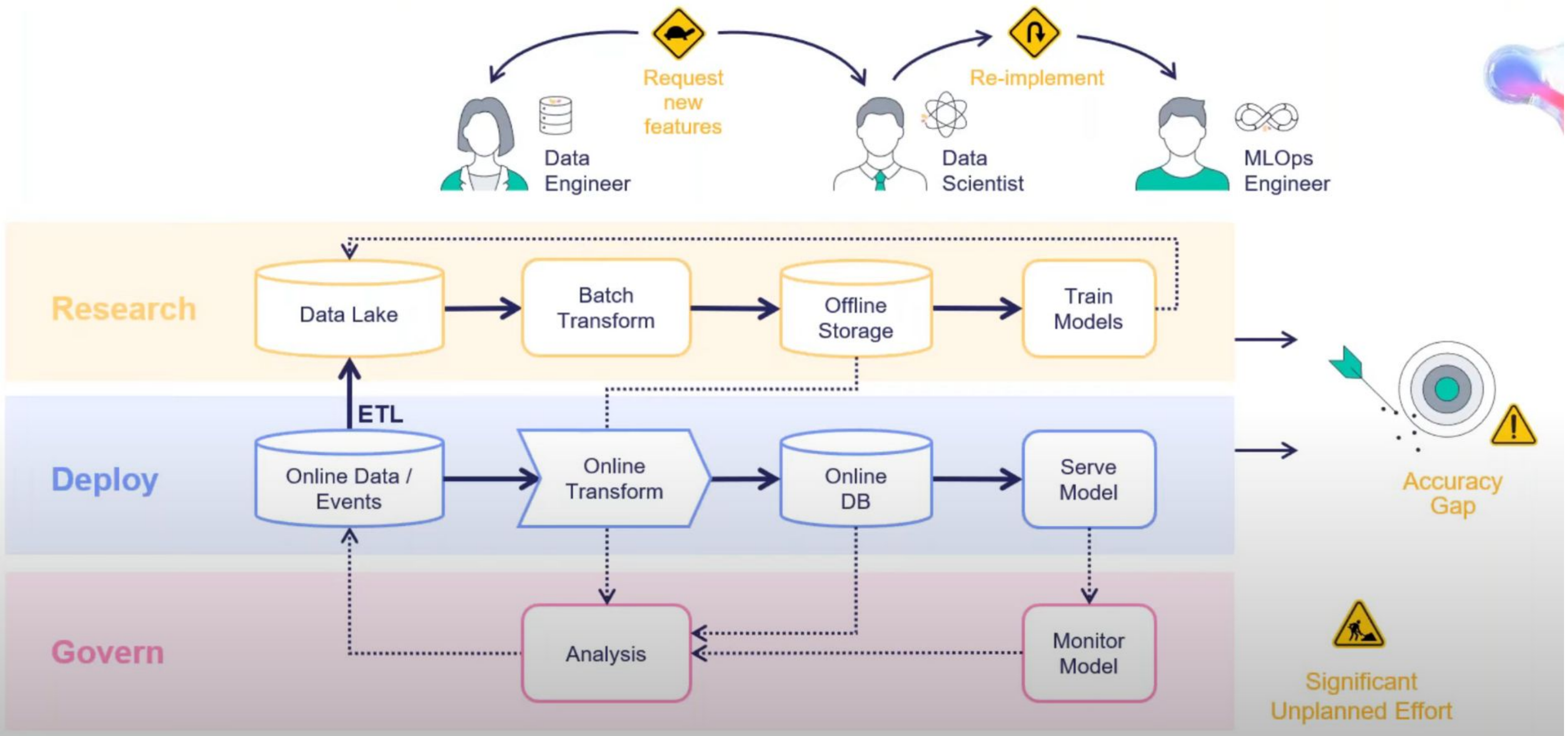
https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations
https://atlan.com/modern-data-stack-101/

- people & platform are separated
  - SQL data pipeline
  - Python data pipeline
- 60-80 % of work is in data engineering / preparation
- source data interpretation is repeated, with possible different outcome

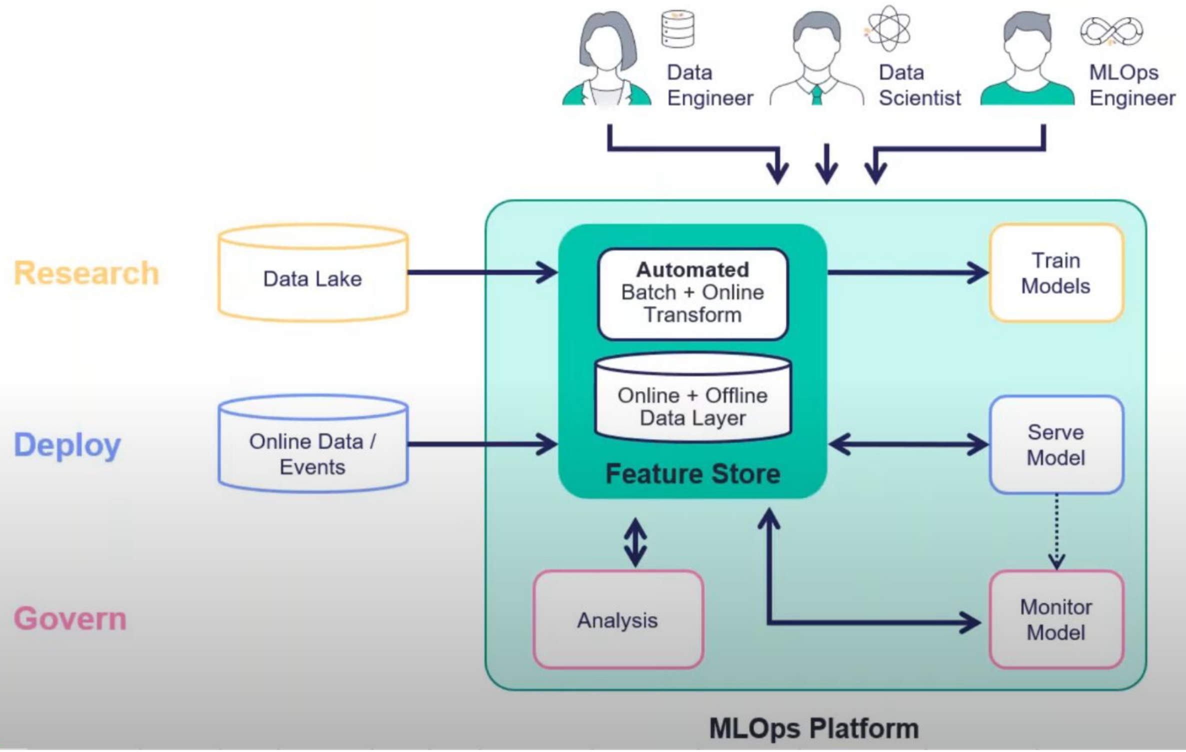
Combining use cases into one model driven platform

- **model driven automation**
- **reuse of existing data assets**
- **SQL vs. Python?**

Most Enterprises Today Suffer from Resource Intensive Processes, Data & Org Silos



MLOps + Feature Stores = Faster Time to Production



MLRun: The Open Source MLOps Orchestration Framework

MLRun

https://mlrun.org

Central metadata management, orchestration, and monitoring

Data ingestion & preparation → Model Training & Testing → Real-time Data + Model Pipeline → Data + Model Monitoring

Elastic Serverless Runtimes + Function Marketplace

Online & offline Feature Store + Data connectors

Using and orchestrating the most common data science and MLOps tools

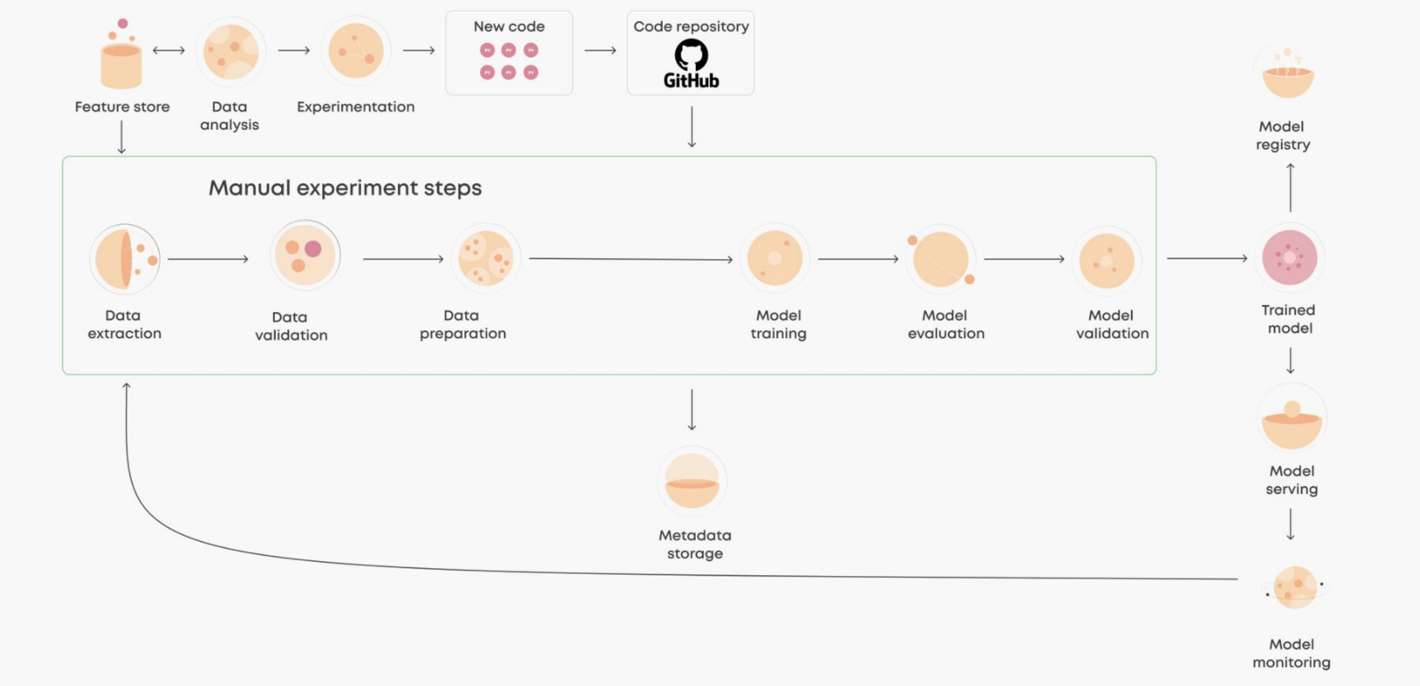
Spark presto HIVE jupyter learn DASK HOROVOD TensorFlow PyTorch nuclio Kubeflow Grafana

Tools und Methoden für eine einheitliche Datenarchitektur

- Common data platform for Infrastructure - horizontal scalability
  - Snowflake - Python Snowpark, SQL Engine
  - Databricks - Spark Engine, SQL Engine on Deltalake
  - Pipeline & Metadata Abstractions & DevOps
    - DBTLabs, MetricFlow
    - SQLMesh, SQLglot
    - Cube.dev
    - Github/Gitlab
  - Scheduling
    - Airflow, Prefect, Dagster, Argo

- Datavault - **model-driven Automation - for Data Management**
  - Datavault Builder
  - AutomateDV
  - Datavault4dbt
  - Vaultspeed
  - Coalesce.io

Automated ML Pipeline (functional view)

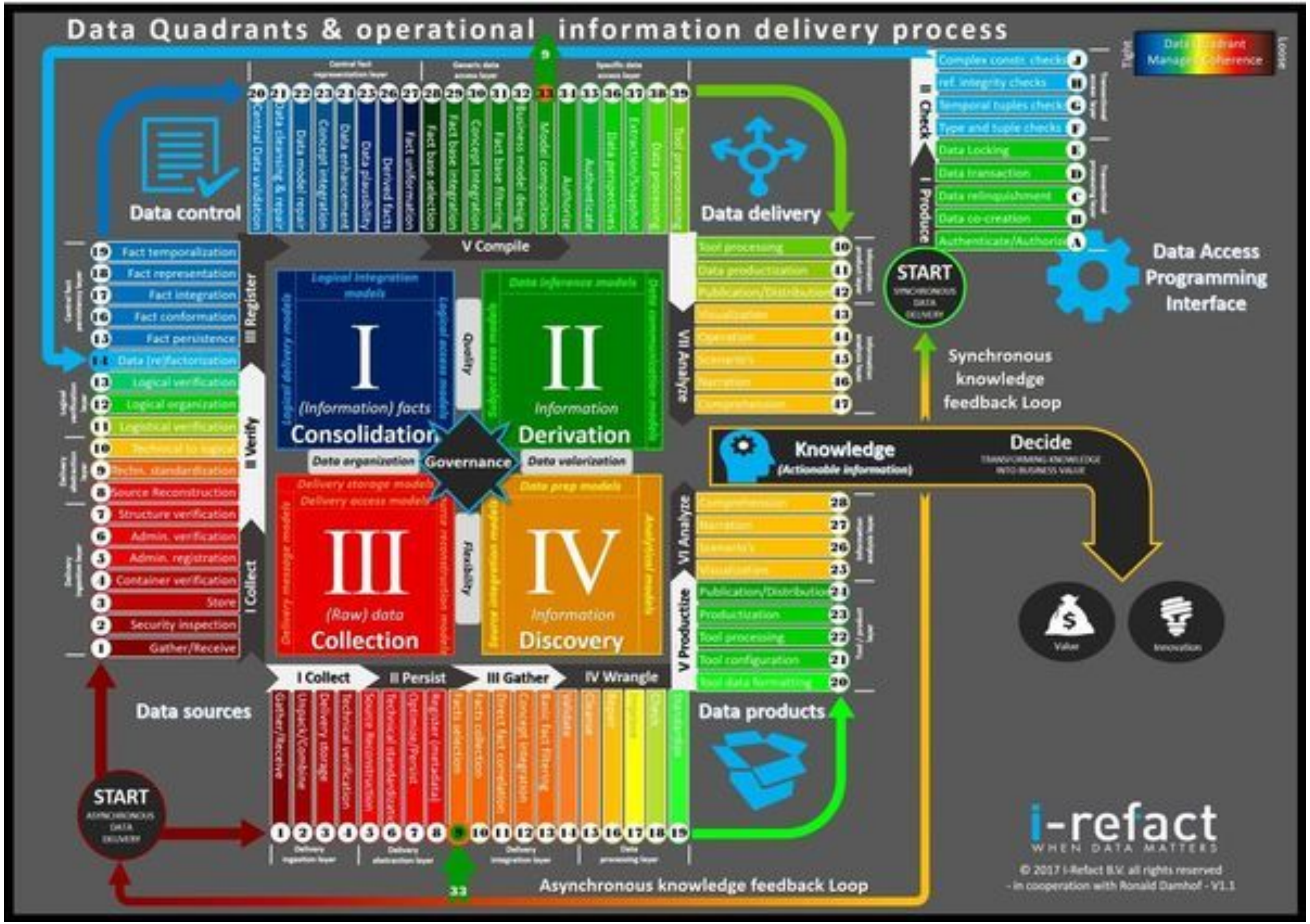


- The pipeline is the product
- Fully automated process
- Co-operation between the data scientist and the engineer
- Fast iteration cycle
- Automated testing and performance monitoring
- Version-controlled

Features for ML Algorithm applications

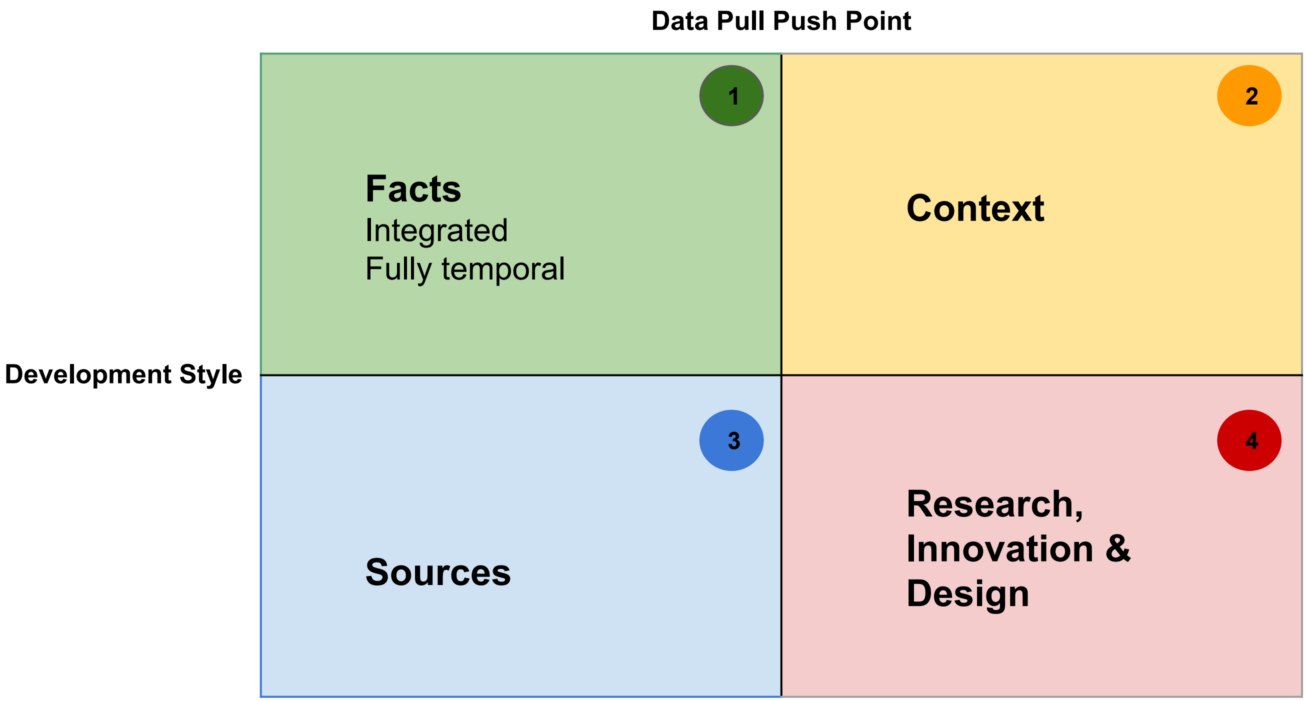
- Entity Resolution
    - Two Customer IDs, Levenshtein Dienstance /Jaro Winkler / Soundex / Cosine Similarity on Names and Emails
- Churn Prediction
- Customer Lifetime Value
    - Customer Demographics (Age, Gender, Income, Education Level, Marital Status)
    - Shopping Behavior (Purchase Frequency, Average Basket Size, Total Spend, Product Category Preferences, Promotion Response Rate)
    - Customer Service Interactions (Support Calls Count, Average Resolution Time, Customer Satisfaction Score)
    - Online Engagement (Website Visits, Average Page Views, Average Session Duration, Feedback Form Submission)
    - Loyalty Program (Loyalty Member Flag, Loyalty Points, Redemption Count)
- Sentiment Analysis

https://www.facebook.com/FullScaleDataArchitects/



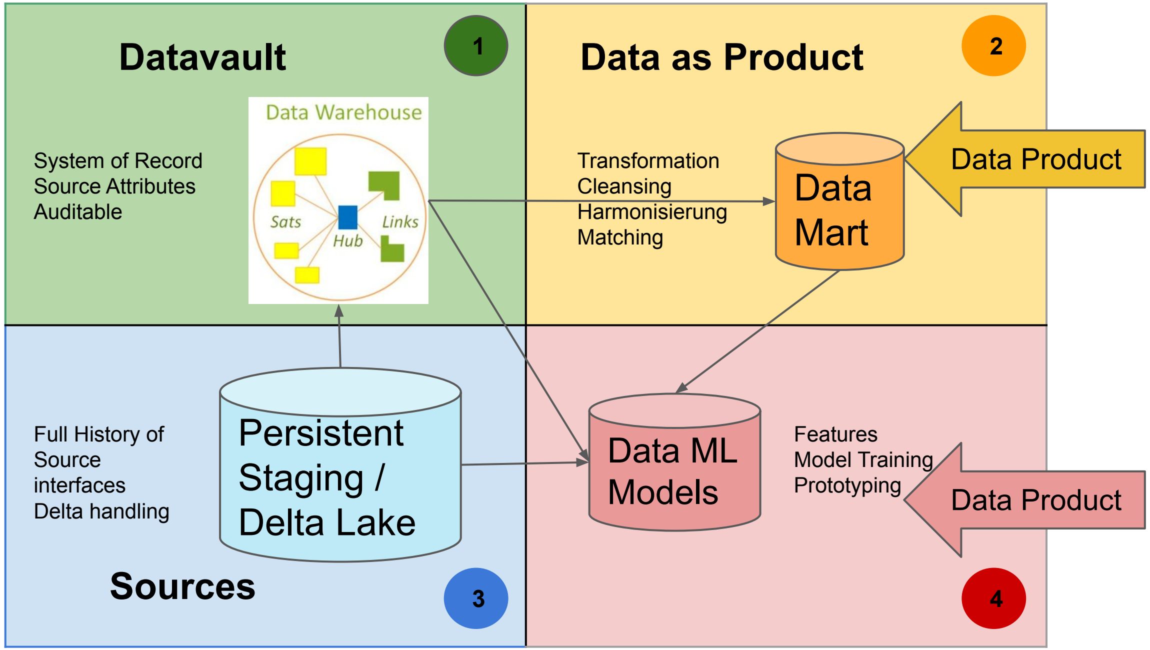
Data Management Quadrant - Ronald Damhof

https://prudenza.typepad.com/dwh/2015/06/make-data-management-a-live-issue-for-discussion-throughout-the-organization.html



https://alligatorcompany.gitlab.io/acs-docs/datamanagement/

Data Management Quadrant - Components - Data Products



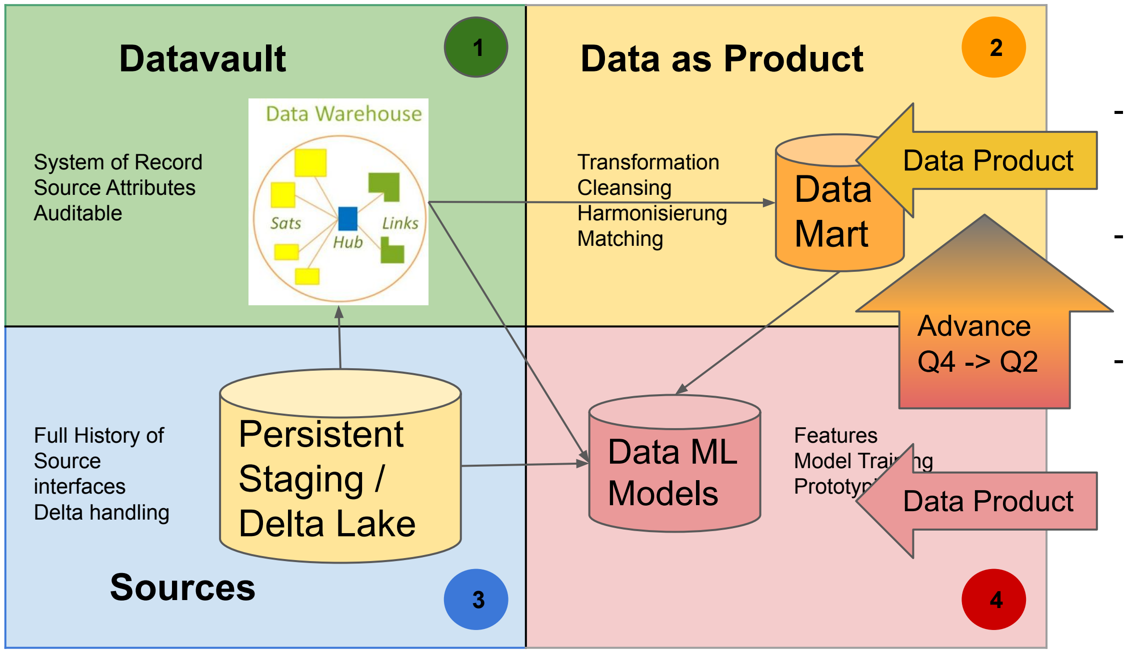
3 Raw data - application driven - no data governance

1 Facts, presented by a business model

2 Multiple truths - sourced by facts & derived data

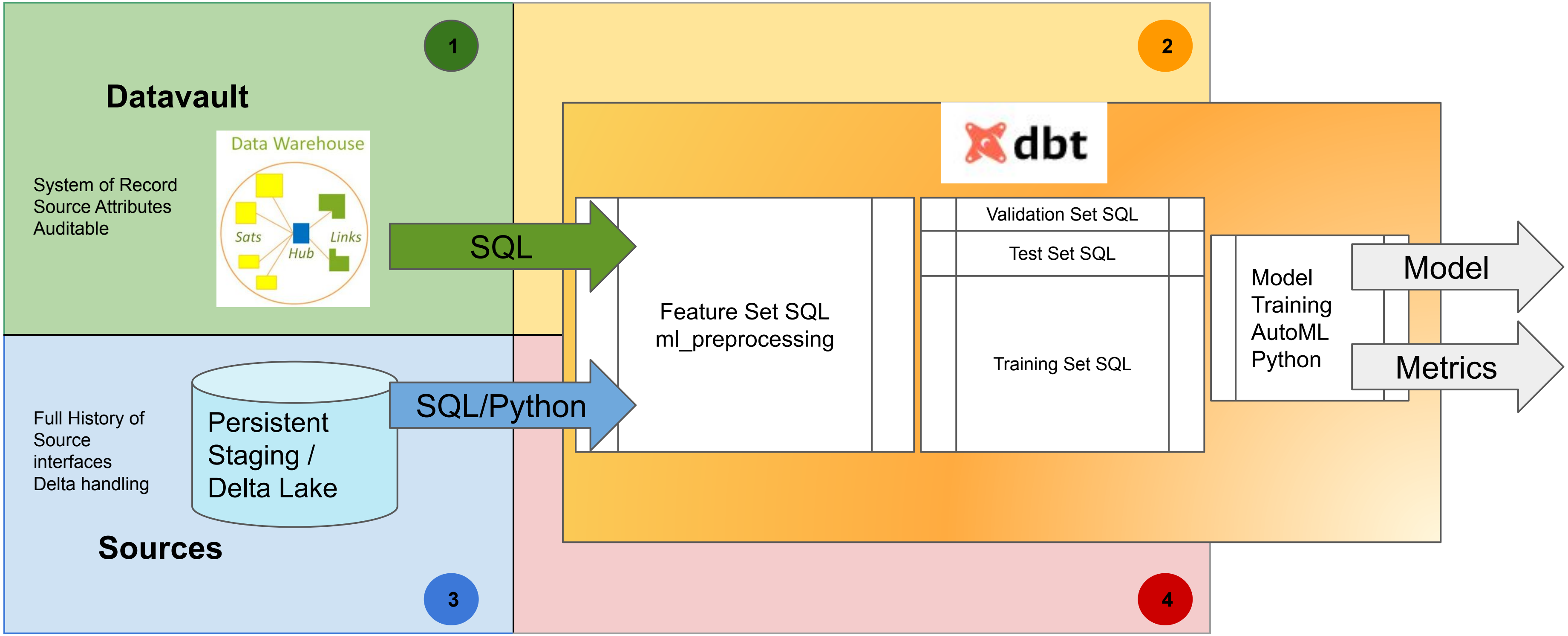
4 Multiple truths - sourced by facts, derived & source data

Data Management Quadrant - Data Governance

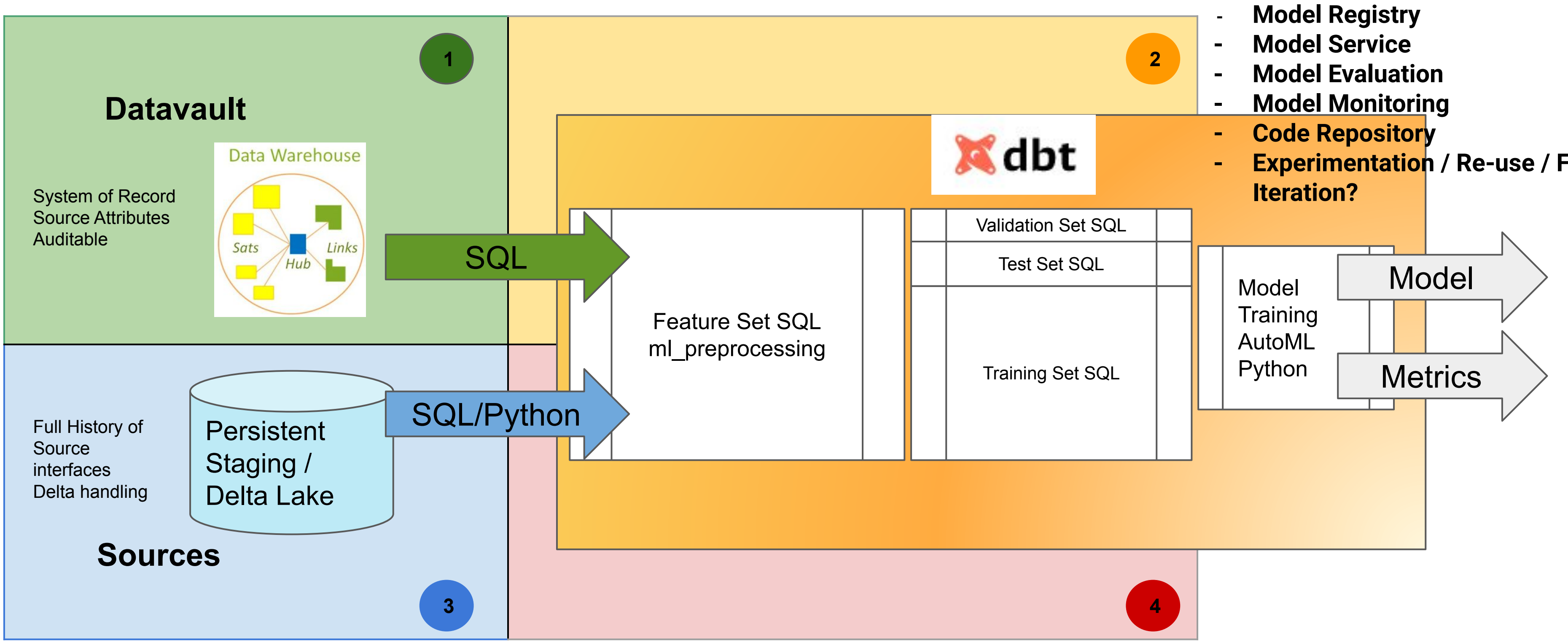


- Data products should be able to easily benefit from all data
- Data products should not be able to advance from Q3 -> Q2
- Governance should differentiate by data source
- Deep Learning usually is source data driven → hard to advance
  - Easier access to all data by using common platform
  - Using more well understood data for interpretation
  - Business Model helps to nurture common understanding

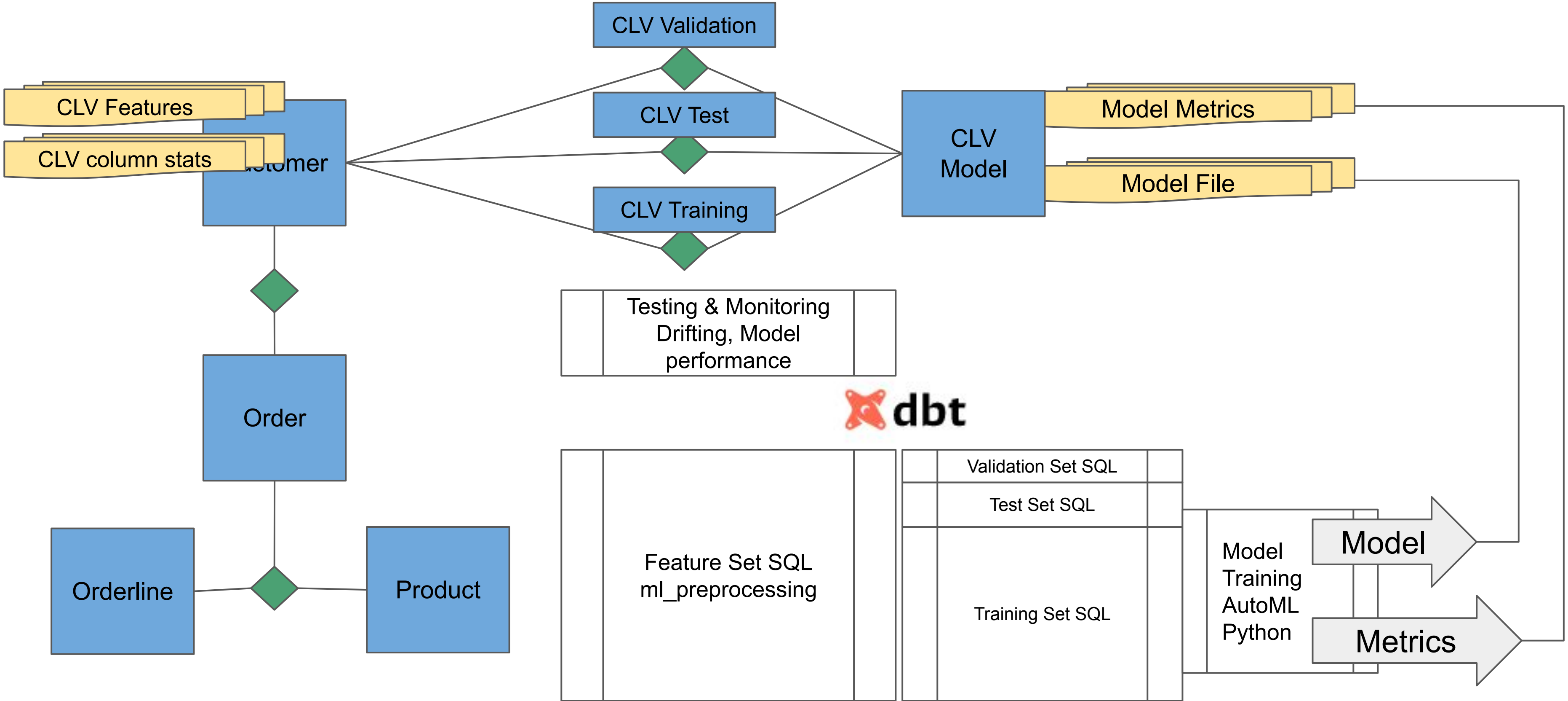
ML Pipeline on Datavault Data Architecture using DBT



ML Pipeline on Datavault Data Architecture using DBT



Are Datavault & DBT a Feature Store & MLOps Replacement?



My Conclusion

- BI & Machine Learning transformations can be used together with DBT
    - Using a common data platform
    - Being flexible in choice of transformation language and execution engine for scaling
    - Re-using common data integration models - potentially saving costs in dev & maintenance
    - Challenges might include
        - Relative young engines for SQL on spark / Python on SQL

- Datavault
    - Helps with data management and governance
    - Can be used as replacement for a feature store, if needed
    - Challenges might include
        - Using the data architecture for transparency needs know-how
        - Sticking to the model driven nature needs know-how
        - Data Storage / physical data challenges with big data in SQL on spark / Python on SQL

An important aspect is model driven and business / domain owner involvement, which could be leveraged using a common data platform with shared model and transformation logic