

Efficient Data Preparation for Better Analytics and AI Models

11/12/2025

Ronny Kober
Senior Data Engineer / Digital Products

Sunfire Company Presentation 250909

SUNFIRE AT A GLANCE

European Electrolysis Excellence

100 % Specialized on Green Hydrogen Technology for Industrial Transformation

15+

Years of H₂ industry
experience

> 650

Employees from
25 nations

800+

Megawatts of
electrolyzer order book




UP
TO 1

Gigawatt of
production capacity¹⁾

1) Actual production output according to
customer demand

 Made in Europe

There Are Three Main Electrolysis Processes for Green Hydrogen Production

	<div>Pressurized Alkaline Electrolysis</div> <div>AEL</div>	<div>Solid Oxide Electrolysis</div> <div>SOEC</div>	<div>Proton Exchange Membrane Electrolysis</div> <div>PEM</div>
	Using deionized water and an alkaline electrolyte solution (potassium hydroxide/KOH)	Using high-temperature steam and a solid oxide ceramic membrane as the electrolyte	Using a gas-tight polymer membrane and a semipermeable exchange membrane
	Most reliable and long-established with best scalability and no reliance on rare or expensive raw materials	Exceptional efficiency with low power consumption and capability for waste heat utilization + synthesis gas production	High response rate and lower space requirements, highest hydrogen purity for special use
	Slightly more space requirements and slower response speed	Early-stage technology and high operation temperatures (600+°C)	Relies on expensive and rare materials (platinum, iridium) and has limited scalability

Sunfire Product: 10 MW AEL Module



Sunfire Product: 10 MW SOEC Module



There are other relevant alkaline electrolysis systems, such as chlor-alkali.

AGENDA

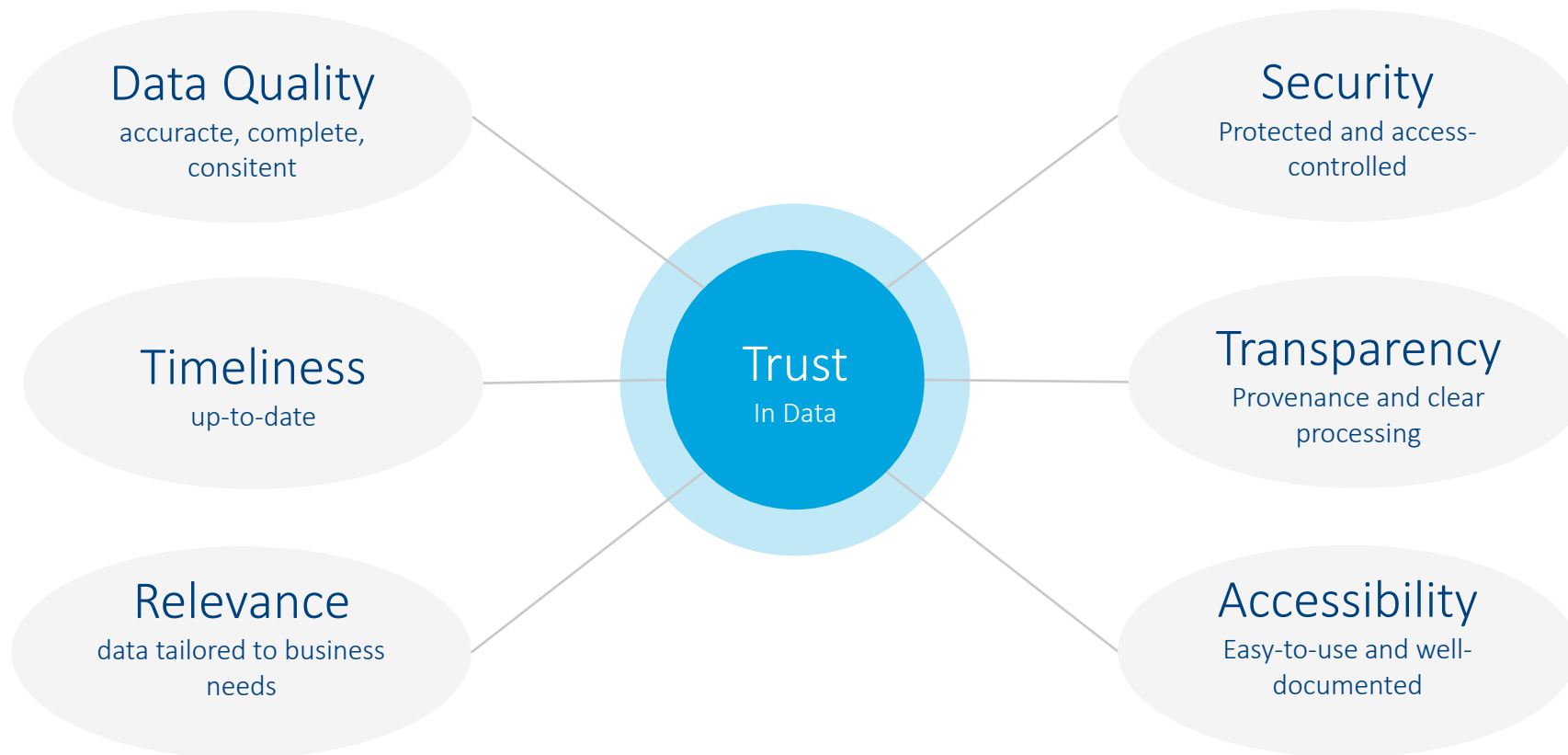
Essential Topics for Efficient Data Preparation

- 1 Top goals in analytics and machine learning
- 2 Key success factors for trust in data – the role of data engineering
- 3 Cutting-Edge technologies
- 4 ELT (Extract, Load, Transform) to update your data products
- 5 dbt core to transform your data over SQL + Demo
- 6 One iteration of the agile software development life cycle
- 7 Summary

Top Goals in Analytics and Machine Learning

“It takes 20 years to build a reputation and five minutes to ruin it.
If you think about that, you'll do things differently.”

Warren Buffet



Key Success Factors for Trust in Data – the Role of Data Engineering

Success factors for trust in data

- Automation and efficiency
- User-friendliness tools
- Flexibility and scalability technologie
- Timeliness and real-time processing
- Quality assurance and monitoring



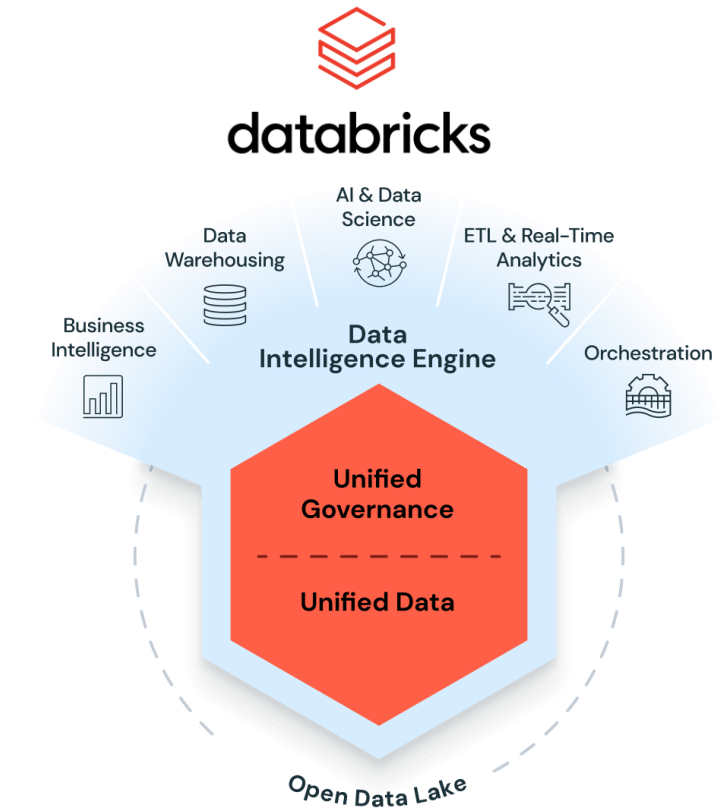
Modern approaches in data engineering

- Cutting-Edge technologies
- ELT (Extract, Load, Transform) and SQL-First approach
- use the agile software development live cycle

Cutting-Edge Technologies

Why cloud data warehouse platform?

- separation of storage and compute
- flexibility and scalability
- strong focus on declarative language SQL
- Continuous updates and innovations
- Additional integrated features (e.g. orchestration, dashboarding, AI features)



ELT (Extract, Load, Transform) to Update your Data Products

Extract

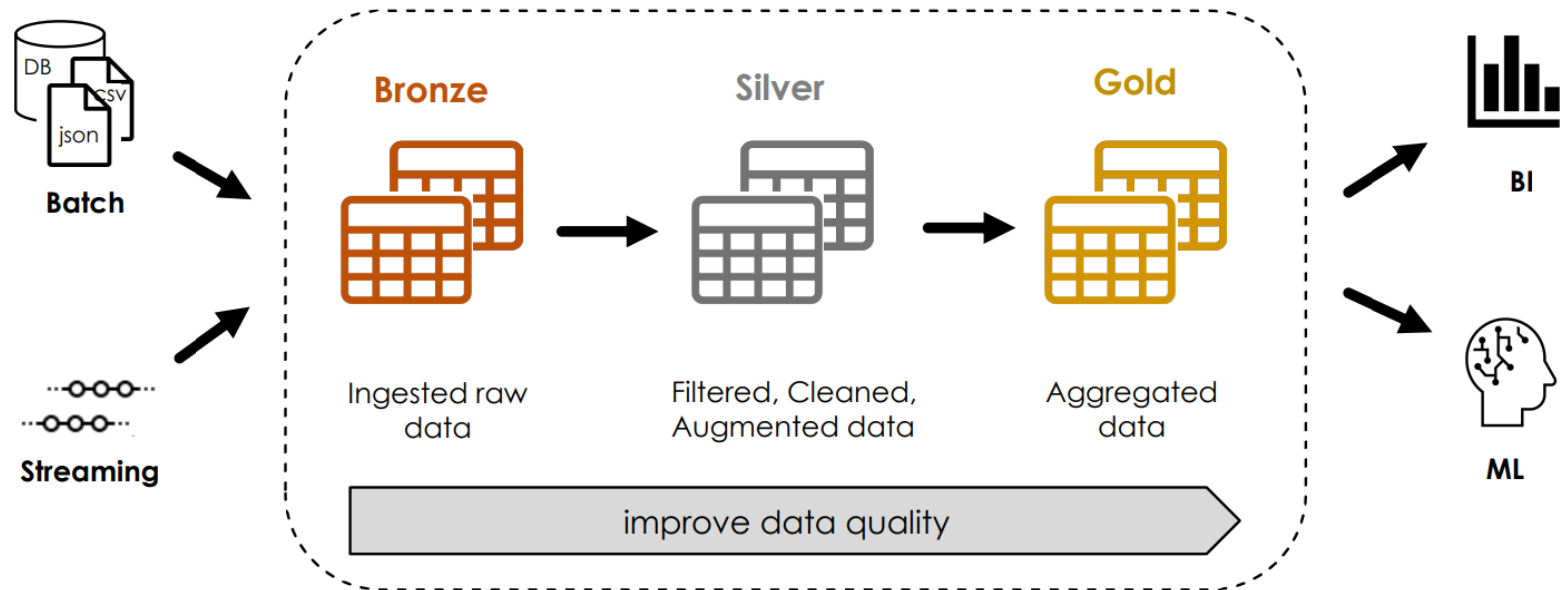
- extract 1:1 copy of raw data

Load

- store raw data in bronze layer of the data warehouse

Transform

- transform your data over SQL with the medallion architecture



dbt Core to Transform your Data over SQL



Create model file “silver_transactions.sql”

- SQL-statement for your model
- references to sources or models
- CTEs improve the structure and readability

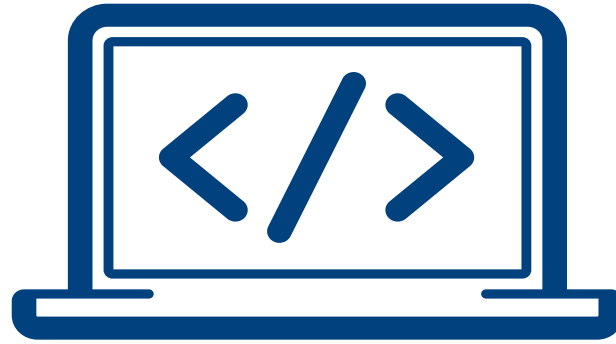
```
1 WITH raw_data AS (  
2     SELECT  
3         *  
4     FROM {{ source('raw_data', 'transactions') }}  
5 )  
6 SELECT  
7     id,  
8     name,  
9     purchase_date,  
10    amount,  
11    CASE  
12        WHEN amount > 100 THEN 'High'  
13        ELSE 'Low'  
14    END AS purchase_category  
15 + FROM raw_data
```

Create schema file for the model “silver_transactions.yml”

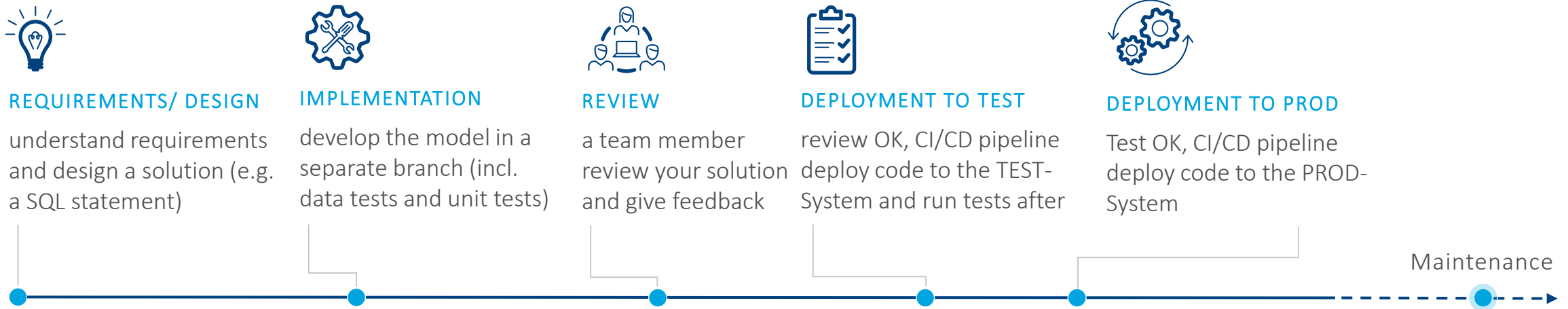
- description of the model/columns for documentation
- data test for model/columns to ensure the data quality
- define a data type to improve the quality

```
1 models:  
  Add documentation or tests  
2  - name: example_model  
3    description: "Cleans and categorizes transaction data."  
4    columns:  
5      - name: id  
6        description: "The unique identifier for the transaction."  
7        data_type: INTEGER  
8        data_tests:  
9          - unique  
10         - not_null  
11      - name: purchase_date  
12        description: "The date the transaction occurred."  
13        data_type: INTEGER  
14        data_tests:  
15          - accepted_values:  
16            values: ['2023-01-01', '2023-12-31']
```

Demo



One Iteration of the Agile Software Development Life Cycle



Versioning and collaboration

- track your code and config changes over git
- use a platform where CI/CD is integrated
- Conflict-free development in a team



Summary



make sure with good data quality that your stakeholder has **trust** in your data products



automate all your data pipelines and save your code in a version control platform



use the simple **ELT** approach to be flexible and fast in modelling



work with the agile **software development workflow** to get iteratively added value



Thank you!

Ronny Kober
Senior Data Engineer / Digital Products
ronny.kober@sunfire.de

[Sunfire.de](https://sunfire.de)

References and Sources

- dbt core to manage your transformations [dbt Labs | Transform Data in Your Warehouse](#)
- git versioning your code [Git](#)
- github [GitHub · Build and ship software on a single, collaborative platform · GitHub](#)
- Gitlab [The most-comprehensive AI-powered DevSecOps platform | GitLab](#)
- databricks cloud data platform [The Data and AI Company — Databricks](#)
- snowflake cloud data platform [The Snowflake AI Data Cloud - Mobilize Data, Apps, and AI](#)
- google big query [BigQuery enterprise data warehouse | Google Cloud](#)
- Book [Fundamentals of Data Engineering\[Book\]](#)
- Zach Wilson collection of links: [GitHub - DataExpert-io/data-engineer-handbook: This is a repo with links to everything you'd ever want to learn about data engineering](#)
- duckdb [DuckDB – An in-process SQL OLAP database management system](#)
- Data Modelling [Things I Learned about Data Modelling | LinkedIn](#)
- One Big Table (OBT) approach [Designing One Big Table \(OBT\). It's a Big A\\$\\$ Table | by Leo Godin | Medium](#)