Professionelle **Datenanalyse** mit **Open Source** Software

Fahrplan

Vorstellung

Einstieg und Motivation

Vorstellung Architekturen und Open Source Tools

Fragen und Diskussion

kielux@msuhr.net

Markus Suhr
*1987
M.Sc Angewandte Informatik / Medizininformatik

Universitätsmedizin Göttingen

2011 Applikationsbetreuer Lehr- u. Lernmanagementsysteme

2014 Systemadministrator, Software Entwickler

2017 Forschungsdatenmanagement, Data Warehousing

NextLytics AG

2022 Consultant, Teamleiter **Data Science & Engineering**

2014 - 2017

2017 - 2019

2019 - 2021
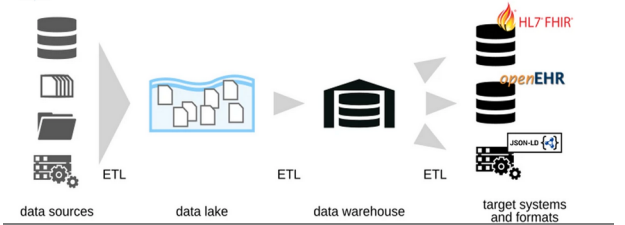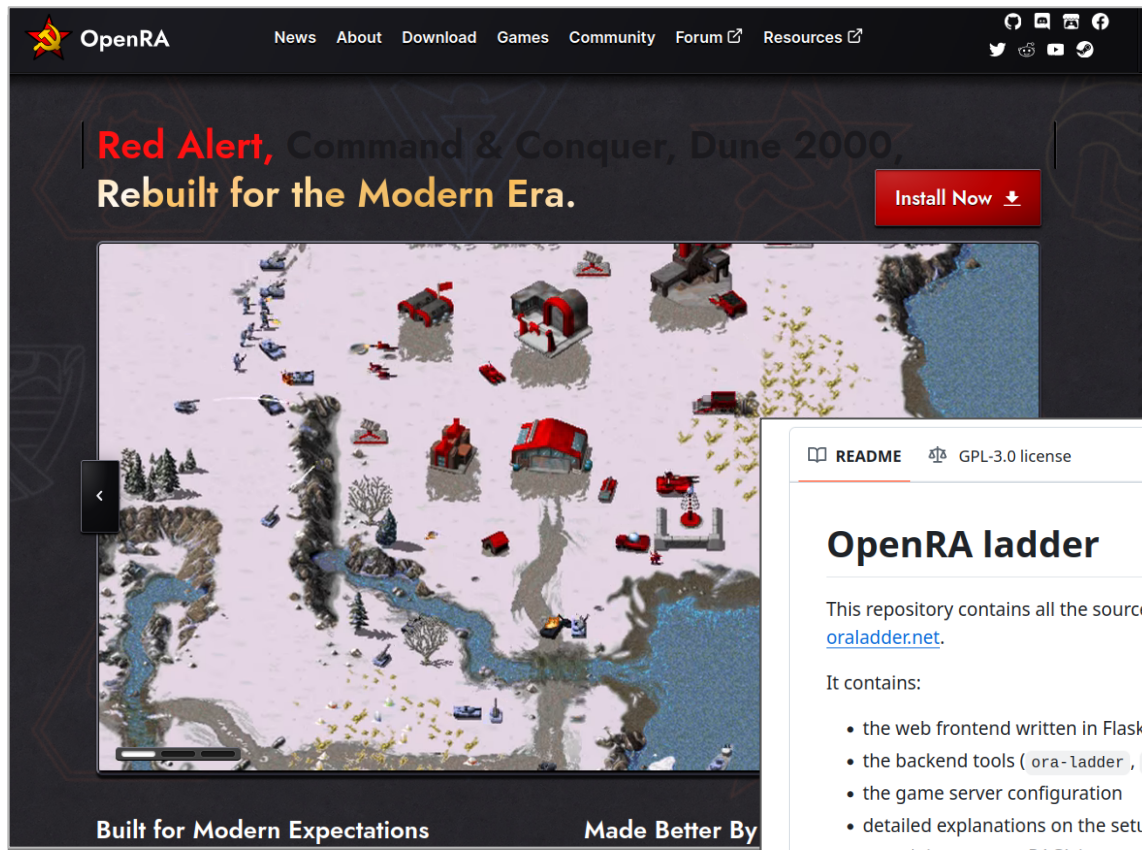
…und Open Source nach Feierabend 🐱

https://www.openra.net

Fahrplan

Vorstellung

Einstieg und Motivation

Vorstellung Architekturen und Open Source Tools

Fragen und Diskussion

kielux@msuhr.net

Vokabeltest zum Einstieg

Datenbanken und Structured Query Language (SQL)

Docker, Git, DevOps

Continuous Integration / Continuous Delivery (CICD)

Data Warehouse

Automatisierung, Orchestrierung

Governance, Data Lineage

kielux@msuhr.net

Wieso *professionell*?

…eher als Qualitätsmerkmal gemeint, denn im Wortsinn:

Verarbeitung sollte automatisiert und reproduzierbar erfolgen,

Änderungen müssen nachvollziehbar sein, Verantwortlichkeiten und Berechtigungen klar geregelt, (...)

Kurz: sämtliche Lehren aus Ingenieurwissenschaften und Software-Entwicklung sollen auch auf Datenverarbeitung und Datenanalyse angewendet werden

kielux@msuhr.net

Grundbausteine der Datenanalyse
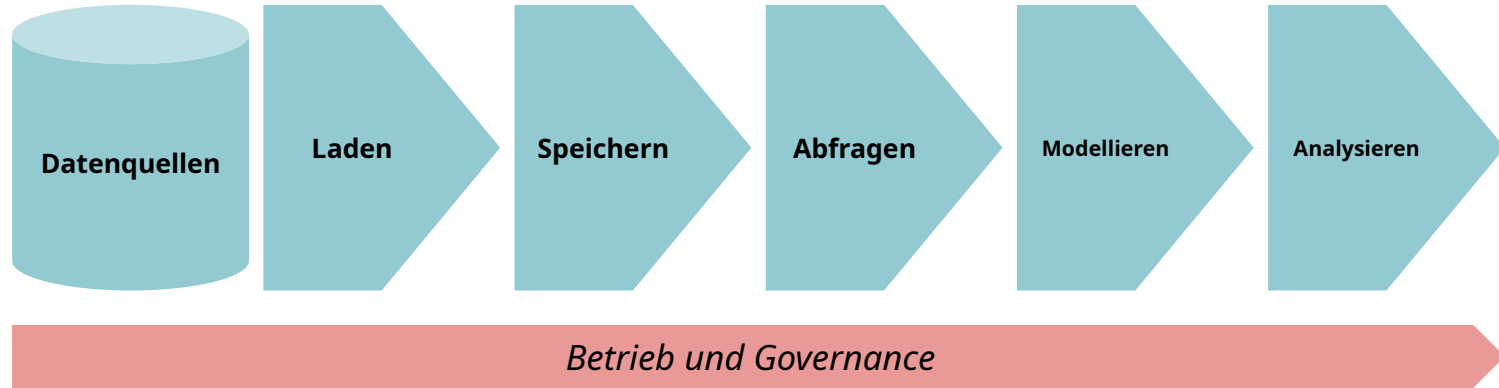
Datenquellen	Laden	Speichern	Abfragen	Modellieren	Analysieren

Betrieb und Governance

kielux@msuhr.net

ellieren Analysieren

	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard
1	KEY	FREQ	CURRENCY	CURRENCY_DENOM	EXR_TYPE	EXR_SUFFIX	TIME_PERIOD	OBS_VALUE	OBS_STATUS
2	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-13	1.02745	A
3	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-14	1.02232	A
4	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-17	1.0087	A
5	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-18	1.00915	A
6	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-19	1.01034	A
7	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-20	1.00873	A
8	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-21	1.00954	A
9	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-24	1.00216	A
10	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-25	1.00064	A
11	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-26	1.00172	A
12	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-27	0.99743	A
13	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-28	0.98464	A
14	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-31	0.97905	A
15	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-01	0.97085	A
16	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-02	0.9715	A
17	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-03	0.9746	A
18	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-04	0.98329	A
19	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-07	0.9761	A
20	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-08	0.99009	A
21	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-09	0.99318	A
22	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-10	0.9866	A
23	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-11	0.9799	A
24	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-14	0.98048	A
25	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-15	0.97759	A
26	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-16	0.9802	A
27	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-02-17	0.9917	A

https://data-api.ecb.europa.eu/service/data/EXR/D..EUR.SP00.A?format=csvdata&start_date=2024-09-19

	Standard	Standard	Standard	Standard	Standard	Standard	Standard
1	KEY	FREQ	CURRENCY	CURRENCY_DENOM	EXR_TYPE	EXR_SUFFIX	TIME_PERIOD
2	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00	A	2000-01-13
3	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
4	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
5	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
6	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
7	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
8	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
9	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
10	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
11	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
12	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
13	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
14	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
15	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
16	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
17	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
18	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
19	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
20	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
21	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
22	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
23	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
24	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
25	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
26	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			
27	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR			

	exchange_rate_date	currency	currency_name	rate	_loadtime
21	2024-09-18	HKD	Hong Kong dollar	8,669	2024-09-19 00:00:04.785
22	2024-09-18	GBP	UK pound sterling	0,842	2024-09-19 00:00:04.785
23	2024-09-18	DKK	Danish krone	7,461	2024-09-19 00:00:04.785
24	2024-09-18	CZK	Czech koruna	25,07	2024-09-19 00:00:04.785
25	2024-09-18	CNY	Chinese yuan renminbi	7,88	2024-09-19 00:00:04.785
26	2024-09-18	CHF	Swiss franc	0,939	2024-09-19 00:00:04.785
27	2024-09-18	CAD	Canadian dollar	1,512	2024-09-19 00:00:04.785
28	2024-09-18	BRL	Brazilian real	6,094	2024-09-19 00:00:04.785
29	2024-09-18	BGN	Bulgarian lev	1,956	2024-09-19 00:00:04.785
30	2024-09-18	AUD	Australian dollar	1,64	2024-09-19 00:00:04.785
31	2024-09-17	ZAR	South African rand	19,642	2024-09-18 00:00:05.051
32	2024-09-17	USD	US dollar	1,114	2024-09-18 00:00:05.051
33	2024-09-17	TRY	Turkish lira	37,914	2024-09-18 00:00:05.051
34	2024-09-17	THB	Thai baht	37,104	2024-09-18 00:00:05.051
35	2024-09-17	SGD	Singapore dollar	1,441	2024-09-18 00:00:05.051
36	2024-09-17	SEK	Swedish krona	11,324	2024-09-18 00:00:05.051
37	2024-09-17	RON	Romanian leu	4,974	2024-09-18 00:00:05.051
38	2024-09-17	PLN	Polish zloty	4,268	2024-09-18 00:00:05.051
39	2024-09-17	PHP	Philippine peso	61,966	2024-09-18 00:00:05.051

https://data-api.ecb.europa.eu/service/data/EXR/D..EUR.SP00.A?format=csvdata&start_date=2024-09-19

"Normalisierte Daten":
Bereinigt und für die
weitere Verarbeitung
aufbereitet; hier:
aktueller Wechselkurs

	Standard	Standard	Standard	Standard	Standard
	KEY	FREQ	CURRENCY	CURRENCY_DENOM	EXR_TYPE
1					
2	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00
3	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00
4	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00
5	EXR.D.ARS.EUR.SP00.A	D	ARS	EUR	SP00
6	EXR.D.ARS.EUR.SP00.A	D			
7	EXR.D.ARS.EUR.SP00.A	D			
8	EXR.D.ARS.EUR.SP00.A	D			
9	EXR.D.ARS.EUR.SP00.A	D			
10	EXR.D.ARS.EUR.SP00.A	D			
11	EXR.D.ARS.EUR.SP00.A	D			
12	EXR.D.ARS.EUR.SP00.A	D			
13	EXR.D.ARS.EUR.SP00.A	D			
14	EXR.D.ARS.EUR.SP00.A	D			
15	EXR.D.ARS.EUR.SP00.A	D			
16	EXR.D.ARS.EUR.SP00.A	D			
17	EXR.D.ARS.EUR.SP00.A	D			
18	EXR.D.ARS.EUR.SP00.A	D			
19	EXR.D.ARS.EUR.SP00.A	D			
20	EXR.D.ARS.EUR.SP00.A	D			
21	EXR.D.ARS.EUR.SP00.A	D			
22	EXR.D.ARS.EUR.SP00.A	D			
23	EXR.D.ARS.EUR.SP00.A	D			
24	EXR.D.ARS.EUR.SP00.A	D			
25	EXR.D.ARS.EUR.SP00.A	D			
26	EXR.D.ARS.EUR.SP00.A	D			
27	EXR.D.ARS.EUR.SP00.A	D			

	exchange_rate_date	currency
21	2024-09-18	HKD
22	2024-09-18	GBP
23	2024-09-18	DKK
24	2024-09-18	CZK
25	2024-09-18	CNY
26	2024-09-18	CHF
27	2024-09-18	CAD
28	2024-09-18	BRL
29	2024-09-18	BGN
30	2024-09-18	AUD
31	2024-09-17	ZAR
32	2024-09-17	USD
33	2024-09-17	TRY
34	2024-09-17	THB
35	2024-09-17	SGD
36	2024-09-17	SEK
37	2024-09-17	RON
38	2024-09-17	PLN
39	2024-09-17	PHP

	currency	currency_name		
1	AUD	Australian dollar		200
2	BGN	Bulgarian lev	4,	2024-09-19 02:00:08.639 +0200
3	BRL	Brazilian real	6,094	2024-09-19 02:00:08.639 +0200
4	CAD	Canadian dollar	1,512	2024-09-19 02:00:08.639 +0200
5	CHF	Swiss franc	0,939	2024-09-19 02:00:08.639 +0200
6	CNY	Chinese yuan renminbi	7,88	2024-09-19 02:00:08.639 +0200
7	CZK	Czech koruna	25,07	2024-09-19 02:00:08.639 +0200
8	DKK	Danish krone	7,461	2024-09-19 02:00:08.639 +0200
9	GBP	UK pound sterling	0,842	2024-09-19 02:00:08.639 +0200
10	HKD	Hong Kong dollar	8,669	2024-09-19 02:00:08.639 +0200
11	HUF	Hungarian forint	394,25	2024-09-19 02:00:08.639 +0200
12	IDR	Indonesian rupiah	17.032,01	2024-09-19 02:00:08.639 +0200
13	ILS	Israeli shekel	4,192	2024-09-19 02:00:08.639 +0200
14	INR	Indian rupee	93,132	2024-09-19 02:00:08.639 +0200
15	ISK	Iceland krona	152,30	2024-09-19 02:00:08.639 +0200
16	JPY	Japanese yen	157,94	2024-09-19 02:00:08.639 +0200
17	KRW	Korean won (Republic)	1.473,65	2024-09-19 02:00:08.639 +0200
18	MSD	Markus Dollar	1,00	2024-09-19 02:00:08.639 +0200
19	MXN	Mexican peso	21,24	2024-09-19 02:00:08.639 +0200
20	MYR	Malaysian ringgit	4,721	2024-09-19 02:00:08.639 +0200
21	NOK	Norwegian krone	11,775	2024-09-19 02:00:08.639 +0200
22	NZD	New Zealand dollar	1,788	2024-09-19 02:00:08.639 +0200
23	PHP	Philippine peso	61,988	2024-09-19 02:00:08.639 +0200
24	PLN	Polish zloty	4,27	2024-09-19 02:00:08.639 +0200
25	RON	Romanian leu	4,975	2024-09-19 02:00:08.639 +0200
26	SEK	Swedish krona	11,324	2024-09-19 02:00:08.639 +0200
27	SGD	Singapore dollar	1,44	2024-09-19 02:00:08.639 +0200
28	THB	Thai baht	37,015	2024-09-19 02:00:08.639 +0200
29	TRY	Turkish lira	37,943	2024-09-19 02:00:08.639 +0200
30	USD	US dollar	1,112	2024-09-19 02:00:08.639 +0200
31	ZAR	South African rand	19,553	2024-09-19 02:00:08.639 +0200

Standard	Standard	Standard	Stan
KEY	FREQ	CURRENCY	CURI
EXR.D.ARS.EUR.SP00.A	D	ARS	EUR
EXR.D.ARS.EUR.SP00.A	D	ARS	EUR
EXR.D.ARS.EUR.SP00.A	D	ARS	EUR
EXR.D.ARS.EUR.SP00.A	D	ARS	EUR
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		

https://data-api.ecb.europa.eu/serv

	date_ym	currency	currency_nar	avg
1	2024-9	AUD	Australian dollar	1,648
2	2024-8	AUD	Australian dollar	1,656
3	2024-7	AUD	Australian dollar	1,626
4	2024-6	AUD	Australian dollar	1,621
5	2024-5	AUD	Australian dollar	1,632
6	2024-4	AUD	Australian dollar	1,647
7	2024-3	AUD	Australian dollar	1,659
8	2024-2	AUD	Australian dollar	1,653
9	2024-1	AUD	Australian dollar	1,642
10	2023-12	AUD	Australian dollar	1,632
11	2023-11	AUD	Australian dollar	1,663
12	2023-10	AUD	Australian dollar	1,664
13	2023-9	AUD	Australian dollar	1,662
14	2023-8	AUD	Australian dollar	1,682
15	2023-7	AUD	Australian dollar	1,642
16	2023-6	AUD	Australian dollar	1,616
17	2023-5	AUD	Australian dollar	1,635

"Analyse": für die menschliche Interpretation aufbereitete Darstellung, hier: Entwicklung durchschnittlicher EUR-USD Wechselkurse pro Quartal

currency	currency_name	rate	update_timestamp
AUD	Australian dollar	1,64	2024-09-19 02:00:08.639 +0200
BGN			3.639 +0200
BRL			3.639 +0200
CAD			3.639 +0200

Standard	Standard	Standard	Stand
KEY	FREQ	CURRENCY	CURR
EXR.D.ARS.EUR.SP00.A	D	ARS	EUR
EXR.D.ARS.EUR.SP00.A	D	ARS	EUR
EXR.D.ARS.EUR.SP00.A	D	ARS	EUR
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		
EXR.D.ARS.EUR.SP00.A	D		

date_ym	currency
2024-9	AUD
2024-8	
2024-7	
2024-6	
2024-5	
2024-4	
2024-3	
2024-2	
2024-1	
2023-12	
2023-11	
2023-10	
2023-9	
2023-8	
2023-7	
2023-6	AUD	Australian dollar	1,616
2023-5	AUD	Australian dollar	1,635
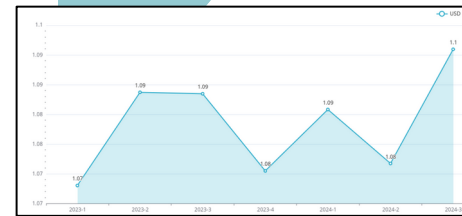
USD

Chart: EUR-USD Wechselkurse pro Quartal
- 2023-1: 1.07
- 2023-2: 1.09
- 2023-3: 1.09
- 2023-4: 1.08
- 2024-1: 1.09
- 2024-2: 1.08
- 2024-3: 1.1

| | USD | US dollar | 1,112 | 2024-09-19 02:00:08.639 +0200 |
| | ZAR | South African rand | 19,553 | 2024-09-19 02:00:08.639 +0200 |

Datenquellen → Laden → Speichern → Abfragen → Modellieren → Analysieren

Datenquellen → **Laden** → **Speichern** → **Abfragen** → **Modellieren** → **Analysieren**
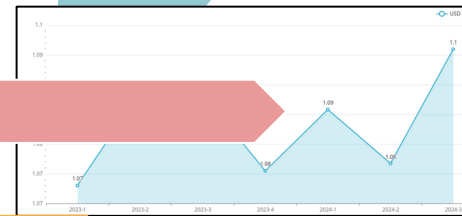
Betrieb und Governance

Alle Lade- und Transformationsschritte sollen automatisiert und reproduzierbar ausgeführt werden - und wenn etwas nicht funktioniert, will ich automatisch informiert werden

Funktionales Spektrum einer maximal ausgebauten Datenanalyse-Plattform

Sources

- SAP Systems
- Non-SAP ERP Systems
- OLTP Databases
- Business Applications
- Event Queues
- Logs
- 3rd Party APIs
- File and Object Storage

Ingestion

- Replication
- Generic Delta Loading
- Change Data Capture
- Event Streaming
- Federation

Storage

Data Warehouse

Cold / Warm / Hot Storage

Data Lakehouse

Interoperable Data Lake Storage

Query/Processing

Data Warehouse

Centralized Query and Processing Engines

Data Lakehouse

Distributed Query and Processing Engines

Stream Processing

Modeling

- Metrics Layer
- Semantic Model
- Feature Store

Consumption

- Dashboards
- Business Planning
- Self-Service Analytics
- Embedded Analytics
- Data Science & Machine Learning Workspace
- Custom Applications
- Reverse ETL

- Governance
- Discoverability
- Collaboration
- Security

Vereinfachte Darstellung, angelehnt an die *Modern Data Architectures* von Andreessen Horowitz

Datenquellen	Laden	Speichern	Abfragen	Modellieren	Analysieren

Betrieb & Governance

THE 2024 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE

- **INFRASTRUCTURE** — Storage, MPP DBs, Data Lakes/Lakehouses, Data Warehouses, Streaming/In-Memory, RDBMS, NoSQL Databases, NewSQL Databases, Real Time Databases, Graph DBs, Vector Databases, Multi-Model Databases & Abstractions, ETL/ELT/Data Transformation, Reverse ETL, Data Integration, Data Governance & Catalog, Orchestration, Data Quality & Observability, Fully Managed, MGMT/Monitoring, Privacy & Security, Compute
- **ANALYTICS** — BI Platforms, Visualization, Data Analyst Platforms, Customer Data Platforms, Product Analytics, Log Analytics, Enterprise Search/Knowledge Analytics
- **MACHINE LEARNING & ARTIFICIAL INTELLIGENCE** — Data Science Notebooks, Data Science Platforms, Enterprise ML/AI Platforms, Data Generation & Labeling, MLOps, AI Observability, Computer Vision, AI Developer Platforms, AI Safety & Security, Nonprofit AI Research, Speech/Voice, NLP, Commercial AI Research, AI Hardware, GPU Cloud/Infra, Edge AI, Closed Source Models
- **APPLICATIONS — ENTERPRISE** — Sales, Marketing, Customer Experience, Human Capital, Automation & Operations, Decision & Optimization, Legal, Partnerships, Regtech & Compliance, Finance
- **APPLICATIONS — HORIZONTAL** — Code & Documentation, Text, Audio & Voice, Image, Presentation & Design, Video Editing, Animation & 3D/Gaming, Search/Conversational AI, Video Generation
- **APPLICATIONS — INDUSTRY** — Finance & Insurance, Healthcare, Life Sciences, Transportation, Agriculture, Industrial & Logistics, Aerospace, Defense & Govt, Cross-Industry
- **OPEN SOURCE INFRASTRUCTURE** — Data Frameworks, Formats, Query/Data Flow, Data Management, Databases, OLAP, Orchestration, Infrastructure, Streaming & Messaging, Stat Tools & Languages, MLOps & AI Infra, AI Frameworks, Tools & Libraries, AI Models, Local AI, Search, Logging & Monitoring, Visualization, Collaboration
- **DATA SOURCES & APIS** — Data Marketplaces & Discovery, Financial & Market Data, Air/Space/Sea, People/Entities, Location Intelligence, ESG
- **DATA & AI CONSULTING**

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

https://mad.firstmark.com/

Open Source Infrastructure

Data Frameworks
Apache Hadoop · Spark · Apache Kafka · Apache Flink · yarn · kubernetes · MESOS · docker · Open Nebula · RAY · dask · HELIX · HighAvailability

Formats
{json} · Parquet · ICEBERG · hudi · delta lake · orc · protobuf · FlatBuffers · CAP'N PROTO · THRIFT · MessagePack · Avro

Query/Data Flow
Spark · Apache Pig · HIVE · presto · trino · GraphQL · Apache DRILL · beam · CloudQuery · HAWQ

Data Management
Amundsen · DataHub · Apache Atlas · MARQUEZ · MARQUEZ · great expectations · lakeFS · elementary · OpenLineage · ckan · Project Nessie · SINGER

Databases
PostgreSQL · MySQL · MongoDB · redis · CockroachDB · SQLite · CASSANDRA · MariaDB · influxdb · HBASE · Firebird · supabase · Timescale · SurrealDB · Milvus · Chroma · weaviate · NEON · RocksDB · SciDB-Py · OpenTSDB · LEVELDB · riak · accumulo · Ignite · FoundationDB · TiKV · spicedb · OCEANBASE

OLAP
ClickHouse · StarRocks · druid · pinot · DuckDB · trino · DORIS

Orchestration
Apache Airflow · PREFECT · dagster · Flyte · METAFLOW · luigi · argo · MAGE · kestra · DolphinScheduler · Cadence

Infrastructure
Apache Ambari · Apache MESOS

Streaming & Messaging
Spark Streaming · kafka · Apache Flink · beam · Apache STORM · PULSAR · RabbitMQ · Apache RocketMQ · samza · nifi · debezium · ESTUARY Flow · ACTIVEMQ · mosquitto · memphis · bytewax · Gazette

Stat Languages
python · R · Scala · NumPy · Julia · SciPy · Octave · datatable · Sas

MLOps & AI Infra
LangChain · PRIVATE GPT · MediaPipe · jina · mindsdb · FlowAI · mlflow · Kubeflow · DVC · Dify · cortex · Pachyderm · BENTOML · Snorkel · Chainlit · polyaxon · ollama · ZenML · deepchecks · mleap · Gorilla · CausalML · LLM

AI Frameworks, Tools & Libraries
PyTorch · TensorFlow · PyTorch · Keras · Hugging Face Candle · fast.ai · OpenCV · XGBoost · DL4J · Caffe2 · deepspeed · SuperAGI · BabyAGI · mxnet · Sonnet · PyCARET · JARVIS · Chainer · ONNX · LUDWIG · OPTUNA · Spark · H2O.ai · CoreNLP · SINGA · Kedro · openNN · NIXTLA TimeGPT

AI Models & Architectures
Meta AI Llama 2 · stability.ai Stable Diffusion · Mistral 7B · Falcon 180B · BLOOM · Hugging Face Vicuna 13B · OpenAI Whisper-v3 · OpenAI Whisper-v3 · Mistral 7B Instruct · Google AI PaLM 2 · EleutherAI Pythia · 01.AI Yi-34B · Together.ai RedPajama · EleutherAI GPT-J-6B · EleutherAI GPT-Neo-2.7B · Together.ai StripedHyena 7B · Together.ai StripedHyena 7B · Together.ai StripedHyena 7B · Hugging Face Zephyr-7B · Microsoft Phi-2 · Microsoft Phi-2 · Google Research Bert · Google Research T5 · Cerebras-GPT · Stanford Alpaca · DeepMind Flamingo · Meta AI BlenderBot · RoBERTa · DistilBERT · Google AI · XLNet · Meta AI ESM · Meta AI Diplomacy Cicero · Meta AI OPT-66B · Google Research Transformer · NVIDIA Megatron · Google Research Dreamix · codium AlphaCodium · Google DeepMind AlphaGeometry

Search
Solr · Apache LUCENE · elasticsearch · meilisearch · Sonic · typesense · OpenSearch · swirl · Toshi Search · tantivy

Local AI
OpenLLM · LLaMA · llamafile · GPT4All · Text generation web UI · LocalGPT · llamafile · LM Studio

Logging & Monitoring
elastic · Grafana · Prometheus · Sentry · opentelemetry · logstash · fluentd · fluentbit · VECTOR · SigNoz · openobserve · graylog · highlight.io · syslog-ng

Visualization
matplotlib · plotly · redash · Superset · TensorFlow TensorBoard · PyGWalker · seaborn · bokeh · Datawrapper · RAWGraphs

Collaboration
BeakerX · Jupyter · Apache Zeppelin · RStudio · Polynote

https://mad.firstmark.com/
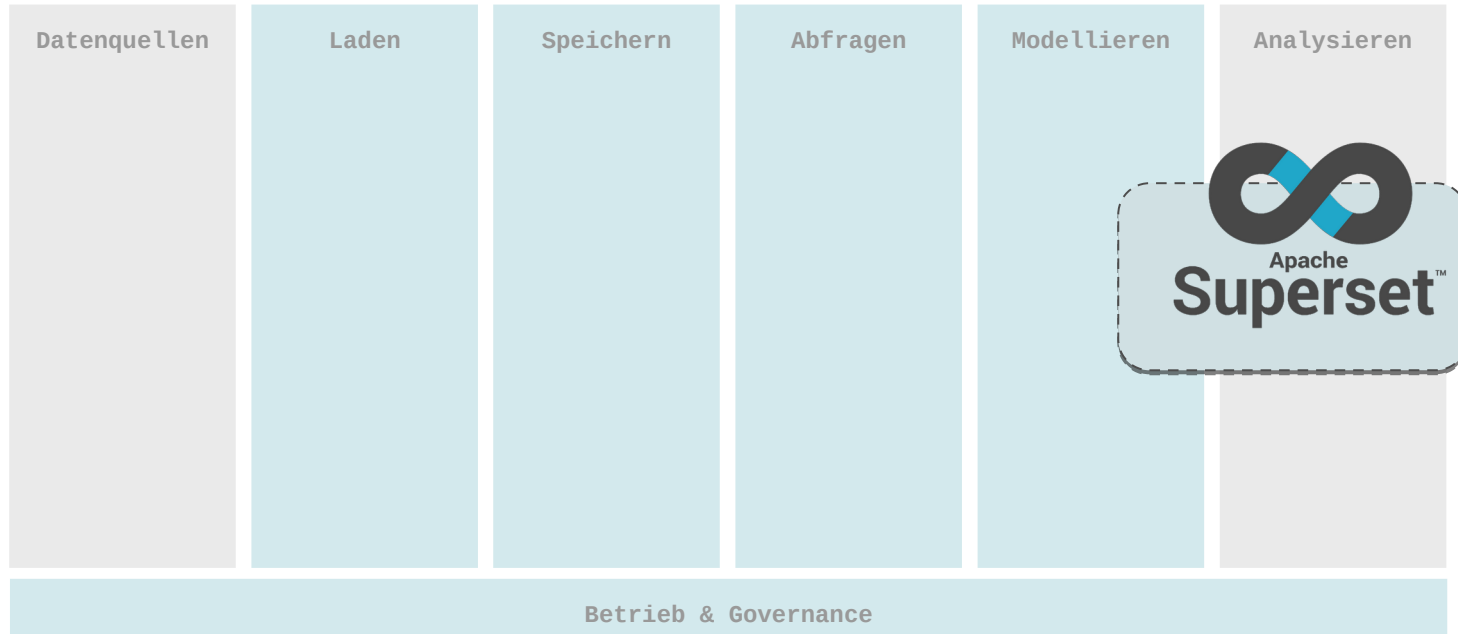
Fahrplan

Vorstellung

Einstieg und Motivation

Vorstellung Architekturen und Open Source Tools

Fragen und Diskussion

kielux@msuhr.net

Datenanalyse und Visualisierung

Datenquellen	Laden	Speichern	Abfragen	Modellieren	Analysieren

Betrieb & Governance

⋮ Untitled Query 1 ⬤ ✕ | ⋮ Untitled Query 2 ✓ ✕ | ⊕

DATABASE

postgresql Nextlytics Lottery Database (... ▾

SCHEMA

public ▾ ↻

SEE TABLE SCHEMA

⊞ eur_exchange_rates ✕ ▾ ↻

eur_exchange_rates ⌃

exchange_rate_date	DATE
currency	VARCHAR
currency_name	VARCHAR
rate	NUMERIC
_loadtime	TIMESTAMP

```
1  SELECT extract(year from exchange_rate_date) || '-' || extract(quarter from exchange_rate_date) as _quarter,
2  currency,
3  AVG(rate)
4  FROM eur_exchange_rates
5  where currency='USD'
6  GROUP BY
7  extract(year from exchange_rate_date),
8  extract(quarter from exchange_rate_date),
9  currency
10 order by 3, 1 desc, 2 desc
```

RUN   LIMIT: 1 000 ▾   00:00:01.18   SAVE ▾   🔗 COPY LINK   ⋯

RESULTS   QUERY HISTORY

📊 CREATE CHART   DOWNLOAD TO CSV   📋 COPY TO CLIPBOARD   Filter results

7 rows returned

_quarter ⇅	currency ⇅	avg ⇅
2023-1	USD	1.073006153846154
2023-4	USD	1.075466129032258
2024-2	USD	1.0767174603174603
2024-1	USD	1.0857873015873016

Apache
Superset™

Leaderboard

current_ranking	name	winner	num_of_rows	dagrun_token
1	dazzling_roentgen	WINNER	181159	0A07EFA
2	tender_pike	WINNER	177829	FB46493
3	dreamy_shockley	WINNER	176793	E78612D
4	fervent_turing	WINNER	173442	F2DB75A
5	epic_mcnulty	WINNER	172051	914F9BD
6	optimistic_borg	WINNER	159020	A493963
7	blissful_bartik	WINNER	155840	A9A55F9
8	mystifying_jennings	WINNER	153215	7F05FC6
9	clever_meitner	WINNER	153058	6960935
10	confident_perlman	WINNER	151276	83B0D4F

Latest Participants

start_time	name	dagrun_token	winner	current_ranking	num_of_rows
2024-05-27 06:49:41	mystifying_jennings	7F05FC6	WINNER	8	153215
2024-05-27 06:49:32	festive_golick	0FAC504	N/A	51	114708
2024-05-23 13:45:17	dreamy_swanson	A42178D	N/A	135	83730
2024-05-23 13:34:54	gracious_payne	D60C48F	N/A	125	87453
2024-05-23 13:19:03	vibrant_hugle	B158B5B	WINNER	28	127575
2024-05-23 13:09:11	trusting_kilby	884449E	N/A	173	45901
2024-05-23 13:09:10	trusting_bohr	E02CE0F	N/A	172	45901
2024-05-23 12:53:55	vigorous_neumann	F8A1D99	N/A	155	66584
2024-05-23 12:53:48	clever_meitner	6960935	WINNER	9	153058
2024-05-23 12:44:05	thirsty_blackwell	C14CCA1	WINNER	11	150966

Progress

win rate

WINNER: 21.98%

Total: 182

no: 78.02%
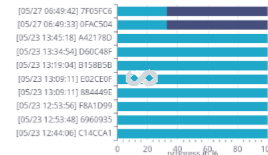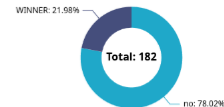
Data points ingested
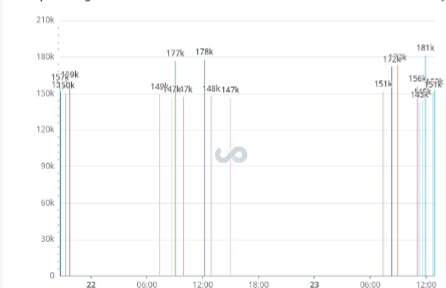
Total rows in database
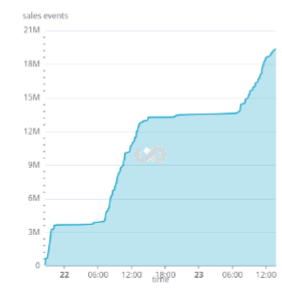
Ingestion speed over time

Sales map
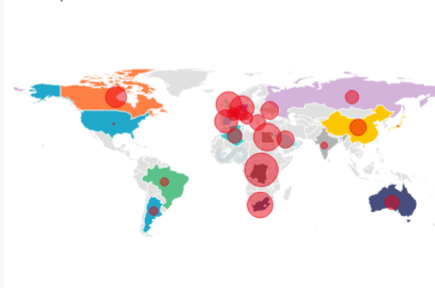
Top selling products

product_id	units sold	Revenue
7363	356419	148M
8615	348177	147M
3175	346782	147M
4796	350129	146M
999	354599	146M
11897	330098	146M
1969	348806	146M
9564	352339	146M
5078	348608	145M
11082	343049	145M
8720	353234	145M
2997	341725	144M
8581	346376	144M
9513	344757	144M

Apache Superset

Klassisches Open Source Data Warehouse

kielux@msuhr.net

https://airflow.apache.org/docs/apache-airflow/stable/ui.html

Ende-zu-Ende Betrieb mit Open Source Tools

source code version control
& CI/CD pipelines

triggering webservice,
data processing logic

distributed,
parallel task execution

analytics
and visualization
(dashboard)

server operating system

containerized
deployment

workflow orchestration,
task distribution

persistent storage
& data aggregation

kielux@msuhr.net

Effizienz auch in großem Maßstab

kielux@msuhr.net
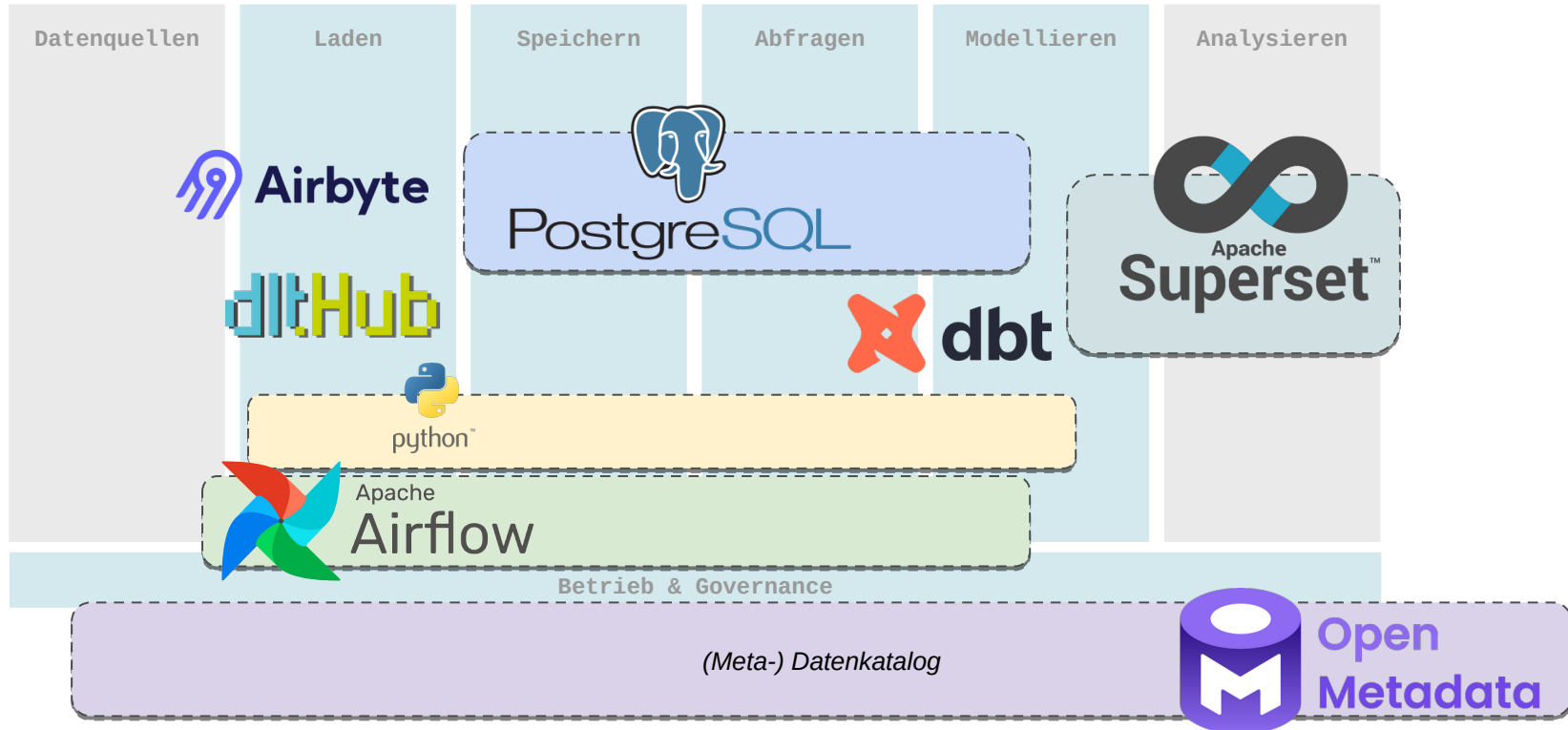
Einsteiger-Architektur (weniger Code-lastig)

Datenquellen	Laden	Speichern	Abfragen	Modellieren	Analysieren

PostgreSQL

Airbyte

Apache Superset

Betrieb & Governance

kielux@msuhr.net

Einsteiger- Data <u>Lake</u>house mit Dremio

Datenquellen	Laden	Speichern	Abfragen	Modellieren	Analysieren

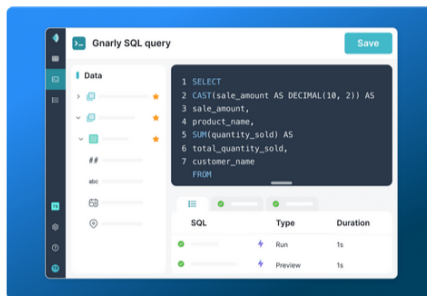
Betrieb & Governance

kielux@msuhr.net
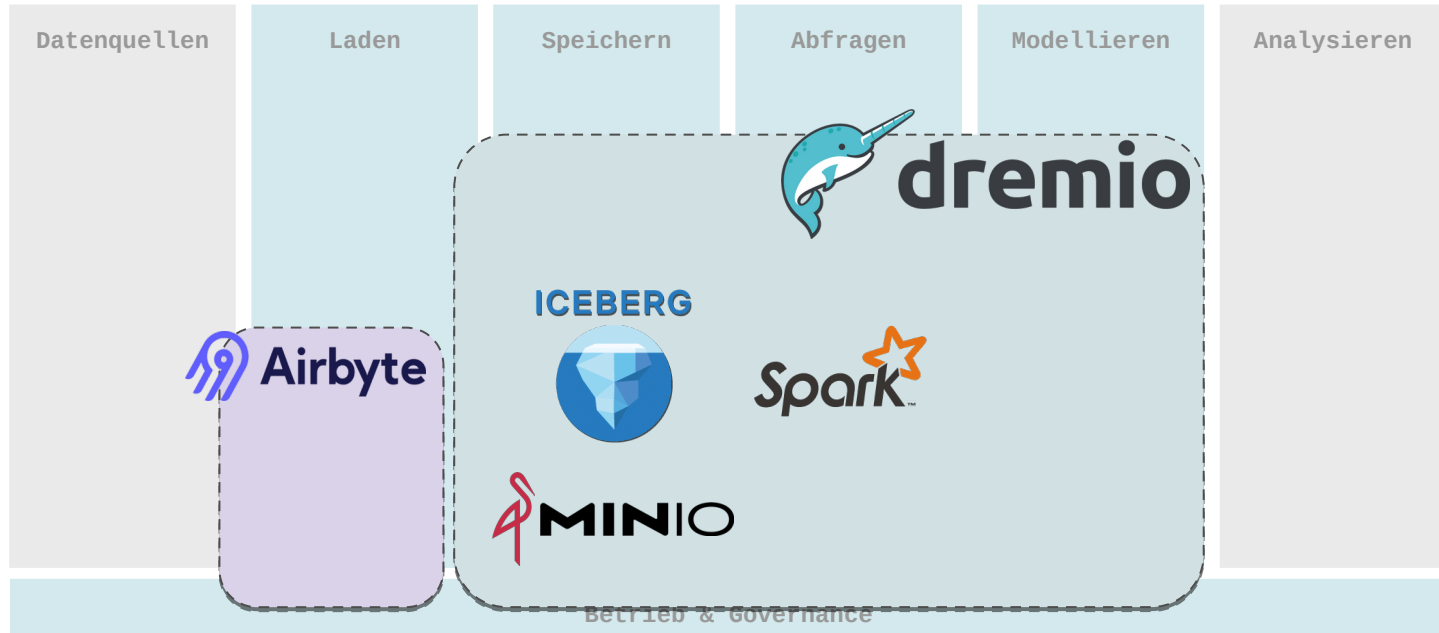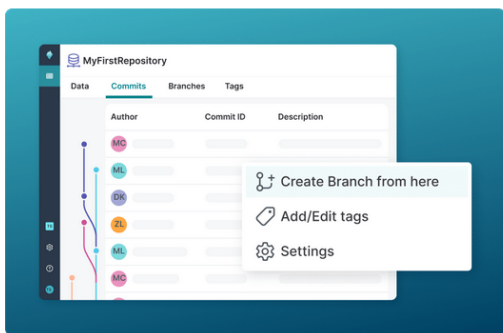
Flexible version control with data history and time-travel

Reproduce models and dashboards with catalog-level time-travel, and use tags to identify milestones in the history of your data. Instantly recover from mistakes by rolling back your lakehouse to a prior state.

Learn more about time-travel →

Zero-copy, zero-movement data

Flexible, fast, lightweight data transformation

Reduce reliance on costly ETL tools and brittle, complex data pipelines. Dremio makes it easy to apply last mile data transformations, including filtering, sorting, aggregating, joining, casting. These transformations can be quickly built as Dremio Views - governed virtual data sets with layered transformations - that can be flexibly shared and modified for downstream data projects.

Learn more about virtual data transformation with Views →

Federated querying for all your data, everywhere

Designed for interactive analytics and DML on the data lake, the Dremio SQL Query Engine makes it easy to analyze all of your data - whether on the data lake or in other data sources. Our connector ecosystem features dozens of integrations with an array of sources, including object storage, metastores, and databases in the cloud and on premises. Because Dremio queries your data at the source, there is no data movement, no data copies, and no complex ETL.

See our connector list →

Fahrplan

Vorstellung

Einstieg und Motivation

Vorstellung Architekturen und Open Source Tools

Fragen und Diskussion

kielux@msuhr.net

Fazit in einem Satz!?

Die **Kombinationen** von

spezifischen Schnittstellen, Anforderungen und Einschränkungen

sowie **verfügbaren Open Source Tools**

für automatisierte, professionelle Datenverarbeitung und -analyse

sind praktisch unendlich!

(es werden ständig mehr)

kielux@msuhr.net

Fazit ~~in einem Satz!?~~

Open Source Software ist ein absoluter **Innovationstreiber**

und ermöglicht einen schnellen **Einstieg** in die Welt **professioneller Datenanalyse**

- auch privat oder fürs Hobby 🐧

…einfach mal ausprobieren

💻 `docker run` und Schluss mit Excel
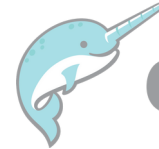
kielux@msuhr.net

Fragen, Diskussion, Ideen, Feedback

Professionelle **Datenanalyse** mit **Open Source** Software