Fontys › FOR SOCIETY

AI4DATA: WHAT IS AN AI-AUGMENTED DATA ENGINEER?

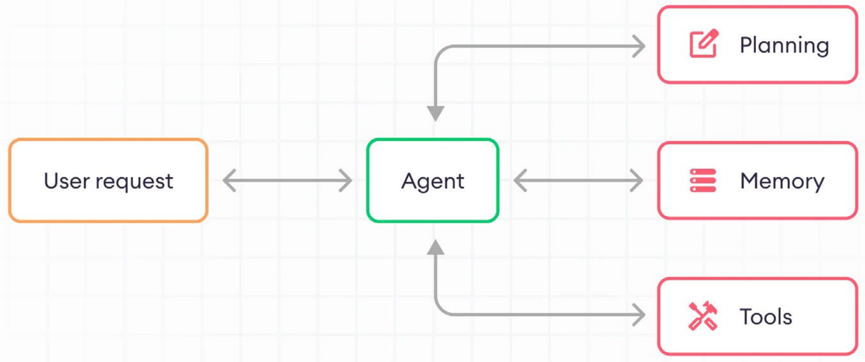
Petra Heck – Associate Professor - p.heck@fontys.nl

About Me

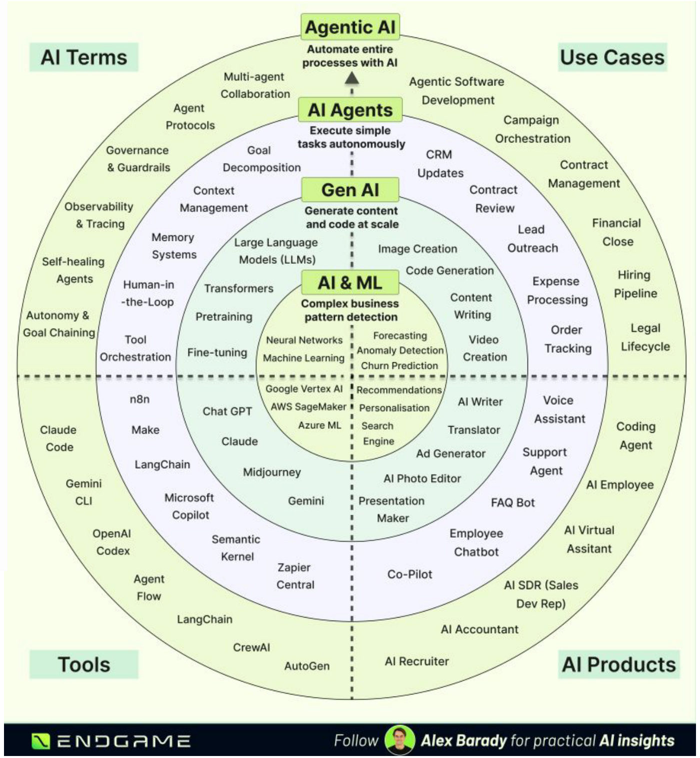
- 2002 **Software engineer** (MSc Computer Science)
- 2004 **Software quality** consultant
- 2012 Lecturer **Fontys ICT**
  - 2016 PhD “Quality of JIT Requirements” (Computer Science, AI4RE)
  - 2019 Senior researcher AI Engineering (SE4AI)
- 2025 **Associate Professor AI & Software Engineering** (AI4SE & SE4AI)

Fontys

AI in 2026 = GenAI & Agentic AI



[LLM agents: The ultimate guide 2025 | SuperAnnotate]



[(9) Post | LinkedIn]

[Software Engineering in the AI Era: from ML to LLMs and Agentic AI]

From MLOps to AI-augmented Engineering

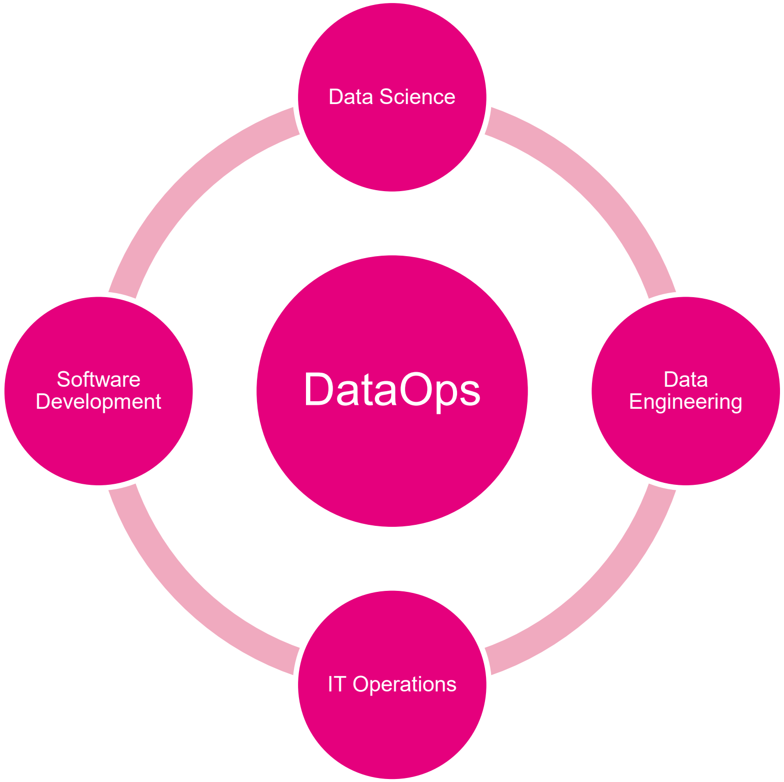
- 2019 AI Engineering and **MLOps**: Building Production-Ready Machine Learning Systems
- 2024 From MLOps to **DataOps**: Data Engineering for AI-based Systems
- 2024 **LLMOps**: Engineering Trustworthy LLM Systems
- 2025 Software Engineering in the AI Era: from ML to LLMs and **Agentic AI**
- 2026 Competences for the **AI-Augmented Software Engineer**

Fontys

DataOps

“an approach that accelerates the delivery of high-quality results by *automation and orchestration* of data life cycle stages.”

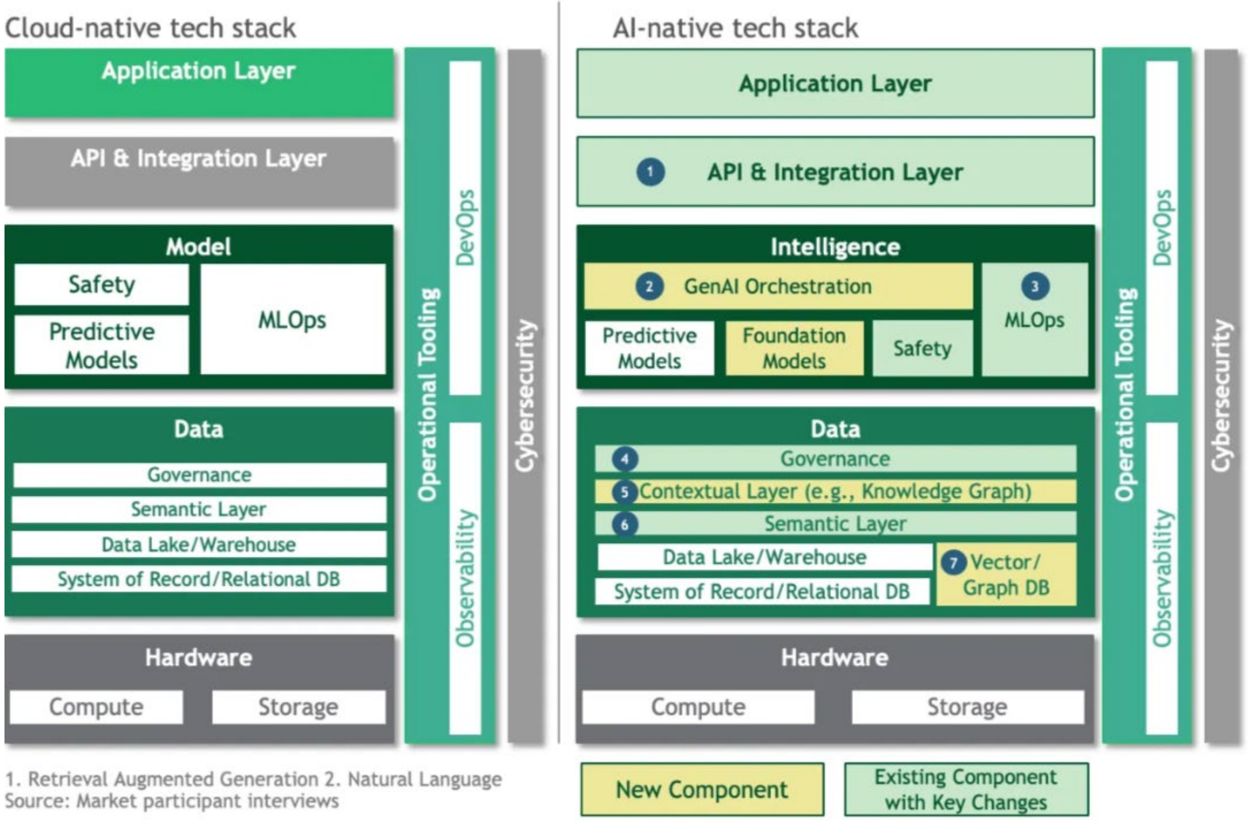
DataOps adopts the best practices, processes, tools and technologies from Agile software engineering and DevOps thereby promoting the culture of *collaboration and continuous improvement*.



Fontys

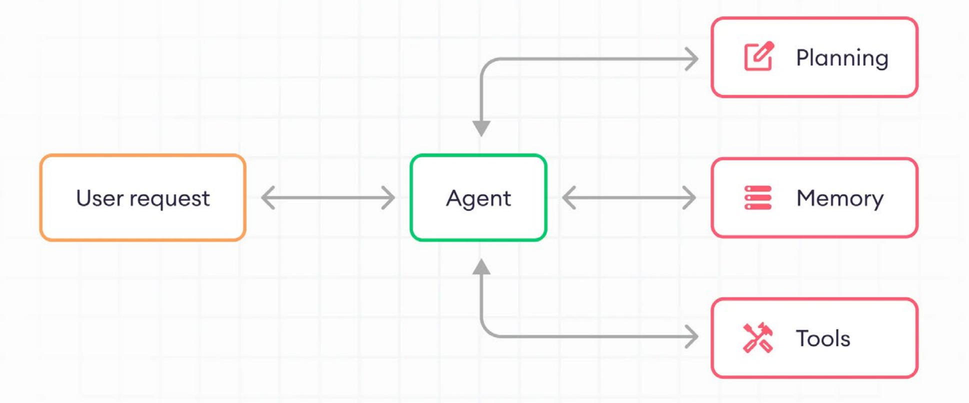
Aiswarya Raj Munappy, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, and Anas Dakkak. 2020.
From ad-hoc data analytics to dataops.
In Proceedings of the International Conference on Software and System Processes. 165–174.

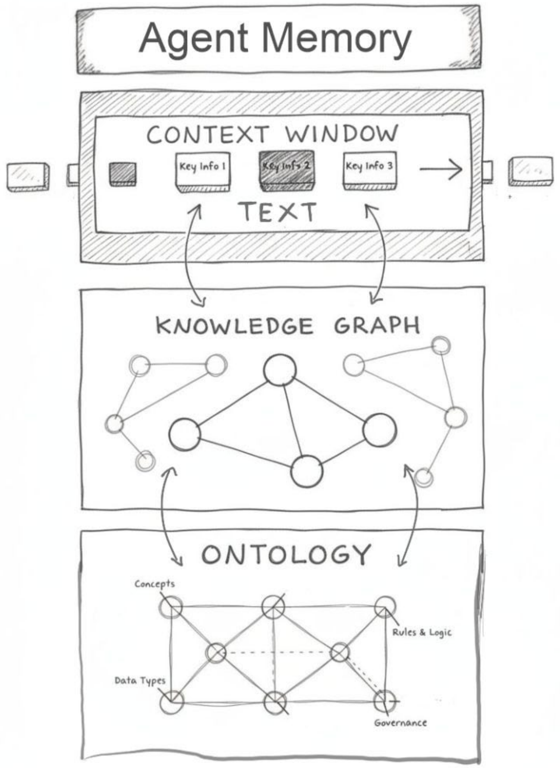
Data4LLMs: GenAI Tech Stack



[From Apps to Agents: How the AI-Native Tech Stack Is Transforming Software | BCGonTech]

Data4Agents





[LLM agents: The ultimate guide 2025 | SuperAnnotate]

[(18) Post | Feed | LinkedIn]

LLMs4Data: AI-Augmented Data Engineer

Voor bezoekers Exposant worden Programma Exposantenlijst Blog & Kennis Contact

Data Governance AI Agents 2 minuten lezen

Augmented analytics verandert de rol van de businessanalist

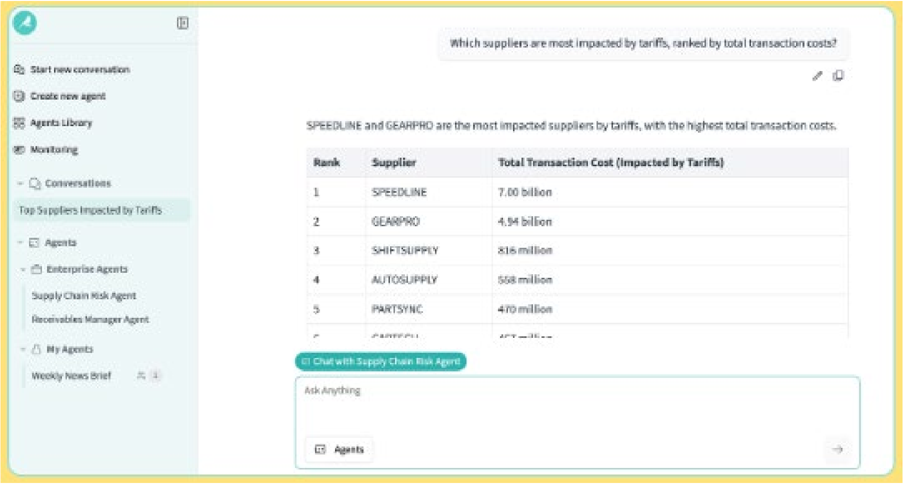
Mijn werk als journalist en tekstschrijver is door de komst van AI ingrijpend veranderd. Waar ik vroeger vaak lange rapporten en andere documenten moest scannen in de hoop dat ene relevante grafiekje of die opvallende conclusie te vinden, gebruik ik nu bijvoorbeeld Perplexity als zoekmachine. Op mijn vraag krijg ik binnen enkele seconden een puntsgewijs antwoord, compleet met bronnen waarin ik de oorspronkelijke gegevens kan terugvinden. Daarnaast helpt ChatGPT me snel overzicht te krijgen van complexe documenten. Ik denk dat iedereen die met data werkt, zo zijn eigen voorbeelden heeft van de impact van AI.

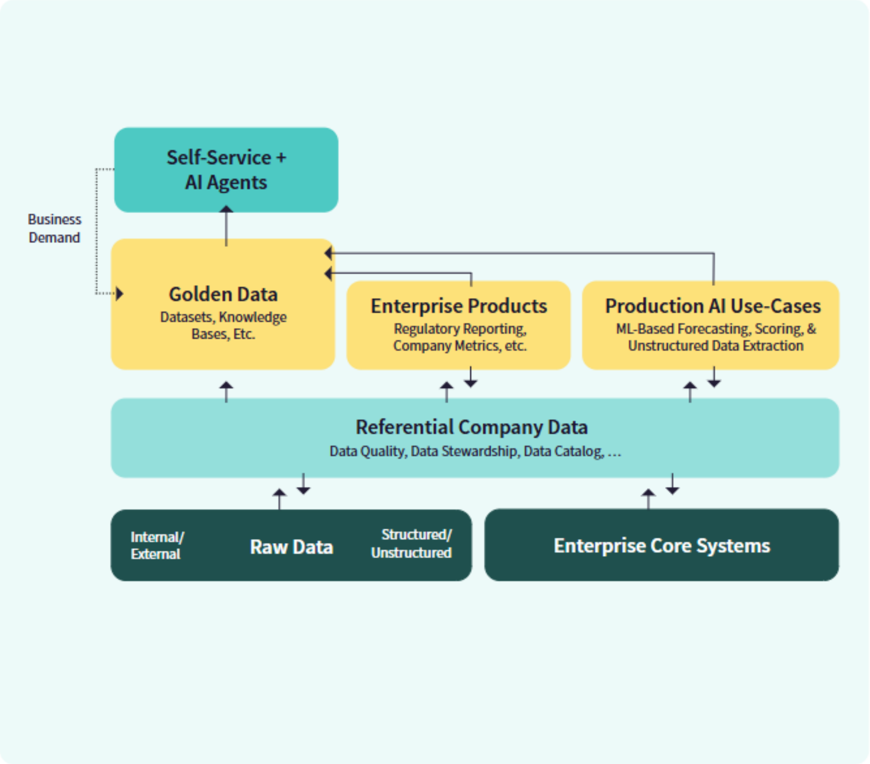
- Pro-active analysis with AI-agents
- “Governed semantic layer”
- Prompt/Requirements engineering
- Check output
- Context / translate output
- Ethics & Law

Fontys

[Augmented analytics verandert de rol van de businessanalist]
[Rutger Rienks: Van data voor AI naar AI voor data | BI-Platform]

GenAI for Self-Service Analytics





[Self-Service Analytics In The Age Of AI Agents]

State of Data and AI Engineering 2025

Explore the table »

Presented by lakeFS

INGEST

Ingest Tech

- Google Cloud Pub/Sub
- Apache Storm
- Amazon Kinesis
- StreamNative
- kafka
- beam
- memphis.dev
- nifi
- Flink
- Azure Data Factory
- Redpanda
- decodable
- WarpStream
- Informatica
- Spark
- Pulsar
- Airbyte
- Confluent
- meltano
- Kaskada
- upsolver
- Materialize
- striim

Ingest SaaS

- Portable
- Airbyte
- Hevo
- Polytomic
- Fivetran
- twilio segment
- Dataddo
- rudderstack
- Estuary
- Integrate.io
- Snowplow
- Trocco
- Keboola
- Rivery
- Matillion

Reverse ETL

- Census
- hightouch
- HeadsUp
- Polytomic
- Seekwell
- Grouparoo
- omnata
- rudderstack
- Octolis
- Trocco

DATA LAKE

Object Storage

- Microsoft Azure Blob Storage
- Google Cloud Storage
- amazon S3
- Oracle Cloud
- IBM Cloud
- Alibaba Cloud
- hadoop
- MinIO
- zadara
- CloudFiles
- DigitalOcean
- Cloudflare
- Storj
- wasabi
- Pure Storage
- StorageGRID
- VAST
- Backblaze
- IDrive
- ceph

DATA VERSION CONTROL

- lakeFS
- Quilt
- DVC
- OXEN.AI
- Underhive
- Nessie
- Git LFS
- Pachyderm

METADATA

Metastores

- databricks Unity Catalog
- Azure Purview
- Tabular
- Hive
- Cloud Dataproc
- AWS Glue
- Cloudera
- OpenHouse
- Gravitino

Open Table Formats

- Onehouse
- Apache Hudi
- Iceberg
- Delta Lake
- Apache ORC

COMPUTE ENGINES

Distributed Compute

- databricks
- Cloudera
- Cloud Dataproc
- Apache Spark
- Ray
- hadoop MapReduce
- Ascend.io
- Google DataFlow
- dask
- Azure HDInsight
- amazon EMR
- trino
- akka
- bodo.ai
- Azure Databricks
- Azure Synapse Analytics

Analytics Engines

- Data Fusion
- Hive
- Google Big Query
- amazon Redshift
- snowflake
- databricks Data Lakehouse
- Apache Impala
- StarRocks
- Firebolt
- Hazelcast
- ClickHouse
- amazon Athena
- LakeSoul
- dremio
- SingleStore
- Materialize
- QuestDB
- Redis
- KX
- pinot
- Polars
- trino
- DuckDB
- Rockset
- imply
- Pentaho
- startree
- Qubole
- druid
- Starburst
- Yellowbrick
- dbt
- Voltron Data
- Hydra
- Doris
- presto

PIPELINES

Orchestration

- Cloud Workflows
- Stitch
- Shipyard
- Astronomer
- Flyte
- union
- Couler
- seqera
- dagster
- Prefect
- Dagworks
- kestra
- Airflow
- sematic
- Mage
- MLRun
- Metaflow
- Luigi
- Kubeflow
- orkes
- Azkaban
- Azure Data Factory

Data Quality / Observability

- Kensu
- WhyLabs
- great expectations
- re_data
- metaplane
- Soda
- lightup
- Datafold
- Bigeye
- acceldata
- Anomalo
- HoloClean
- Union Pandera
- Monte Carlo
- Apache Griffin
- unravel
- Databand
- elementary
- awslabs/deequ
- foundational.
- timeseer.AI
- Datorios

PRACTITIONER APPS

MLOps End-to-End

- Abacus.AI
- mlflow
- Metaflow
- baseten
- AIBLE
- DataRobot
- Vertex AI
- cnvrg.io
- Hugging Face
- Seldon
- Wallaroo.AI
- Domino
- kedro
- Weights & Biases by CoreWeave
- graft
- Kubeflow
- Feast
- ClearML
- H2O.ai
- dotData
- SaturnCloud
- neptune.ai
- comet
- ZenML
- Modular
- Azure Machine Learning
- BentoML
- DagsHub
- data iku
- Qwak
- Valohai
- databricks Data Intelligence Platform
- Uber Michelangelo
- tecton
- Hopsworks
- DVC
- Amazon SageMaker

Data Centric AI/ML

- datachain
- scale
- Graviti
- Snorkel
- Falcon (Lectzo.com)
- dataloop
- gretel
- StackAI
- Pachyderm
- Voxel51
- dstack
- DVID
- DataLad
- activeloop
- tasq.ai
- hudi
- Labelbox
- V7
- Toloka
- YData

ML/AI Observability & Monitoring

- deepchecks.
- mona
- Superwise
- Arthur
- fiddler
- FirstEigen
- truera
- census
- WhyLabs
- Evidently AI
- Galileo
- HoneyHive
- AiMon
- hydrosphere.io
- Giskard
- HOW/SO
- aporia
- nannyML
- Robust Intelligence
- LatticeFlow
- Patronus AI
- distributional
- braintrust
- arize

Vector DBs

- Pinecone
- MongoDB.
- milvus
- Chroma
- marqo
- deeplake by activeloop
- FalkorDB
- qdrant
- elastic
- Vespa
- Google ScaNN
- pgvector
- KX
- FAISS
- ClickHouse
- Vald
- Weaviate
- zilliz
- Cassandra
- OpenSearch
- tigris

Notebooks

- jupyter
- Apache Zeppelin
- HEX
- count
- Deepnote
- cocalc
- kaggle
- CloudBank
- databricks Notebooks
- nteract
- Google colab
- Datalore

Analytics Workflows

- Querybook
- databricks Delta Live Tables
- Google Cloud Dataform

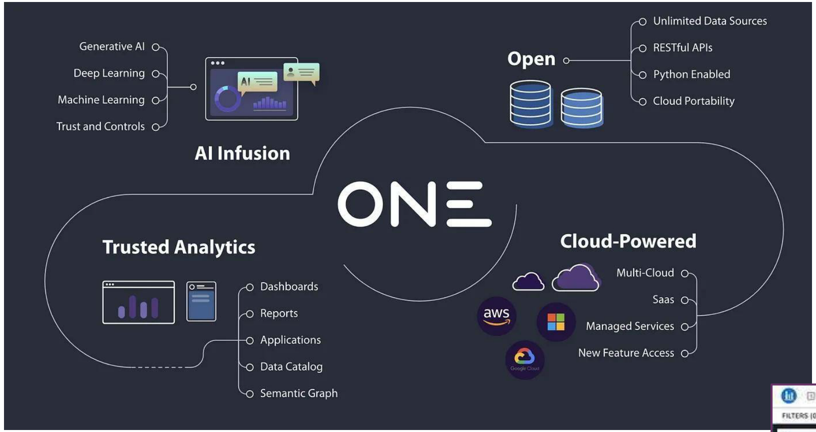
GOVERNANCE

Data Catalog / Governance

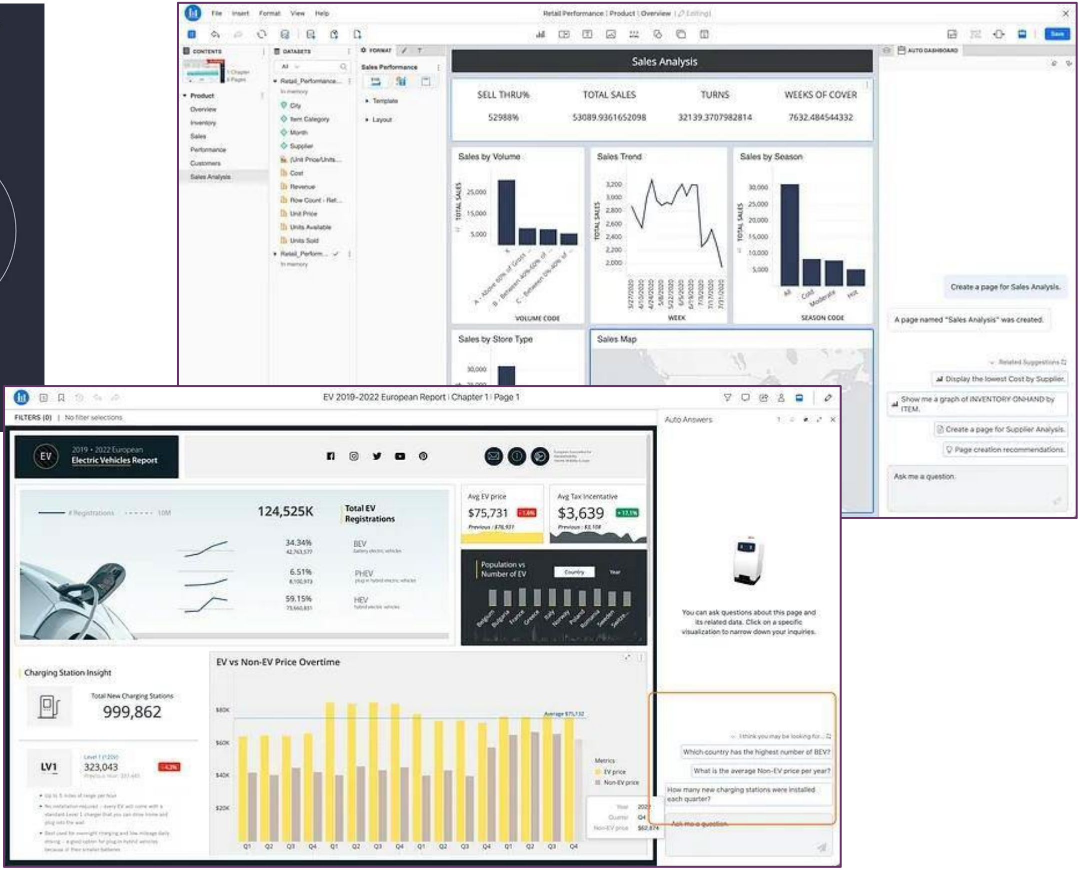
- Meta Cat
- ckan
- Open Metadata
- iguazio
- Nemo
- Amundsen
- DataHub
- atlan
- Alation
- boomi
- big id
- magda.
- data.world
- Collibra
- zeenea
- satori
- Platform
- raito
- Marquez
- Apache Atlas
- Google Cloud Dataproc Metastore
- databricks Unity Catalog
- Immuta
- IBM Knowledge Catalog
- Acryl Data
- Datastrato
- SELECT STAR

The State of Data and AI Engineering 2025

GenAI for Dashboards

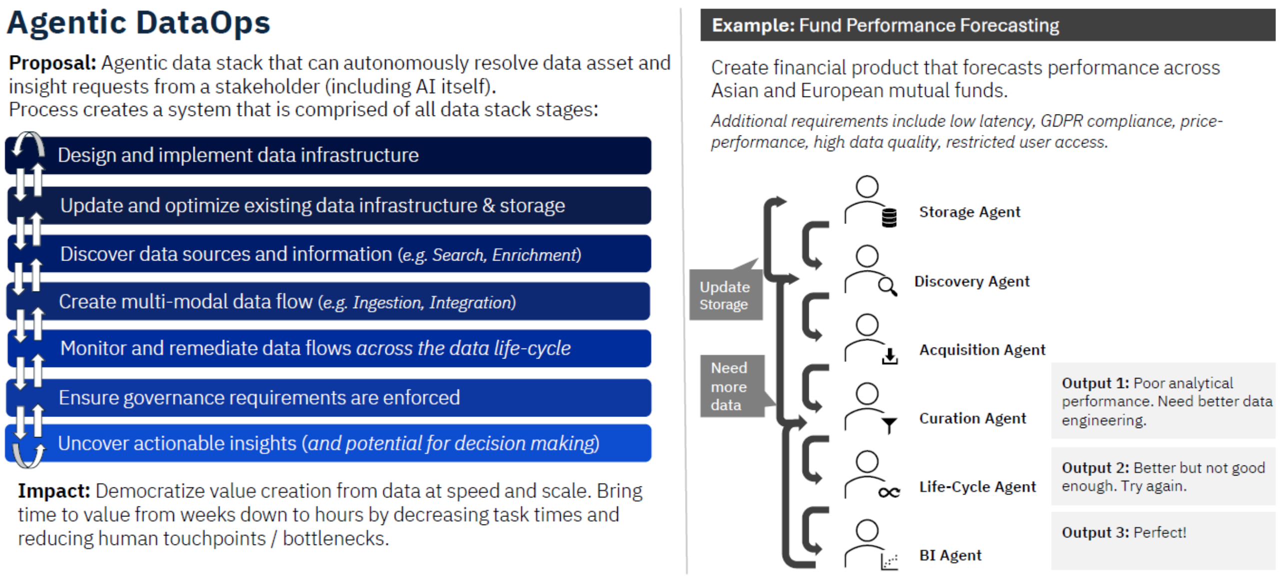


[AI en Design Thinking in de Dashboardmeesterschap van MicroStrategy]()



Fontys

Agentic DataOps

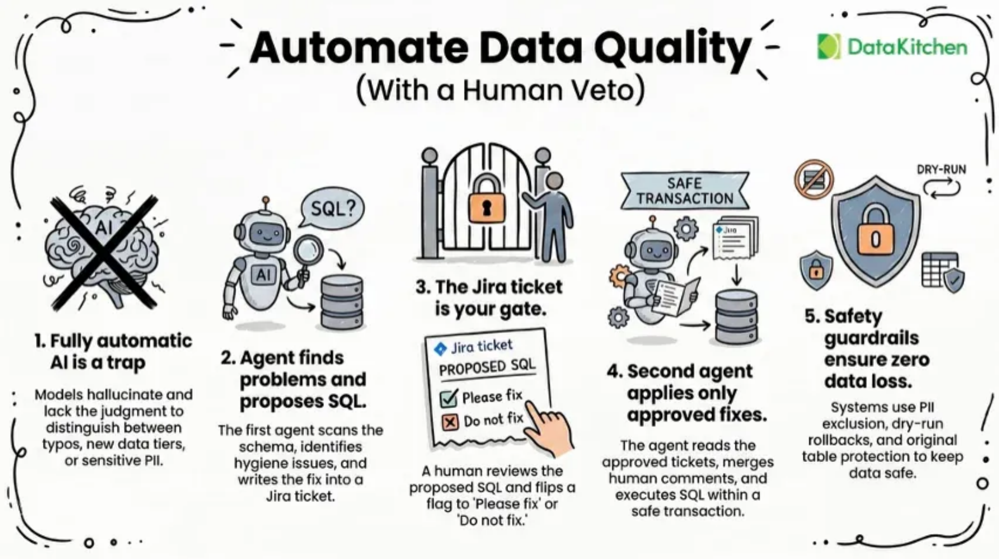


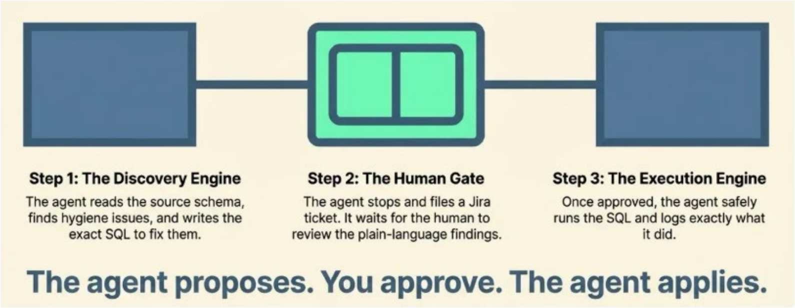
IBM Research

Fontys

[Keynote Lisa Amini-What's Next in AI for Data and Data Management--Pydata Global 2025](#)
[[2512.07926] Can AI autonomously build, operate, and use the entire data stack?](#)

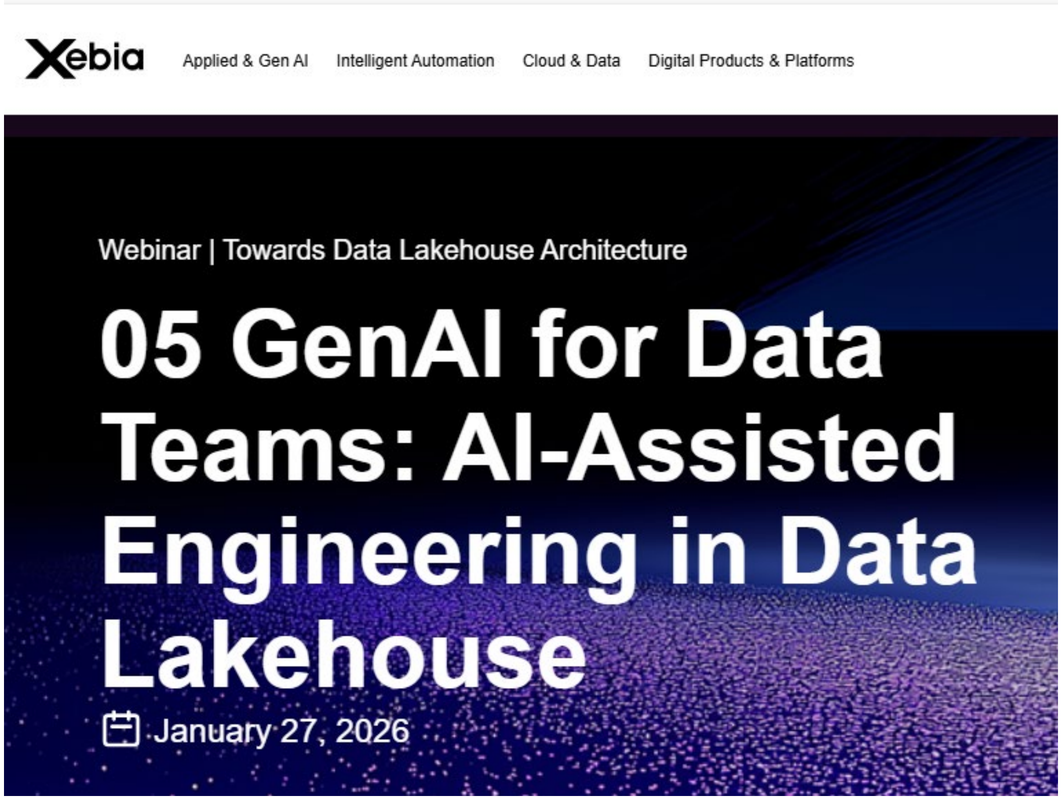
Human-Gated Agents





[Agents to Fix Data Quality Automatically With Control: A Design Pattern | DataKitchen]()

Agentic AI for Data Lakehouse



"how **Agentic AI** shifts data teams from hand-built pipelines to **outcome-driven data products** guided by clear intent and intelligent assistants.

You'll learn how **MCP integrations** enable AI assistants to tap your data catalogs, databases and transformation tools like dbt - dramatically accelerating development while preserving governance and trust.

We'll show how **spec-driven development** turns business needs and technical requirements into unambiguous, testable data products - boosting delivery speed and confidence.

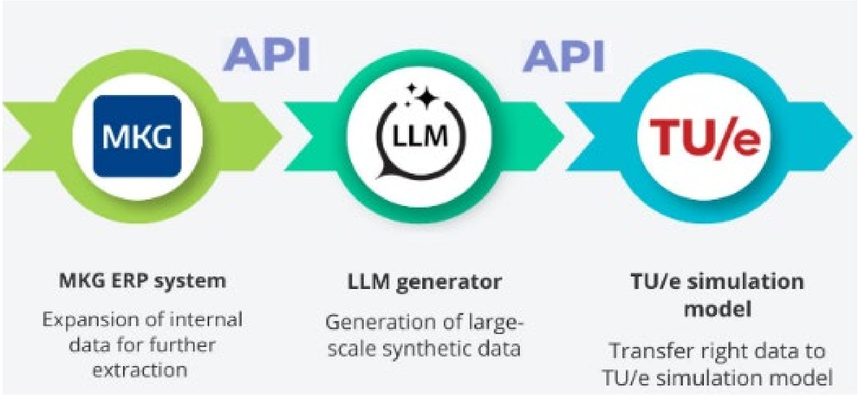
Finally, we'll explore **conversational data interfaces** that allow both technical and business users to interact with data products using natural language."

https://events.xebia.com/5-towards-data-lakehouse-architecture-webinar-ai-assisted-engineering

Other Examples GenAI for Data

- AI Agents With Dataiku | Dataiku
- Introducing Genie Code | Databricks Blog
- Agent Bricks | Databricks
- Actian introduceert Data Observability Agents | BI-Platform
- IBM Db2 Genius Hub voegt autonome mogelijkheden toe aan Db2 | BI-Platform
- Qlik maakt agentic analytics beschikbaar in Qlik Cloud | BI-Platform
- LLMs4OM: Matching Ontologies with Large Language Models
- Oracle introduceert zelfstandige AI-agents in Database 26ai
- MongoDB opent cloud-database Atlas voor AI-agents | BI-Platform
- While You Slept, an Agent Fixed 14 Data Quality Failures | DataKitchen

Applied Research: Supply Chain Resilience



NXT GEN hightech

Goal
Development of a data pipeline between MKG (ERP) and TU/e simulation model

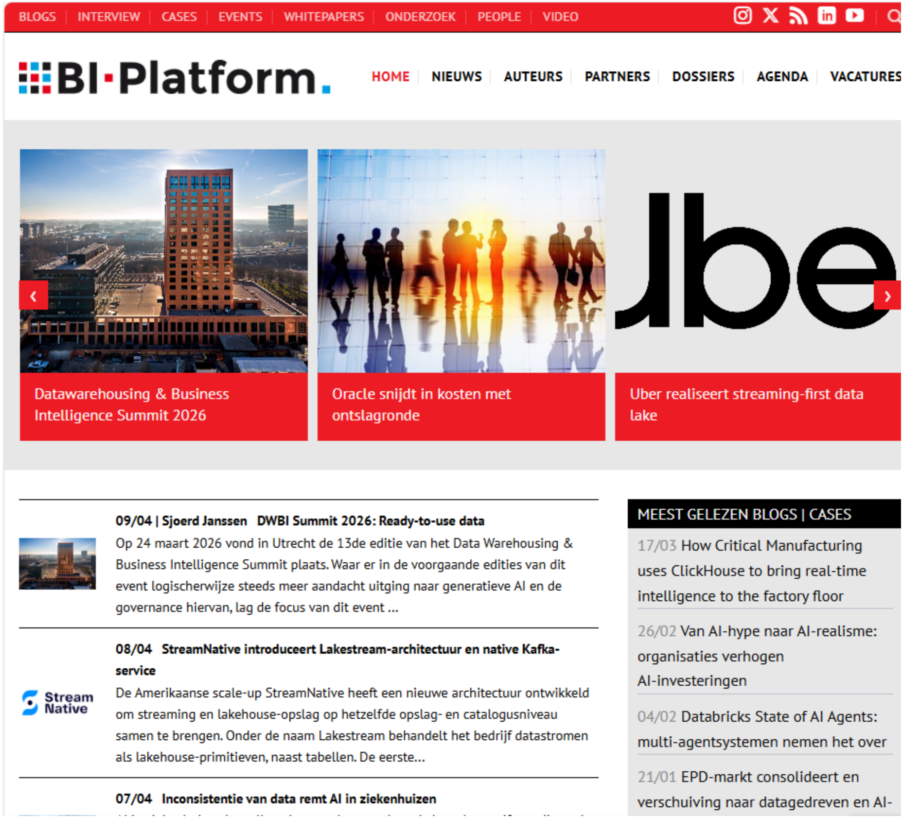
Synthetic data
Deployment of LLM models for data enrichment and synthesis prior to simulation.

Why?
Real ERP data is confidential; TU/e simulation requires much larger datasets than a single real order can provide.

Fontys

Sources

- [BI-Platform | Business Intelligence & Analytics]



Fontys

FOR SOCIETY

CASE STUDY: F1 RACING SIMULATOR

Fontys

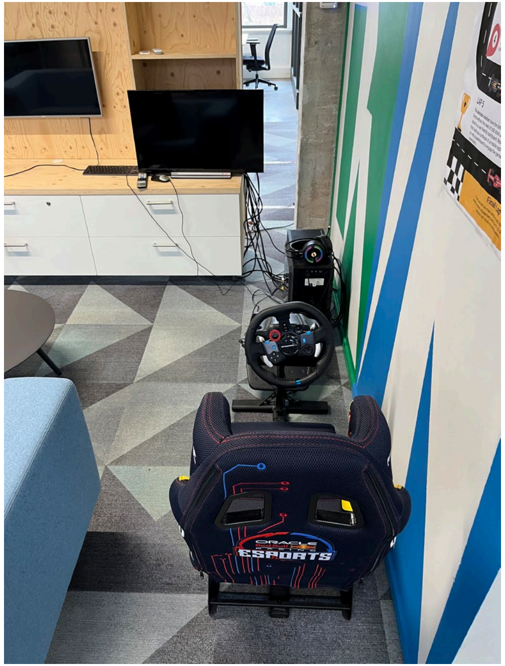
F1 setup

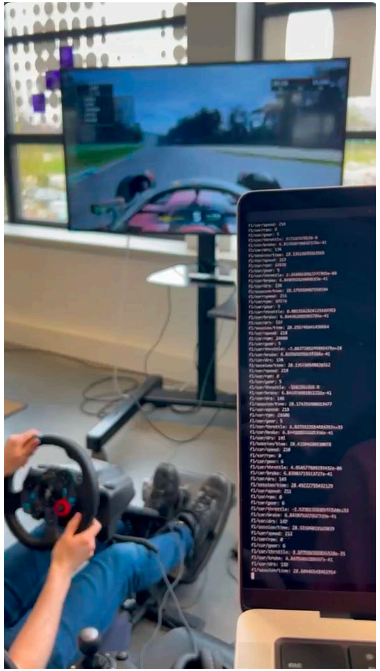
Students have access to an F1 setup where ”own” data can be generated

Claude

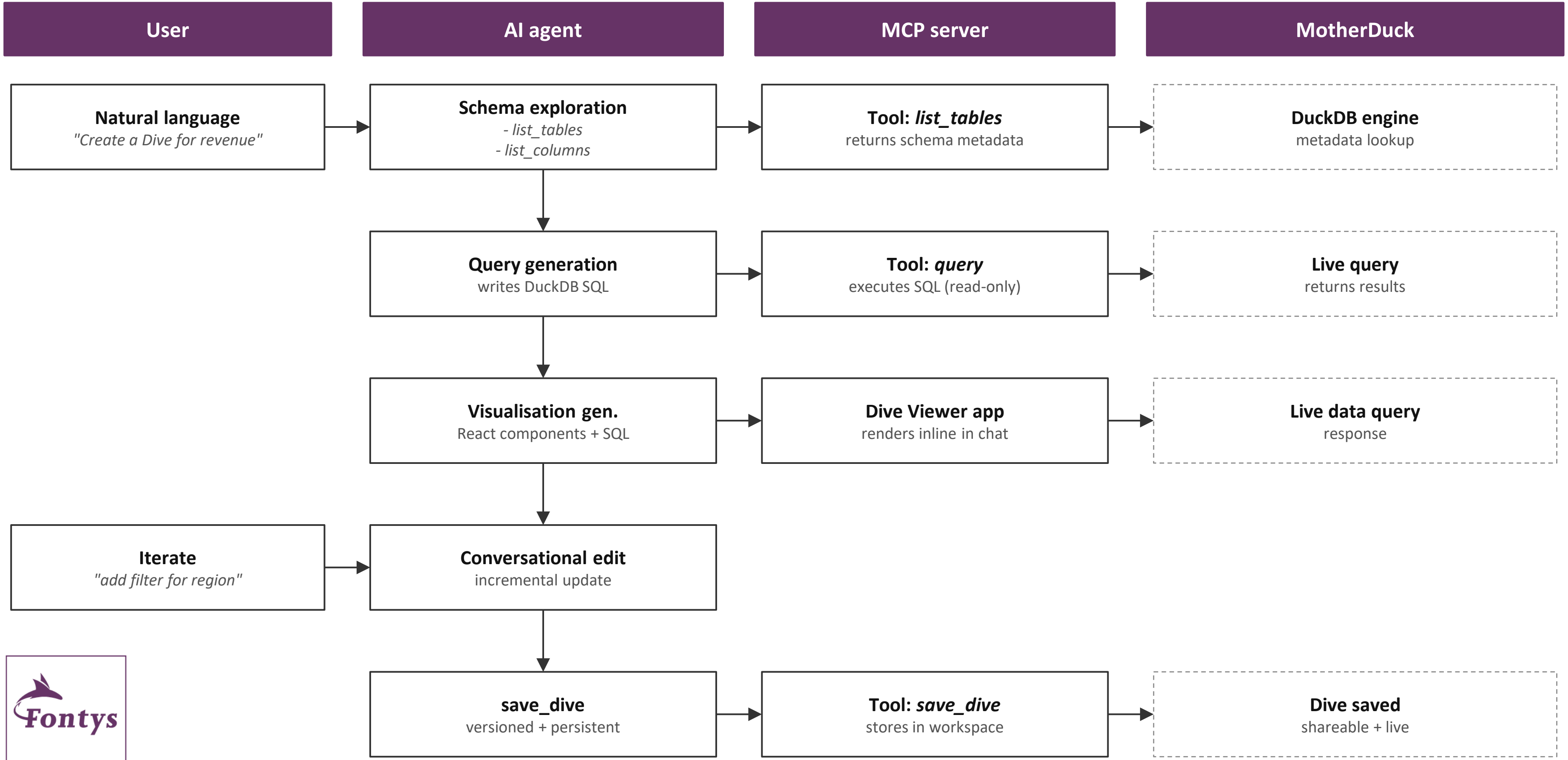
MotherDuck

Model Context Protocol





MotherDuck MCP Request Flow



Querying in Natural Language

“What data is available in the F1 database?”

MotherDuck Query

Request


```
{
  "database": "f1",
  "sql": "\nSELECT 'spanjeangelo' AS table_name, COUNT(*) AS rows, COUNT(DISTINCT carId) AS cars,
COUNT(DISTINCT trackId) AS tracks, COUNT(DISTINCT lapNum) AS laps, MIN(trackLength) AS
track_length, string_agg(DISTINCT trackId, ', ') AS track_ids, string_agg(DISTINCT carId, ', ') AS
car_ids\nFROM main.spanjeangelo\nUNION ALL\nSELECT 'spanjejordi', COUNT(*), COUNT(DISTINCT carId),
COUNT(DISTINCT trackId), COUNT(DISTINCT lapNum), MIN(trackLength), string_agg(DISTINCT trackId, ',
'), string_agg(DISTINCT carId, ', ')\nFROM main.spanjejordi;\n"
}
```


The `f1` database has two tables — `spanjeangelo` and `spanjejordi` (uploaded 3/3/2026 and 2/27/2026) — both with identical schemas: detailed per-frame telemetry, presumably F1 24/25 game UDP telemetry data.

Each table has:

- 27,900 rows, 6 laps, one car (Red Bull Racing), one track: Catalunya (track length 4650 m — the Circuit de Barcelona-Catalunya, matching "Spanje"/Spain in the filenames)
- 193 columns covering: lap info (lap number, lap time, distance, validity flags), car position/orientation (world position/forward/right vectors, yaw/pitch/roll), physics (velocity, g-force, angular velocity/acceleration), driver inputs (throttle, brake, clutch, steering, gear, DRS), engine data (RPM, fuel, engine temp), per-wheel data (suspension, tyre temps/pressures/wear/damage, brake temps/damage, wheel speed/slip), car setup (wings, camber, toe, springs, ARBs, ride height, brake bias/pressure, differential), assists (traction control, ABS, ERS, racing line), damage model (wings, floor, sidepods, gearbox, engine components), and session context (weather, track/air temp, safety car status, pit status, race position, penalties).

So essentially it's raw telemetry from two separate sessions (by "angelo" and "jordi", likely two drivers) run at Catalunya in Red Bull Racing — good for comparing driving lines, braking points, throttle application, tyre/brake temps, or setup differences lap by lap. Want me to dig into a specific comparison (e.g., lap times, throttle/brake traces, or setup diffs between the two drivers)?

Creating Dashboards

Create a simple dashboard containing a bar chart which shows the lap times of both drivers vertically so I can easily see which driver was the fastest.

Working on it 1 m 10 s

MotherDuck Query

Request


```
{
  "database": "f1",
  "sql": "\nSELECT 'angelo' AS driver, lapNum, MAX(lap_time) AS lap_time, MAX(lap_time_invalid) AS invalid\nFROM main.spanjeangelo\nGROUP BY lapNum\nUNION ALL\nSELECT 'jordi', lapNum, MAX(lap_time), MAX(lap_time_invalid)\nFROM main.spanjejordi\nGROUP BY lapNum\nORDER BY driver, lapNum;\n"
}
```


Response

MotherDuck Query

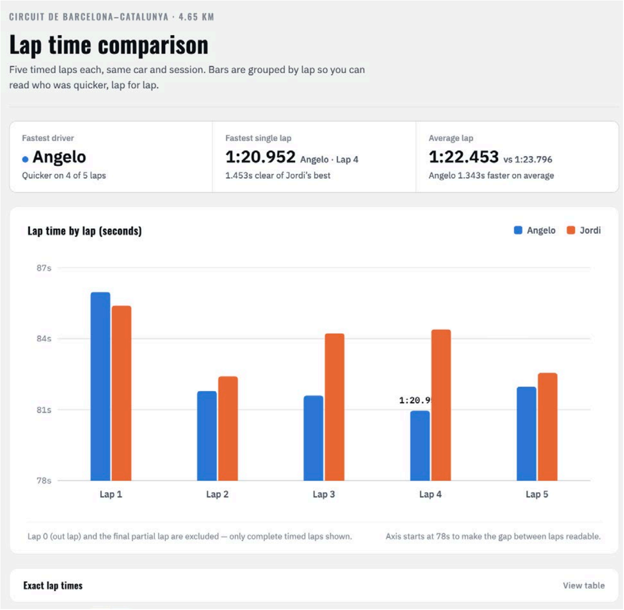
Request


```
{
  "database": "f1",
  "sql": "\nSELECT 'angelo' AS driver, lapNum, MAX(lap_distance) AS max_dist, COUNT(*) AS n\nFROM main.spanjeangelo GROUP BY lapNum\nUNION ALL\nSELECT 'jordi', lapNum, MAX(lap_distance), COUNT(*)\nFROM main.spanjejordi GROUP BY lapNum\nORDER BY driver, lapNum;\n"
}
```


Response

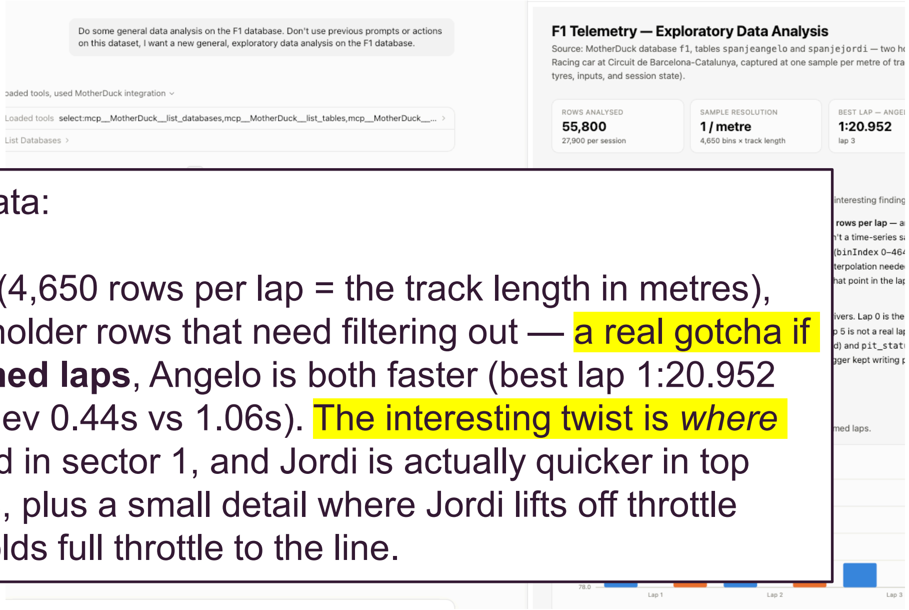

```
{"success":true."columns":["driver"."lapNum"."max dist"."n"]."columnTypes":
```


Used a skill dataviz



Open question

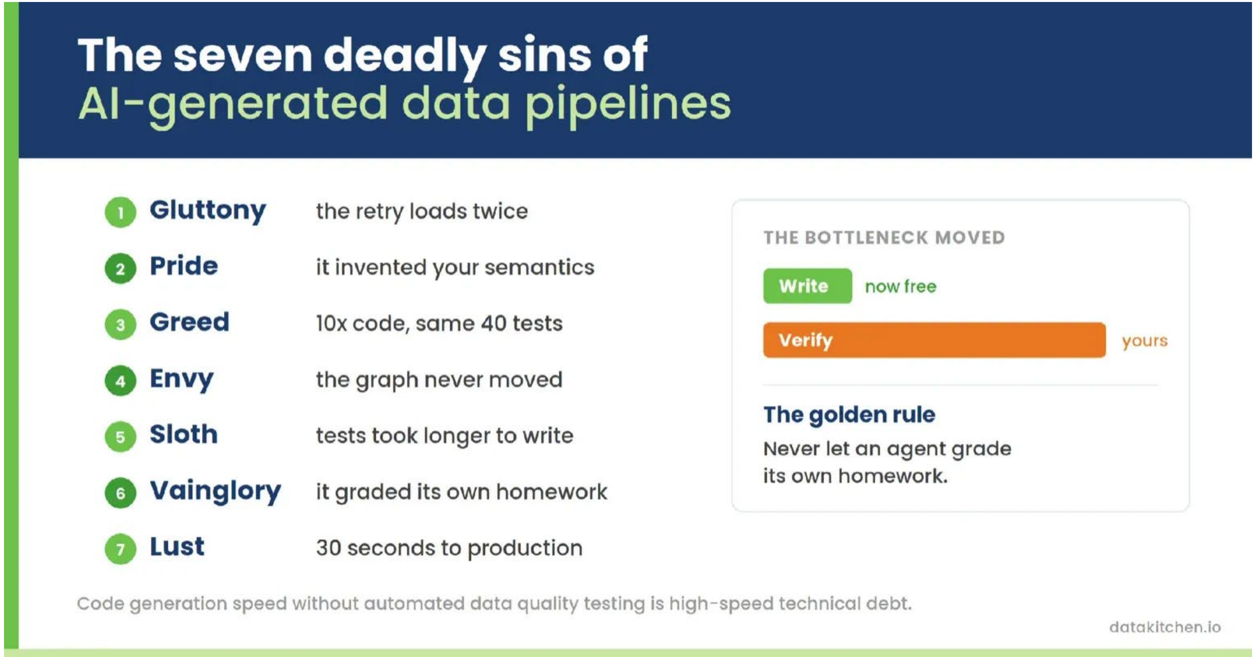
“Do some general data analysis on the F1 database”



The main findings, all worked out fresh from the raw data:

The data is distance-binned rather than time-sampled (4,650 rows per lap = the track length in metres), and lap 5 in both tables turned out to be garage placeholder rows that need filtering out — a real gotcha if queried naively. Once cleaned to the **four genuine timed laps**, Angelo is both faster (best lap 1:20.952 vs 1:22.554) and much more consistent (lap-time std dev 0.44s vs 1.06s). The interesting twist is *where* the gap comes from: the two drivers are essentially tied in sector 1, and Jordi is actually quicker in top speed — the whole deficit shows up in sectors 2 and 3, plus a small detail where Jordi lifts off throttle slightly in the last 150m of his best lap while Angelo holds full throttle to the line.

Quality Matters



[The Seven Deadly Sins of AI-Generated Data Pipelines | DataKitchen]

What Could Go Wrong?

Students get used to AI tools quickly, but have a hard time understanding architecture / know where to look for errors

Bugs I caught and corrected in Claude's drafts

These are the specific moments where Claude's first draft was wrong and I caught it before it shipped. Each one is worth naming because it shows the editorial role I played as the prompter, not just the prompts themselves.

The Success Rate substring bug

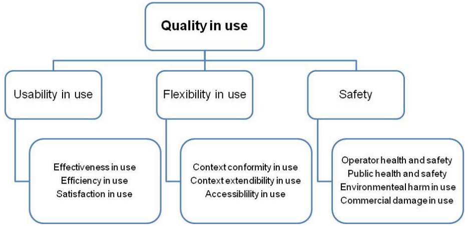
This is the most important one because it was caught during validation with [redacted] Claude's first version of the Success Rate SQL used a substring match (ILIKE '%success%') which accidentally counted Unsuccessful jobs as successful too, because the word 'success' is a substring of 'Unsuccessful'. At the 17/04 demo [redacted] pointed out that every system showed 100 percent success which is obviously wrong. I went back, replaced the substring match with an exact match (LOWER(TRIM(j.successful)) = 'successful') and the rate started reading realistic values. This story is also documented in the master research document as one of the moments where the Showroom strategy directly fed back into the build.

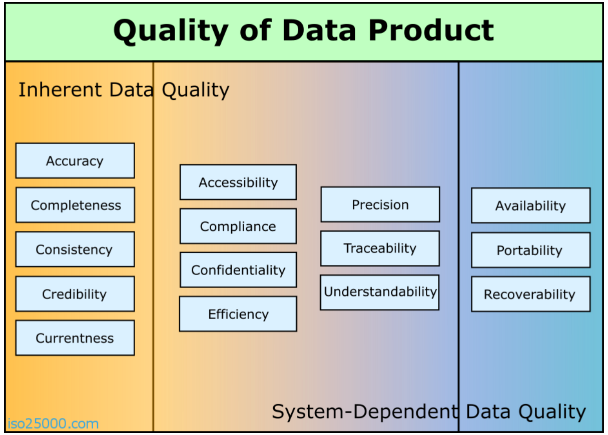
› FOR SOCIETY

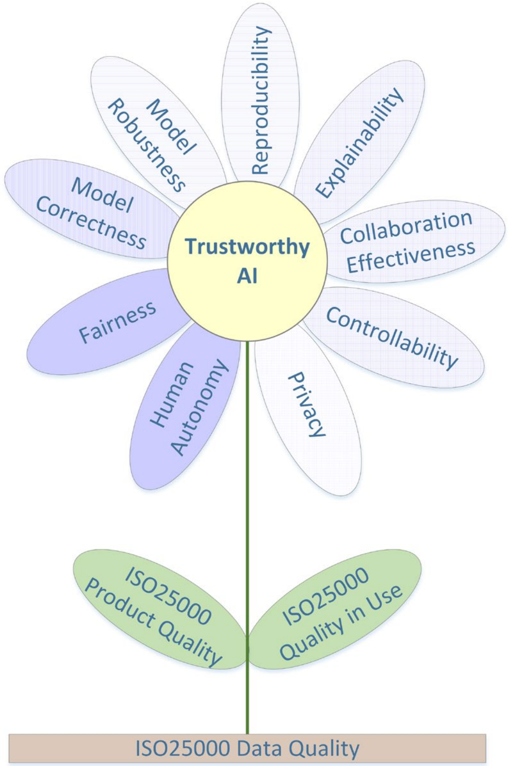
SE4DATA – PRODUCT QUALITY

Software Quality: Data, Models and Code

SOFTWARE PRODUCT QUALITY								
FUNCTIONAL SUITABILITY	PERFORMANCE EFFICIENCY	COMPATIBILITY	INTERACTION CAPABILITY	RELIABILITY	SECURITY	MAINTAINABILITY	FLEXIBILITY	SAFETY
FUNCTIONAL COMPLETENESS	TIME BEHAVIOUR	CO-EXISTENCE	APPROPRIATENESS RECOGNIZABILITY	FAULTLESSNESS	CONFIDENTIALITY	MODULARITY	ADAPTABILITY	OPERATIONAL CONSTRAINT
FUNCTIONAL CORRECTNESS	RESOURCE UTILIZATION	INTEROPERABILITY	LEARNABILITY	AVAILABILITY	INTEGRITY	REUSABILITY	SCALABILITY	RISK IDENTIFICATION
FUNCTIONAL APPROPRIATENESS	CAPACITY		OPERABILITY	FAULT TOLERANCE	NON-REPUDIATION	ANALYSABILITY	INSTALLABILITY	FAIL SAFE
			USER ERROR PROTECTION	RECOVERABILITY	ACCOUNTABILITY	MODIFIABILITY	REPLACEABILITY	HAZARD WARNING
			USER ENGAGEMENT		AUTHENTICITY	TESTABILITY		SAFE INTEGRATION
			INCLUSIVITY		RESISTANCE			
			USER ASSISTANCE					
			SELF-DESCRIPTIVENESS					
iso25000.com								







https://fontysblogt.nl/a-quality-model-for-trustworthy-ai-systems/

LLM-enabled Systems: Quality Model

Trust & Interaction Quality
- Transparency
- Collaboration Effectiveness
- Resistance

Operational Quality
- Reproducibility
- Replaceability
- Operational Constraint

Model Quality
- Model Correctness
- Functional Correctness
- Model Robustness

Quality Characteristic	Probabilistic generation	Natural Language as primary interface	Generative open-ended output	Model training opacity and scale	Continuous evolution	Human-AI co-creating paradigm	New interaction-based attack surface	Example LLM-specific Failure Modes
Model Correctness	x		x	x				- Fabricated facts presented as true - Plausible but incorrect statements - Invented references or citations
Functional Correctness	x	x				x		- Misinterpretation of user intent - Partially correct responses that miss the goal - Ignoring implicit constraints
Model Robustness	x	x						- Small prompt changes leading to significantly different outputs
Reproducibility	x							- Different answers to identical prompts - Variation in reasoning or conclusions across runs
Operational Constraint	x	x	x					- Generation of toxic, biased, or unsafe content - Harmful advice or inappropriate language
Transparency		x		x				- Inability to trace why outputs are produced - Hidden influence of training data - Opaque relation between prompt and output
Replaceabilty	x				x			- Behavior drift after model updates - Regression in performance or style - Loss of prior capabilities
Collaboration Effectiveness						x		- Over-reliance on model output - Goal-drift
Resistance		x	x				x	- Prompt injection attacks - Jailbreaks bypassing safety constraints - Unintended data leakage

Fontys

[Leon Schrijvers - Fontys Hogeschool]

Fontys > FOR SOCIETY

GENAI & DATA @ FONTYS ICT

Fontys ICT – New Professional Profiles

🏗 GenAI Architect

Design AI strategies, integrated systems, and ethical frameworks that align with business needs.

⚙ GenAI Engineer

Build robust prototypes, explore frameworks, and create high-performance AI solutions.

🎨 Experience Designer

Craft intuitive interfaces and smooth AI interactions through empathetic user journeys.

[Study Guide Fontys ICT - Applied Generative AI]

Fontys ICT – Study Programmes

Study programme Software Engineering

- 1st year: Start Semester; Software Design & Engineering
- 2nd year: Full Stack Development; Applied Generative AI
- 3rd year: Internship; Complex Software Systems
- 4th year: Enhanced AI Techniques; Graduation Semester

Study programme Business IT & Management

- 1st year: Start Semester; Business IT & Data Analysis
- 2nd year: Business IT & Data Analysis; Applied Generative AI
- 3rd year: Internship; Business IT & Data Analysis
- 4th year: Industry 5.0; Graduation Semester

Fontys

[Study Guide Fontys ICT - Programmes]

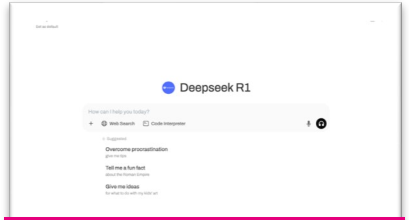
GenAI Usage Policy

We view Generative AI as both a **productivity tool and a learning tool**. This means we expect you to:

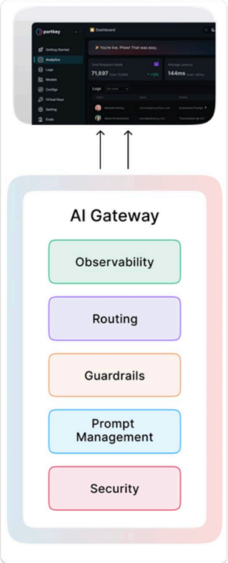
- **Achieve more** in the same timeframe: conduct deeper research, create higher-quality professional products, and engage in more profound learning
- Understand that GenAI will not reduce your study time but will **elevate the standards** we expect you to meet
- Recognize that your **future employers** will expect enhanced productivity with GenAI availability, and our educational standards reflect this reality

FHICT Beleidswiki - Het Fontys ICT AI Manifesto

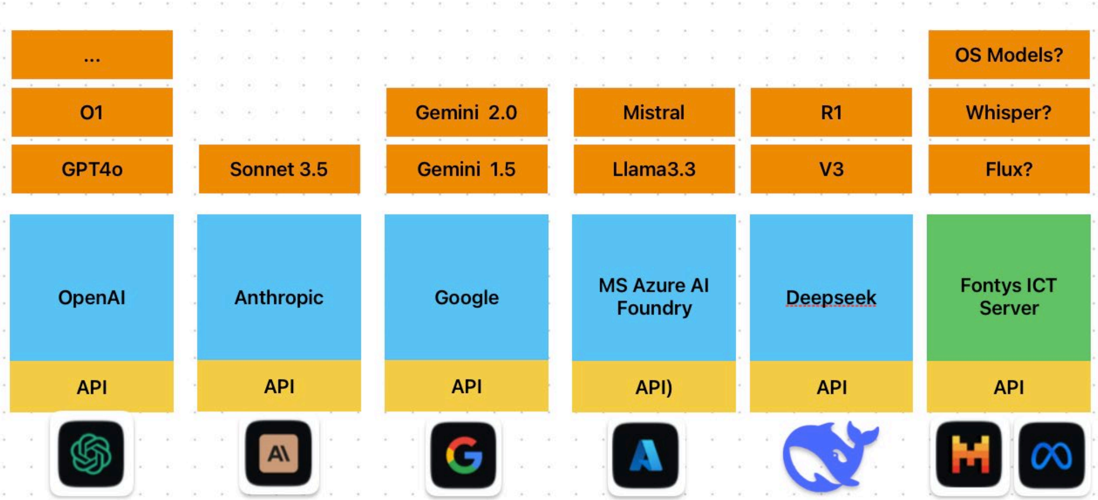
Fontys ICT – Infrastructure



1. Front end



2. Gateway



3. Modelhosting

https://chat.fontysict.nl/

CONCLUSION

Lessons We Are Learning

- AI > Agentic AI > Agents > GenAI > LLMs
- Data engineering ~ Software engineering
- AI-augmented Data Engineer
  - Human-in-the-lead Automation
  - Verification-driven Engineering
- Start experimenting, prepare for flexibility
- It is not just about technology

Let's learn together! p.heck@fontys.nl

Fontys

FontysBlogt

Questions?

Petra Heck, auteur op Fontys, p.heck@fontys.nl, https://www.linkedin.com/in/petraheck/