getindata
Part of Xebia

Introduction to Analytics Engineering

with dbt and Microsoft Fabric

Cracow PowerBI & Fabric UG
2024-12-11

Speaker

getindata
Part of Xebia

Tomasz Kostyrka, PL

- Data Platform Architect **@GetInData | Part of Xebia**
- 12 years in Data
- Azure/Databricks/Snowflake
- Data Engineering, Cloud Engineering, DevOps/DataOps, Data Platform Architecture
- Databricks Solution Architect Champion
- Community speaker

- https://www.linkedin.com/in/tomasz-kostyrka/
- https://sessionize.com/tomasz-kostyrka/
- https://pl.seequality.net/

Session

getindata
Part of Xebia

Plan:

- ETL vs. ELT
- PowerBI vs Fabric
- Analytics Engineering
- dbt [DEMO!]
- Going Prod!
- Costs

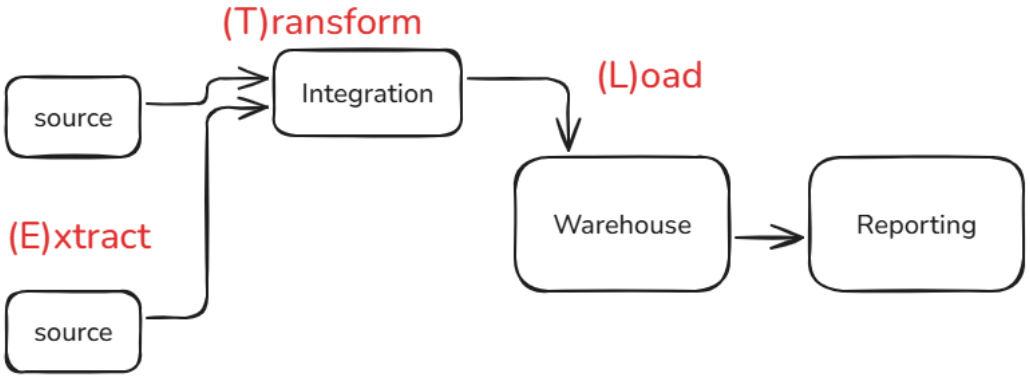
Disclaimers:

- **This is an intro session.** If there's interest, ping the organizer and we'll do a deep dive.
- This **is not** a data modeling session.
- The goal **is not** to convince that dbt is the cure for all data problems.

ETL vs. ELT

Extract > Transform > Load

getindata
Part of Xebia



- SQL Server Integration Services (DataStage, AbInitio)
- SQL Server (Oracle, Teradata, Exadata)
- SQL Server Reporting Services (Qlik, Tableau)

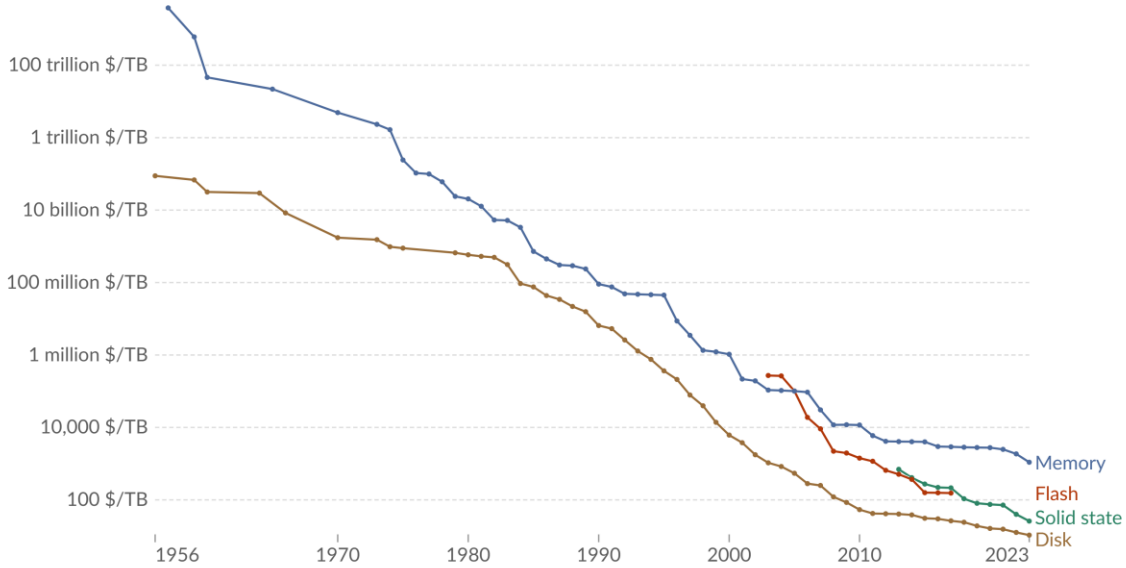
Price of computer memory and storage.

getindata
Part of Xebia

Historical price of computer memory and storage

Our World in Data

This data is expressed in US dollars per terabyte (TB), adjusted for inflation. "Memory" refers to random access memory (RAM), "disk" to magnetic storage, "flash" to special memory used for rapid data access and rewriting, and "solid state" to solid-state drives (SSDs).



Data source: John C. McCallum (2023); U.S. Bureau of Labor Statistics (2024) OurWorldinData.org/technological-change | CC BY

Note: For each year, the time series shows the cheapest historical price recorded until that year. This data is expressed in constant 2020 US$.

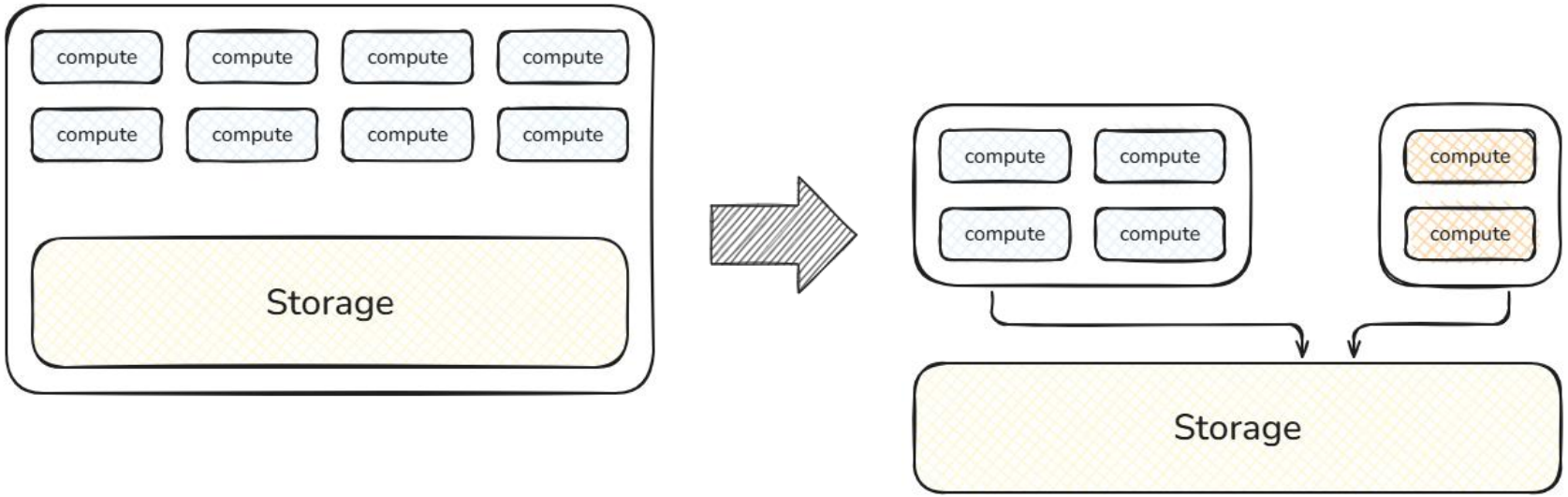
Cloud Computing!

getindata
Part of Xebia

2006 – AWS
2008 – GCP
2010 – Azure

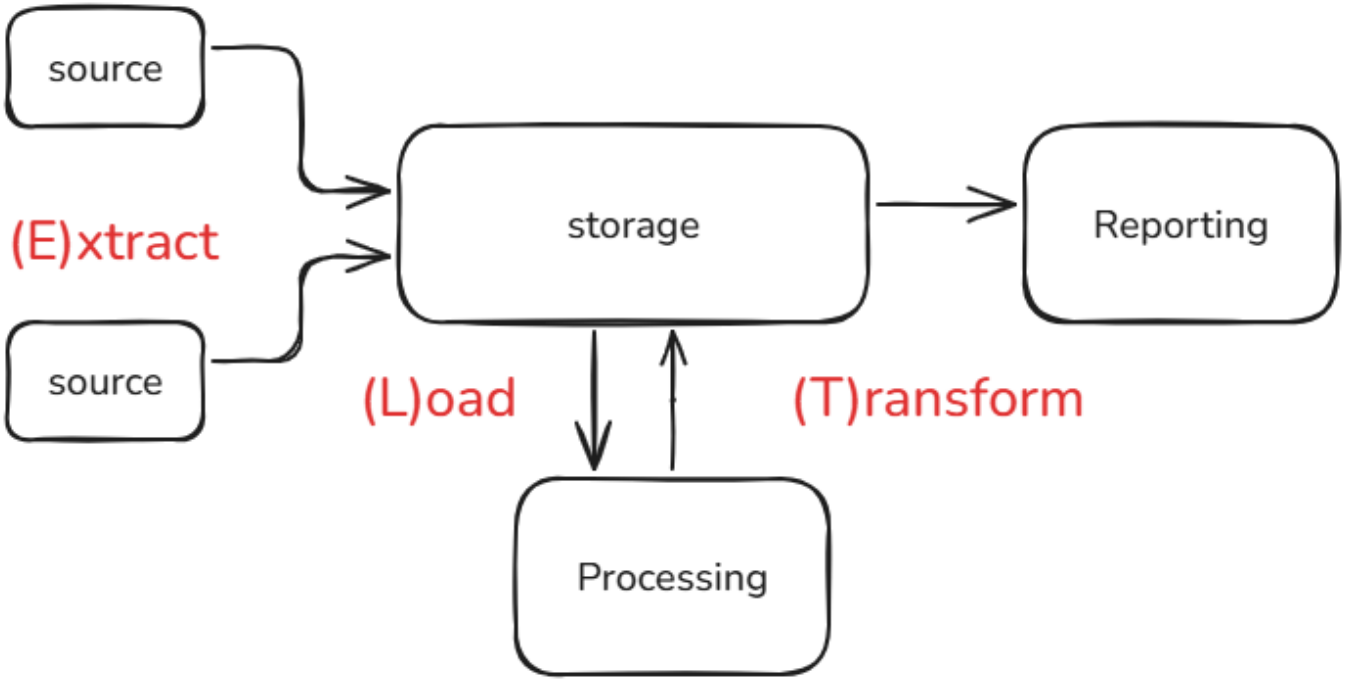
Decoupling storage from compute

getindata
Part of Xebia



Extract > Load > Transform

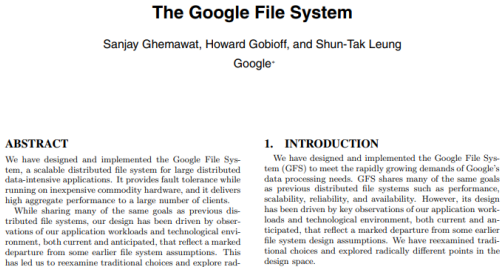
getindata
Part of Xebia



Big Data!

getindata
Part of Xebia

2003 – Google File System
2006 – Hadoop
2008 – Apache Pig
2010 – Apache Hive
2011 – Apache Spark
2012 – Snowflake
2013 – Databricks
2023 – Fabric



The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*

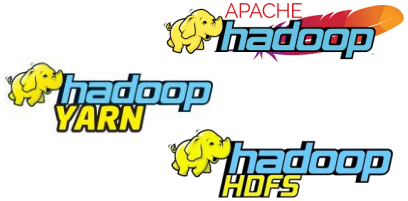
ABSTRACT

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore rad-

1. INTRODUCTION

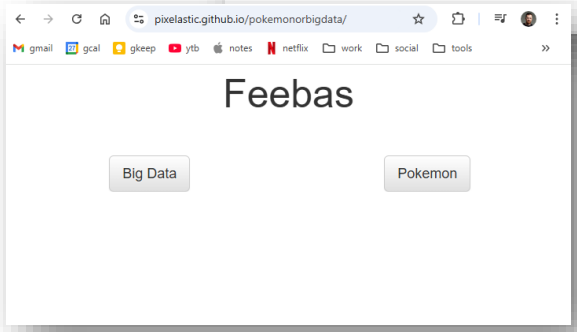
We have designed and implemented the Google File System (GFS) to meet the rapidly growing demands of Google's data processing needs. GFS shares many of the same goals as previous distributed file systems such as performance, scalability, reliability, and availability. However, its design has been driven by key observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system design assumptions. We have reexamined traditional choices and explored radically different points in the design space.

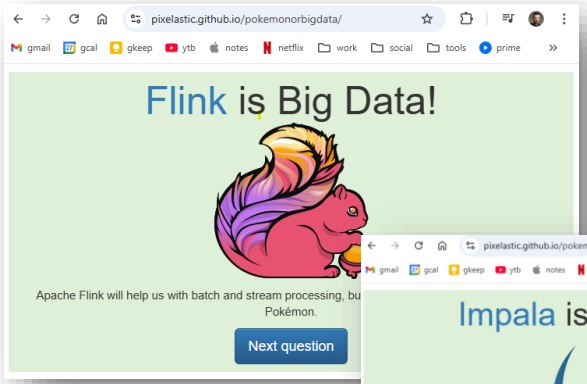


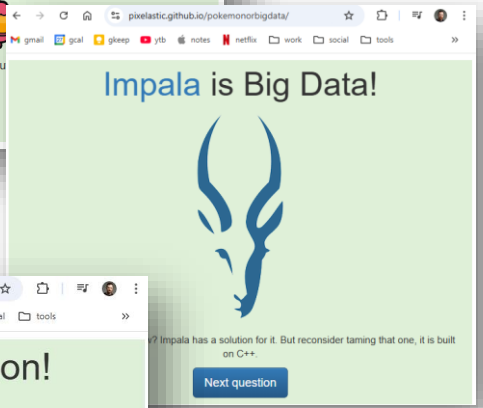
Apache Pig

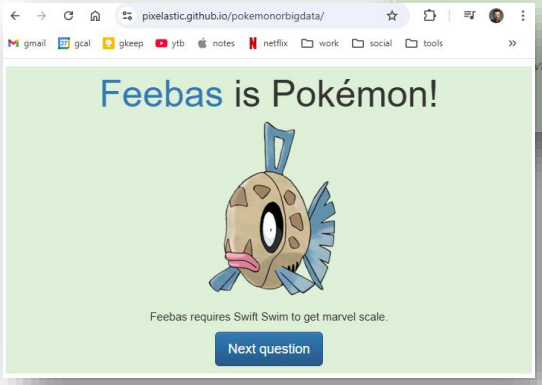
Pokemon or Big Data?

getindata
Part of Xebia

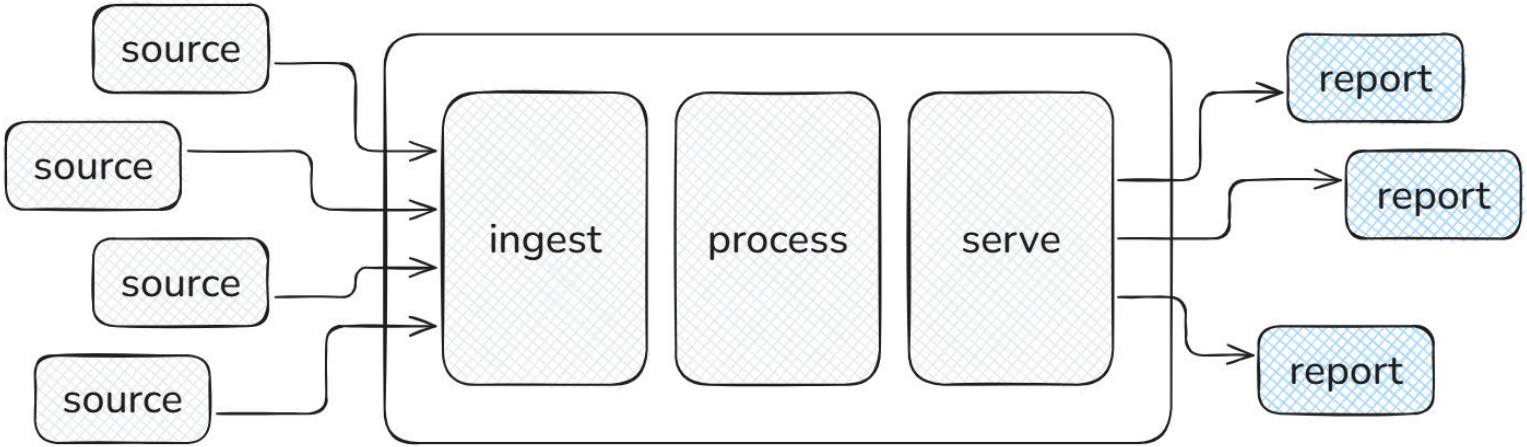






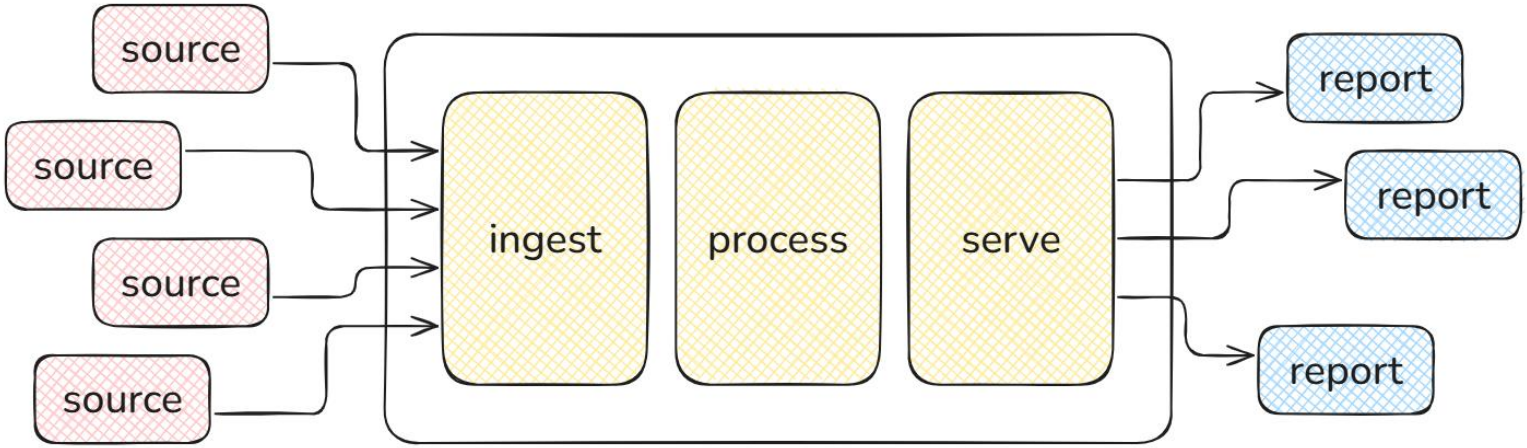


PowerBI vs Fabric


source
source
source
source
ingest
process
serve
report
report
report

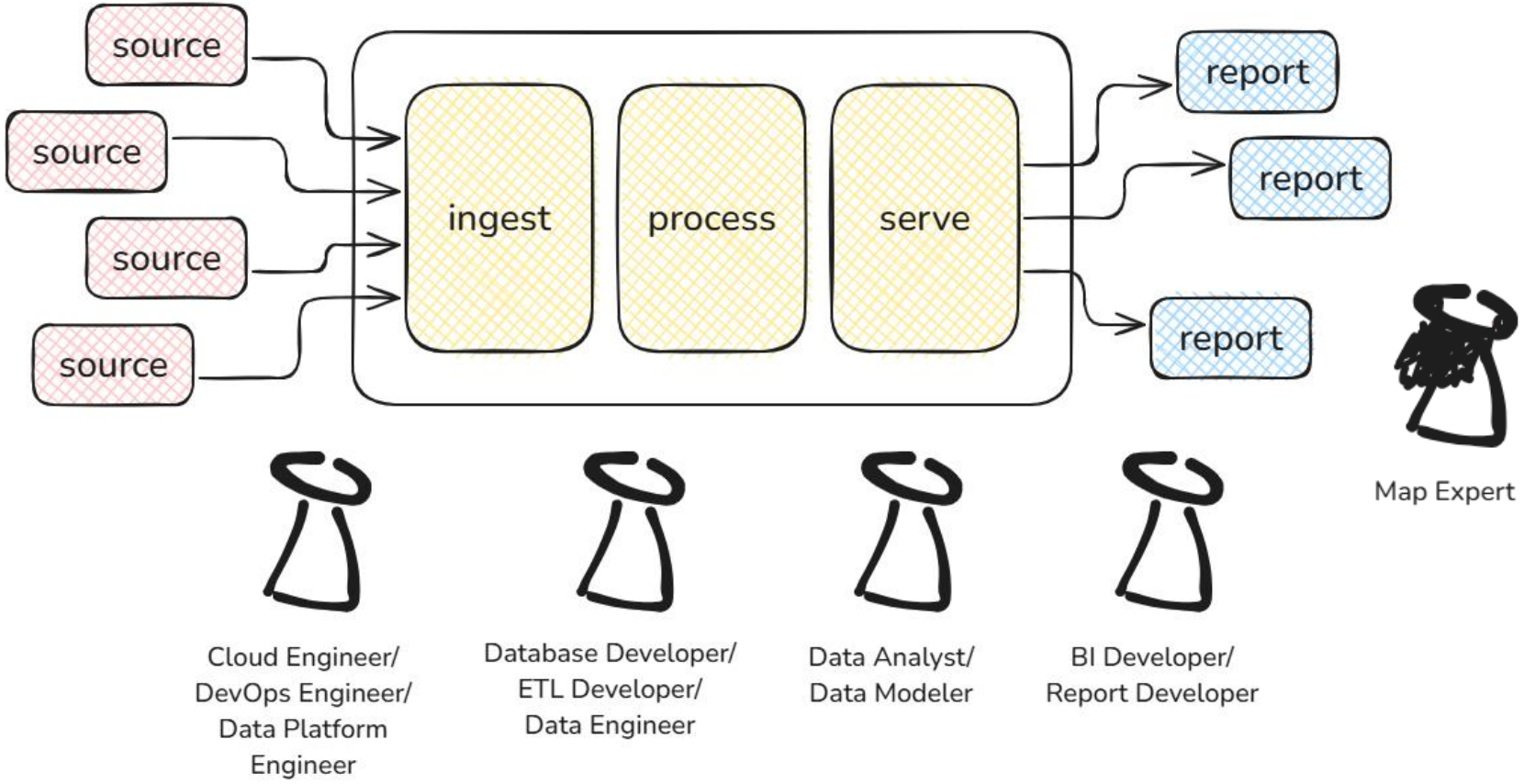
Fabric

getindata
Part of Xebia



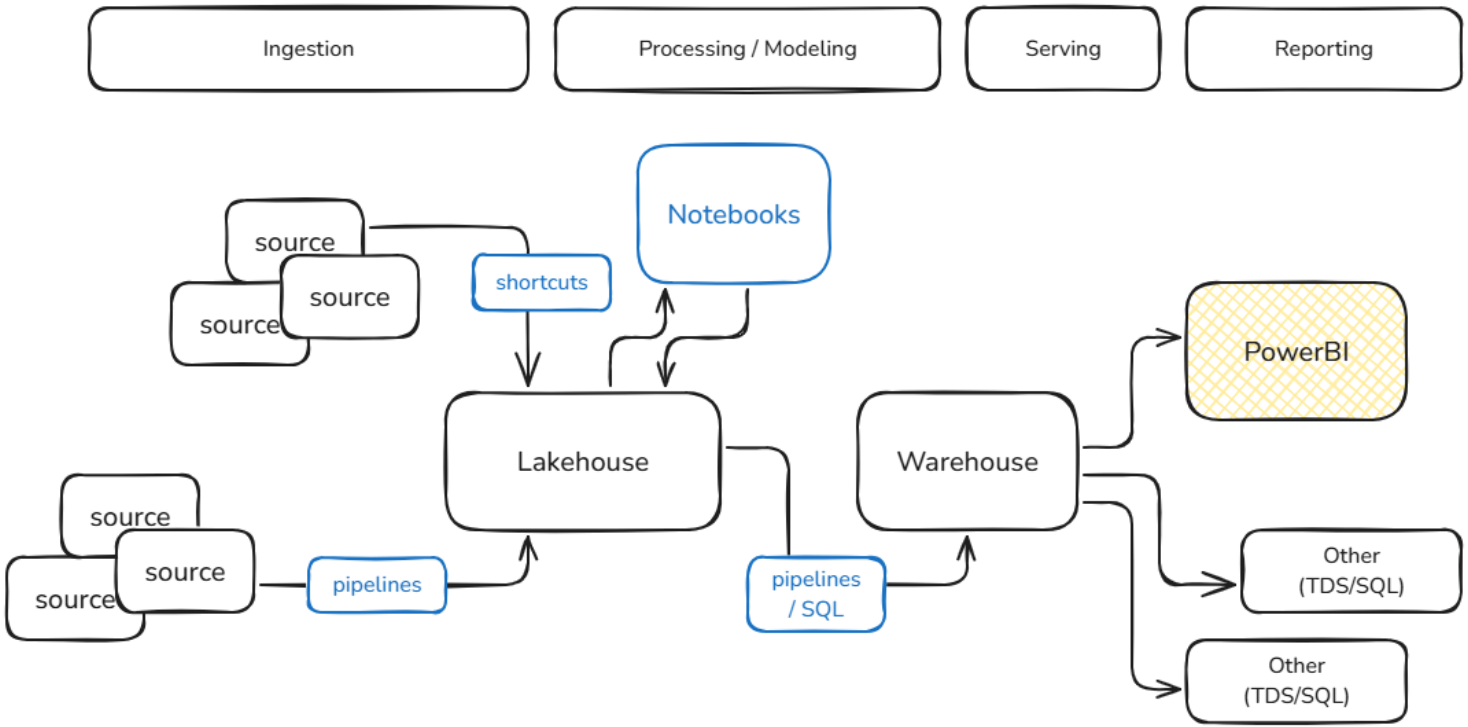
Roles

getindata
Part of Xebia



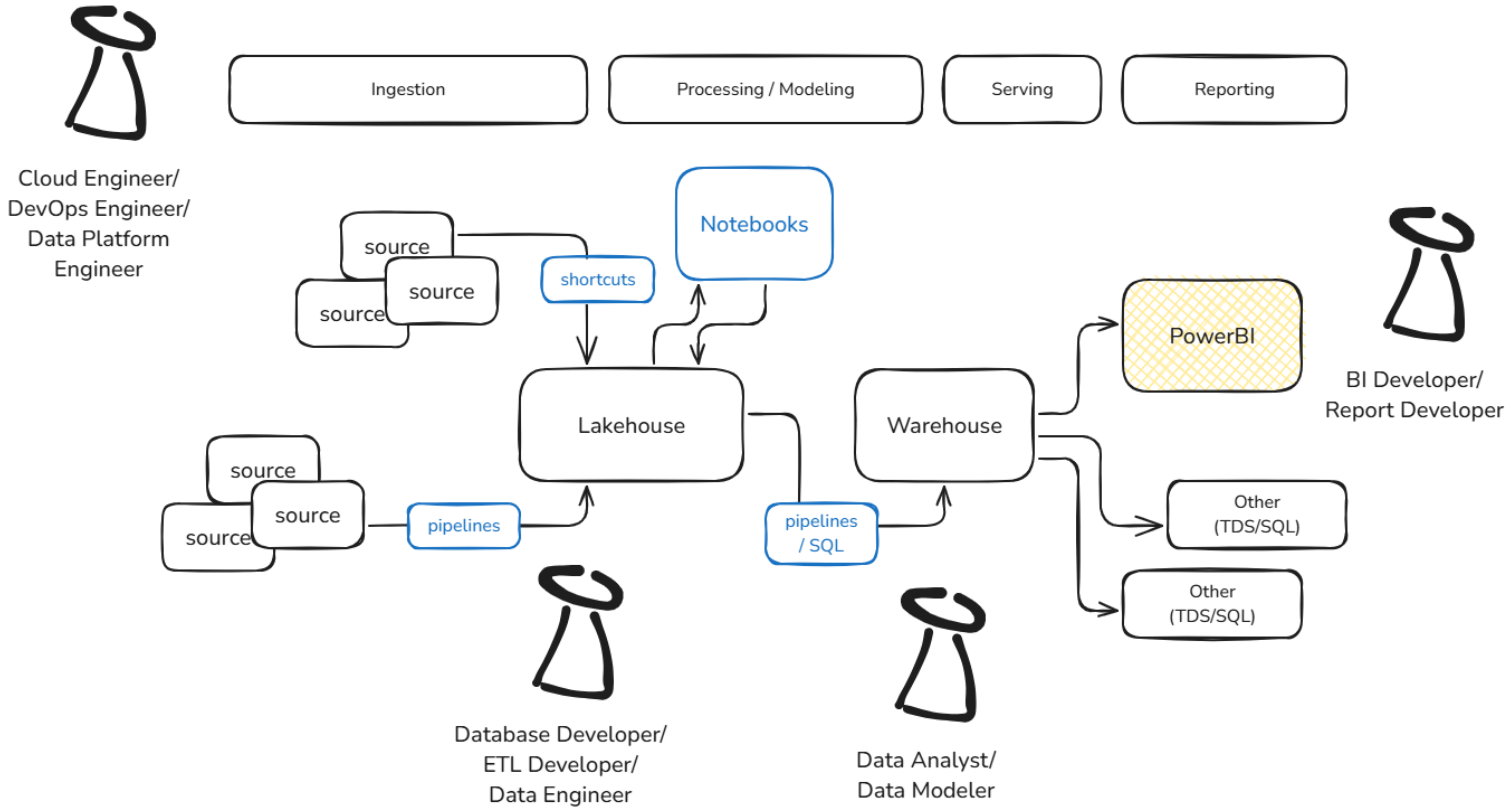
Fabric Components

getindata
Part of Xebia



Roles

getindata
Part of Xebia



It can't be difficult.

getindata
Part of Xebia

We'd like to do more!
We'd like to become Data Engineers!

But we only know SQL!

We are ok with learning new stuff!

Data Analyst/
Data Modeler

BI Developer/
Report Developer

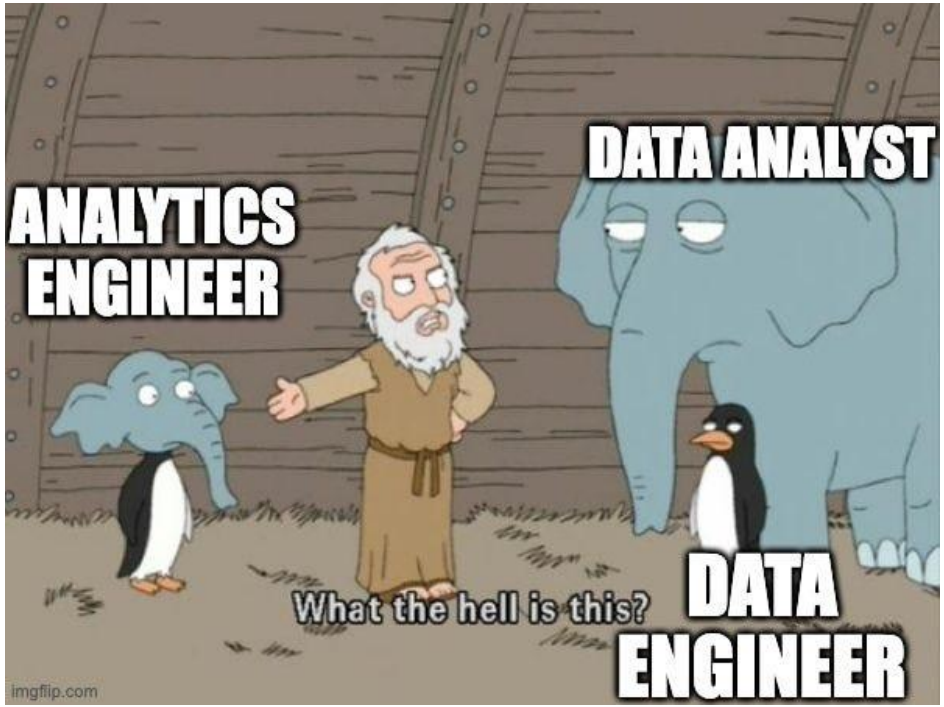
DevOps, DatOps, CI/CD, Tests,
Environments, Cloud, Programming,
Performance Tuning, Optimization,
Open Table Formats, Spark, Python,
Streaming, ..

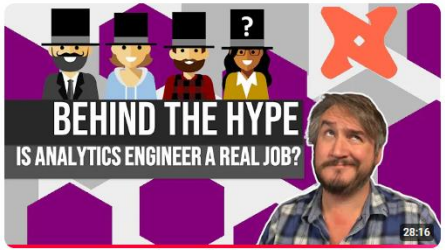
Oh, fu*k... :(

Analytics Engineering

Analytics Engineer

getindata
Part of Xebia





Behind The Hype - Is Analytics Engineer a Real Job
6,2 tys. wyświetleń • 1 rok temu
Advancing Analytics
We've been hearing more and more about the "**Analytics Engineer**", pushed hea
4K
Rozdziały: 4 Setting the Scene | Defining Engineering | The Analy

Analytics Engineer

getindata
Part of Xebia

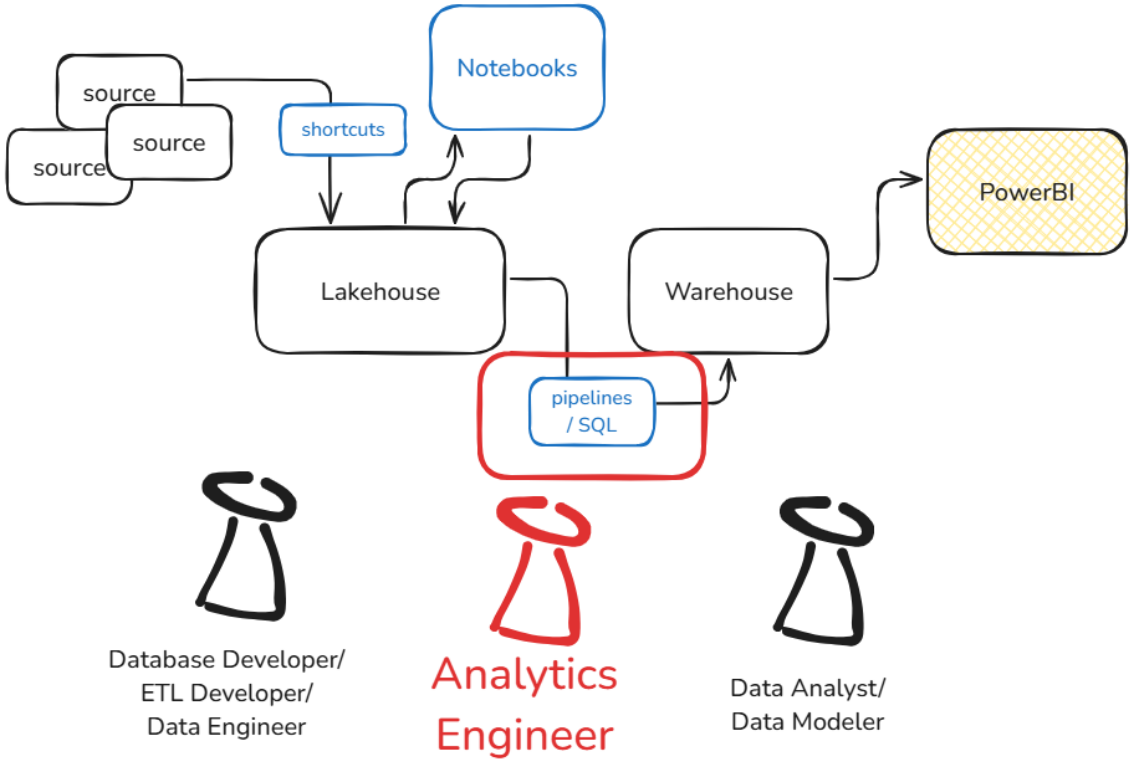
"Analytics engineers ***provide clean data sets to end users****, modeling data in a way that empowers end users to answer their own questions.*

[...], an analytics engineer spends their time ***transforming, testing, deploying, and documenting data****.*

Analytics engineers ***apply software engineering best practices*** *like VC and CI/CD to the analytics code base."*

DE / AE / DA

getindata
Part of Xebia



DE / AE / DA

getindata
Part of Xebia

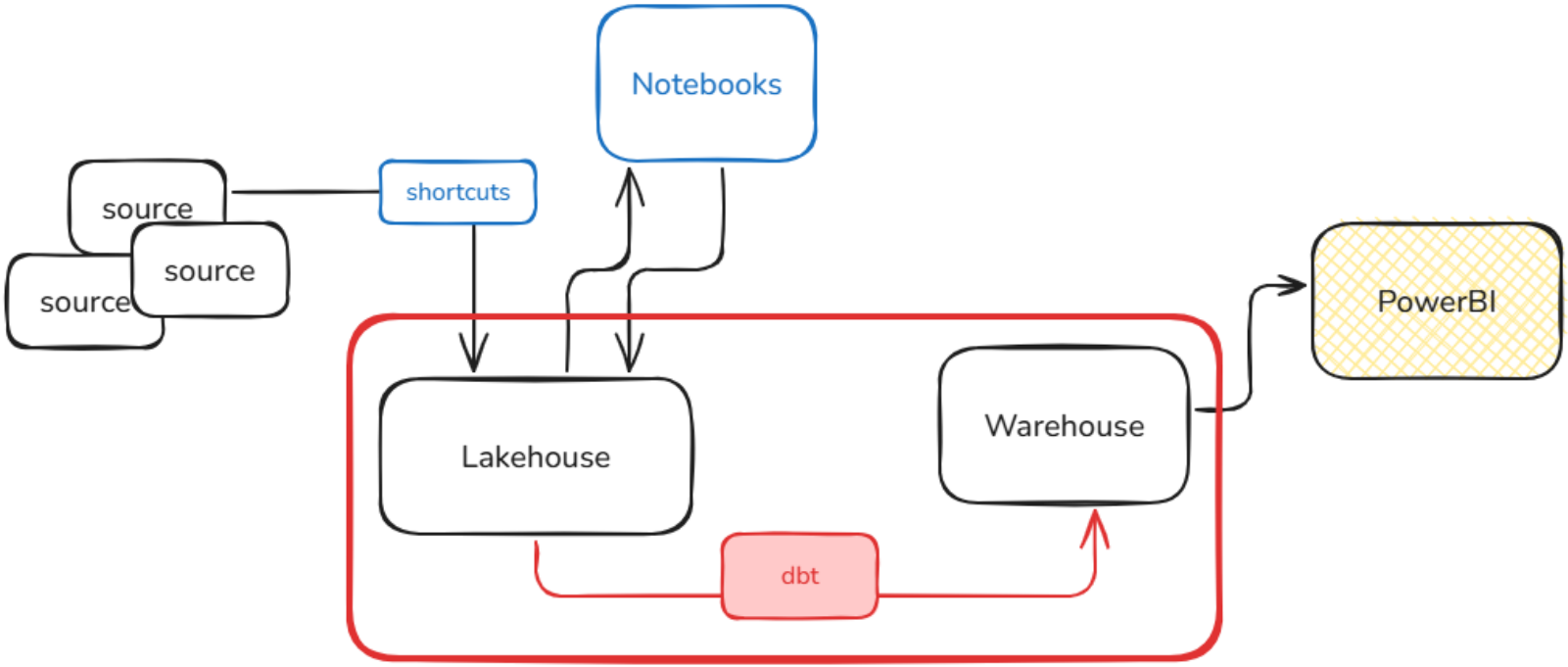
Data Engineer	Analytics Engineer	Data Analyst
• Build custom data integrations • Manage overall pipeline orchestration • Develop & deploy machine learning endpoints • Build and maintain the data platform • Data warehouse performance optimizations	• Provide clean, transformed data ready for analysis • Apply software engineering best practices to analytics code (ex: version control, testing, continuous integration • Maintain data documentation & definitions • Train business users on how to use data visualization tools	• Deep insights work (ex: why did churn spike last month? what are the best acquisition channels?) • Work with business users to understand data requirements • Build critical dashboards • Forecasting

https://www.getdbt.com/what-is-analytics-engineering

dbt [DEMO!]

goal

getindata
Part of Xebia



action plan

getindata
Part of Xebia

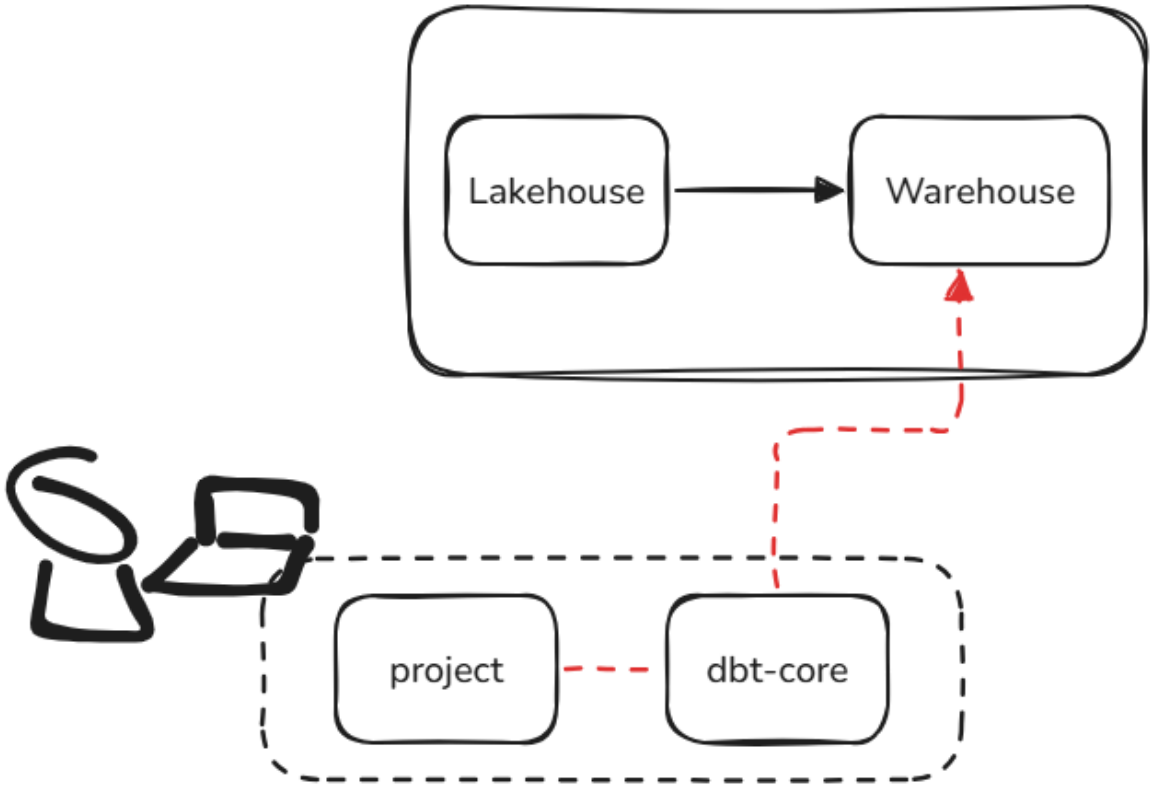
- Installation:
  - ODBC
  - Python max 3.11
  - pip install dbt-fabric
- Configuration:
  - dbt init
  - dbt_project.yml
  - profiles.yml
- Components:
  - sources
  - seeds
  - models
- Commands:
  - dbt compile
  - dbt run
  - dbt test
  - dbt build
- Documentation:
  - docs
  - lineage

getindata
Part of Xebia

[DEMO!]

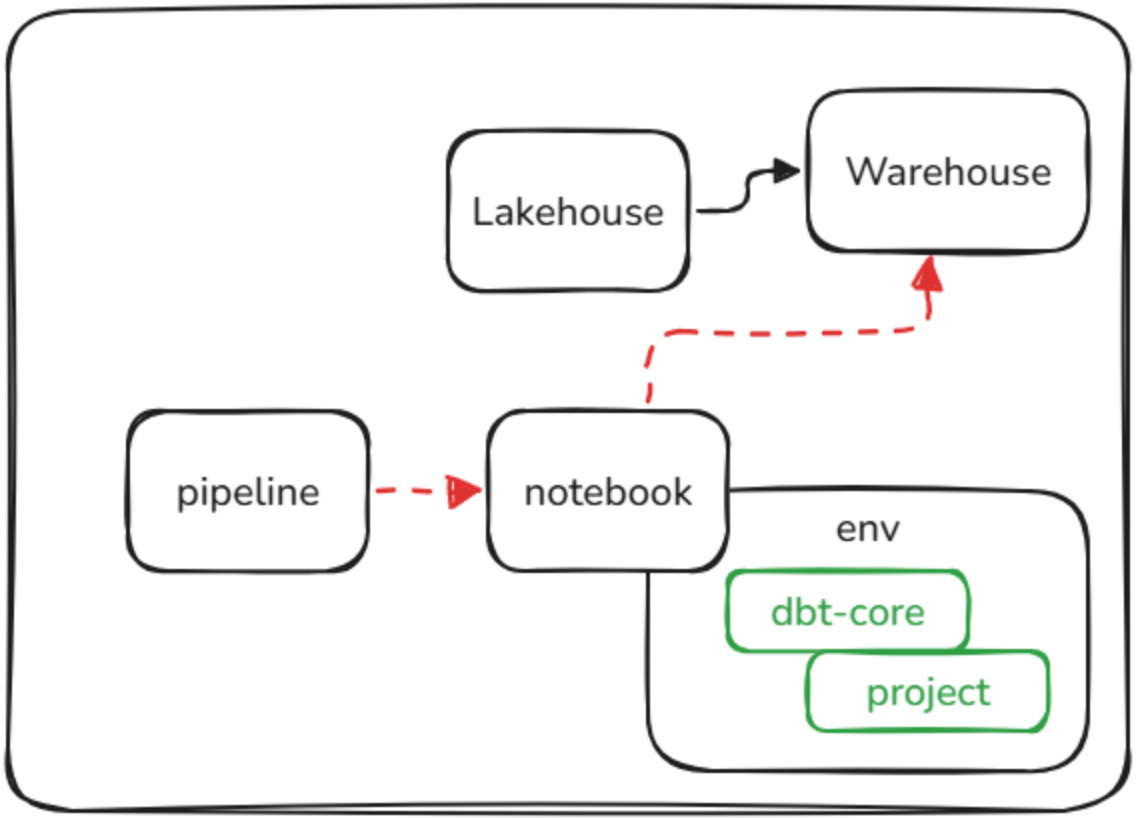
Going Prod!

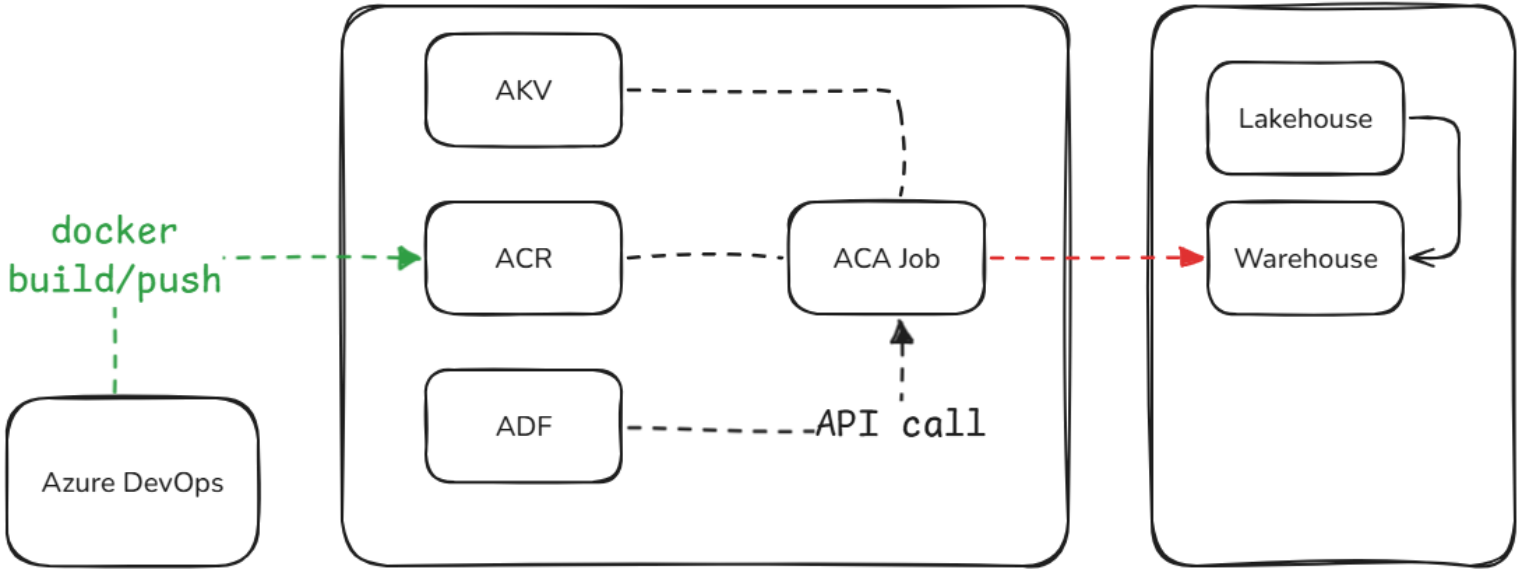
Local Execution



Pure Fabric

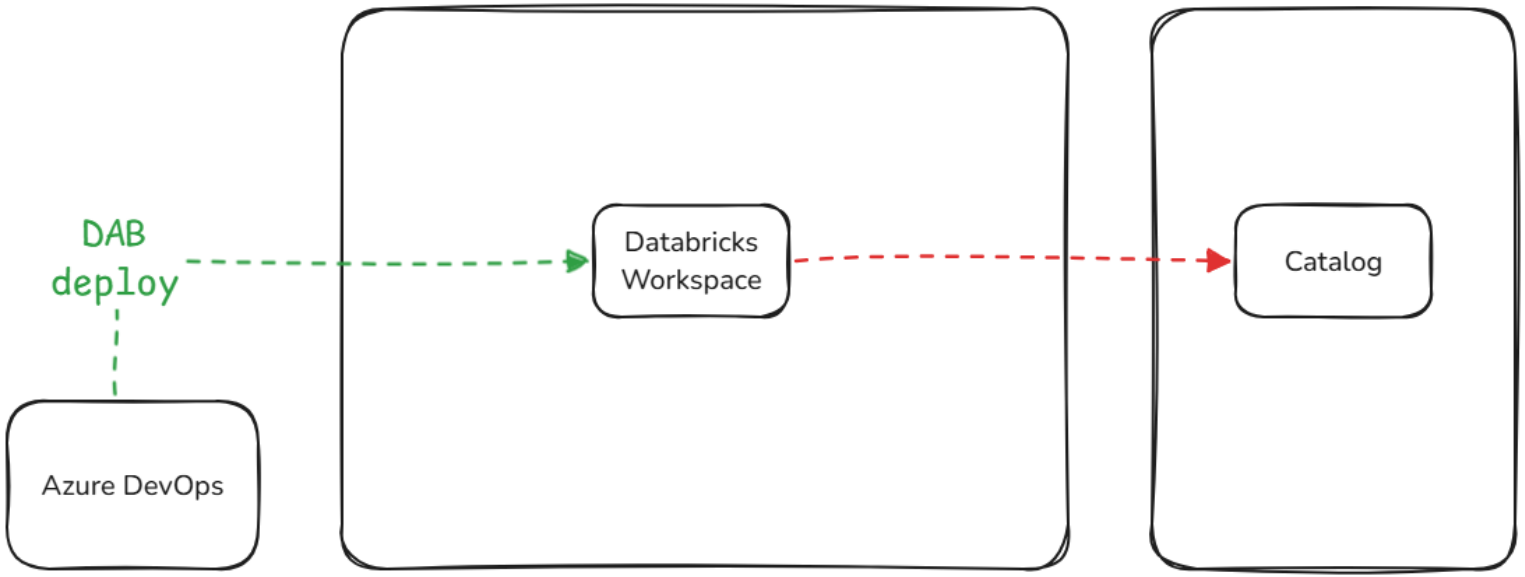
getindata
Part of Xebia




AKV
ACR
ADF
ACA Job
API call
docker
build/push
Azure DevOps
Lakehouse
Warehouse

Azure Databricks

getindata
Part of Xebia



Costs

it's free*.

* You still need to host it somewhere.

dbt Cloud pricing

getindata
Part of Xebia

dbt Cloud pricing

Get started with dbt Cloud today and get one step closer to reliable data that you and your team can trust.

Developer

The fastest way to get started with dbt Cloud

Free

One developer seat

3,000 successful models built per month

Sign up

Features

- 14 day free trial of Teams plan
- Browser-based IDE
- Job scheduling
- One project

Team

Pay as you go pricing for emerging dbt Cloud teams

$100/mo/seat

Up to 8 developer seats

15,000 successful models built per month

Start my free trial

Features

- All features in Developer
- One project
- 5 read-only seats
- Unlimited concurrent running jobs

Enterprise

Scale dbt Cloud to the changing needs of your business

Custom pricing

Custom number of developers

Custom per successful model built pricing

Book a demo

Features

- All features in Team
- Unlimited projects
- Single Sign On (SSO)
- Multiple deployment regions

getindata

Part of Xebia

QA & Thanks!