17th edition
SQLDay Conference

12-14 May 2025, WROCŁAW + ONLINE

SQL Day

Platinum sponsors

DATUMO

Woodler

Gold sponsors

lingaro

VOLVO

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

software one

Dataedo

Silver sponsors

Capgemini

DATA RISE LAB

elitmind

C&F

kursy FABRIC

Going Live with dbt-core on Databricks and MS Fabric

SQLDay Wrocław 2025-05-14

Speaker

Tomasz Kostyrka, PL

- Data Platform Architect @GetInData | Part of Xebia
- ~12 years in Data
- Azure/Databricks/PowerPoint
- Databricks Architect Champion

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

Session

Plan:

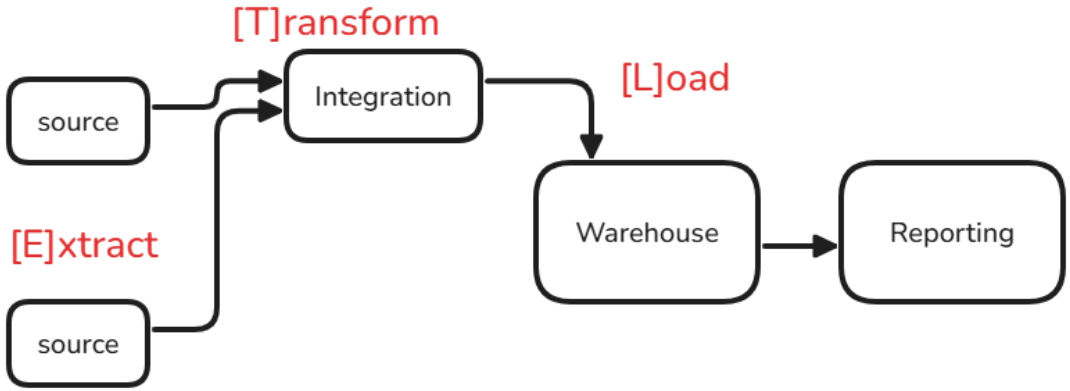
- extremely quick introduction to dbt
- Go Prod!
- Databricks/Microsoft Fabric
- Documentation

Disclaimers:

- Opinions are personal
- I'll oversimplify a lot
- Let's have some fun

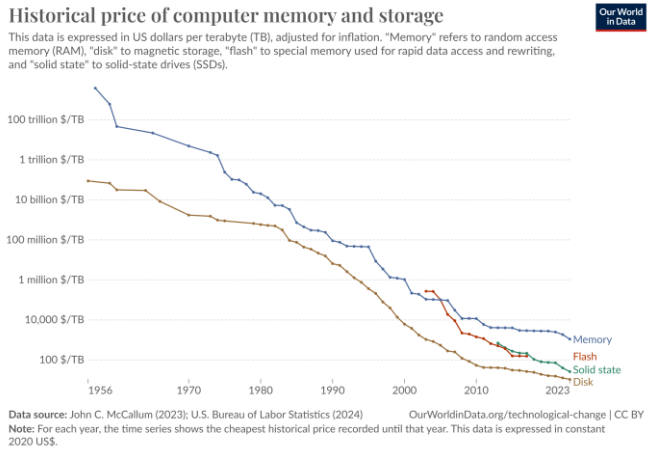
extremely quick introduction to dbt

Extract-Transform-Load

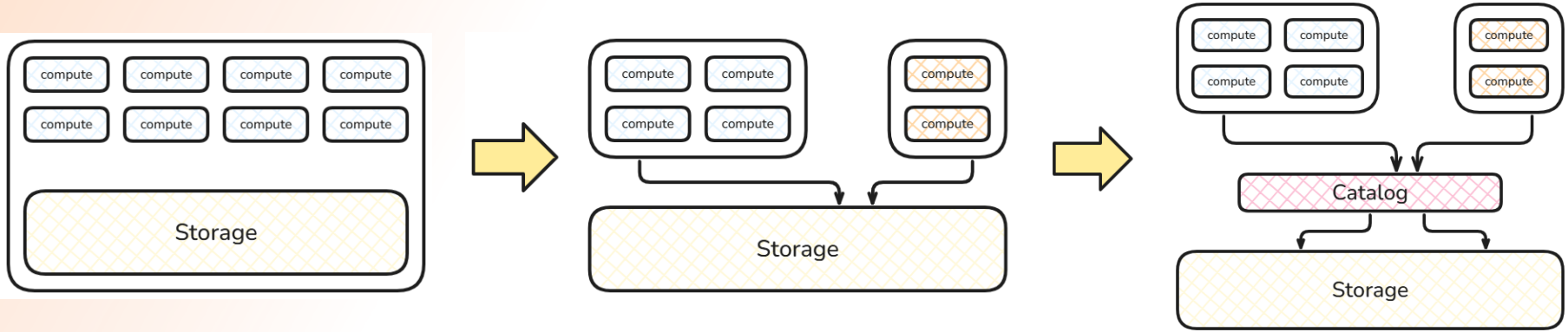


- SQL Server Integration Services (DataStage, AbInitio, Talend)
- SQL Server (Oracle, Teradata, Exadata)
- SQL Server Reporting Services (Qlik, Tableau)

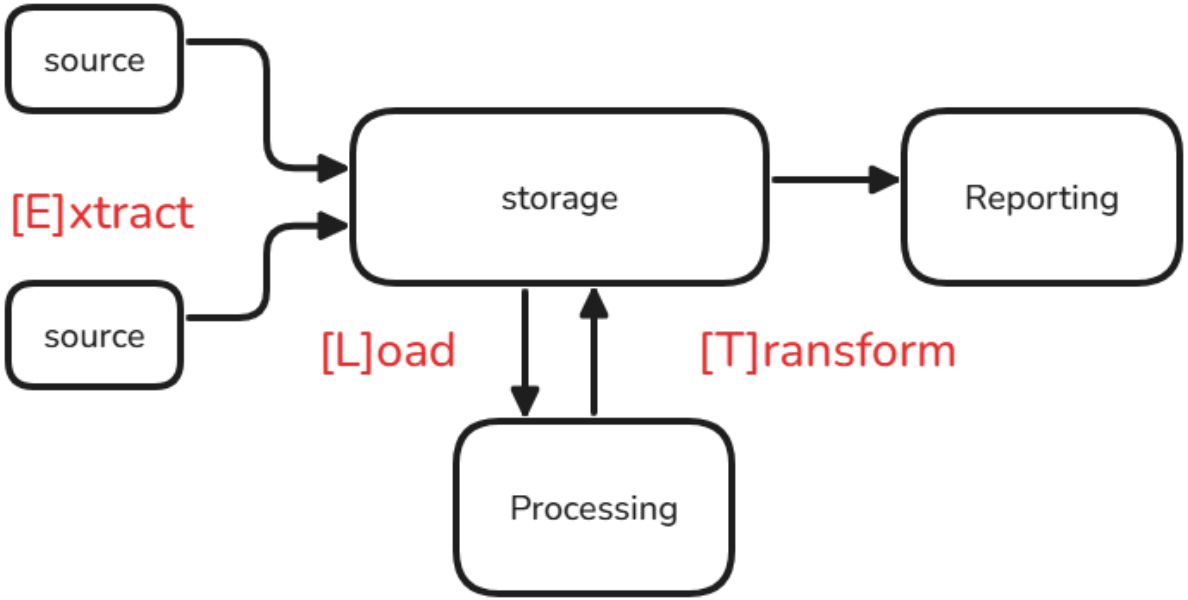

EVERY
1 MINUTE
OF THE
DAY
PRESENTED BY DOMO
TWITTER
USERS POSTS
575k
TWEETS
TIKTOK
USERS WATCH
167M
VIDEOS
GOOGLE
CONDUCTS
5.7M
SEARCHES
DISCORD
USERS SEND
668k
MESSAGES
12M
PEOPLE SEND AN
IMESSAGE
INSTAGRAM
USERS SHARE
65k
PHOTOS
CLUBHOUSE
CREATES 208 ROOMS
FACEBOOK
USERS SHARE 240k PHOTOS
SNAPCHAT
USERS SEND 2M SNAPCHATS
FACEBOOK
LIVE RECEIVES 44M VIEWS
AMAZON
CUSTOMERS SPEND $283k
YOUTUBE
USERS STREAM
694k
HOURS
6M PEOPLE SHOP
ONLINE
NETFLIX
USERS STREAM
452k
HOURS
STRAVA
ATHLETES SHARE
1.5k
ACTIVITIES
TEAMS
CONNECTS
100k
USERS
ZOOM
HOSTS
856
MINUTES OF WEBINARS
SLACK
USERS SEND
148k
MESSAGES
VENMO
USERS SEND
$304k
INSTACART
USERS SPEND
$67k


Historical price of computer memory and storage
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).
Our World in Data
100 trillion $/TB
1 trillion $/TB
10 billion $/TB
100 million $/TB
1 million $/TB
10,000 $/TB
100 $/TB
1956
1970
1980
1990
2000
2010
2023
Memory
Flash
Solid state
Disk
Data source: John C. McCallum (2023); U.S. Bureau of Labor Statistics (2024)
OurWorldinData.org/technological-change | CC BY
Note: For each year, the time series shows the cheapest historical price recorded until that year. This data is expressed in constant 2020 US$.

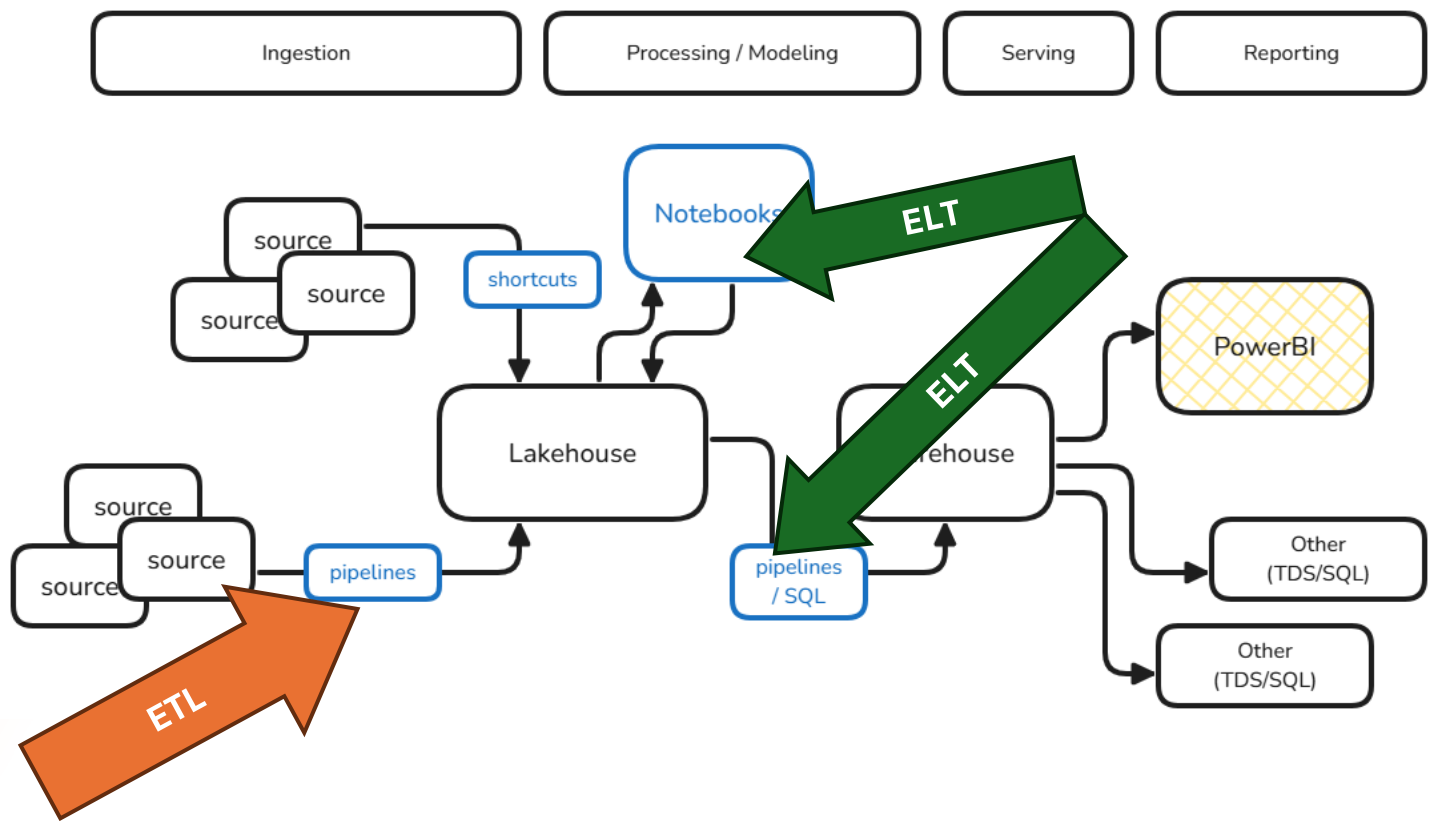
From Warehouses to Lakehouses in 30 seconds



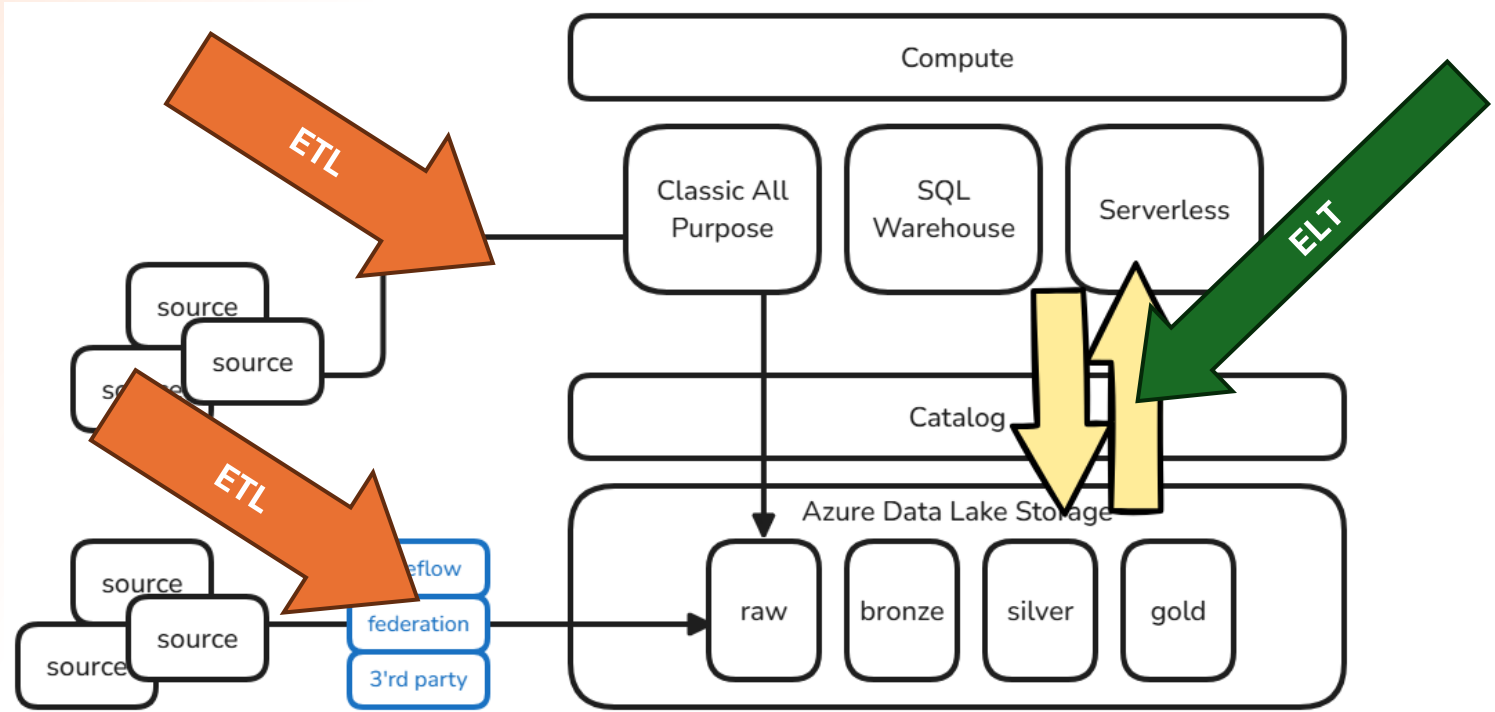
Extract-Load-Transform



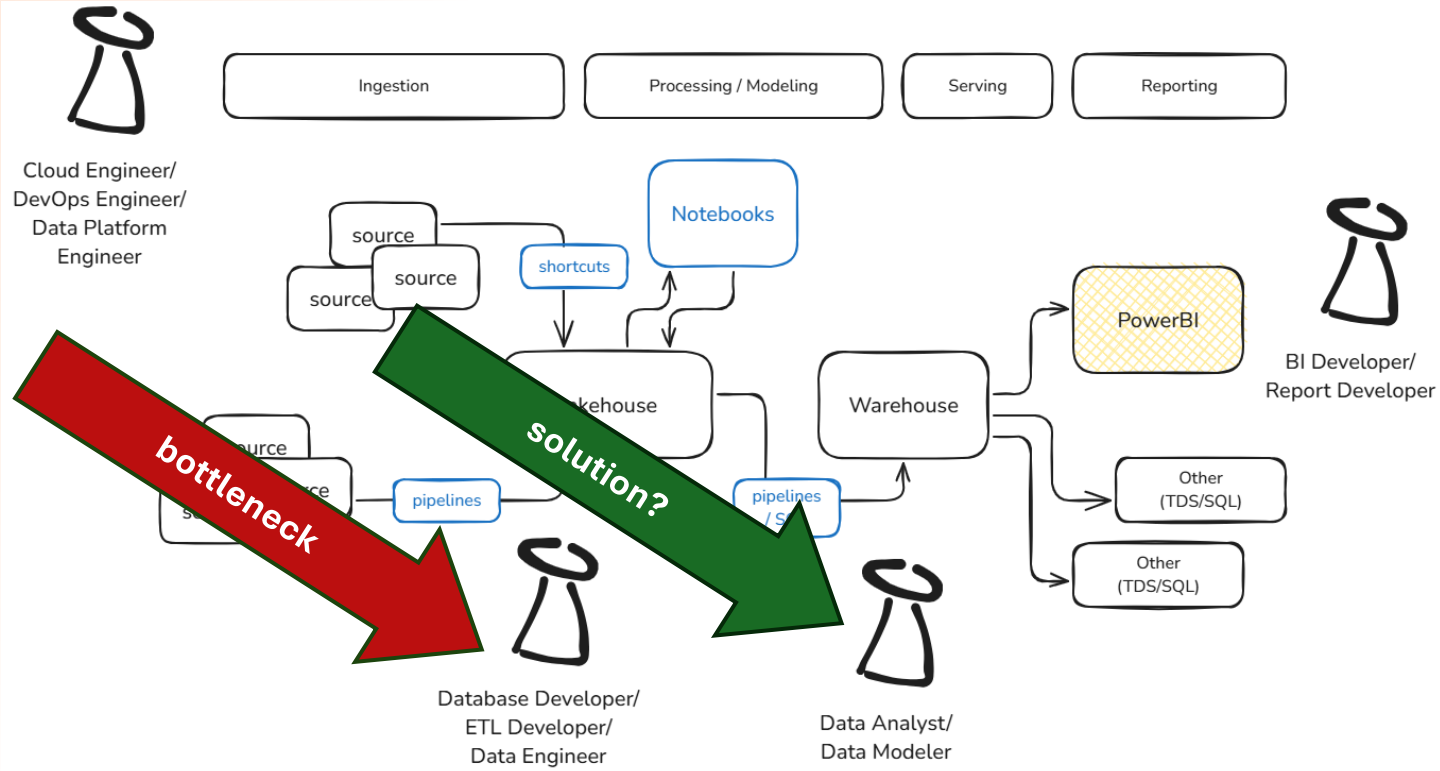
Microsoft Fabric

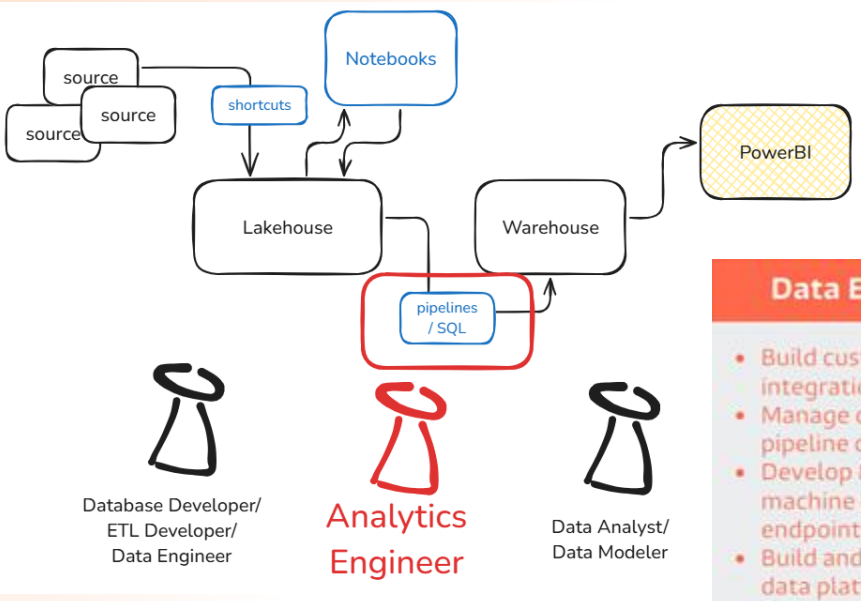


Databricks

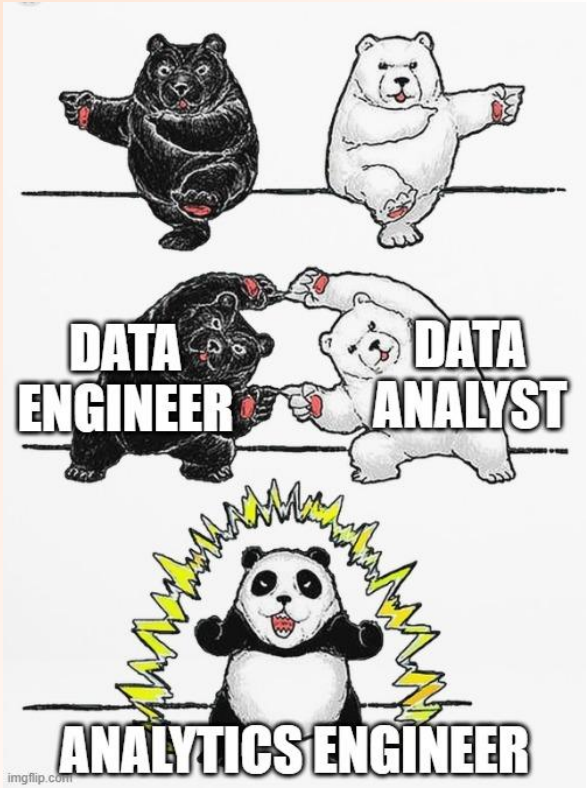

Compute
ETL
Classic All Purpose
SQL Warehouse
Serverless
ELT
source
source
Catalog
ETL
Azure Data Lake Storage
source
source
source
federation
3'rd party
raw
bronze
silver
gold

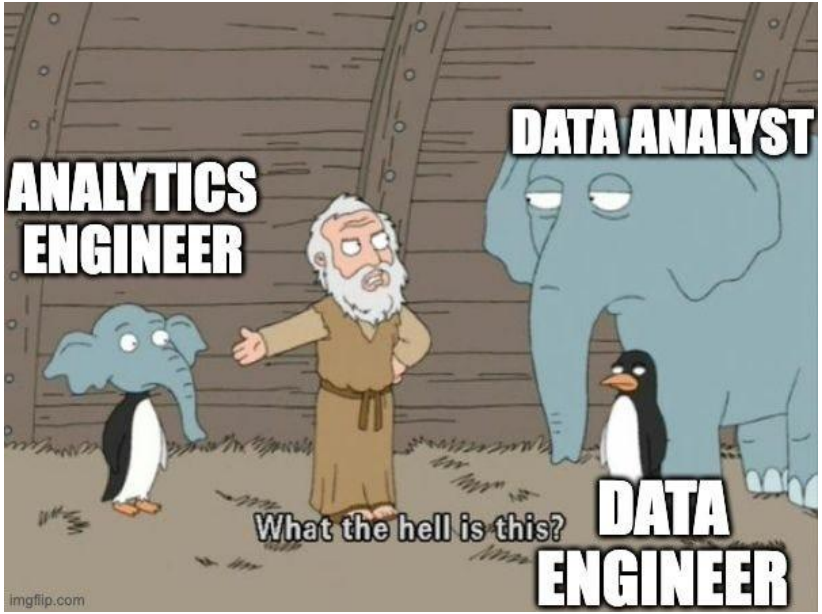
Roles


Ingestion
Processing / Modeling
Serving
Reporting
Cloud Engineer/
DevOps Engineer/
Data Platform
Engineer
source
source
source
shortcuts
Notebooks
kehouse
Warehouse
PowerBI
BI Developer/
Report Developer
pipelines
pipelines
Other
(TDS/SQL)
Other
(TDS/SQL)
bottleneck
solution?
Database Developer/
ETL Developer/
Data Engineer
Data Analyst/
Data Modeler



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


DATA ENGINEER
DATA ANALYST
ANALYTICS ENGINEER
imgflip.com


ANALYTICS ENGINEER
DATA ANALYST
What the hell is this?
DATA ENGINEER
imgflip.com

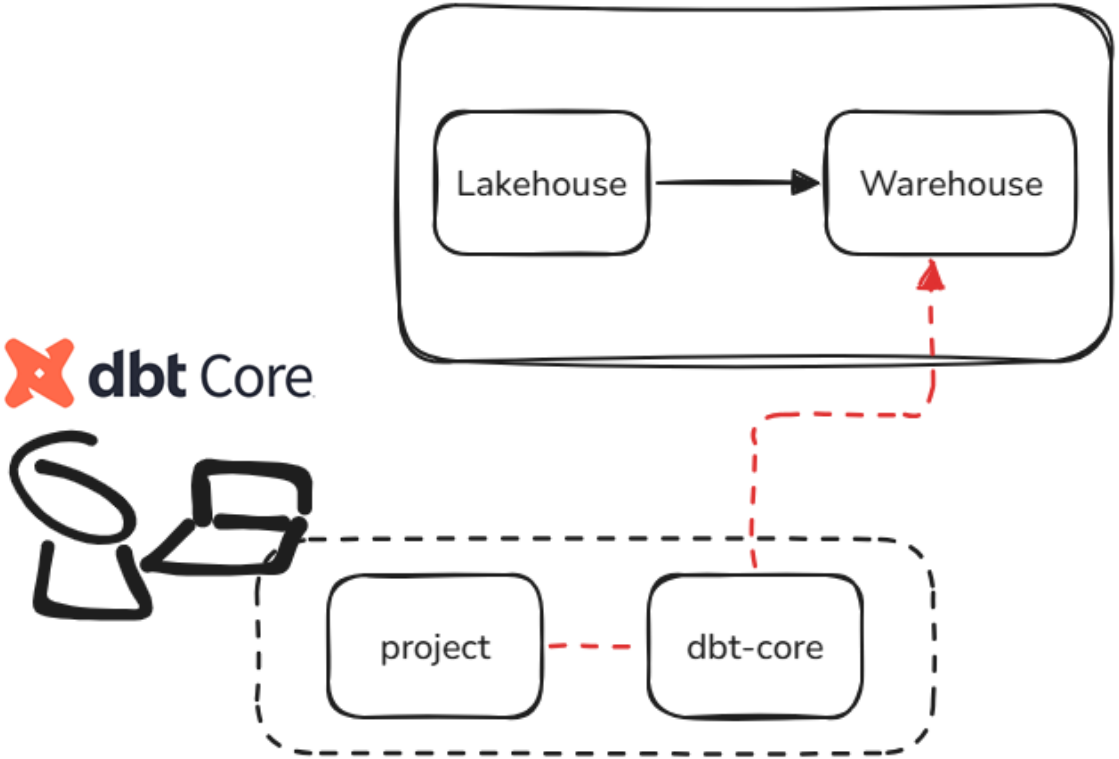

dbt™


dataform


coalesce®


SQLMesh

Local



Go Prod!

SQL Day

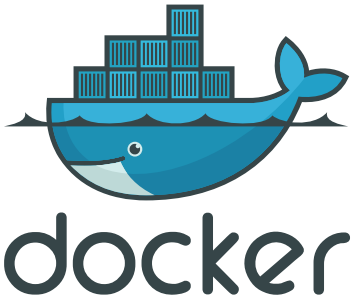
dbt Cloud

dbt Core

Host

We need a **compute environment**.

- A developer's laptop.
- A virtual machine (e.g., on-prem or cloud VM).
- A container (e.g., Docker).

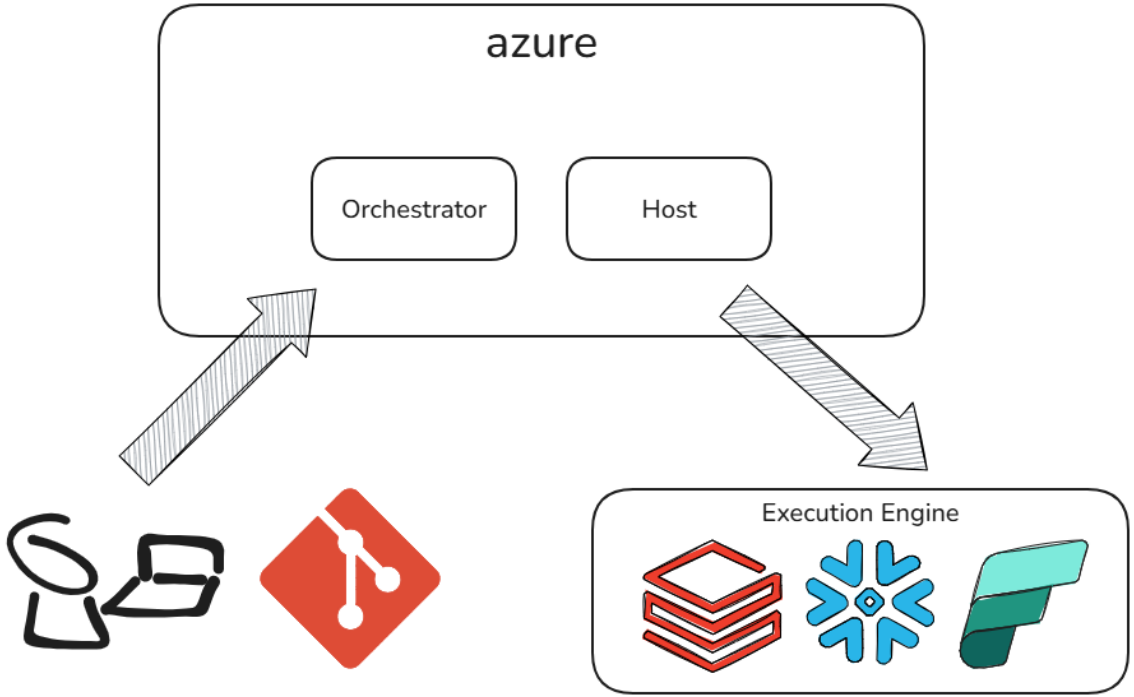


Scheduler

We need a **scheduler**.

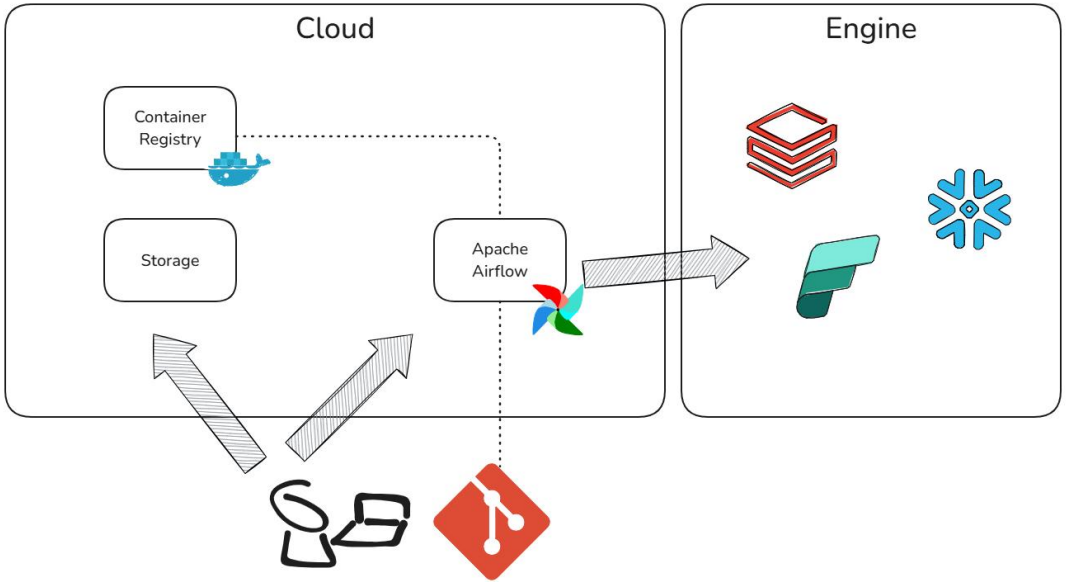
- cron jobs
- Apache Airflow
- Azure Data Factory
- Databricks Workflows

Apache Airflow

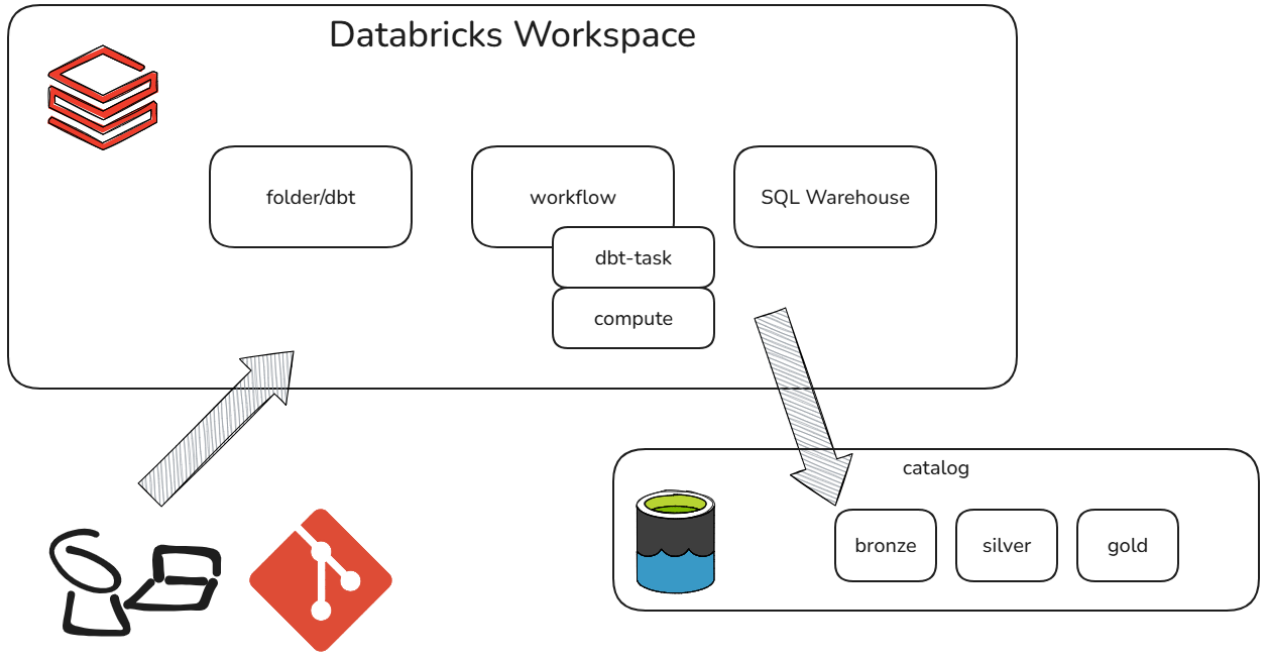

azure
Orchestrator
Host
Execution Engine

Apache Airflow

- Astronomer
- Google Cloud Composer
- Amazon MWAA (Managed Workflows for Apache Airflow)
- Azure Data Factory with Airflow integration



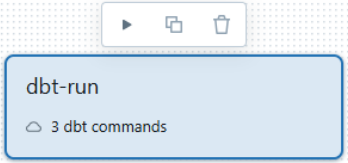
Databricks


Databricks Workspace
folder/dbt
workflow
dbt-task
compute
SQL Warehouse
catalog
bronze
silver
gold

sqlday-2025-demo ☆

Send feedback | Run now

Runs | Tasks



Task name* dbt-run

Type* dbt

Source* Workspace

Project directory* /Workspace/Users/tomasz.kostyrka@getindata.com

dbt commands*
1 `dbt deps`
2 `dbt seed`
3 `dbt run`
Add

SQL warehouse 2xsmall-tkostyrka (2XS)

Warehouse catalog sqlday2025

Warehouse schema dbo

dbt CLI compute* Serverless Autoscaling

Environment and Libraries Default

Cancel | Create task

Job details

Job ID 159261289714592

Creator Tomek Kostyrka (GID)

Run as Tomek Kostyrka (GID)

Tags Add tag

Description Add description

Lineage No lineage information for this job.
Learn more

Schedules & Triggers

None

Add trigger

Job parameters

No job parameters are defined for this job

Edit parameters

Job notifications

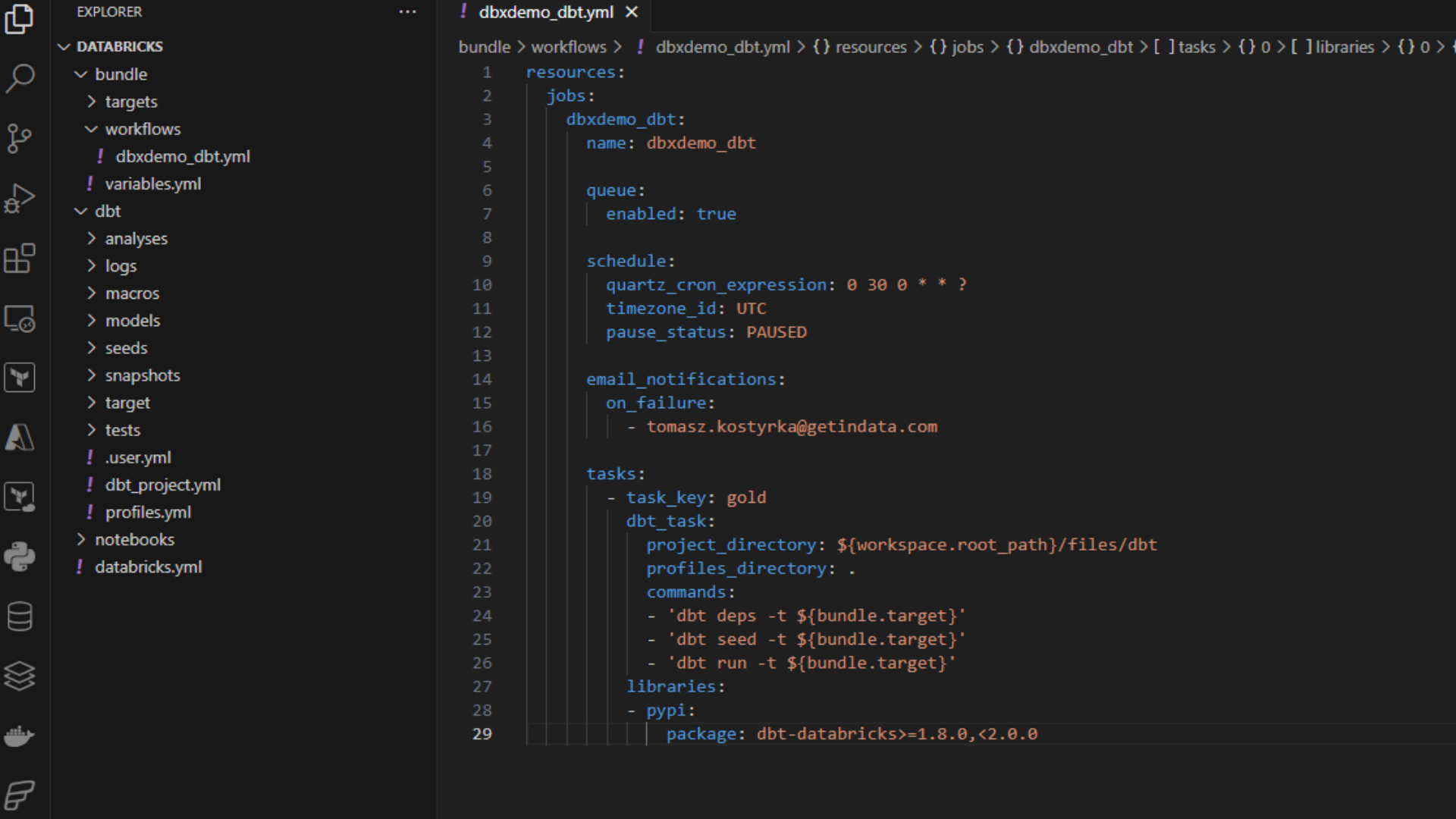
No notifications

Edit notifications

Duration and streaming backlog thresholds

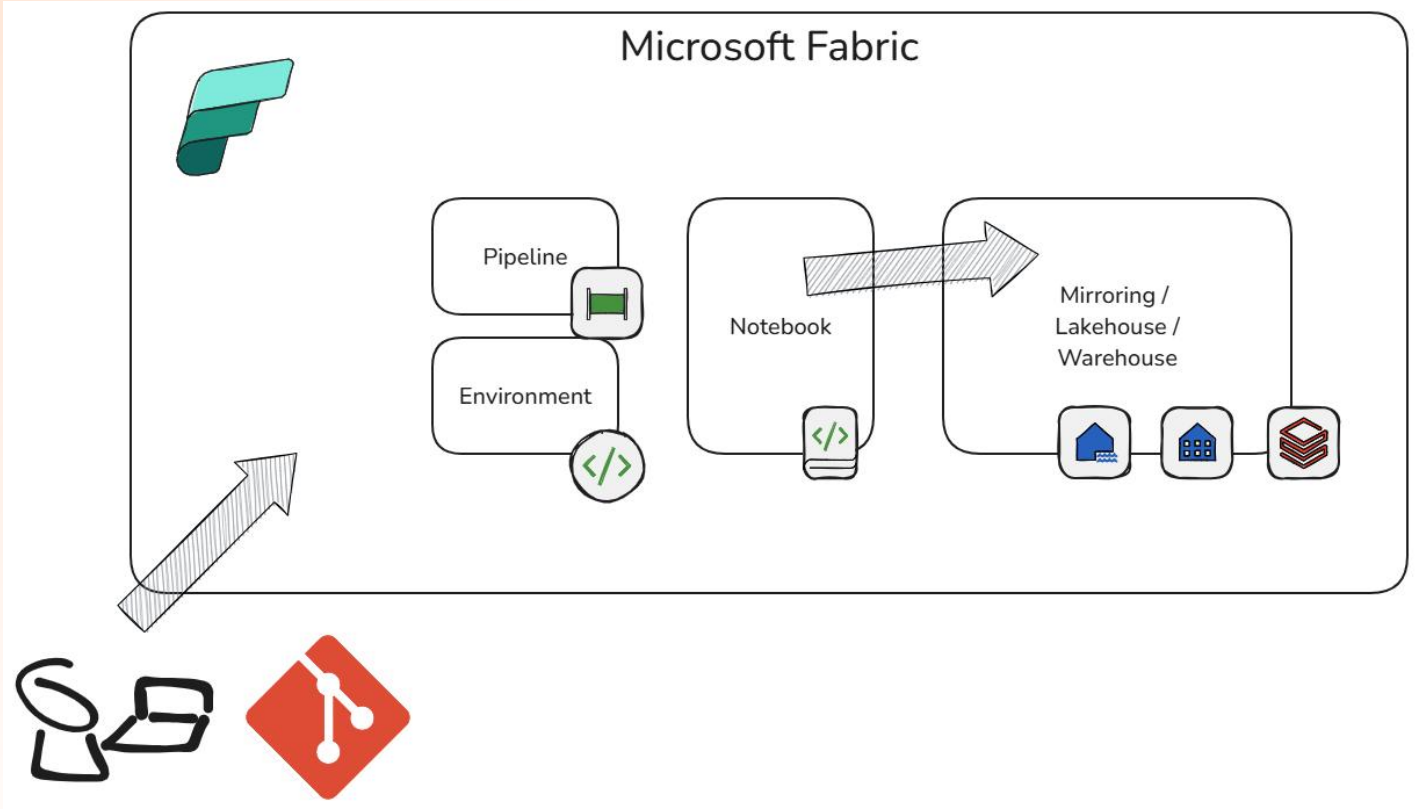
No thresholds defined

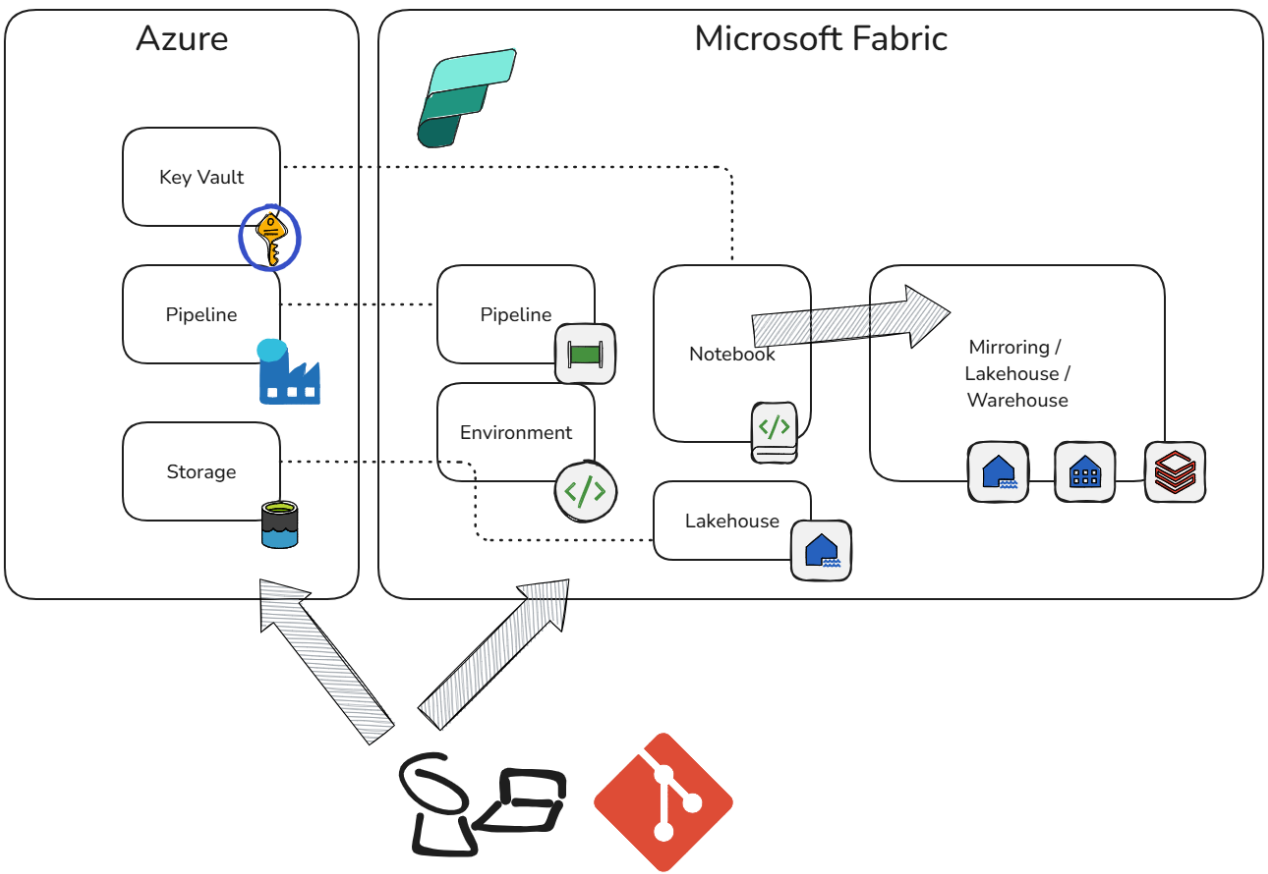
Add metric thresholds

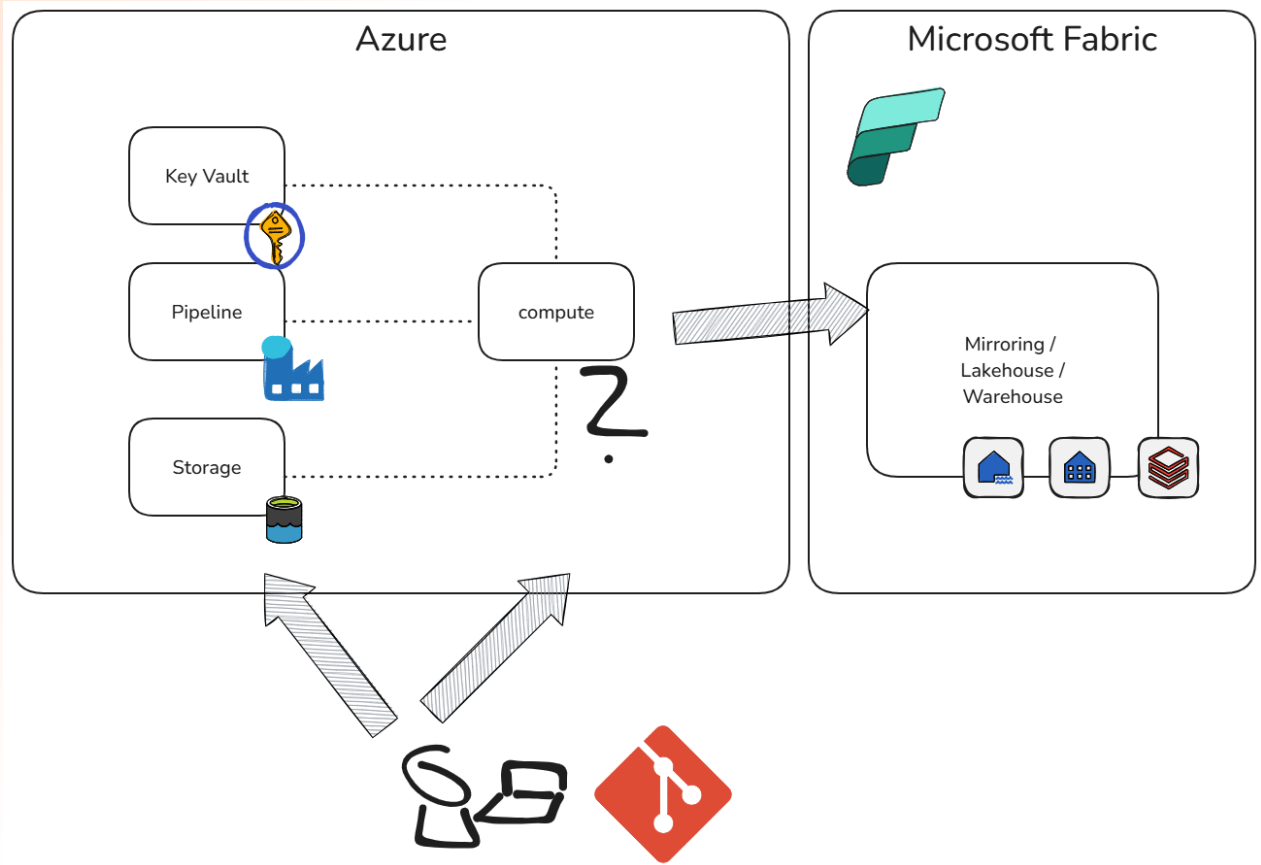

EXPLORER
DATABRICKS
bundle
targets
workflows
dbxdemo_dbt.yml
variables.yml
dbt
analyses
logs
macros
models
seeds
snapshots
target
tests
.user.yml
dbt_project.yml
profiles.yml
notebooks
databricks.yml
dbxdemo_dbt.yml
bundle > workflows > dbxdemo_dbt.yml > {} resources > {} jobs > {} dbxdemo_dbt > [ ] tasks > {} 0 > [ ] libraries > {} 0 >
resources:
jobs:
dbxdemo_dbt:
name: dbxdemo_dbt
queue:
enabled: true
schedule:
quartz_cron_expression: 0 30 0 * * ?
timezone_id: UTC
pause_status: PAUSED
email_notifications:
on_failure:
- tomasz.kostyrka@getindata.com
tasks:
- task_key: gold
dbt_task:
project_directory: ${workspace.root_path}/files/dbt
profiles_directory: .
commands:
- 'dbt deps -t ${bundle.target}'
- 'dbt seed -t ${bundle.target}'
- 'dbt run -t ${bundle.target}'
libraries:
- pypi:
package: dbt-databricks>=1.8.0,<2.0.0

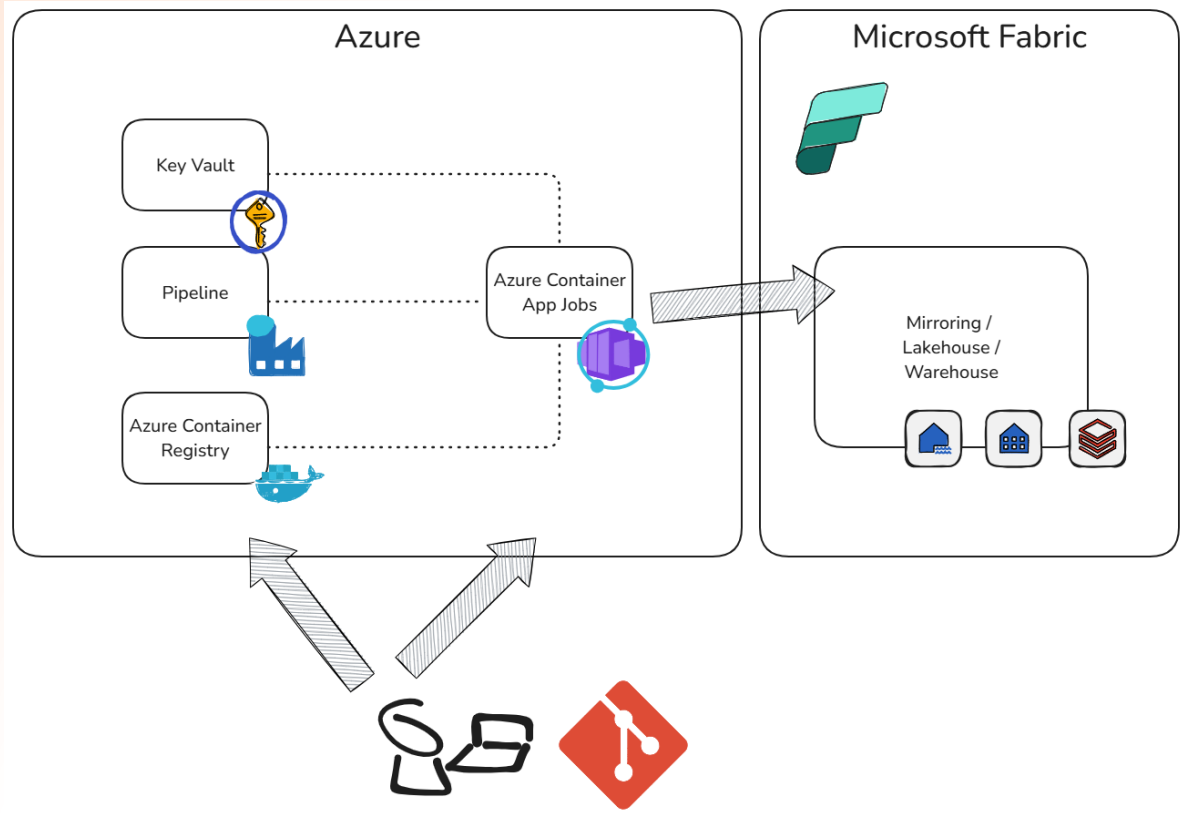
```yaml
strategy:
  runOnce:
    deploy:
      steps:
      - checkout: self

      - task: UsePythonVersion@0
        inputs:
          versionSpec: '3.11'
        displayName: 'Use Python 3.11'

      - script: |
          curl -fsSL https://raw.githubusercontent.com/databricks/setup-cli/main/install.sh | sh
          databricks -v
        displayName: install databricks

      - script: |
          curl -sSL https://install.python-poetry.org | python3 -
          poetry --version
        displayName: install poetry

      - task: AzureCLI@2
        displayName: deploy bundle
        inputs:
          azureSubscription: ${{ parameters.serviceconn }}
          scriptType: bash
          scriptLocation: inlineScript
          addSpnToEnvironment: true
          workingDirectory: ${{ parameters.workingDirectory }}
          inlineScript: |
            export ARM_CLIENT_ID=$servicePrincipalId
            export ARM_CLIENT_SECRET=$servicePrincipalKey
            export ARM_TENANT_ID=$tenantId

            databricks bundle deploy -t ${{ parameters.target }}
```

Microsoft Fabric


Microsoft Fabric
Pipeline
Environment
Notebook
Mirroring /
Lakehouse /
Warehouse


Azure
Microsoft Fabric
Key Vault
Pipeline
Storage
Pipeline
Environment
Notebook
Lakehouse
Mirroring /
Lakehouse /
Warehouse

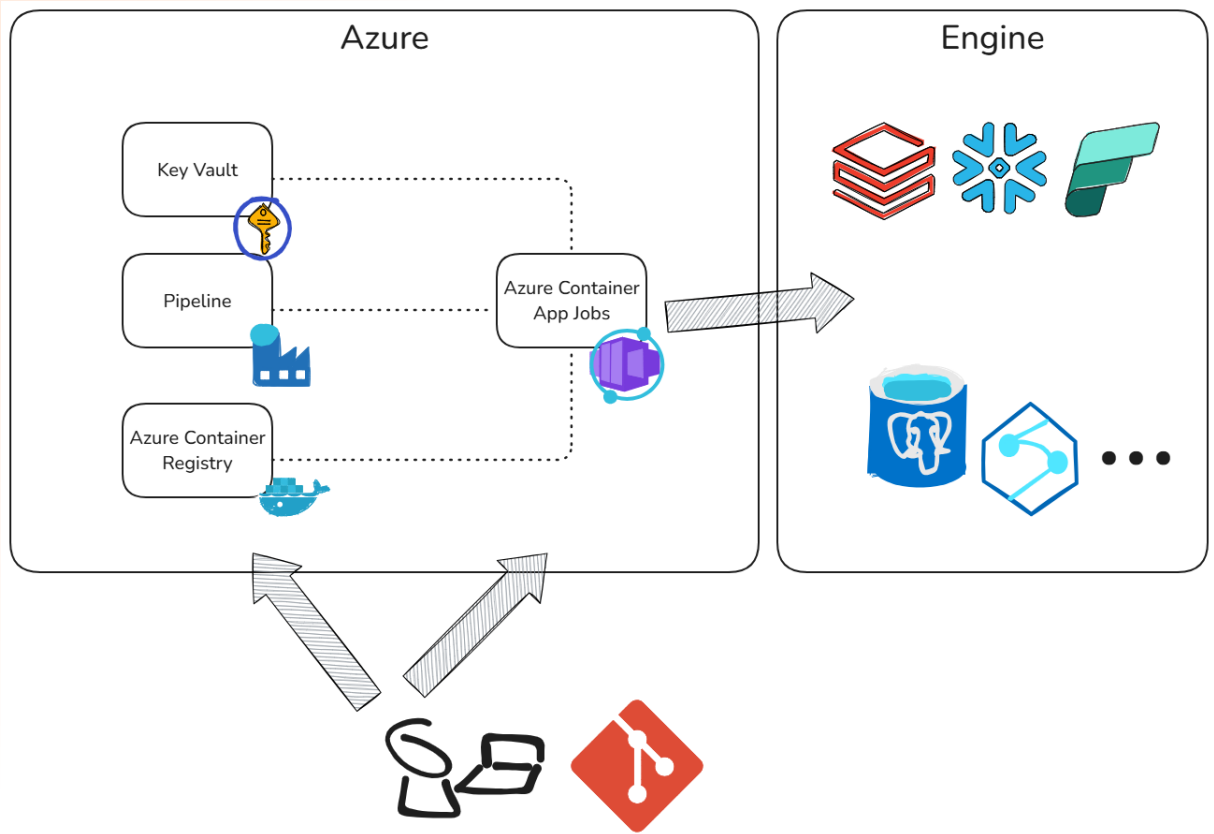

Azure
Microsoft Fabric
Key Vault
Pipeline
Storage
compute
Mirroring /
Lakehouse /
Warehouse


Azure
Key Vault
Pipeline
Azure Container Registry
Azure Container App Jobs
Microsoft Fabric
Mirroring / Lakehouse / Warehouse

Dockerfile

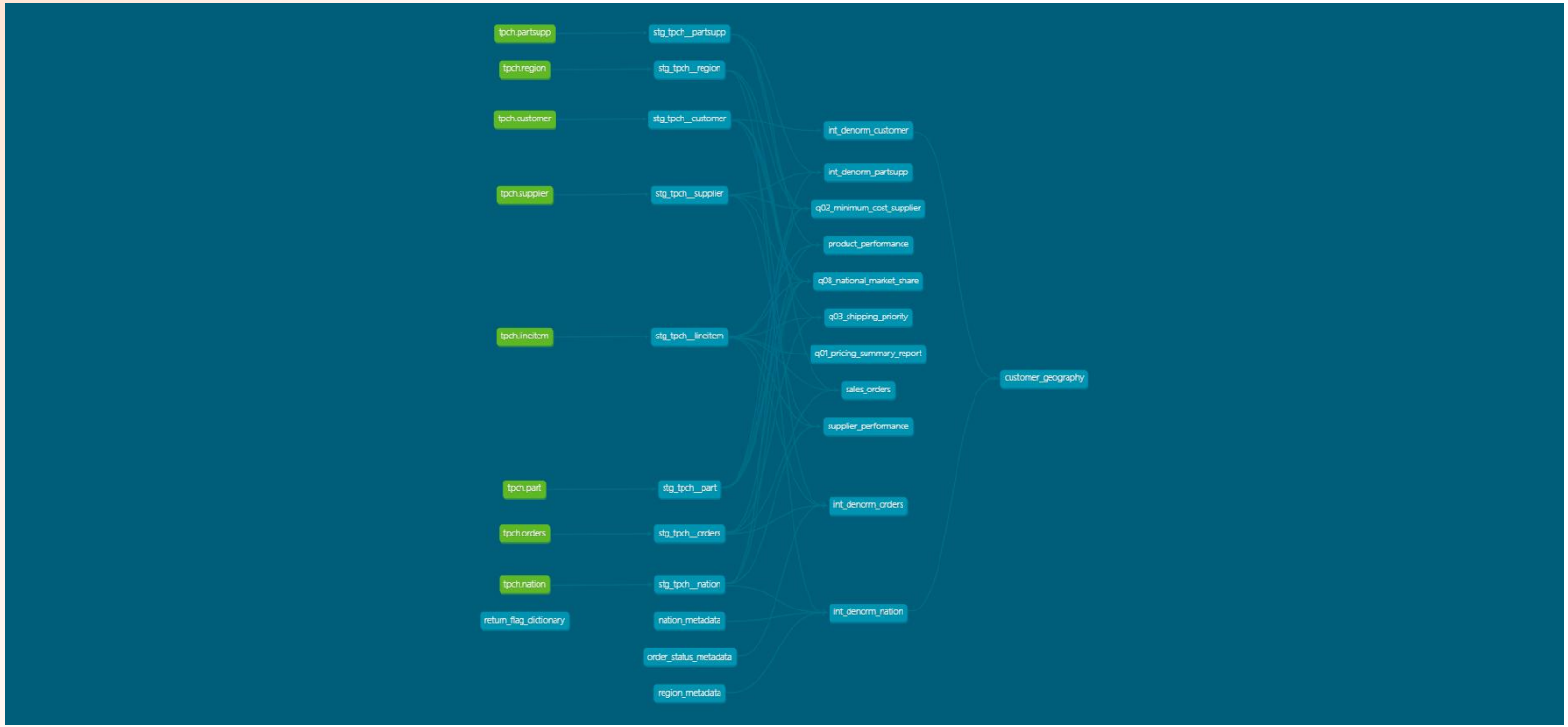
```
FROM ghcr.io/dbt-labs/dbt-core:1.9.2

RUN apt-get update && \
    apt-get install -y --no-install-recommends unixodbc-dev curl gnupg2 && \
    rm -rf /var/lib/apt/lists/*

RUN curl https://packages.microsoft.com/keys/microsoft.asc | tee /etc/apt/trusted.gpg.d/microsoft.asc && \
    curl https://packages.microsoft.com/config/debian/11/prod.list | tee /etc/apt/sources.list.d/mssql-release.list && \
    apt-get update && \
    ACCEPT_EULA=Y apt-get install -y msodbcsql18 && \
    apt-get clean && \
    rm -rf /var/lib/apt/lists/*

RUN pip install dbt-fabric~=1.9.2

COPY . /usr/app
COPY profiles.yml /root/.dbt/

RUN dbt deps
CMD ["run"]
```

```
        # Download Artifact
        # ----------------------------------------
        - task: DownloadPipelineArtifact@2
          inputs:
            buildType: current
            artifactName: ${{ parameters.artifactname }}
            targetPath: $(Pipeline.Workspace)/drop/dbt

        # Download Artifact
        # ----------------------------------------
        - task: Docker@2
          displayName: buildAndPush to ACR
          inputs:
            command: buildAndPush
            repository: ${{ parameters.repository }}
            dockerfile: $(Pipeline.Workspace)/drop/dbt/Dockerfile
            containerRegistry: ${{ parameters.registry }}
            tags: |
              latest
              $(Build.BuildId)
```

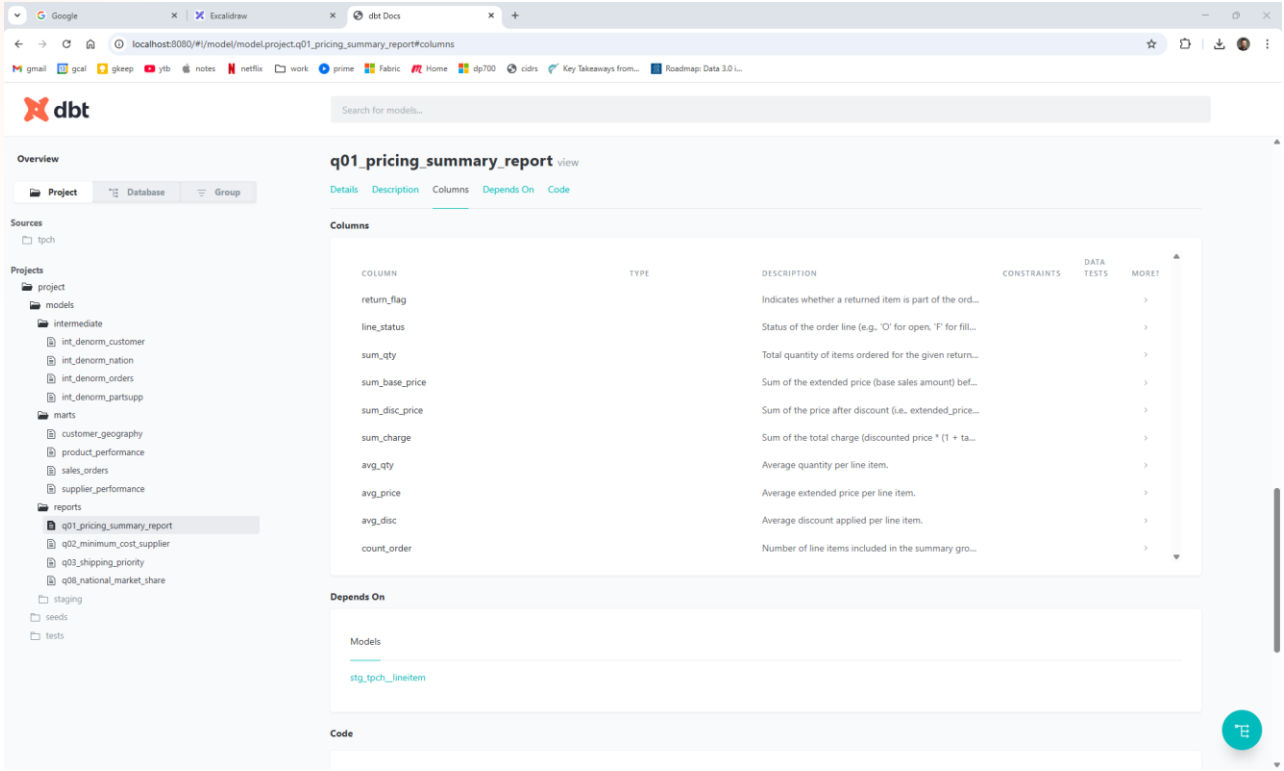

Azure
Key Vault
Pipeline
Azure Container Registry
Azure Container App Jobs
Engine

Documentation

Lineage


tpch.partsupp
tpch.region
tpch.customer
tpch.supplier
tpch.lineitem
tpch.part
tpch.orders
tpch.nation
return_flag_dictionary
stg_tpch__partsupp
stg_tpch__region
stg_tpch__customer
stg_tpch__supplier
stg_tpch__lineitem
stg_tpch__part
stg_tpch__orders
stg_tpch__nation
nation_metadata
order_status_metadata
region_metadata
int_denorm_customer
int_denorm_partsupp
q02_minimum_cost_supplier
product_performance
q08_national_market_share
q03_shipping_priority
q01_pricing_summary_report
sales_orders
supplier_performance
int_denorm_orders
int_denorm_nation
customer_geography

Documentation



Hosting

Azure Blob Storage

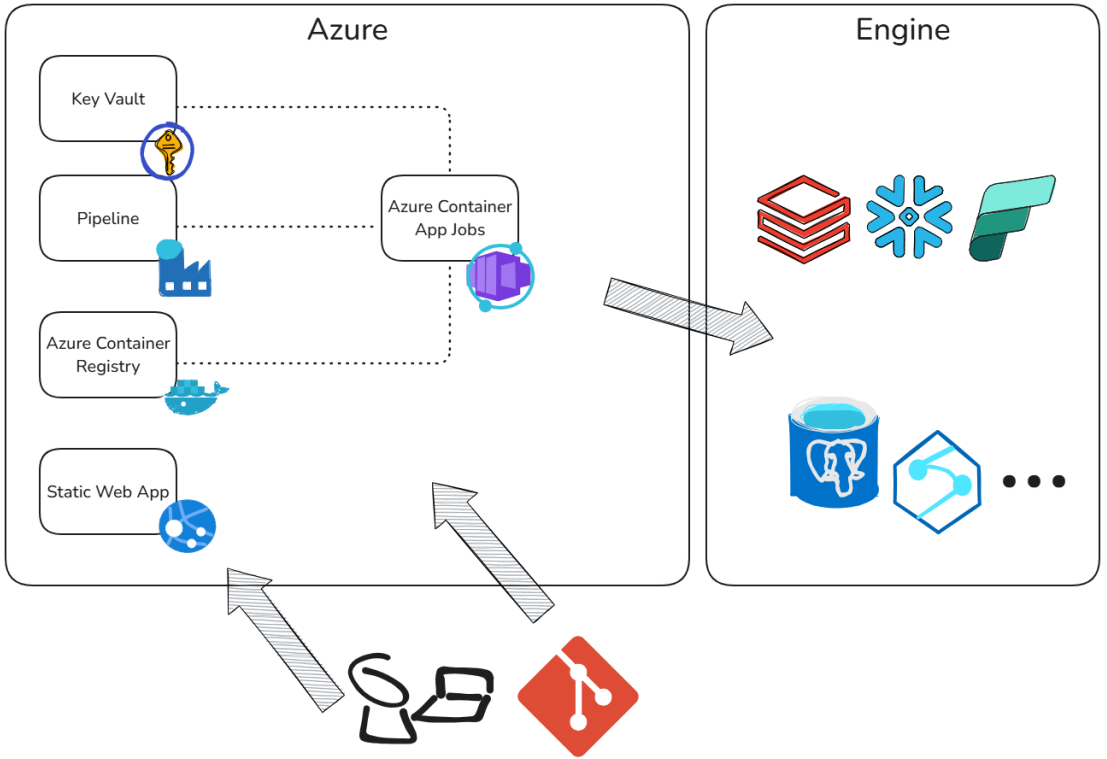
- Store static files in Blob Storage.
- Low cost based on storage and data transfer.

Azure Static Web Apps

- Free tier with 1 custom domain, SSL, and CDN.
- Easy deployment from GitHub/Azure DevOps.
- Ideal for small static sites.

```
# install dbt-fabric
# --------------
- script: |
    pip install dbt-fabric
  displayName: Install dbt-fabric

# Microsoft ODBC 18
# The following sections explain how to install the
# Microsoft ODBC driver 18 from the bash shell for different Linux distributions.
# https://learn.microsoft.com/en-us/sql/connect/odbc/linux-mac/installing-the-microsoft-odbc-driver-for-sql-server?
# --------------
- task: Bash@3
  displayName: Install ODBC 18
  inputs:
    targetType: 'inline'
    script: |
      curl https://packages.microsoft.com/keys/microsoft.asc | sudo tee /etc/apt/trusted.gpg.d/microsoft.asc
      curl https://packages.microsoft.com/keys/microsoft.asc | sudo gpg --dearmor -o /usr/share/keyrings/microsoft-prod.gpg
      curl https://packages.microsoft.com/config/ubuntu/$(lsb_release -rs)/prod.list | sudo tee /etc/apt/sources.list.d/mssql-release.list

      sudo apt-get update
      sudo ACCEPT_EULA=Y apt-get install -y msodbcsql18

# generate documentation
# --------------
- task: AzureCLI@2
  displayName: dbt docs generate
  inputs:
    azureSubscription: ${{ parameters.serviceconn }}
    scriptType: bash
    scriptLocation: inlineScript
    addSpnToEnvironment: true
    workingDirectory: ${{ parameters.workingDirectory }}
    inlineScript: |
      export AZURE_TENANT_ID=$tenantId
      export AZURE_CLIENT_ID=$servicePrincipalId
      export AZURE_CLIENT_SECRET=$servicePrincipalKey

      mv profiles.yml.sample profiles.yml

      dbt deps
      dbt docs generate -t ${{ parameters.env }}
```

The Poor Man's Setup



Summary

- If you haven't done so already, take some time to explore concepts like ELT, Analytics Engineering, and the tools that support these roles - such as dbt.
- A production-level setup can be built with dbt Core.
- We can build a generic mechanism that works across different execution engines.
- Databricks makes this much easier for us by providing support for dbt through Databricks Asset Bundles.

QA & Thanks!