Building a Data Platform from scratch

Rodel van Rooijen

PyData Amsterdam 2024

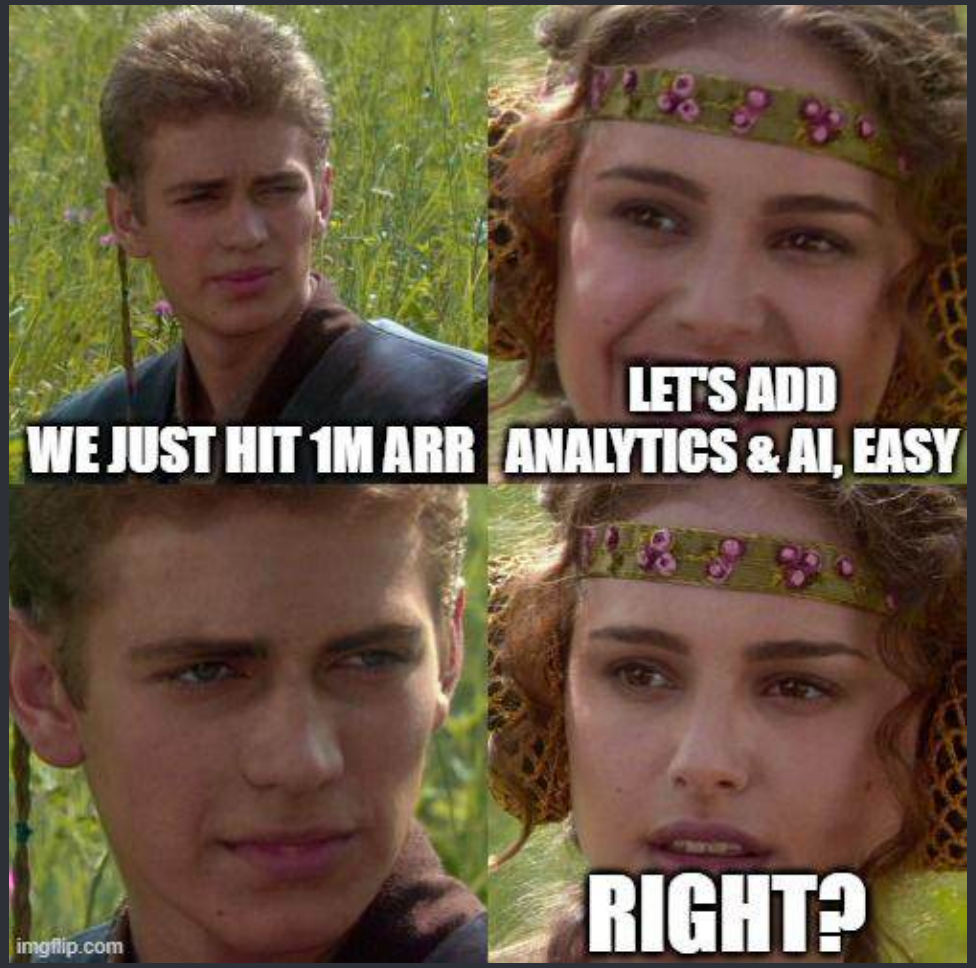
PyData

About me

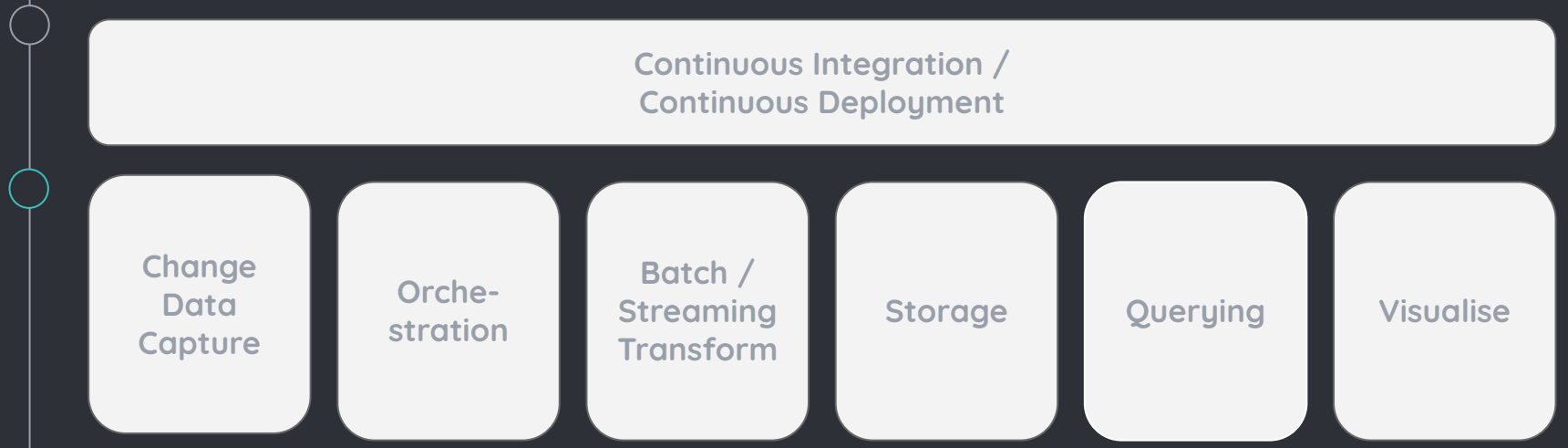
Over 8 years of experience in building data infrastructure and data products.

- **ING**
  - Data Scientist
- **Adyen**
  - Senior ML Scientist
  - Tech Lead (Manager)
  - Engineering Lead (Director)
- **Solvimon**
  - Founding Engineer, Data
- **Achmea**
  - Senior ML Engineer

https://rodel.dev/


WE JUST HIT 1M ARR
LET'S ADD
ANALYTICS & AI, EASY
RIGHT?
imgflip.com

Dissecting a data platform



Azure DevOps

GitLab

Continuous Integration / Continuous Deployment

argo

GitHub

Airbyte

Fivetran

debezium

Change Data Capture

Kubeflow

MAGE

Azure Data Factory

PREFECT

Apache Airflow

Orche-stration

Apache Spark

polars

Kafka Streams

beam

Flink

Batch / Streaming Transform

Google BigQuery

amazon REDSHIFT

PostgreSQL

snowflake

databricks

hadoop

DELTA LAKE

Storage

trino

druid

Querying

Apache Superset

Metabase

Power BI

tableau

Looker

Visualise


NOT SURE WHICH TOOL I SHOULD USE
SO MANY CHOICES
memegenerator.net

Where to start?

Azure

Google Cloud

aws

Cloud Platform

GitLab

Apache Spark

Apache Airflow

Looker

My team's knowledge

Airbyte

Metabase

Use open-source

1. Pick cloud platform (or use existing).
2. Start with what you know.
3. Use open-source solutions where possible.

What are the (hosting) options?

Self-hosting

++ Full control
++ High degree of flexibility
+ Lowest immediate cost

-- Long implementation time
-- Maintenance & expertise

Managed (open-source)

++ Less expertise required
+ Decent flexibility
+ Low implementation time

- Less control
- Can be more expensive

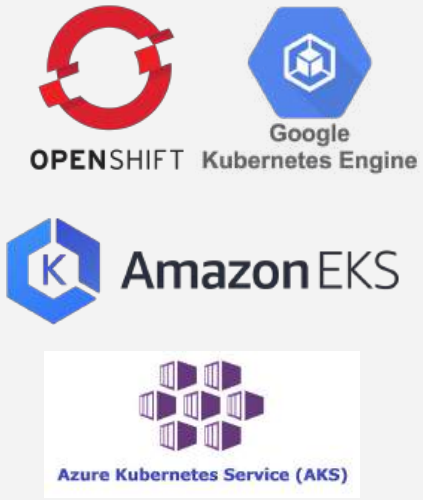
Proprietary

+ Low implementation time

+- Can be more cost efficient

-- Least amount of flexibility
-- Least level of control

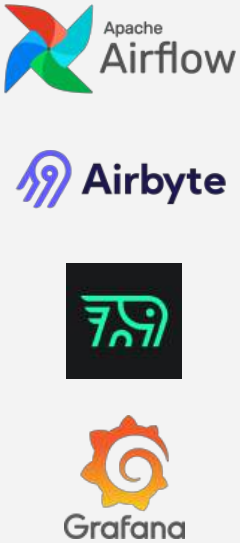
How to self-host?



Managed Kubernetes

Github repo

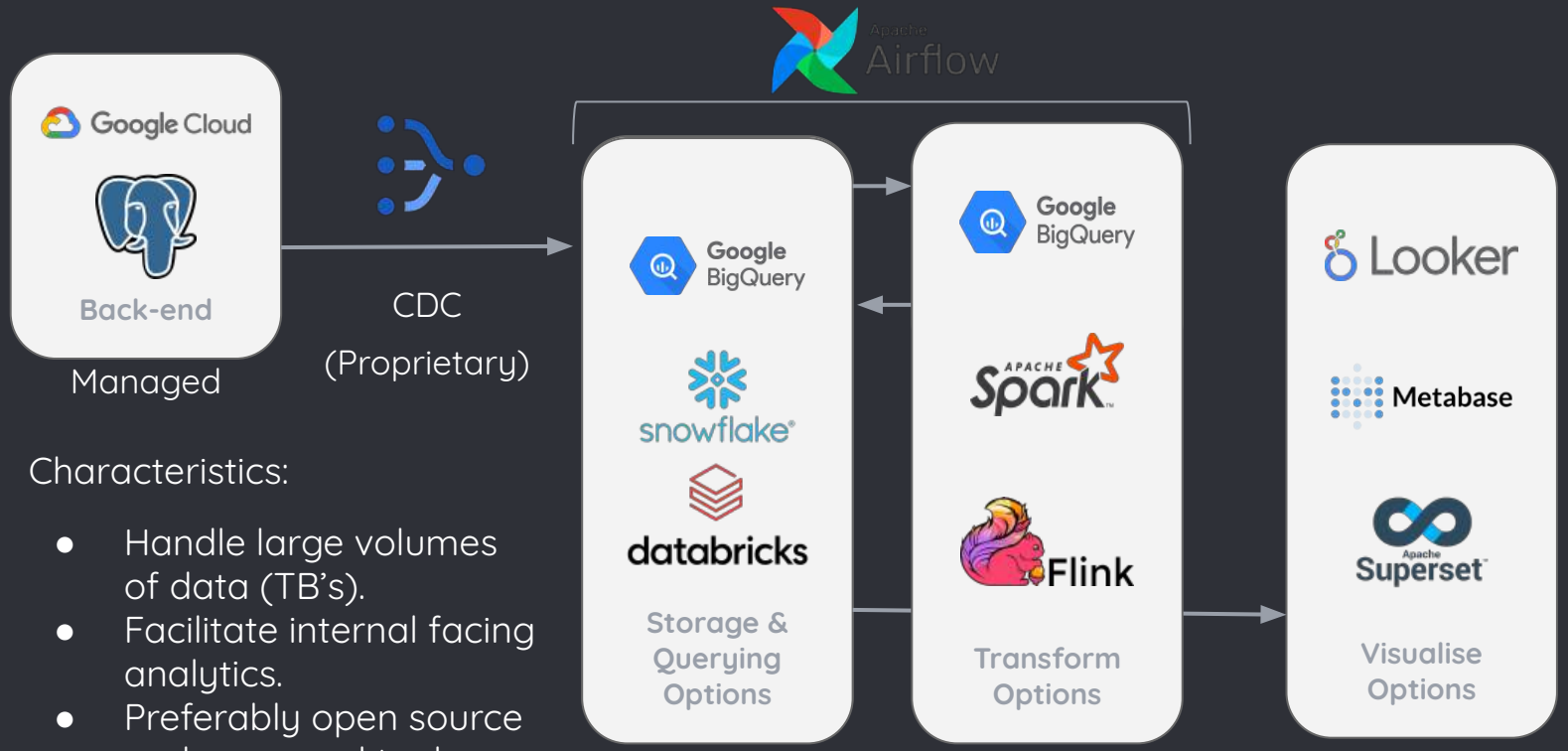
Set-up infrastructure



Deploy using helm charts

1. Pick a managed Kubernetes platform
2. Set-up cluster with auto-scaling
3. Use pre-configured helm charts and customise values

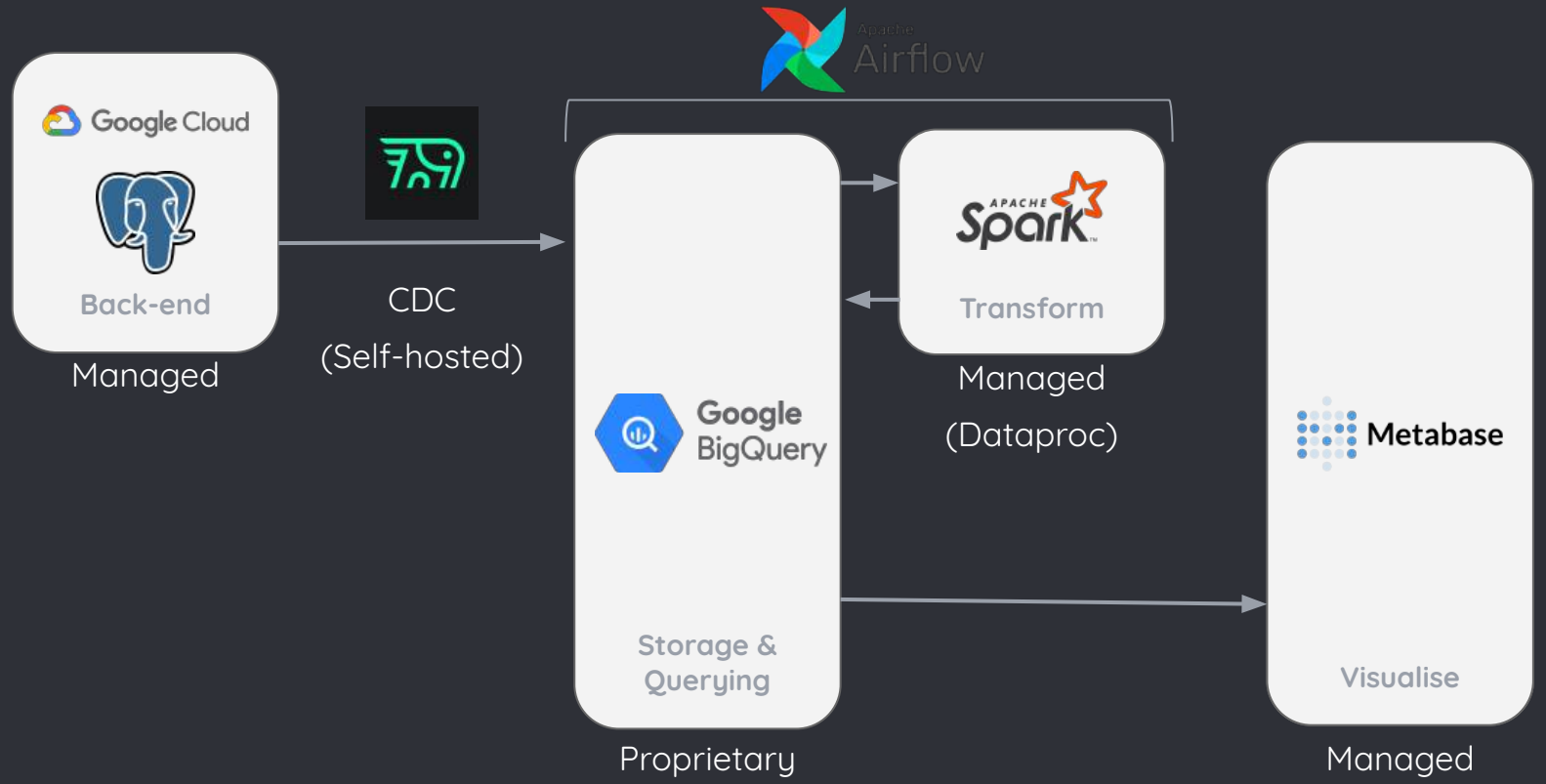
Building the data platform (starting options)



Characteristics:

- Handle large volumes of data (TB's).
- Facilitate internal facing analytics.
- Preferably open source and managed tools.
- Keep costs under control.

Building the data platform (conclusion)



Medium blogpost

Value-add data products

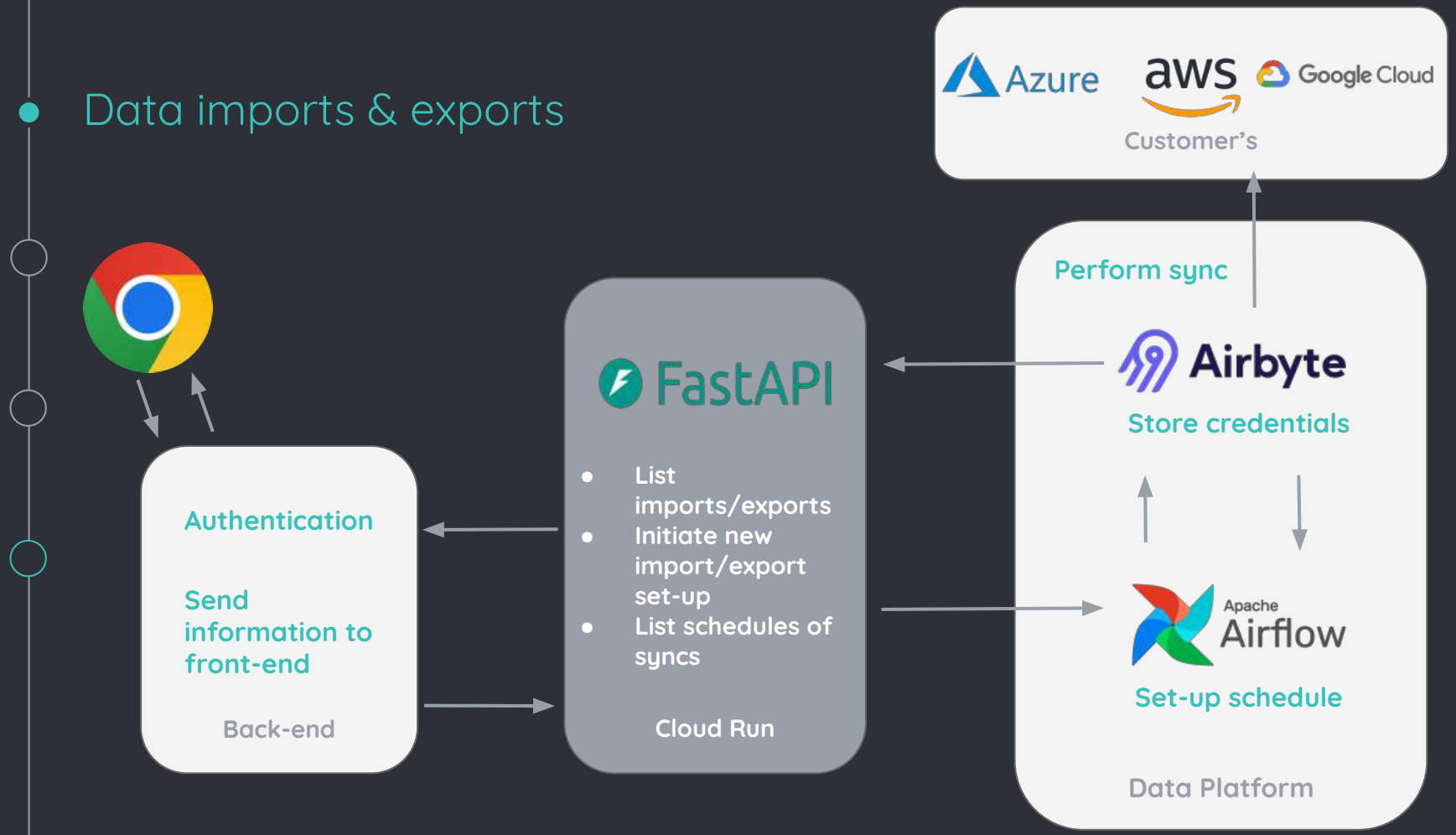
Embedded analytics

Metabase

Embeddable

Medium blogpost

Data imports & exports



Summary

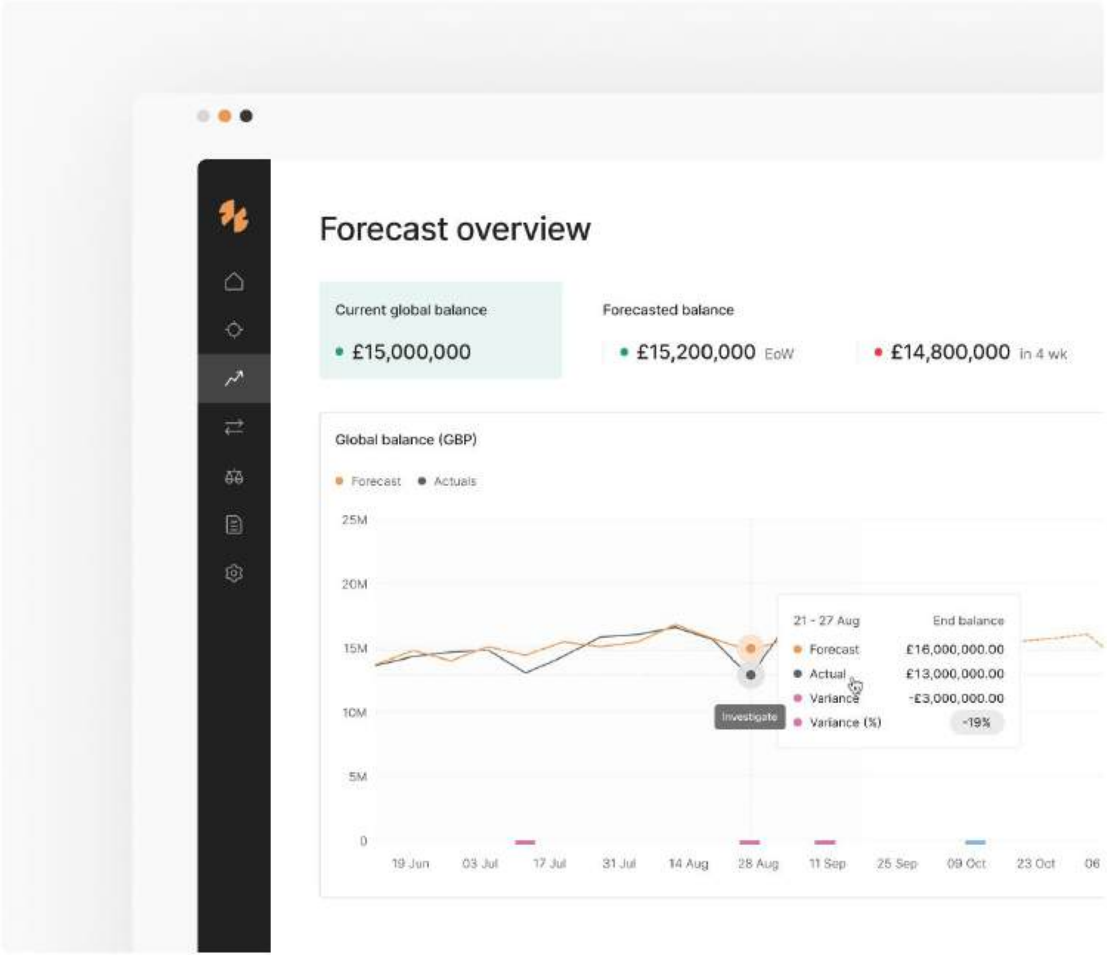
Start early, don’t wait	Use what you are familiar with	Design with value-add products in mind

I am Joining

Palm

Rodel

Data & AI Lead



We are Hiring

Back-end Engineers

ML Scientists

usepalm.com