SUMMIT

BERLIN

SessionID

The Scout24 Data Platform
A Technical Deep Dive

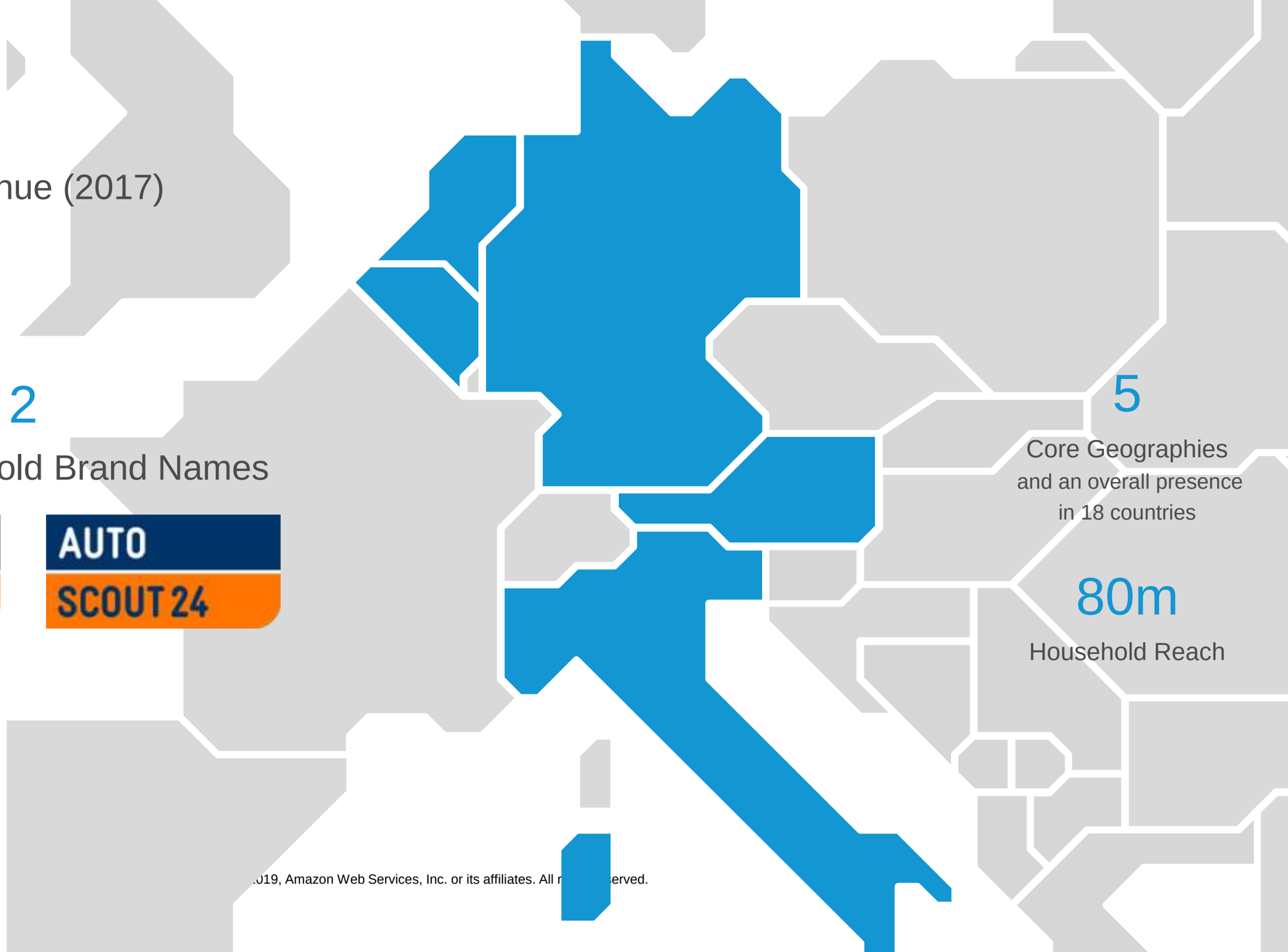
Sean Gustafson
Senior Technical Product Manager
Scout24

Raffael Dzikowski
Senior Data Engineer
Scout24

aws SUMMIT

Scout24 AG

- SDAX
- € 489 million revenue (2017)
- ~1500 employees



2

Major Household Brand Names

IMMOBILIEN SCOUT 24

AUTO SCOUT 24

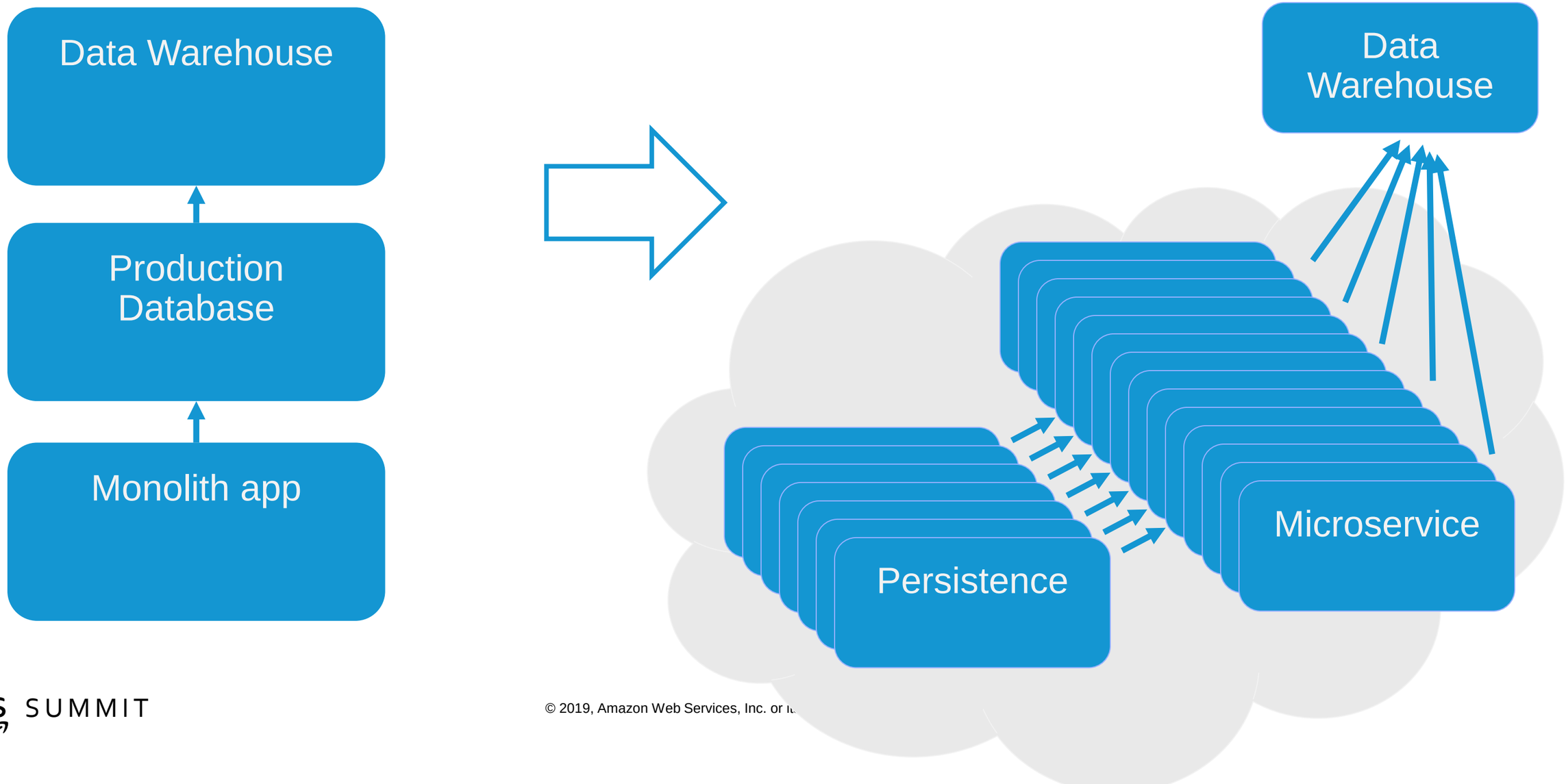
5

Core Geographies
and an overall presence
in 18 countries

80m

Household Reach

Our technical evolution



Our data warehouse was a bottle neck

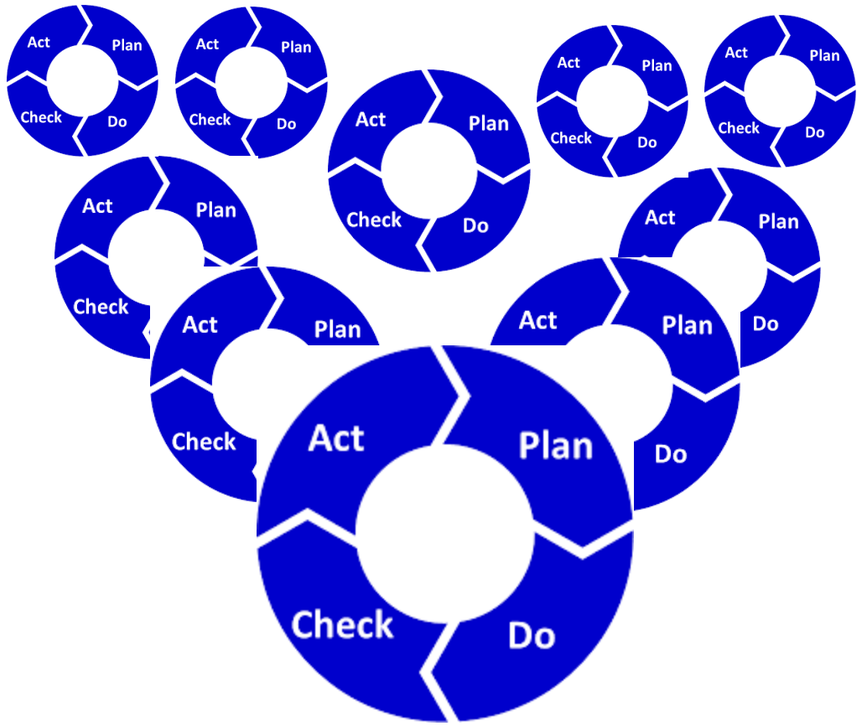
Scout24 wants to become a truly data-driven company

Fast & easy data-driven product development…

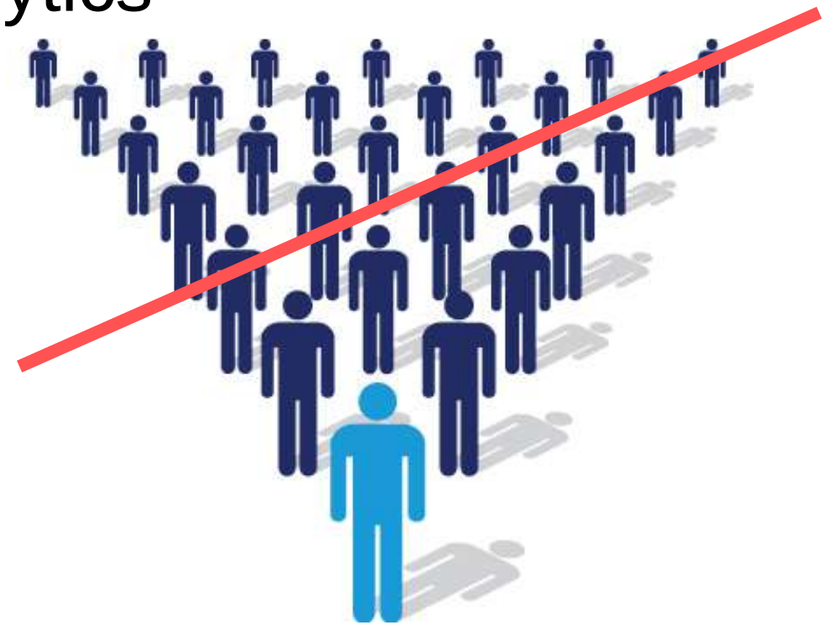
…supported by Data & Analytics

Scout24 wants to become a truly data-driven company

Everywhere in the company...



...without bloating up Data & Analytics



Our solution:

Build an internal “platform” for data

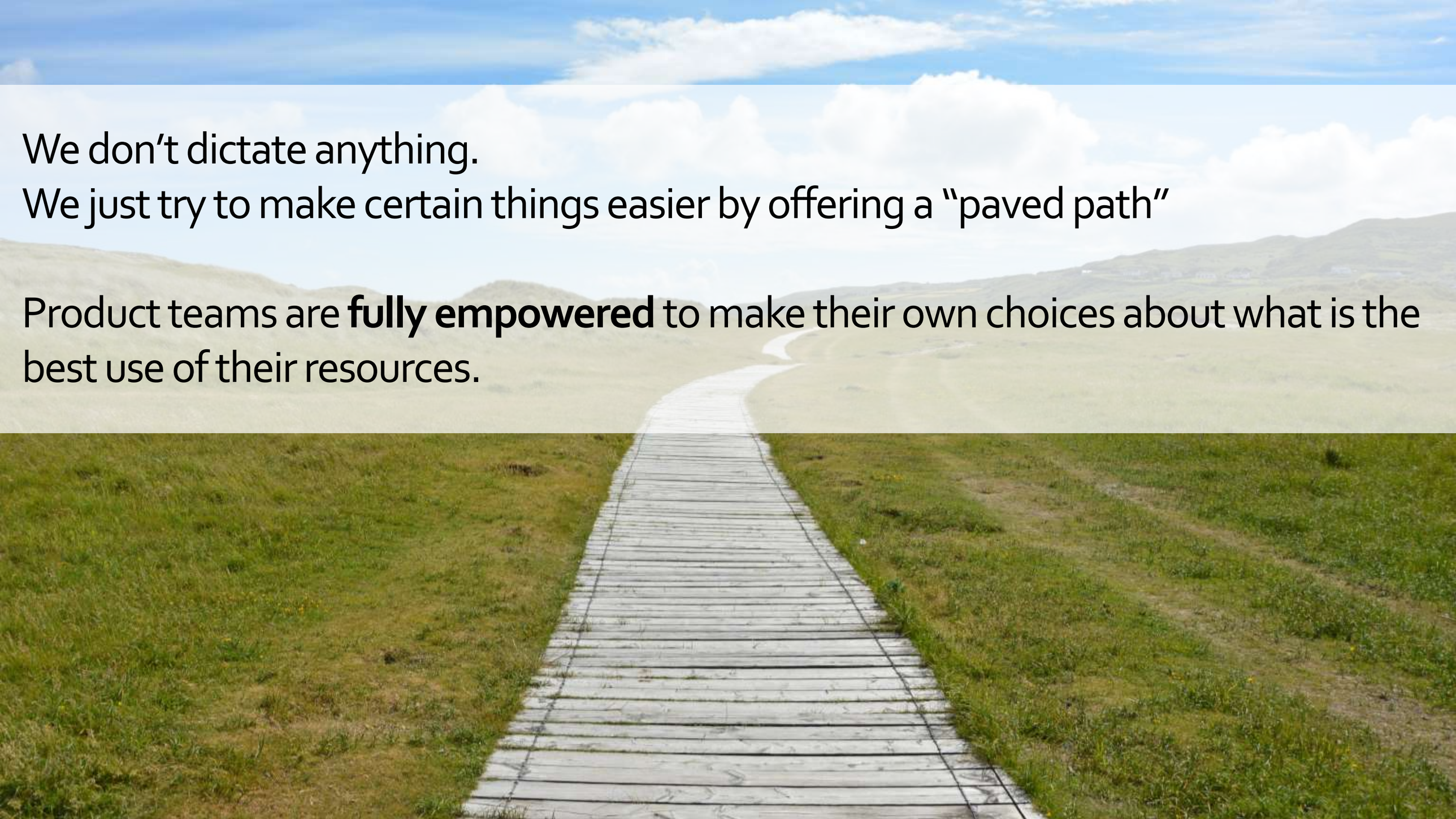
What is the Scout24 Data “Platform”?

We think of our Data Platform as a **Product**

Just like AWS, Salesforce, etc. – the platform is a **generic layer** upon which Scout24's products can be built

BUT, we have a very, very small number of customers.

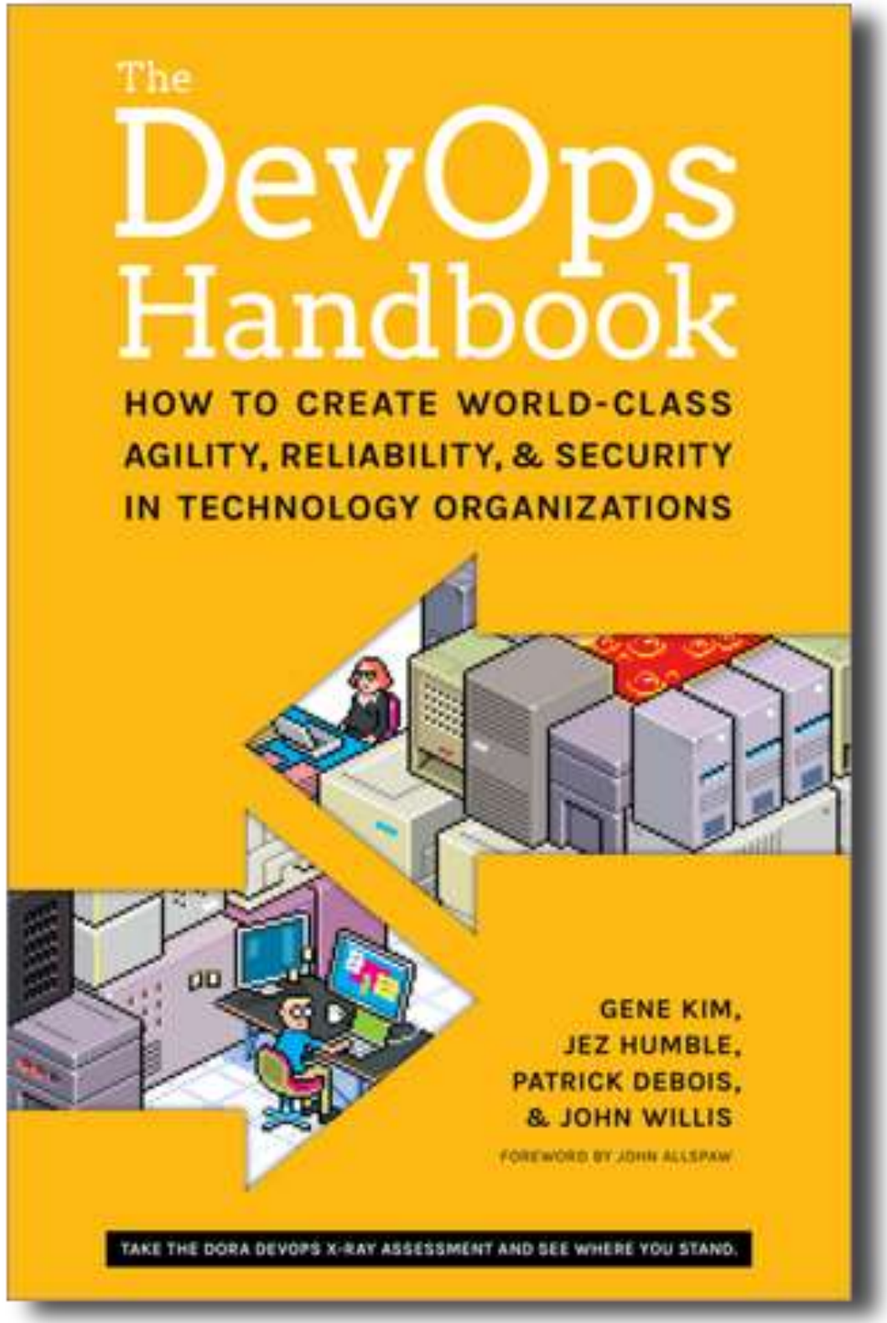
That means, product teams get **personalized support** and there is lots of **opportunity for collaboration.**



We don’t dictate anything.
We just try to make certain things easier by offering a “paved path”

Product teams are **fully empowered** to make their own choices about what is the best use of their resources.

"In almost all cases, we will not mandate that internal team use these platforms and services—**these platform teams will to win over and satisfy their internal customers**, even competing with external vendors."



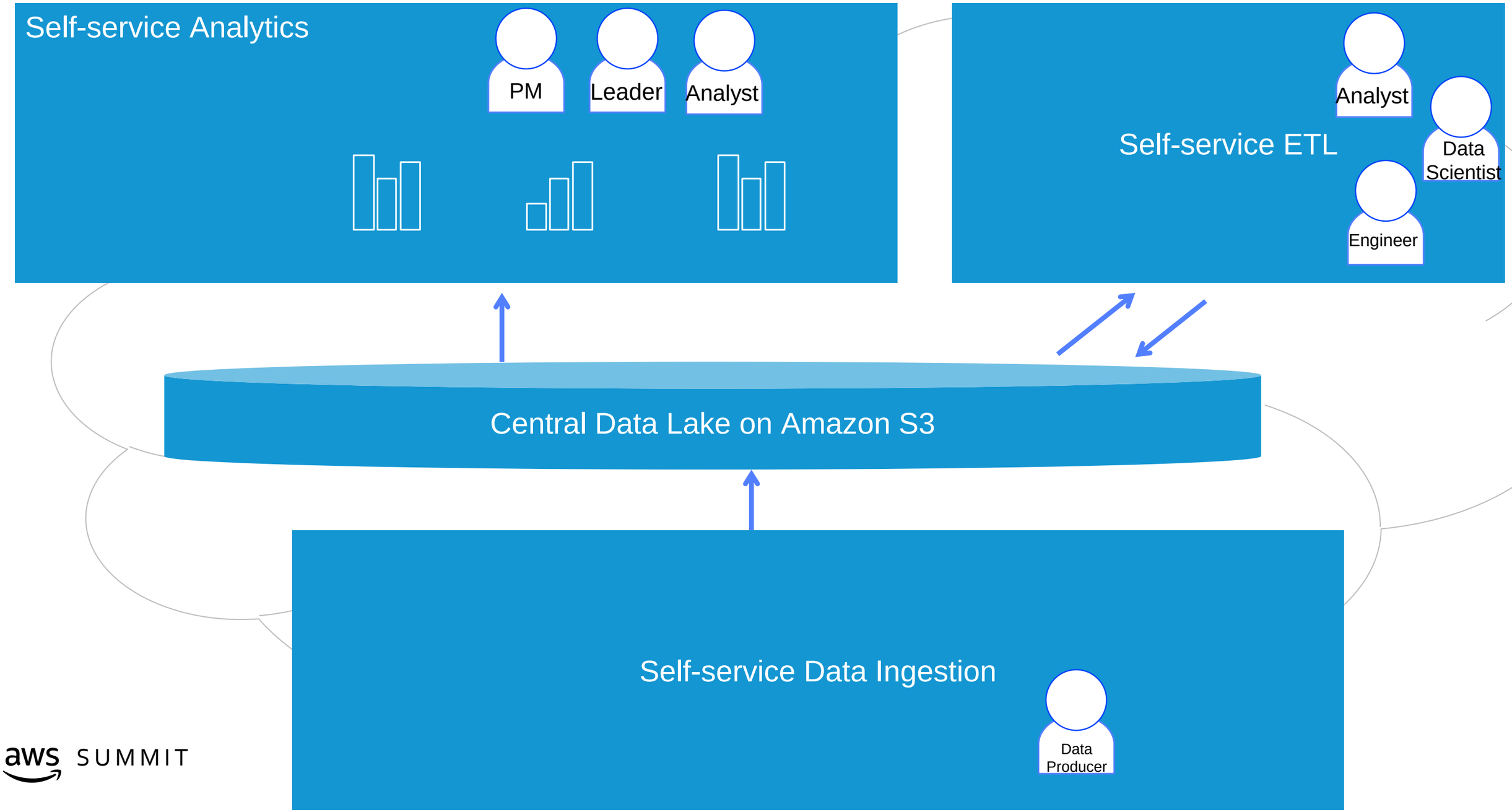
Guiding principle of the platform

Autonomy for producers and consumers

Self-service Analytics

Self-service Data Ingestion

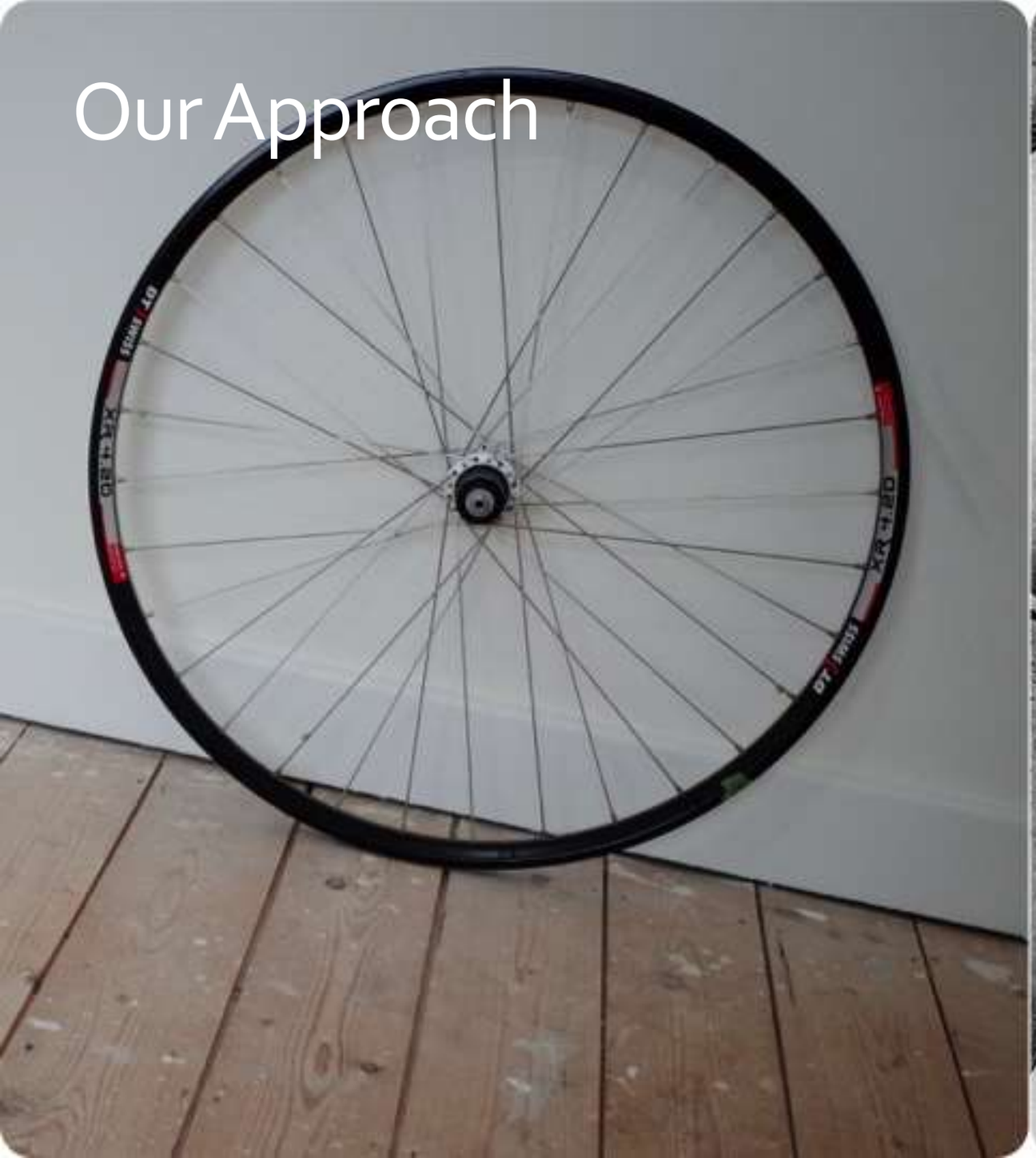
Self-service ETL


Self-service Analytics
PM
Leader
Analyst
Self-service ETL
Analyst
Data Scientist
Engineer
Central Data Lake on Amazon S3
Self-service Data Ingestion
Data Producer

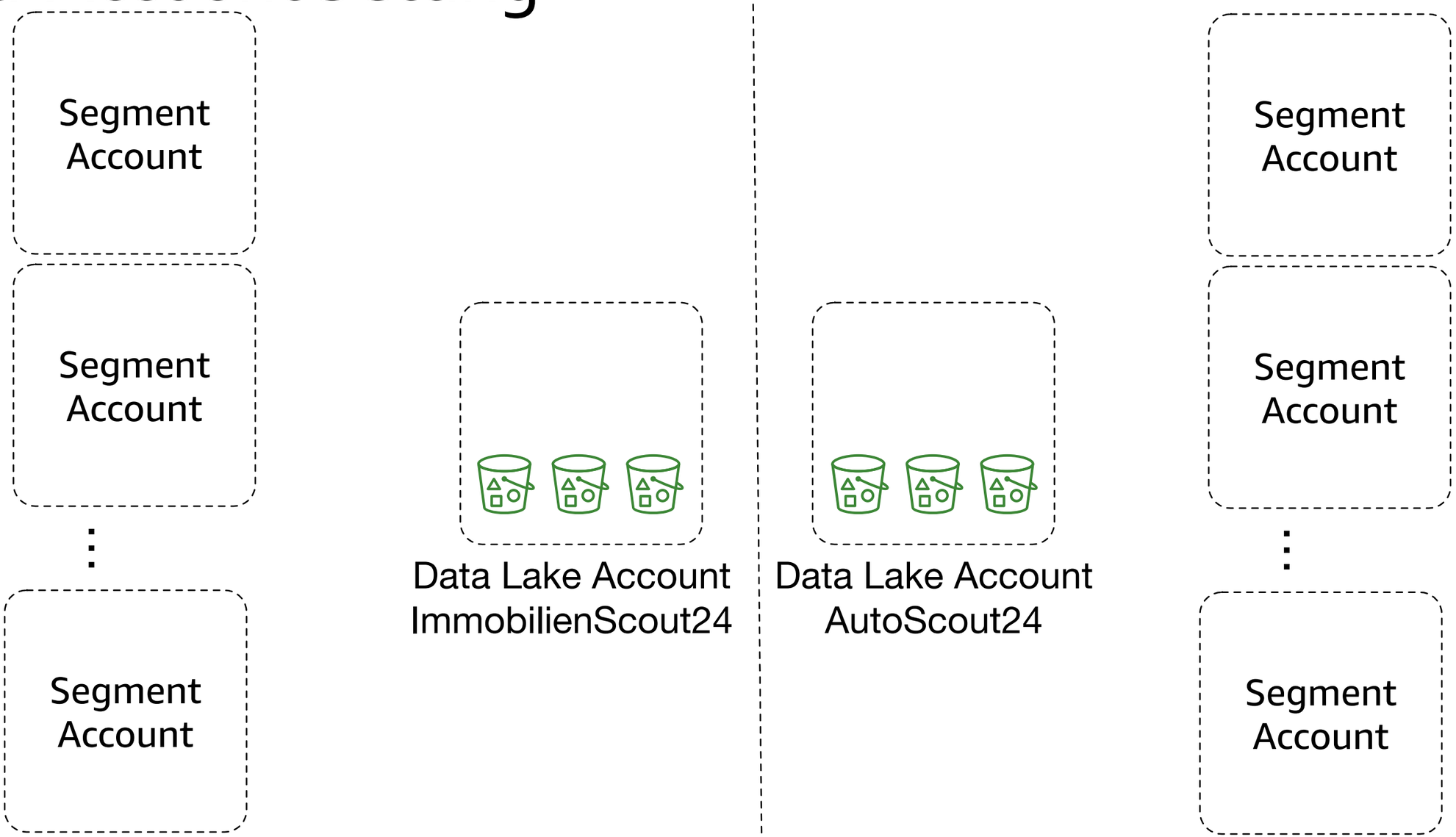
Self-Service Ingestion

aws SUMMIT

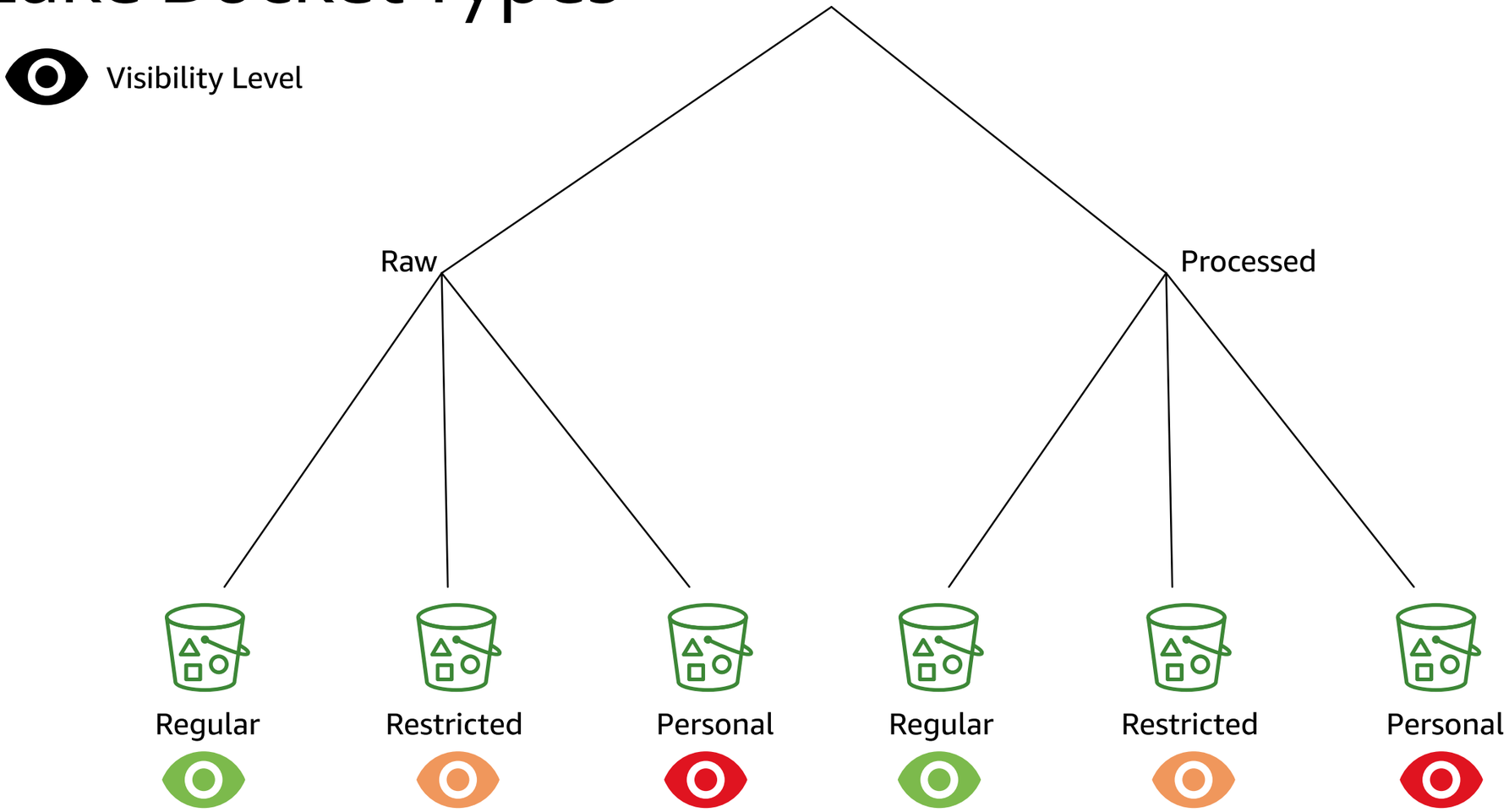
Our Approach



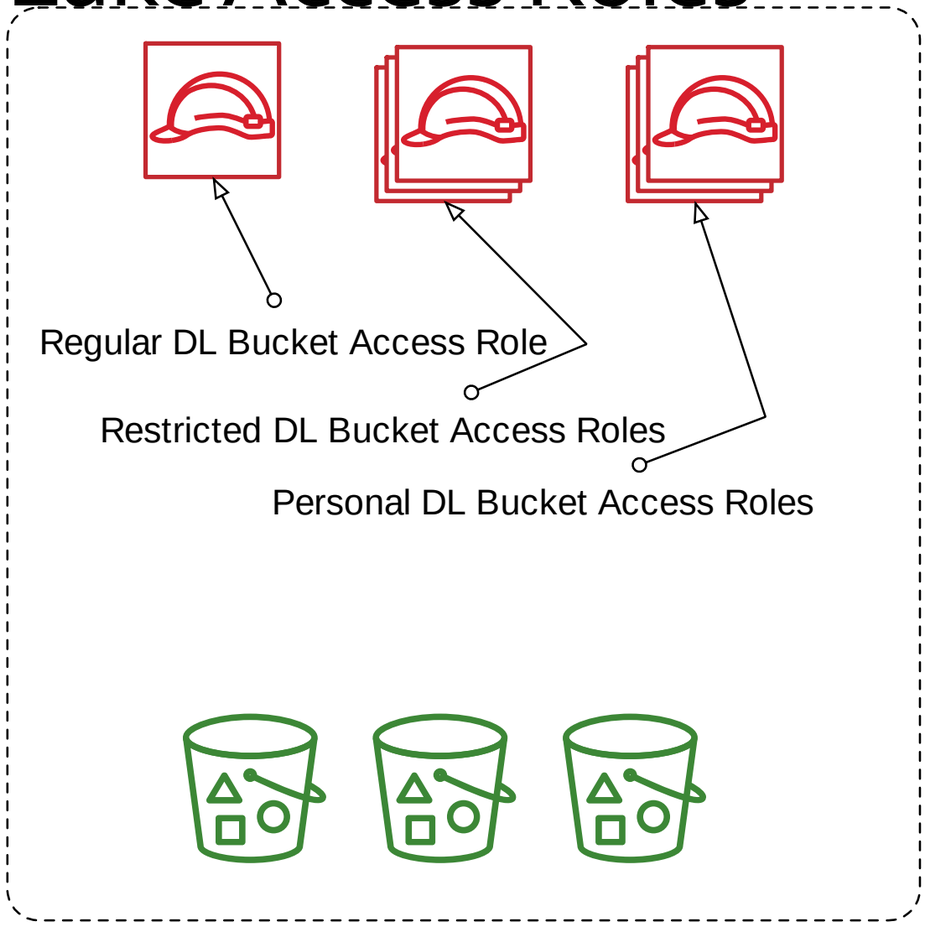
Multi-Account Setting



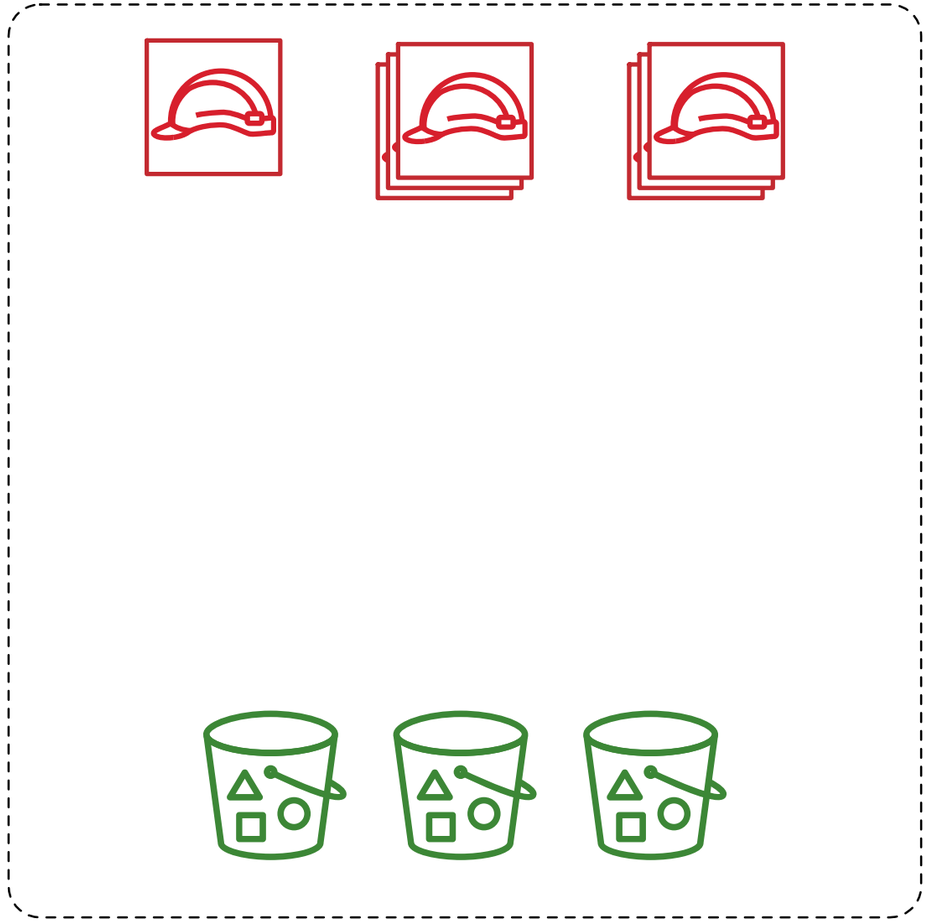
Data Lake Bucket Types



Data Lake Access Roles

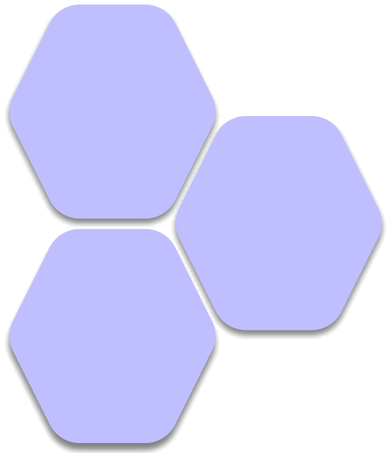


Data Lake Account
ImmobilienScout24

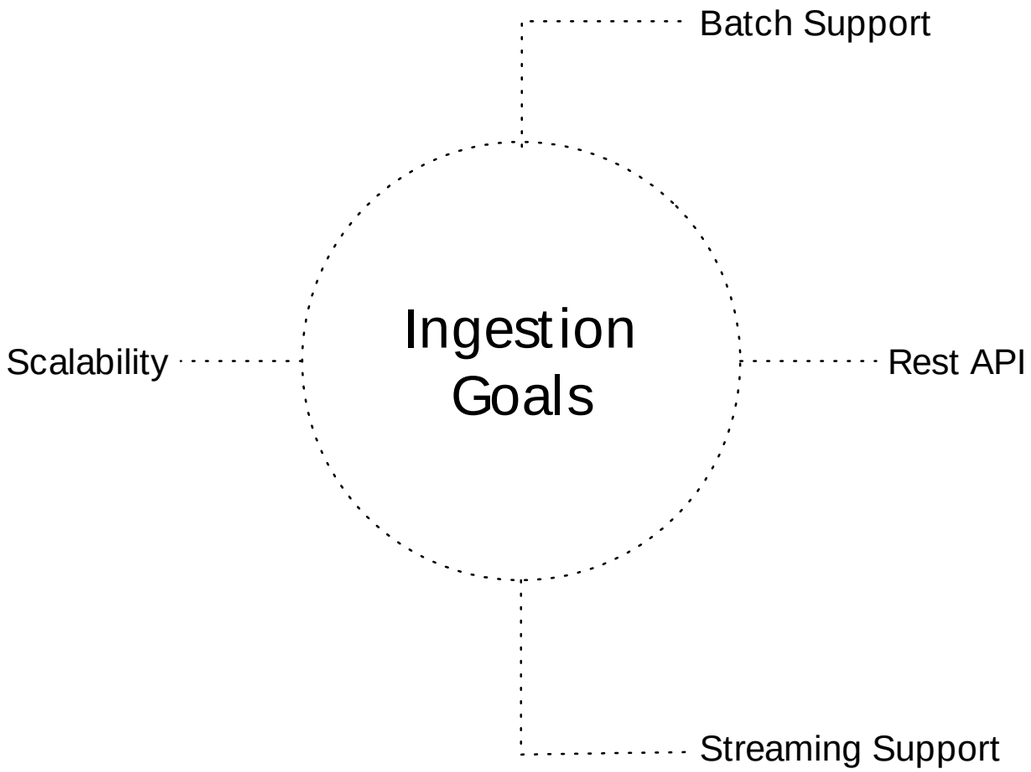


Data Lake Account
AutoScout24

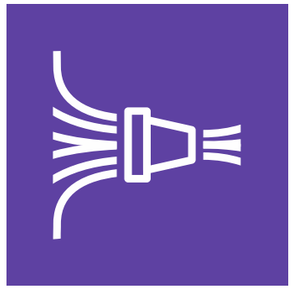
Ingestion Goals



Microservices



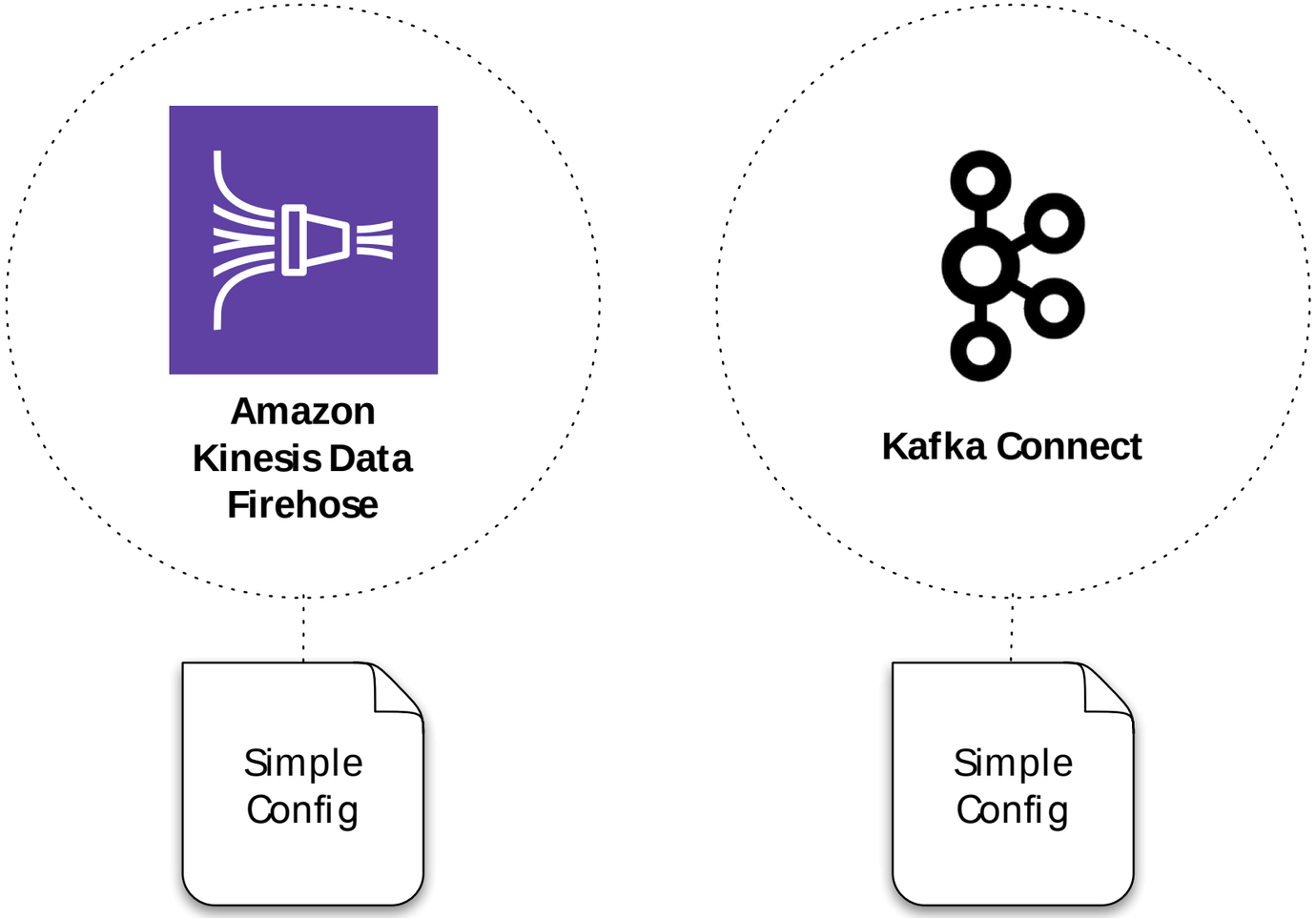
Ingestion Options



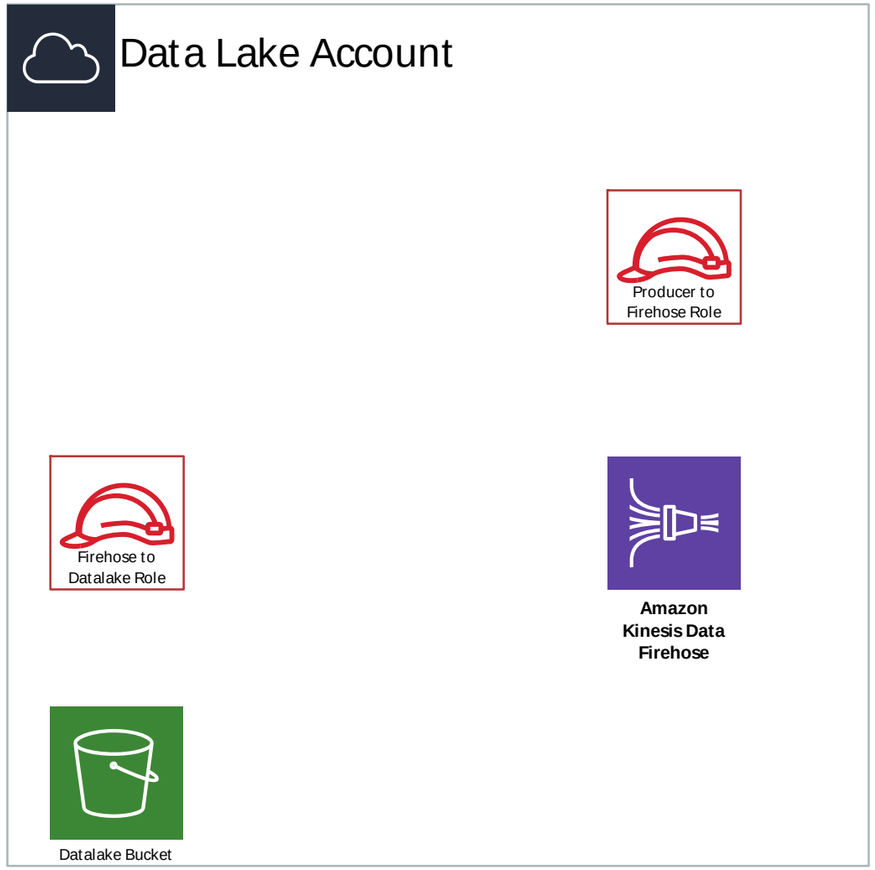
Amazon Kinesis Data Firehose

Kafka Connect

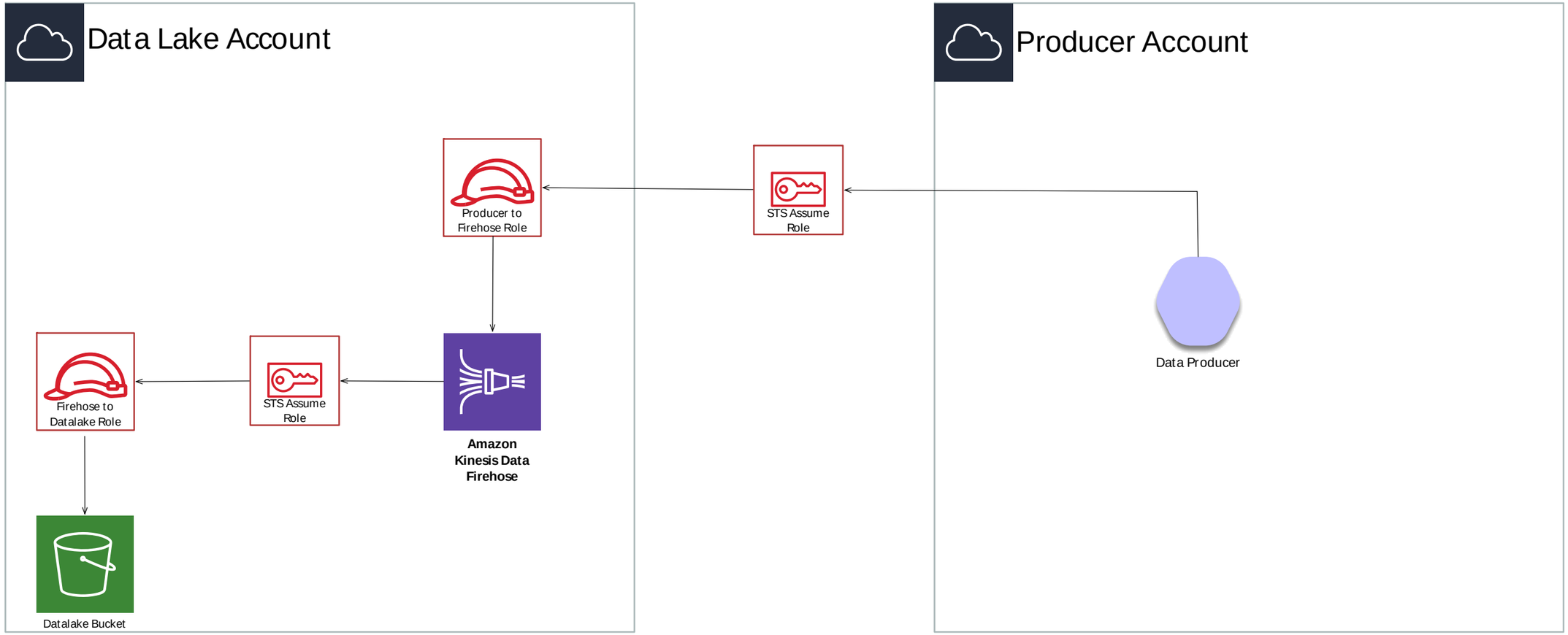
Ingestion Options



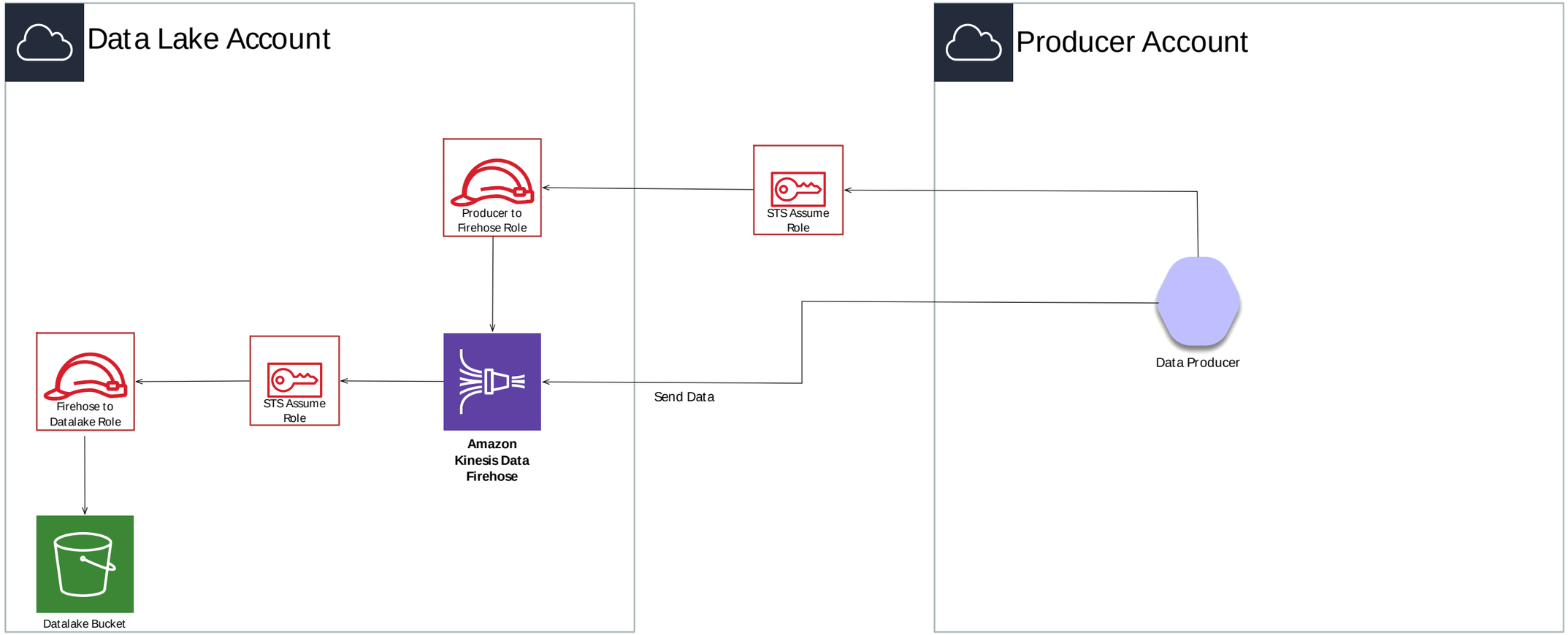
Firehose Ingestion Architecture



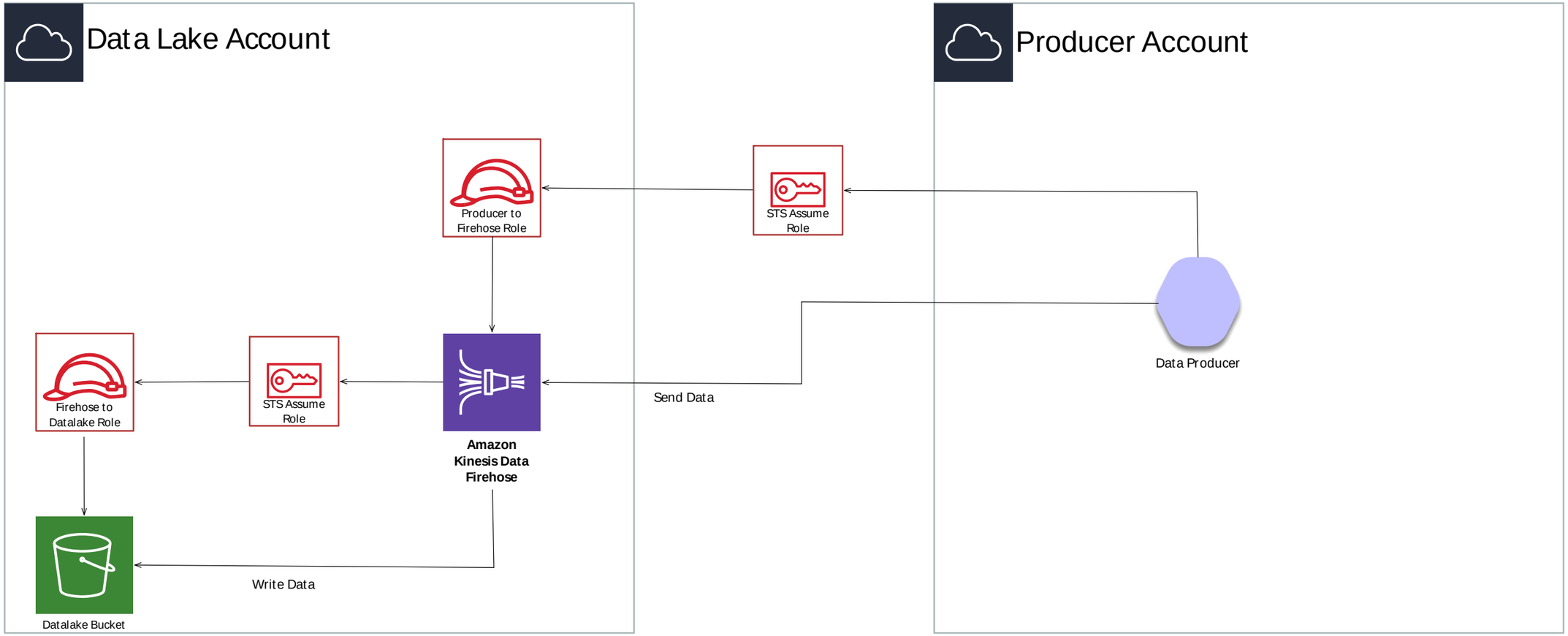
Firehose Ingestion Architecture



Firehose Ingestion Architecture

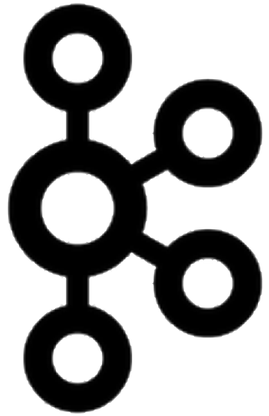


Firehose Ingestion Architecture



Kafka Connect

Kafka Connect



ActiveMQ

Kafka

Elasticsearch

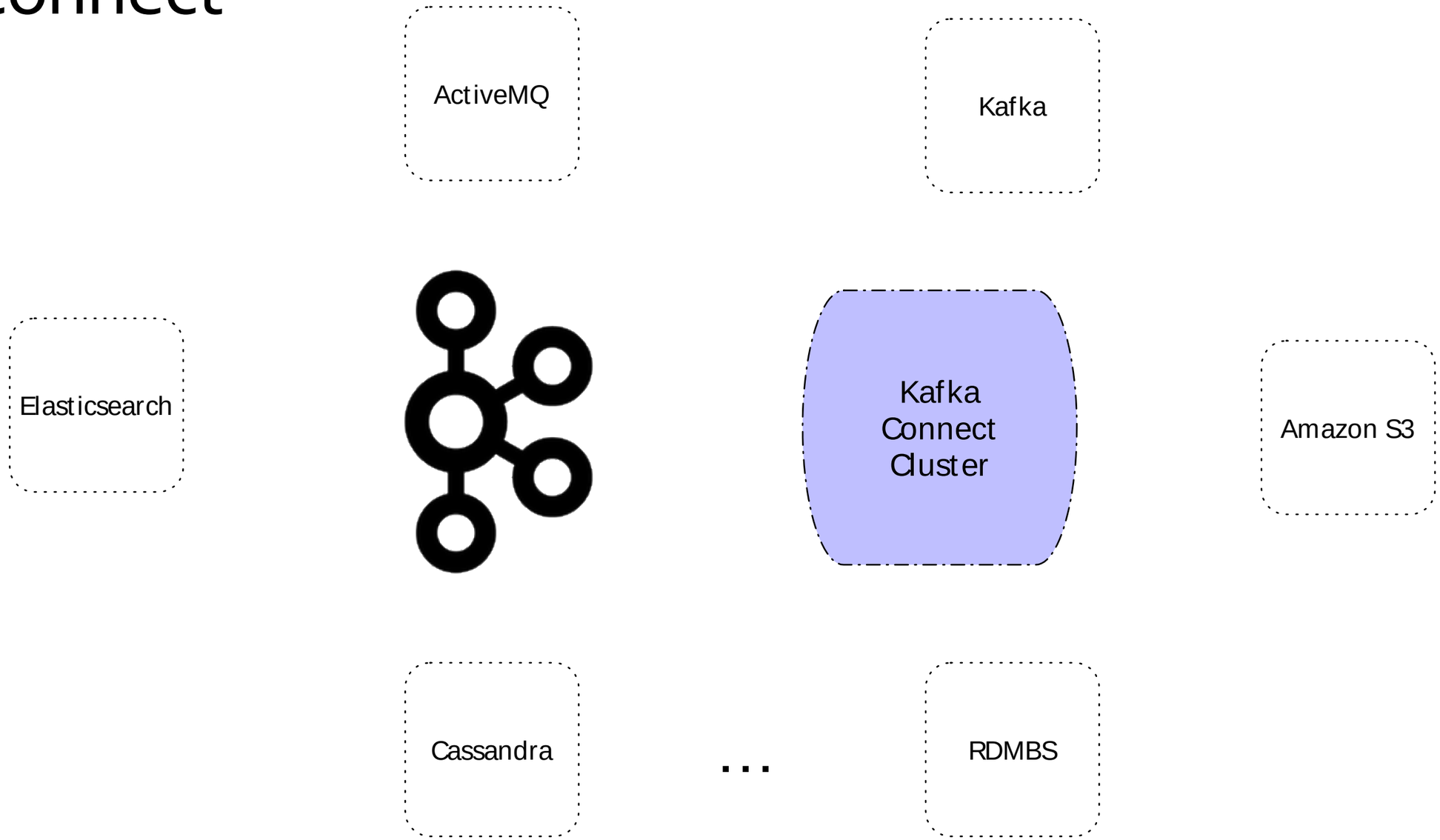
Amazon S3

Cassandra

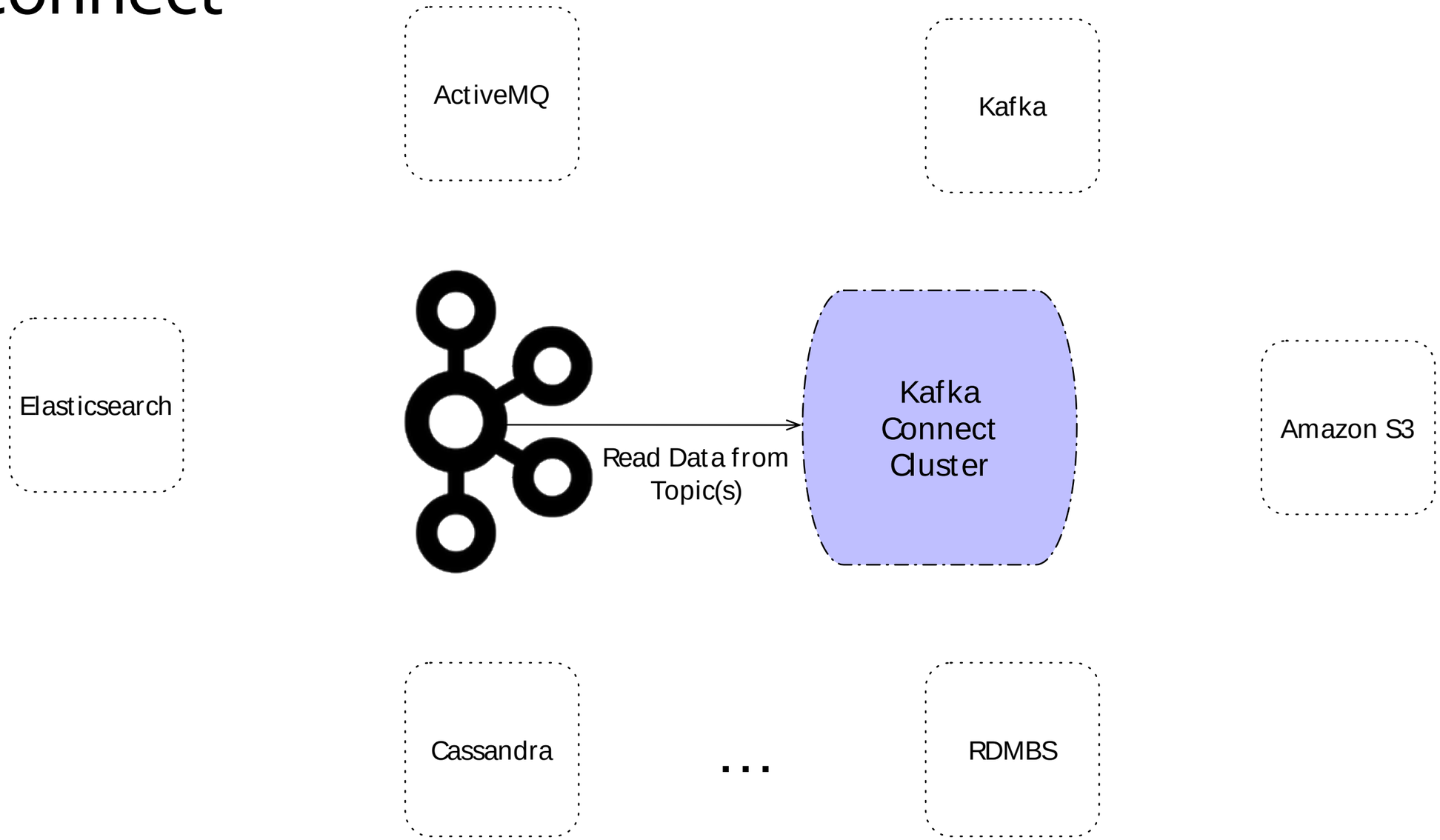
...

RDMBS

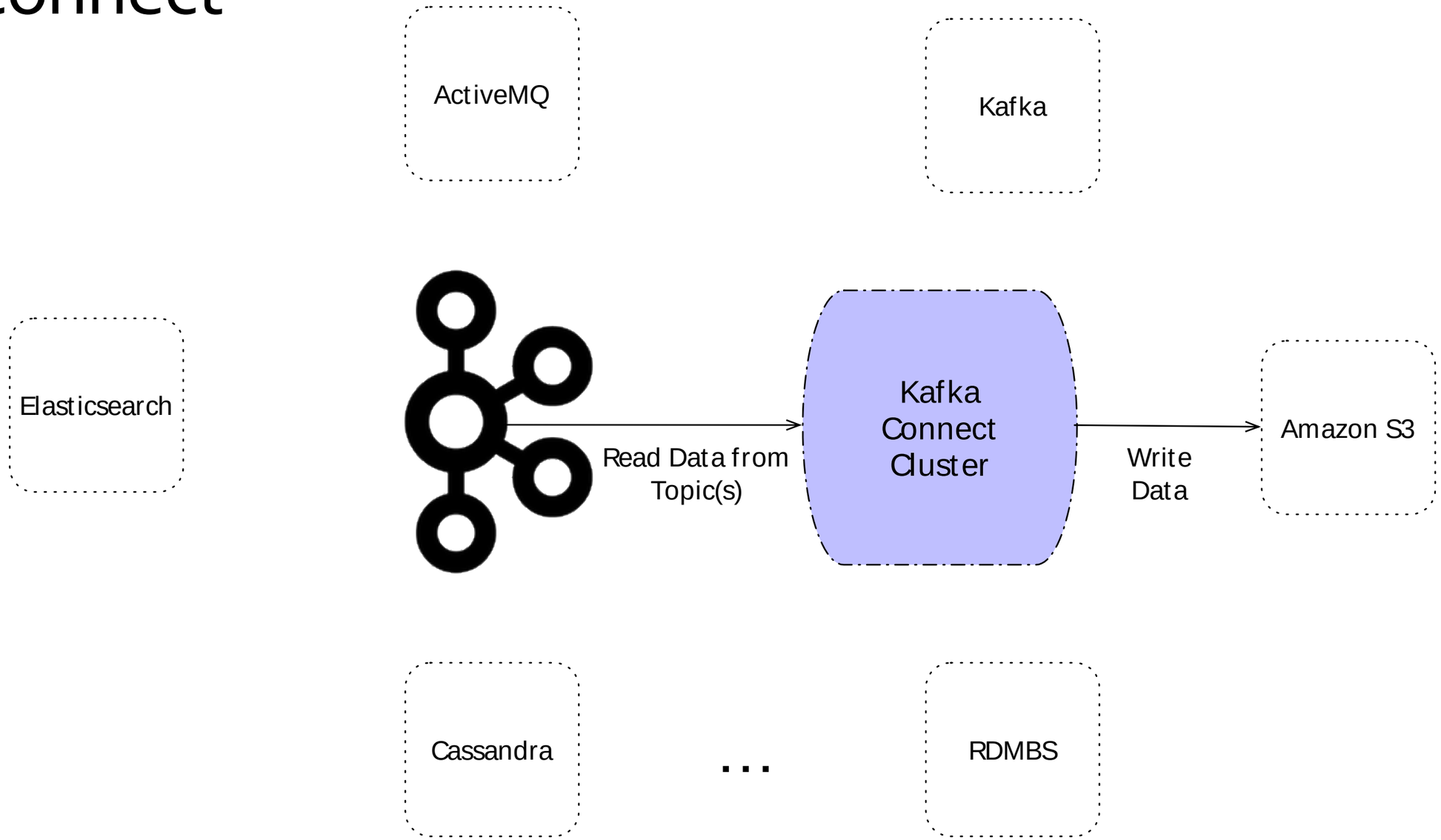
Kafka Connect



Kafka Connect



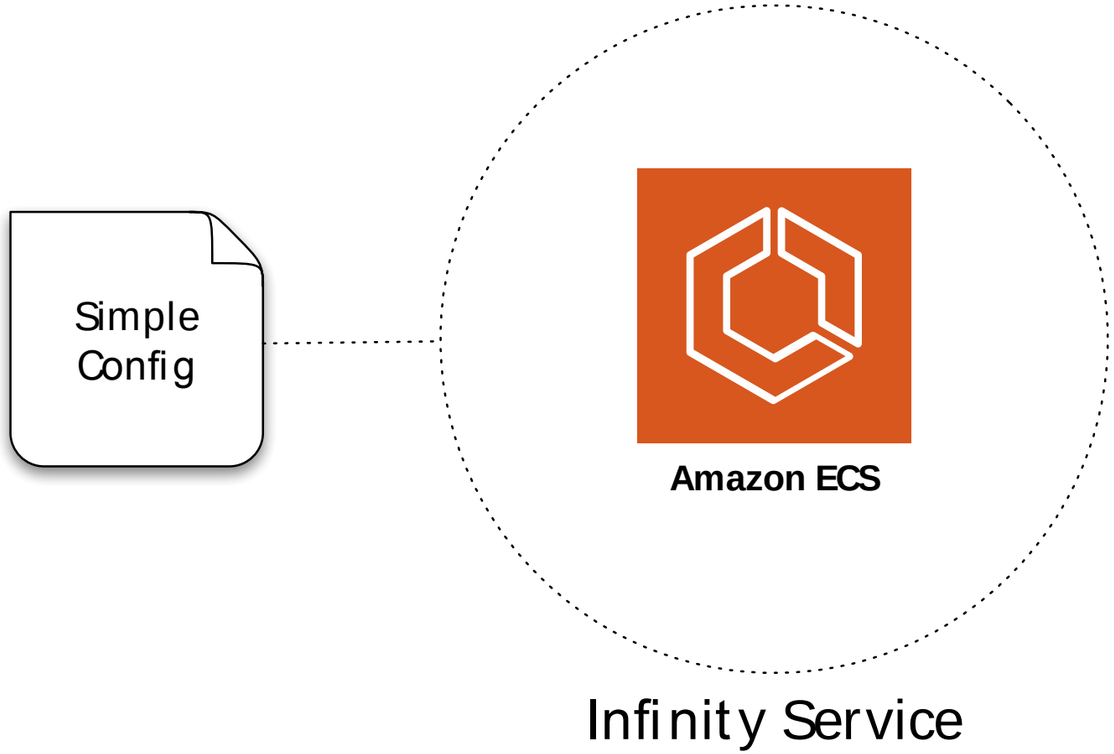
Kafka Connect



Scout24 Infinity Cluster

Infinity Service

Scout24 Infinity Cluster



Scout24 Infinity Cluster

Related Breakouts

15:00 in Hall 1

To Infinity and Beyond – Handling Heterogeneous Container Clusters in AWS

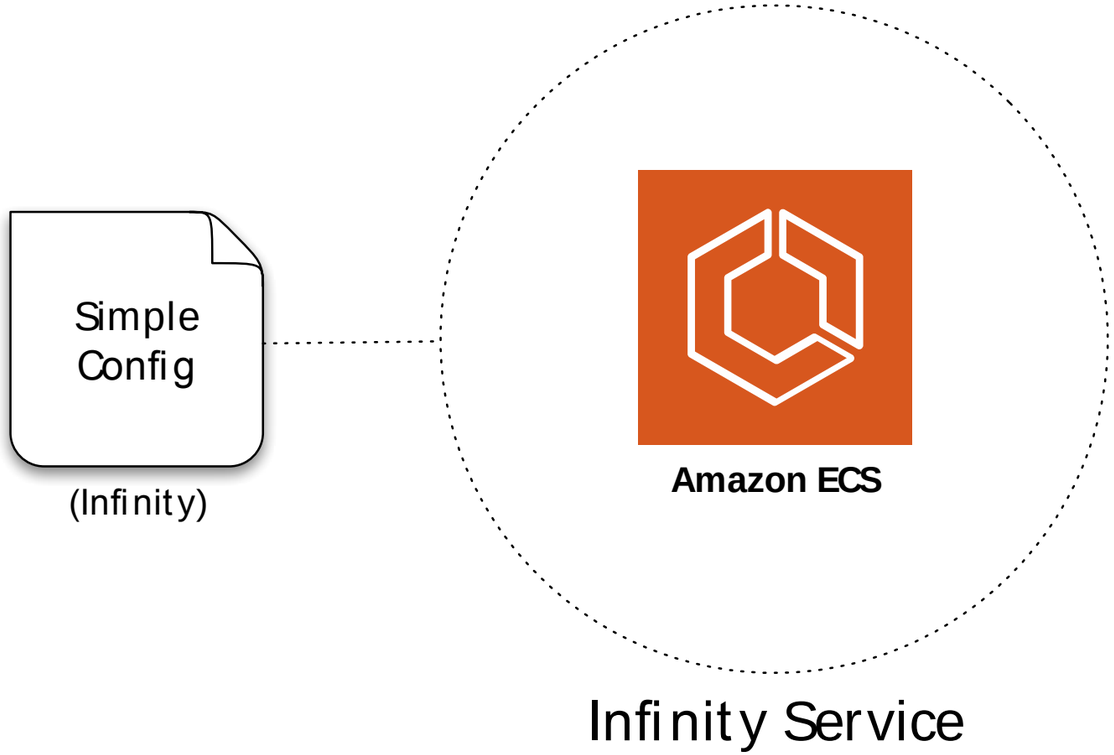
Christine Trahe, Platform Engineer @ Scout24

16:00 in Hall 1

Boost your AWS Infrastructure

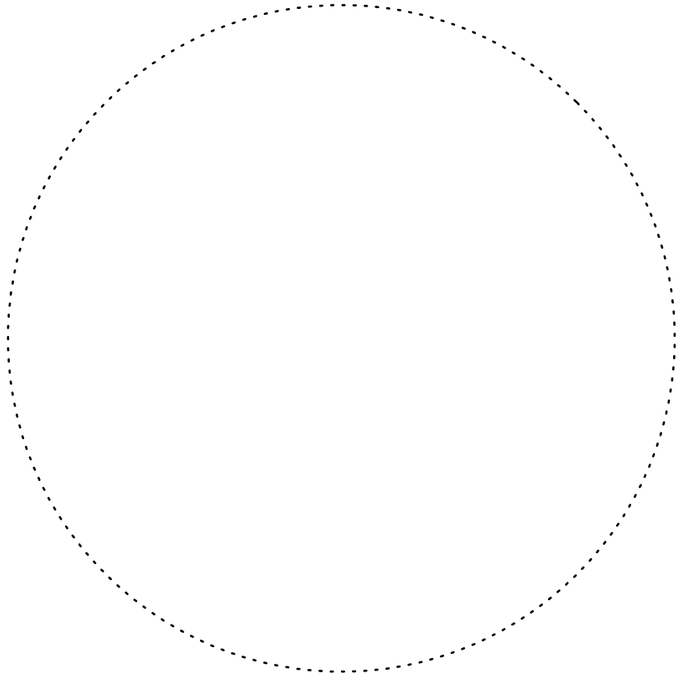
Philipp Garbe, AWS Container Hero @ Scout24

Kafka Connect on Infinity

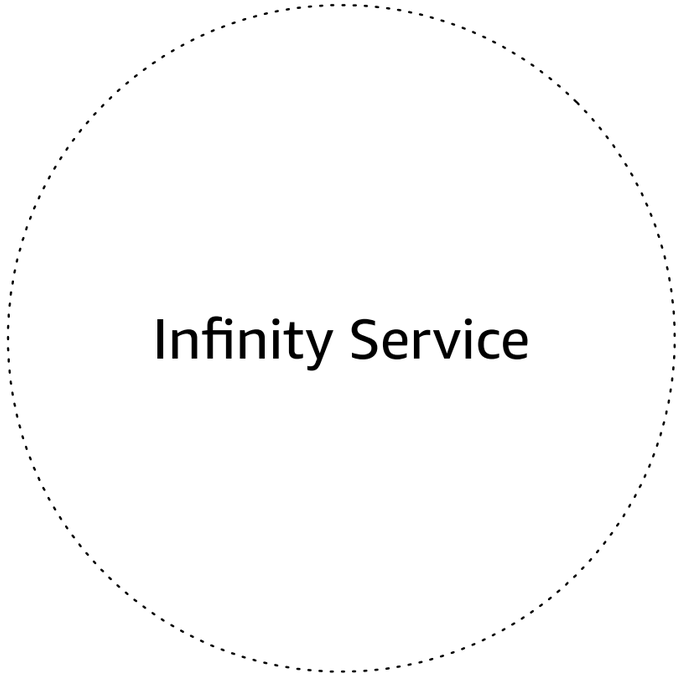


Kafka Connect on Infinity

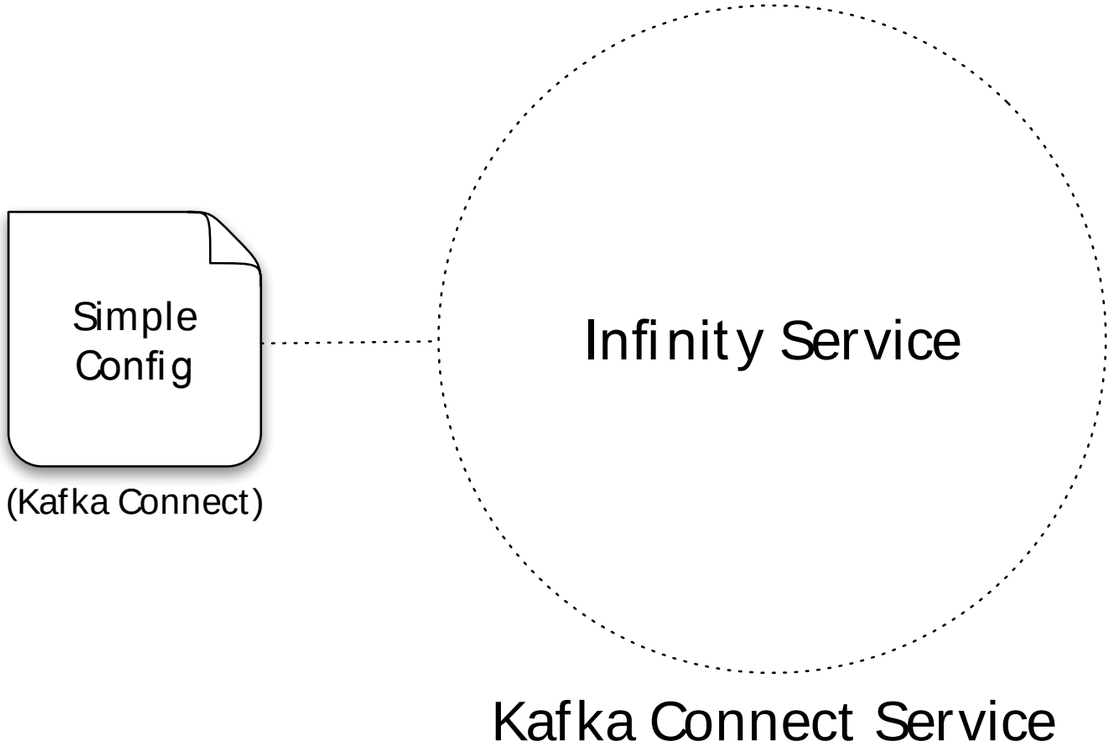
Kafka Connect on Infinity



Kafka Connect on Infinity



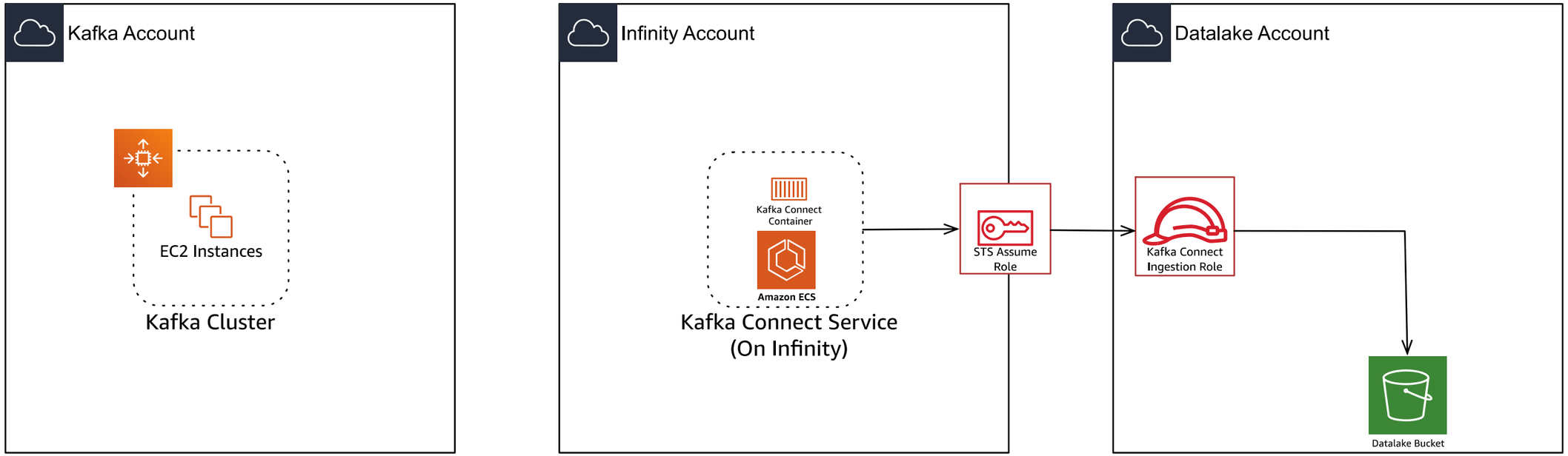
Kafka Connect on Infinity



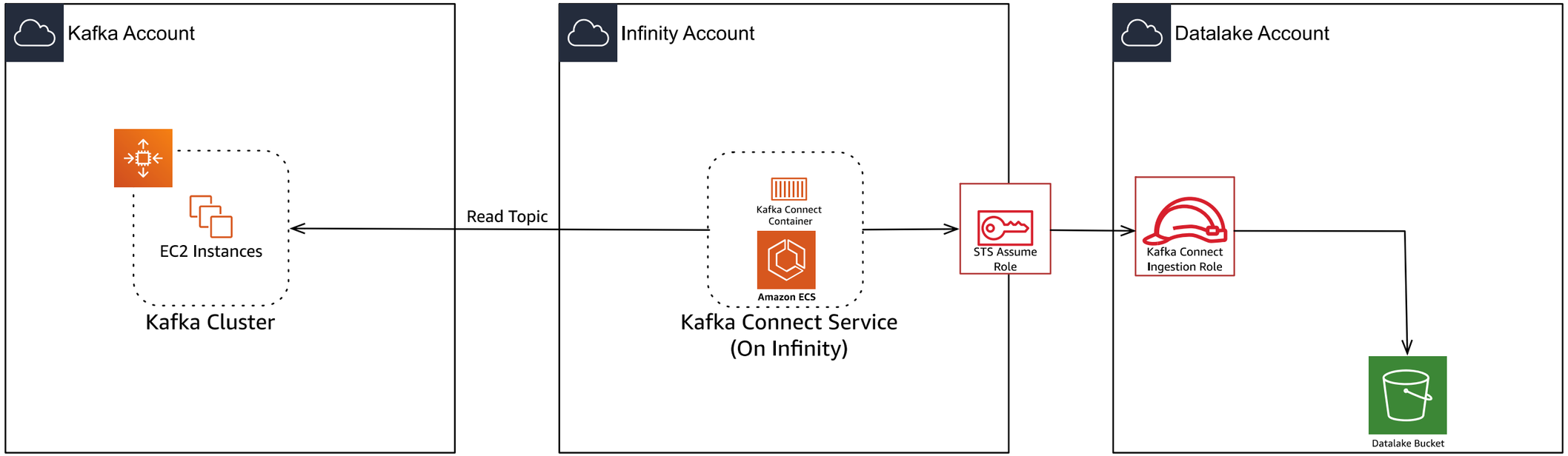
Kafka Connect Deployment

Kafka Connect Deployment

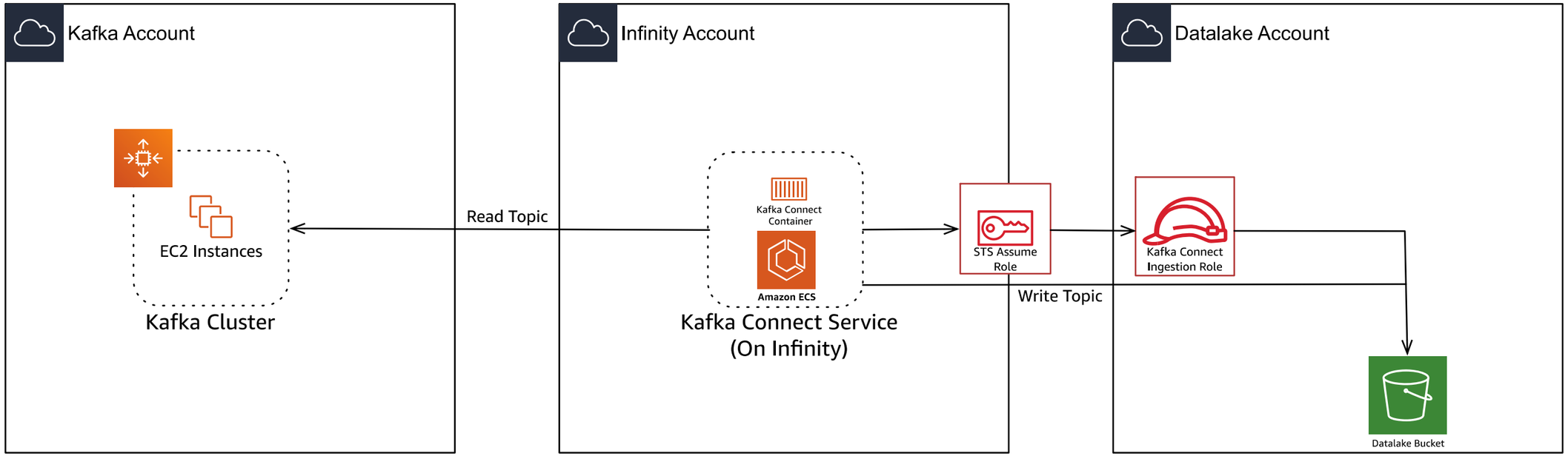
Kafka Connect Deployment



Kafka Connect Deployment



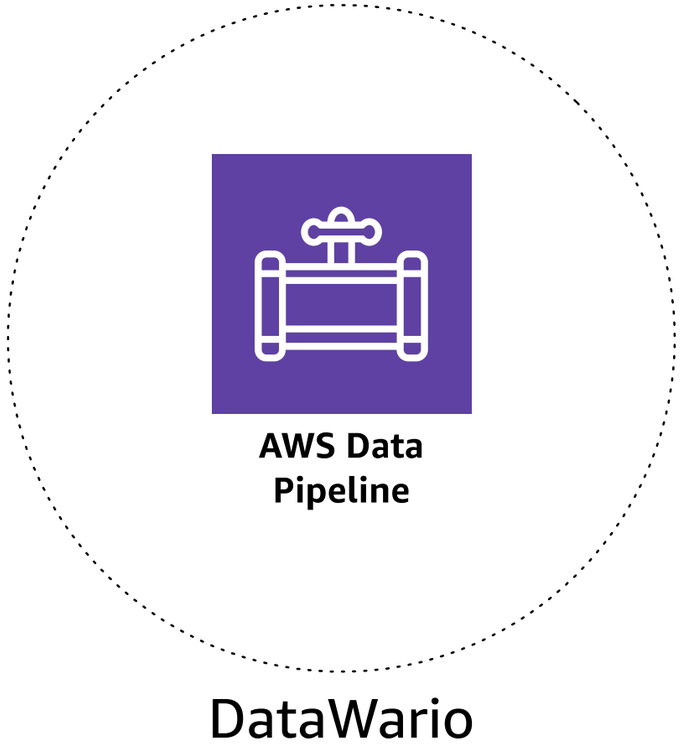
Kafka Connect Deployment



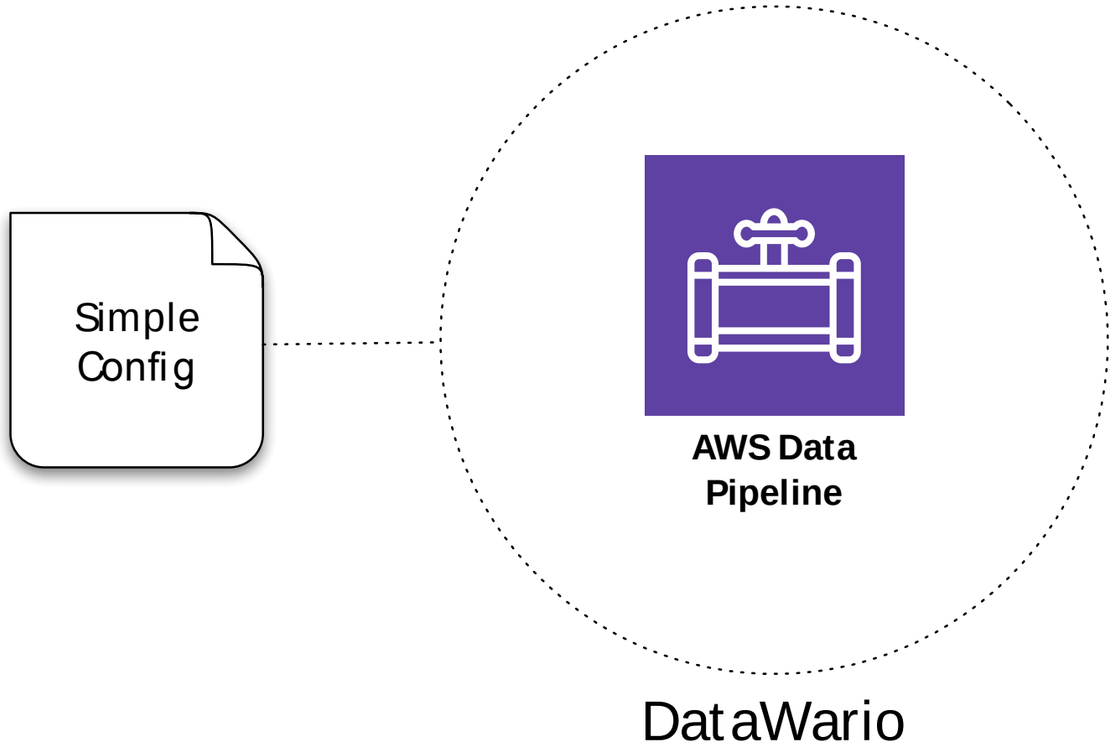
Self-Service ETL

aws SUMMIT

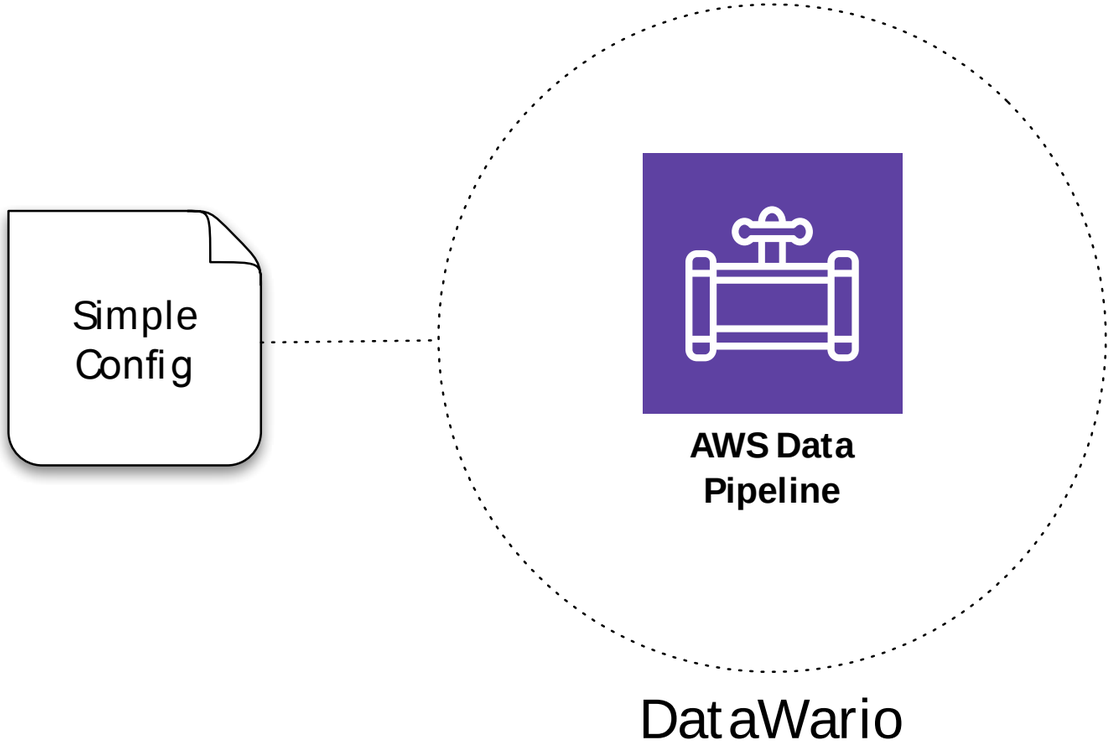
DataWario – Our Wrapper for AWS Data Pipeline



DataWario – Our Wrapper for AWS Data Pipeline

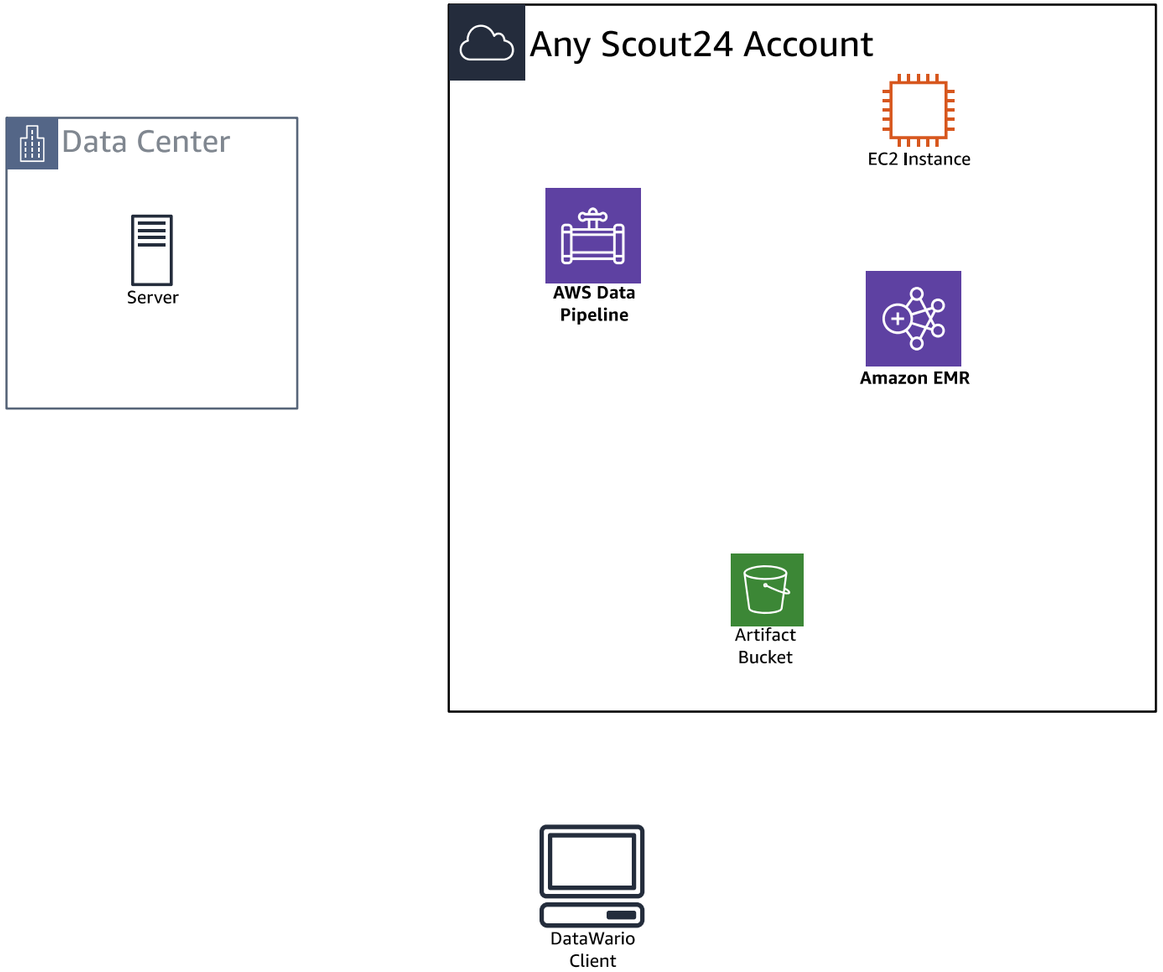


DataWario – Our Wrapper for AWS Data Pipeline

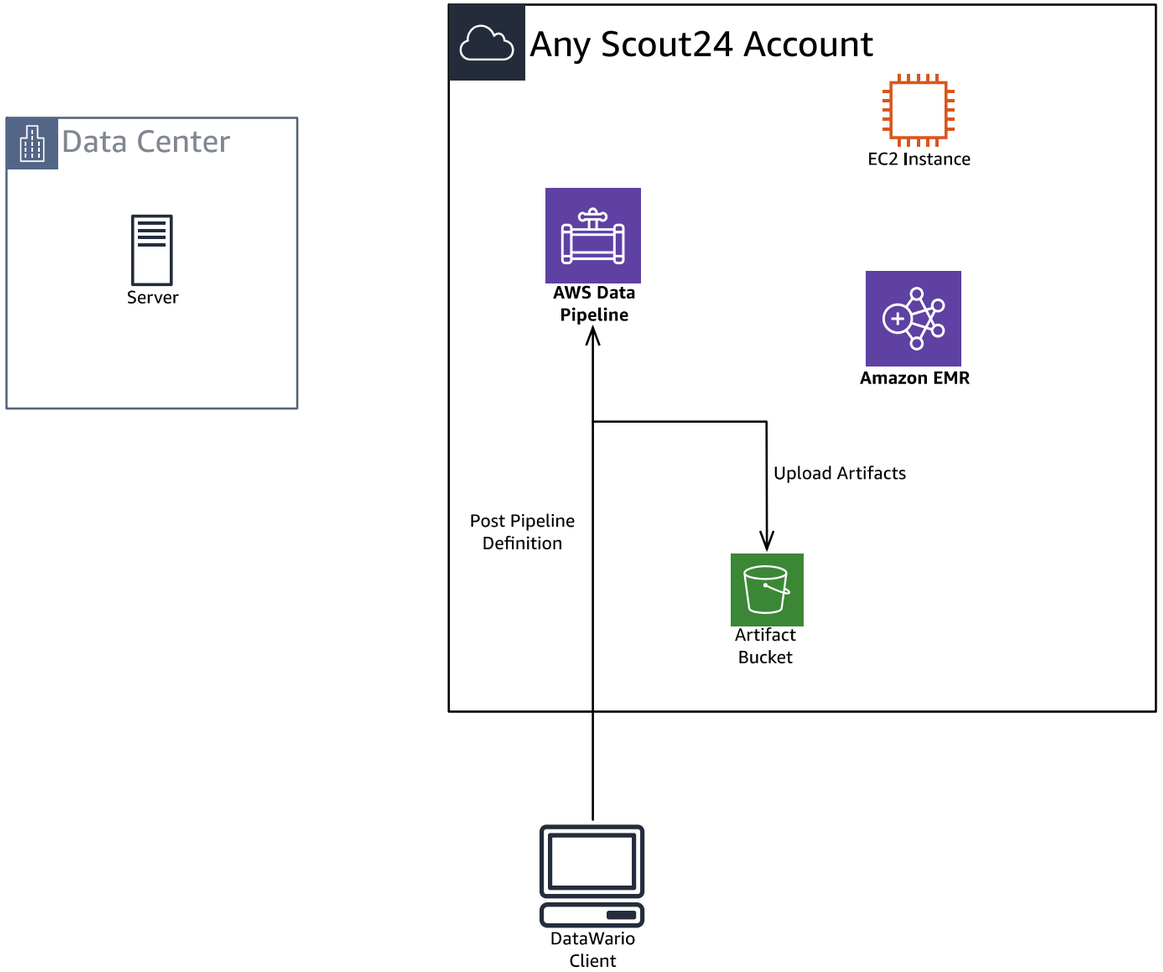


- Builtin Support for Scout24 Ecosystem
- Shortens Development Cycles
- Automatically Manages Artifacts
- Only Exposes Configuration Essentials
- Introduces Custom Step Types

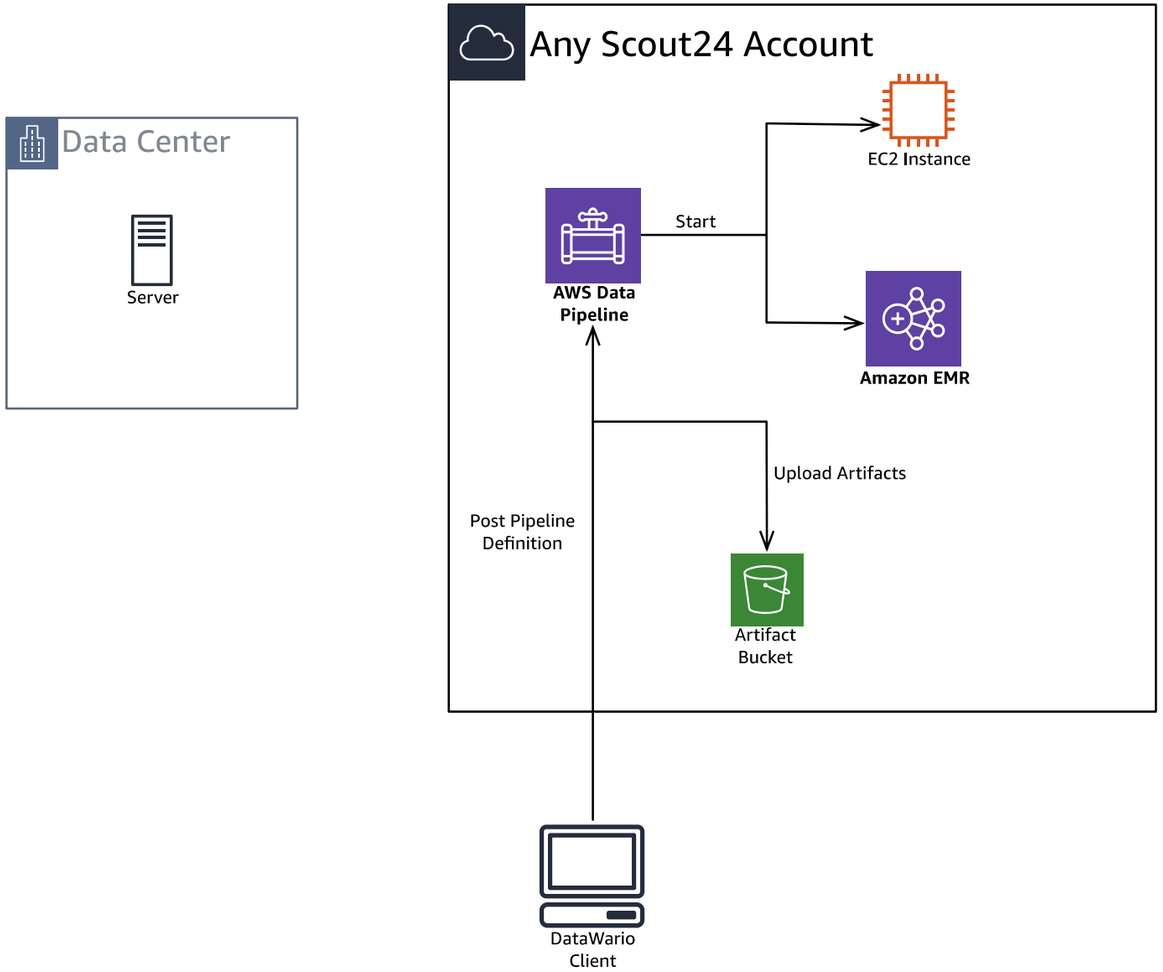
DataWario Architecture



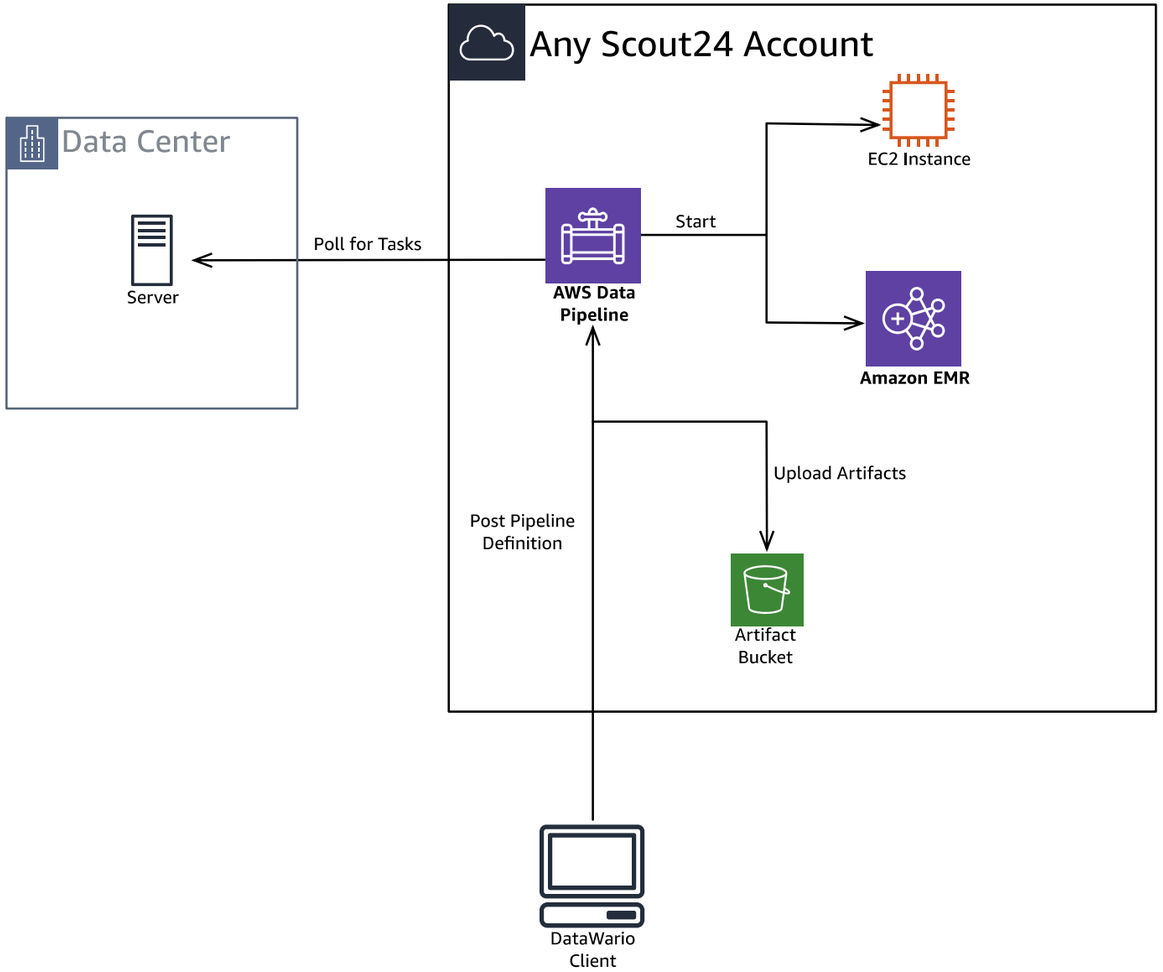
DataWario Architecture



DataWario Architecture



DataWario Architecture



Library of Common Data Transformations

Spark library for:

- Aggregating
- Snapshotting
- Filtering
- Materializing

Self-Service Analytics

aws SUMMIT

Query Challenges

What's Ahead

Unlock the Datalake for Scout24's Toolset and Users with Different Skillsets

Data Analysis for Various User Groups

Provide a Timely and Accurate Update of the Metadata Layer

What's Ahead

Unlock the Datalake for Scout24's Toolset and Users with Different Skillsets

OneScout Hive Metastore

Data Analysis for Various User Groups

Provide a Timely and Accurate Update of the Metadata Layer

What's Ahead

Unlock the Datalake for Scout24's Toolset and Users with Different Skillsets

OneScout Hive Metastore

Data Analysis for Various User Groups

Personal Analytics Cluster

Provide a Timely and Accurate Update of the Metadata Layer

What's Ahead

Unlock the Datalake for Scout24's Toolset and Users with Different Skillsets

OneScout Hive Metastore

Data Analysis for Various User Groups

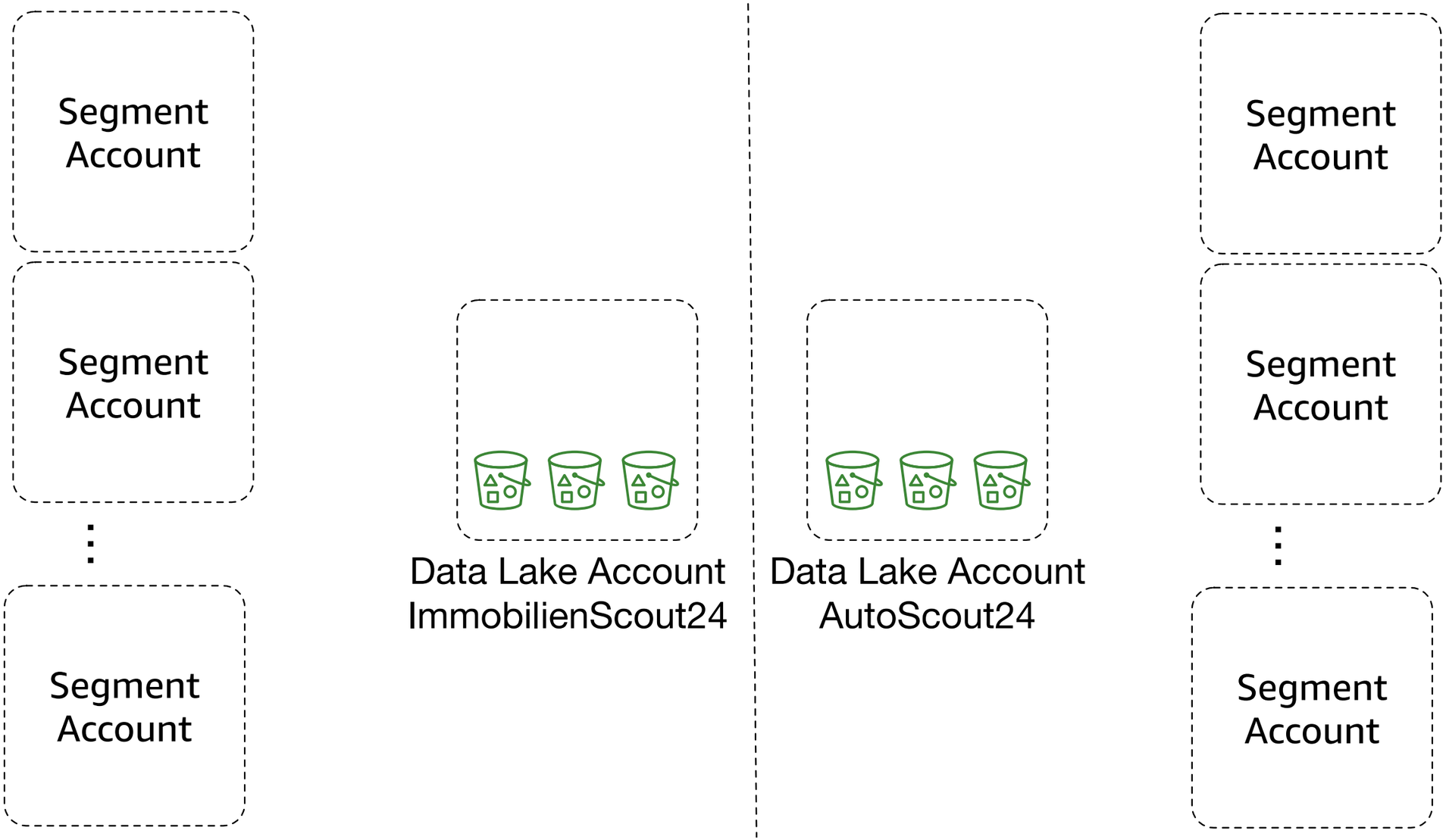
Personal Analytics Cluster

Provide a Timely and Accurate Update of the Metadata Layer

Automatic Hive Partition Detection

OneScout Hive Metastore – A Schematic View

OneScout Hive Metastore – Recap of Ecosystem



EMR Metastore Configuration Options

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

External Metastore in RDS

External Metastore in Glue Data Catalog

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

External Metastore in RDS

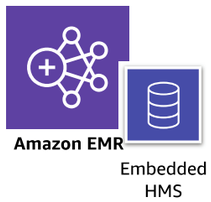
External Metastore in Glue Data Catalog

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

External Metastore in RDS

External Metastore in Glue Data Catalog



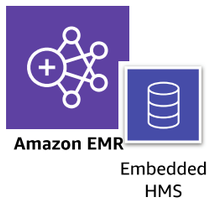
Metadata Lost at Shutdown

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

External Metastore in RDS

External Metastore in Glue Data Catalog



Metadata Lost at Shutdown

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

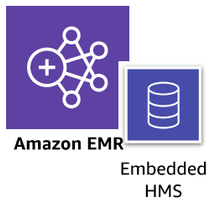
Metadata Lost at Shutdown

External Metastore in RDS

External Metastore in Glue Data Catalog

EMR Metastore Configuration Options

Metastore Embedded on Master Instance



Metadata Lost at Shutdown

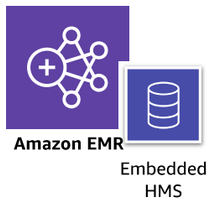
External Metastore in RDS

Metadata Persisted and Usable by all Clusters within the Account

External Metastore in Glue Data Catalog

EMR Metastore Configuration Options

Metastore Embedded on Master Instance



Metadata Lost at Shutdown

External Metastore in RDS

Metadata Persisted and Usable by all Clusters within the Account

External Metastore in Glue Data Catalog

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

Metadata Lost at Shutdown

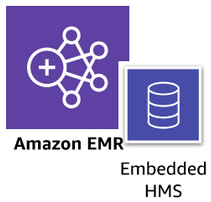
External Metastore in RDS

Metadata Persisted and Usable by all Clusters within the Account

External Metastore in Glue Data Catalog

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

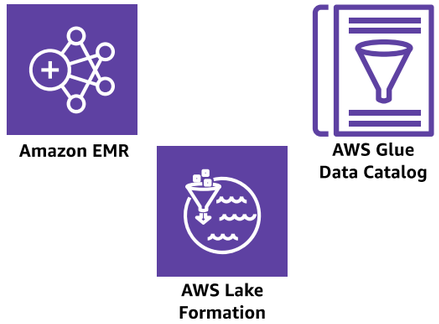


Metadata Lost at Shutdown

External Metastore in RDS

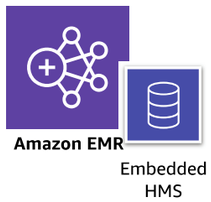
Metadata Persisted and Usable by all Clusters within the Account

External Metastore in Glue Data Catalog



EMR Metastore Configuration Options

Metastore Embedded on Master Instance

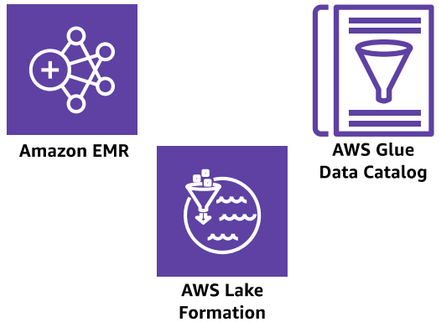


Metadata Lost at Shutdown

External Metastore in RDS

Metadata Persisted and Usable by all Clusters within the Account

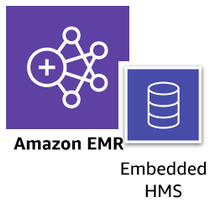
External Metastore in Glue Data Catalog



Future-Proof Solution with Cross-Account Support and Service Integrations (e.g. Lake Formation)

EMR Metastore Configuration Options

Metastore Embedded on Master Instance

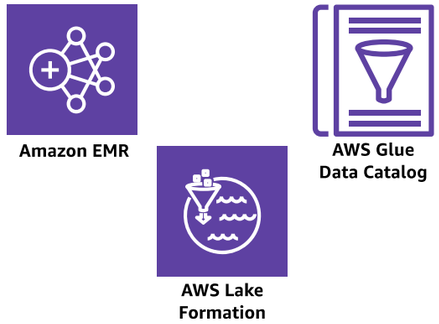


Metadata Lost at Shutdown

External Metastore in RDS

Metadata Persisted and Usable by all Clusters within the Account

External Metastore in Glue Data Catalog



Future-Proof Solution with Cross-Account Support and Service Integrations (e.g. Lake Formation)

The Scout24 Hive Metastore Proxy – A Motivation

Goal: One Long-Running Metastore DB every EMR connects to

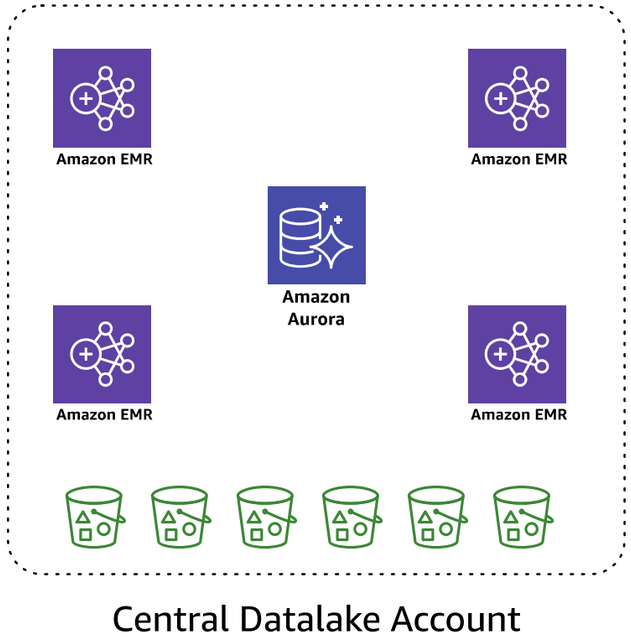
Ideal Situation

Solution for Multi-Account-Setting

The Scout24 Hive Metastore Proxy – A Motivation

Goal: One Long-Running Metastore DB every EMR connects to

Ideal Situation



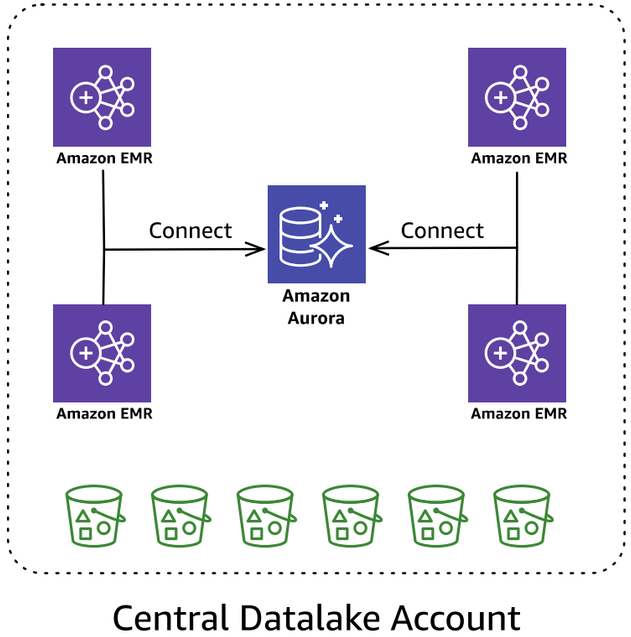
Central Datalake Account

Solution for Multi-Account-Setting

The Scout24 Hive Metastore Proxy – A Motivation

Goal: One Long-Running Metastore DB every EMR connects to

Ideal Situation



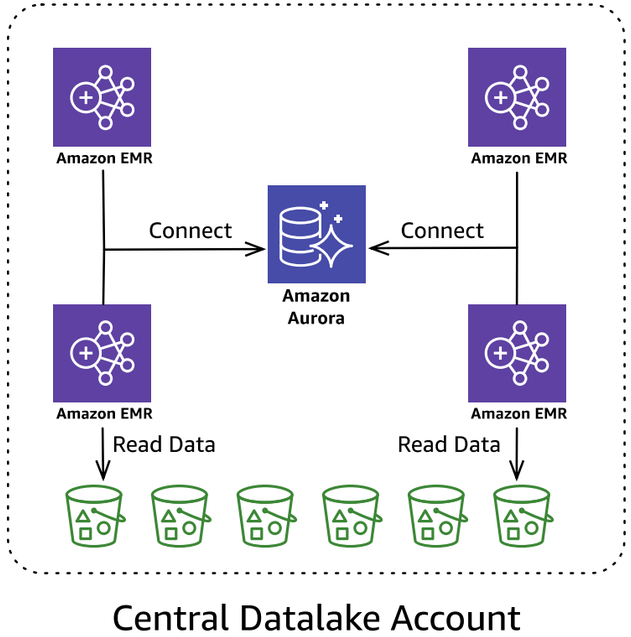
Central Datalake Account

Solution for Multi-Account-Setting

The Scout24 Hive Metastore Proxy – A Motivation

Goal: One Long-Running Metastore DB every EMR connects to

Ideal Situation



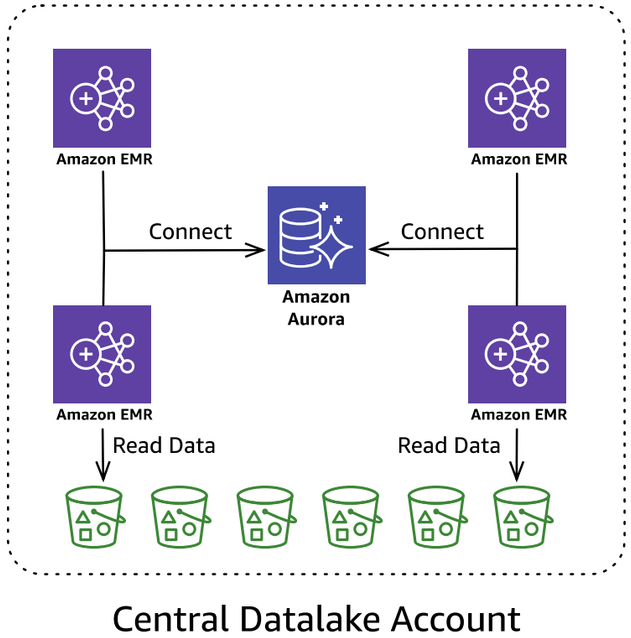
Central Datalake Account

Solution for Multi-Account-Setting

The Scout24 Hive Metastore Proxy – A Motivation

Goal: One Long-Running Metastore DB every EMR connects to

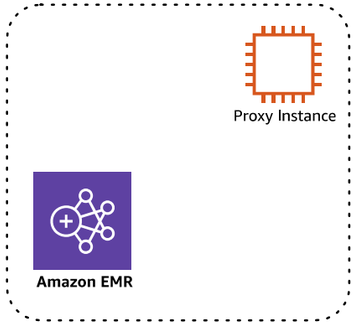
Ideal Situation



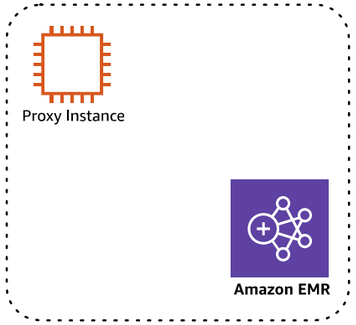
Central Datalake Account

Solution for Multi-Account-Setting

Metastore Account



Segment Account

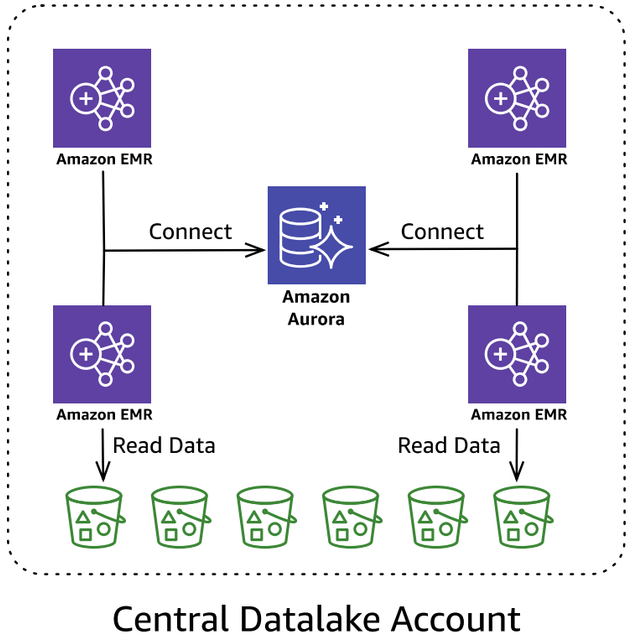


Segment Account

The Scout24 Hive Metastore Proxy – A Motivation

Goal: One Long-Running Metastore DB every EMR connects to

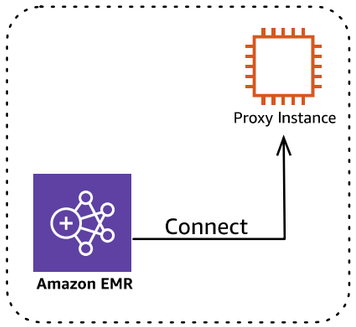
Ideal Situation



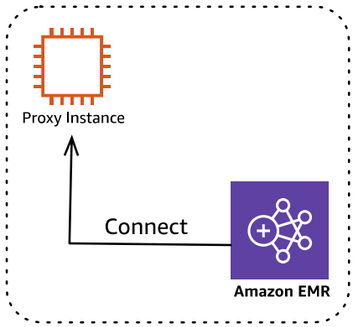
Central Datalake Account

Solution for Multi-Account-Setting

Metastore Account



Segment Account

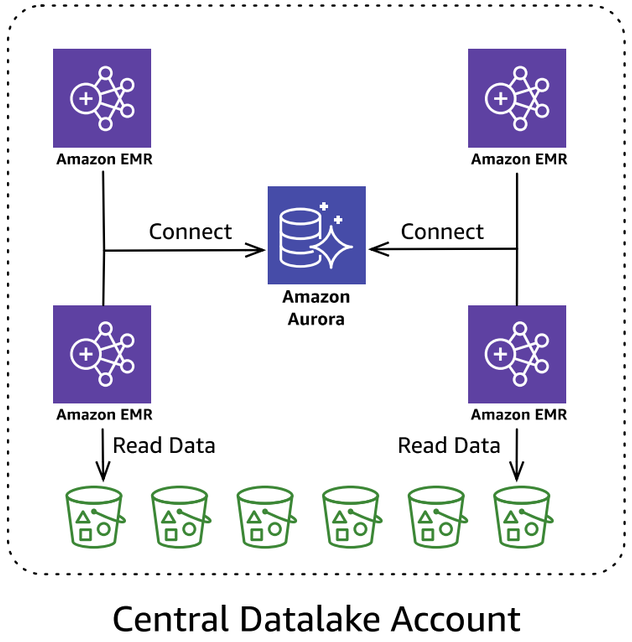


Segment Account

The Scout24 Hive Metastore Proxy – A Motivation

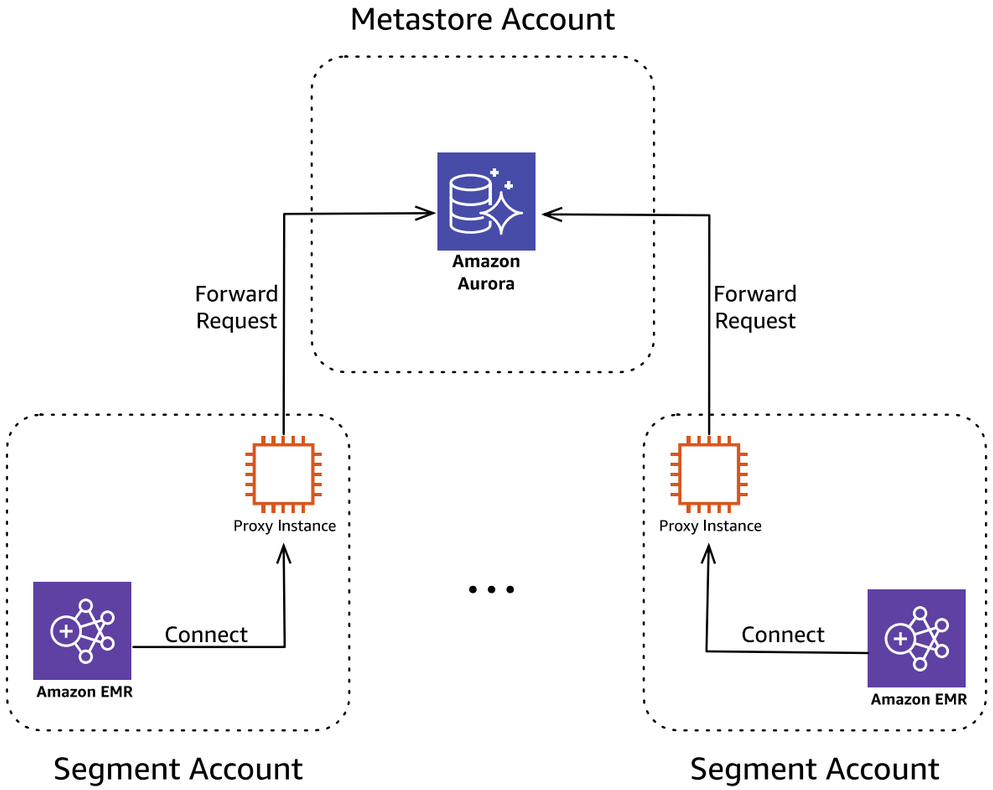
Goal: One Long-Running Metastore DB every EMR connects to

Ideal Situation



Central Datalake Account

Solution for Multi-Account-Setting

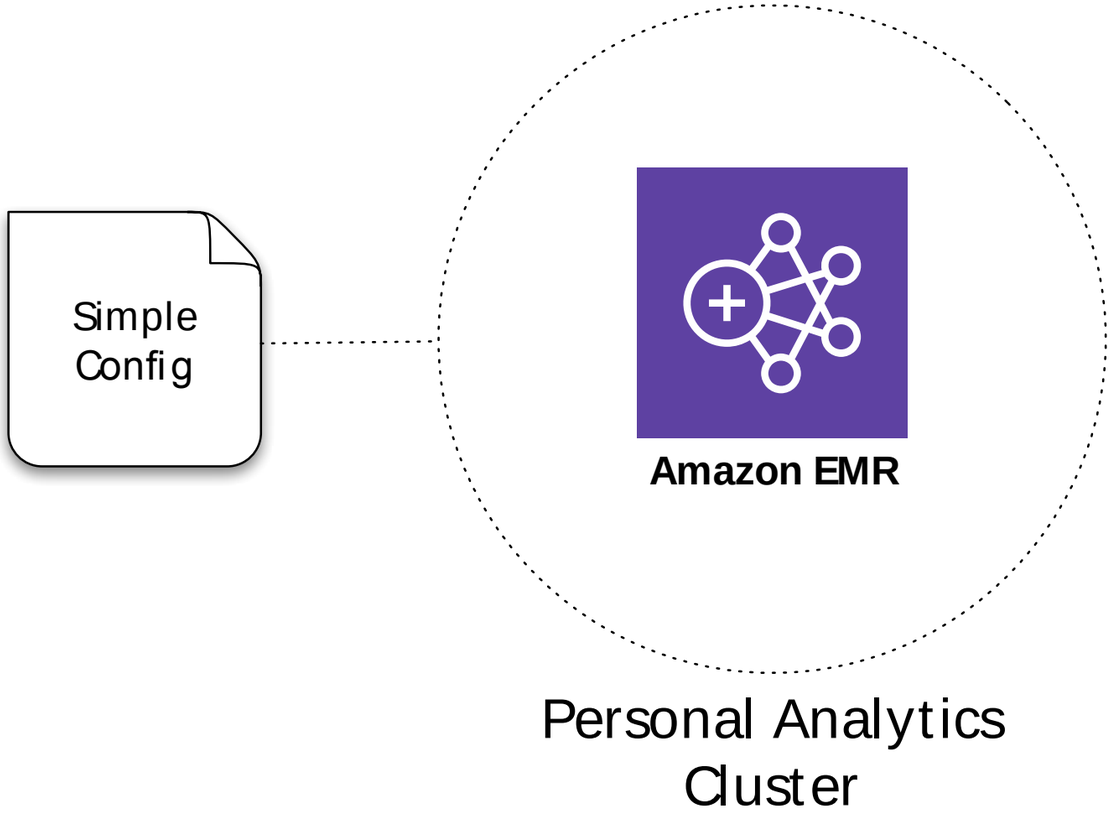


Segment Account

Segment Account

The Personal Analytics Cluster – An Overview

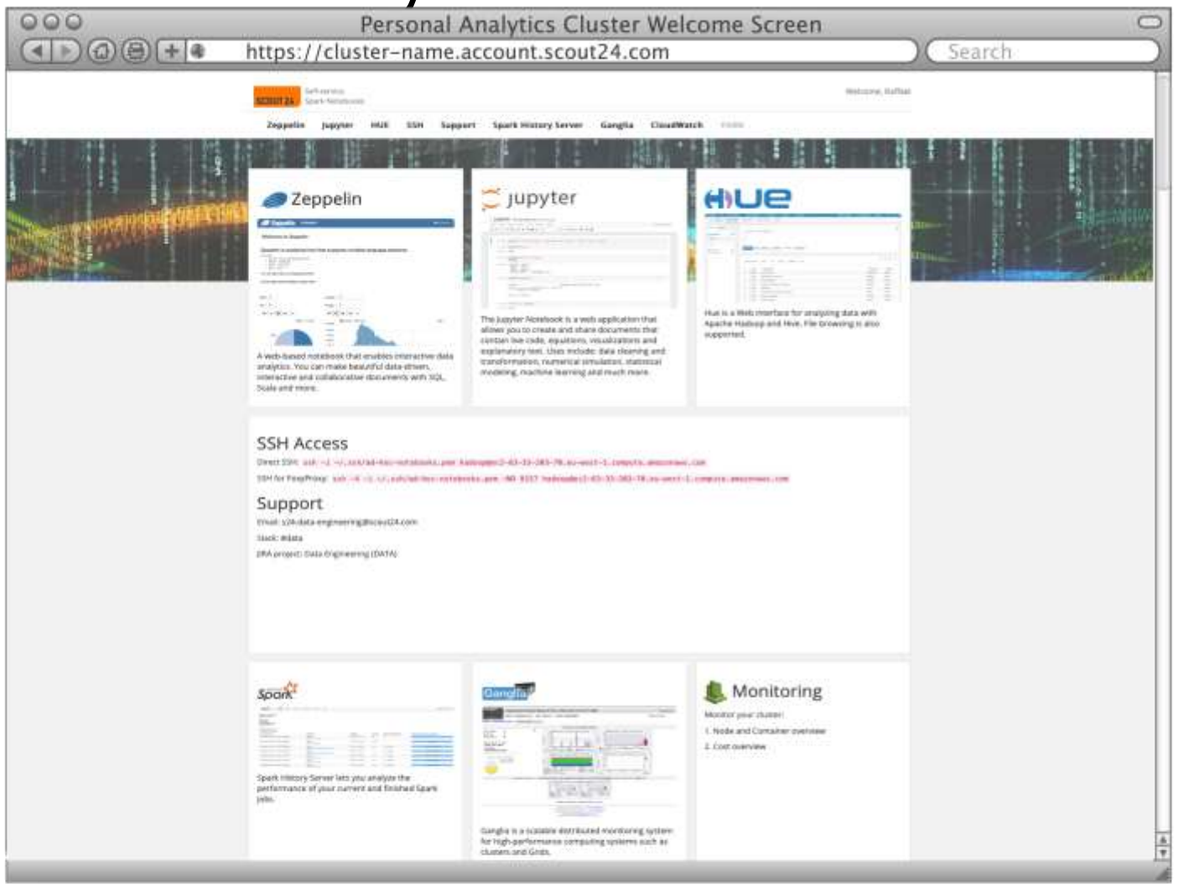
The Personal Analytics Cluster – An Overview



The Personal Analytics Cluster – An Overview

- Easy Access via Web Interface
- Zeppelin and Jupyter Notebook Restore
- OneClick Deployment
- Managed Scaling and Shutdown
- Support for Pre-baked AMIs and Configs

The Personal Analytics Cluster – An Overview

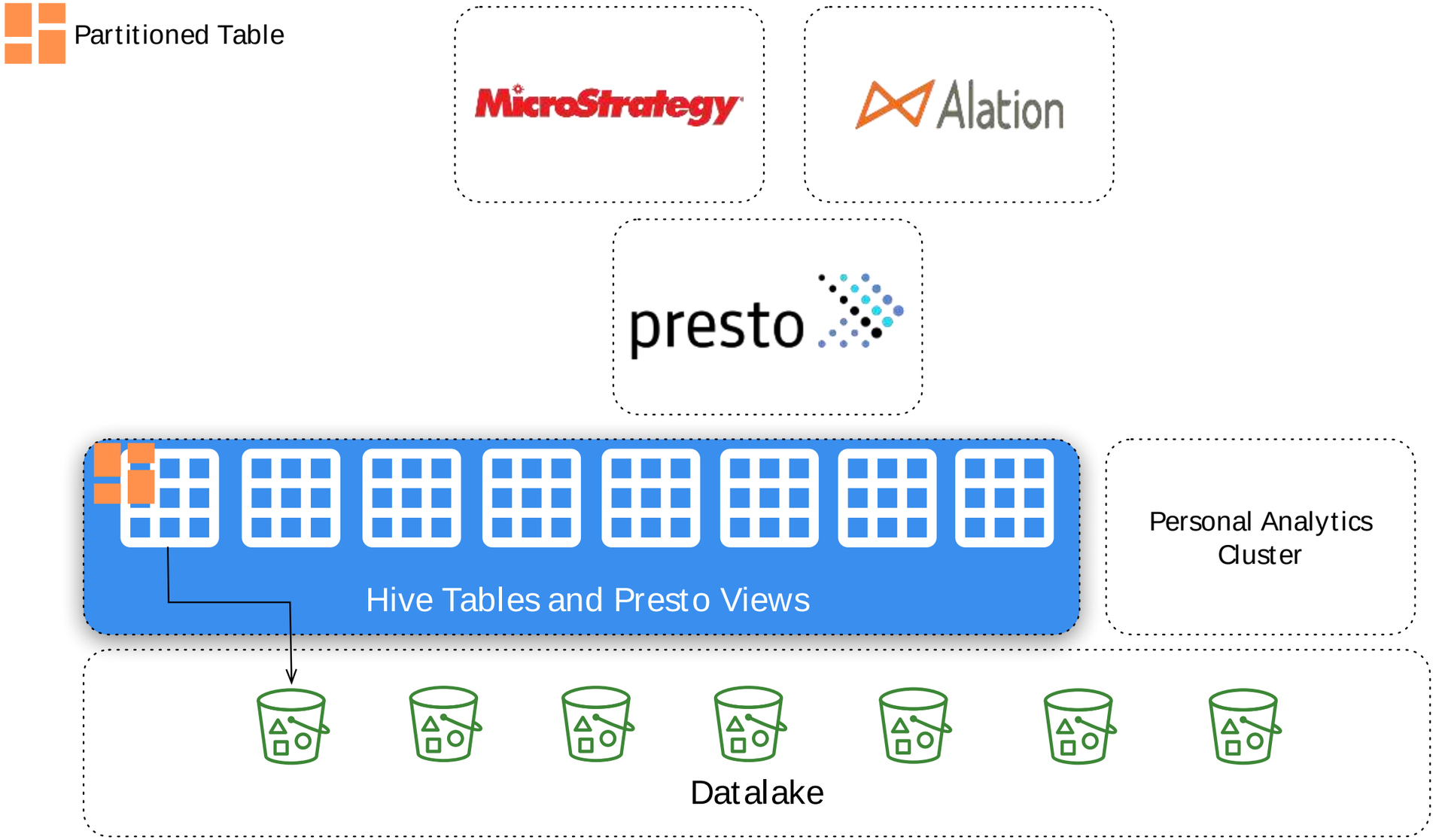


Personal Analytics Cluster

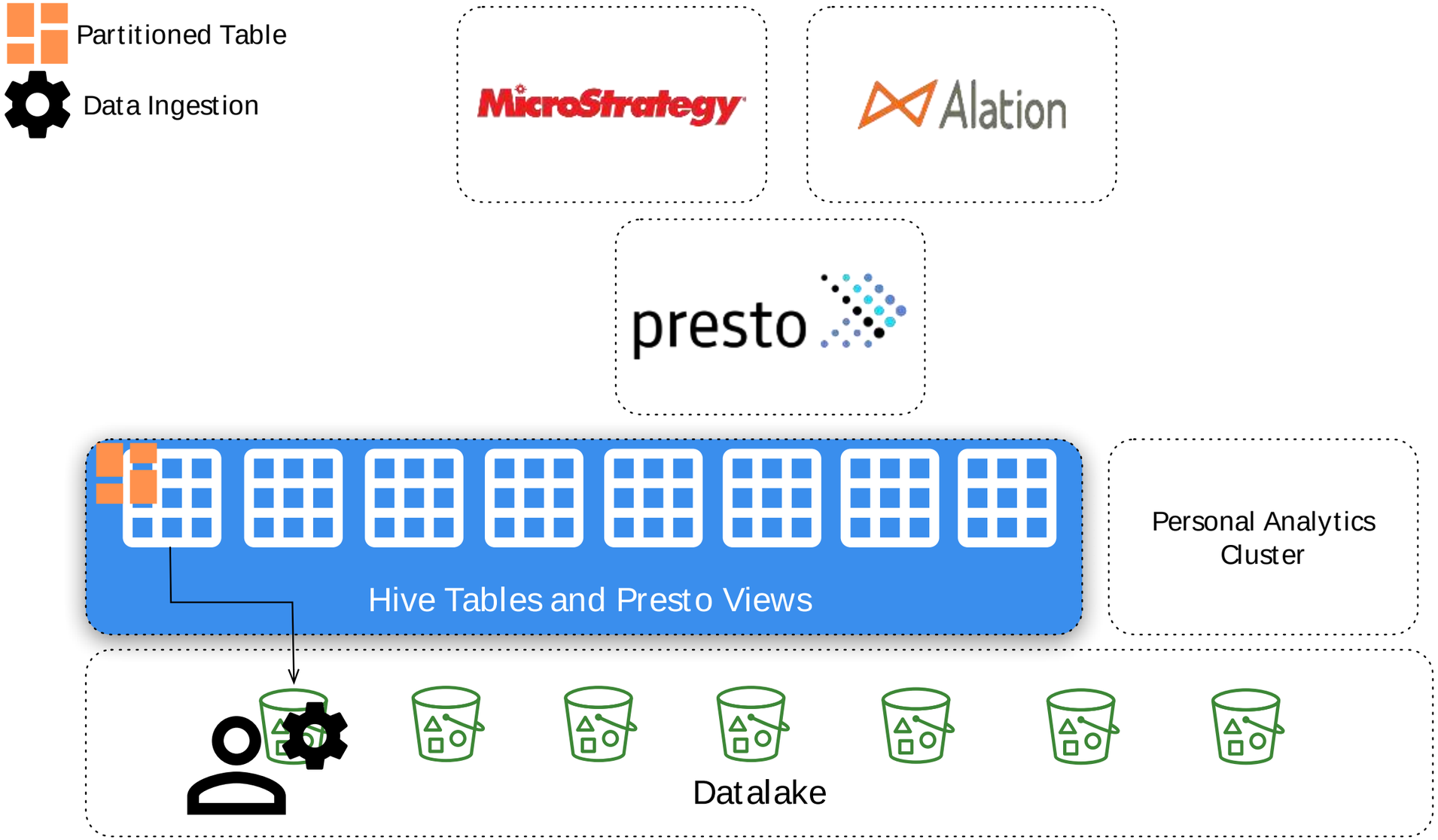
- Easy Access via Web Interface
- Zeppelin and Jupyter Notebook Restore
- OneClick Deployment
- Managed Scaling and Shutdown
- Support for Pre-baked AMIs and Configs

Automated Partition Detection – A Motivation

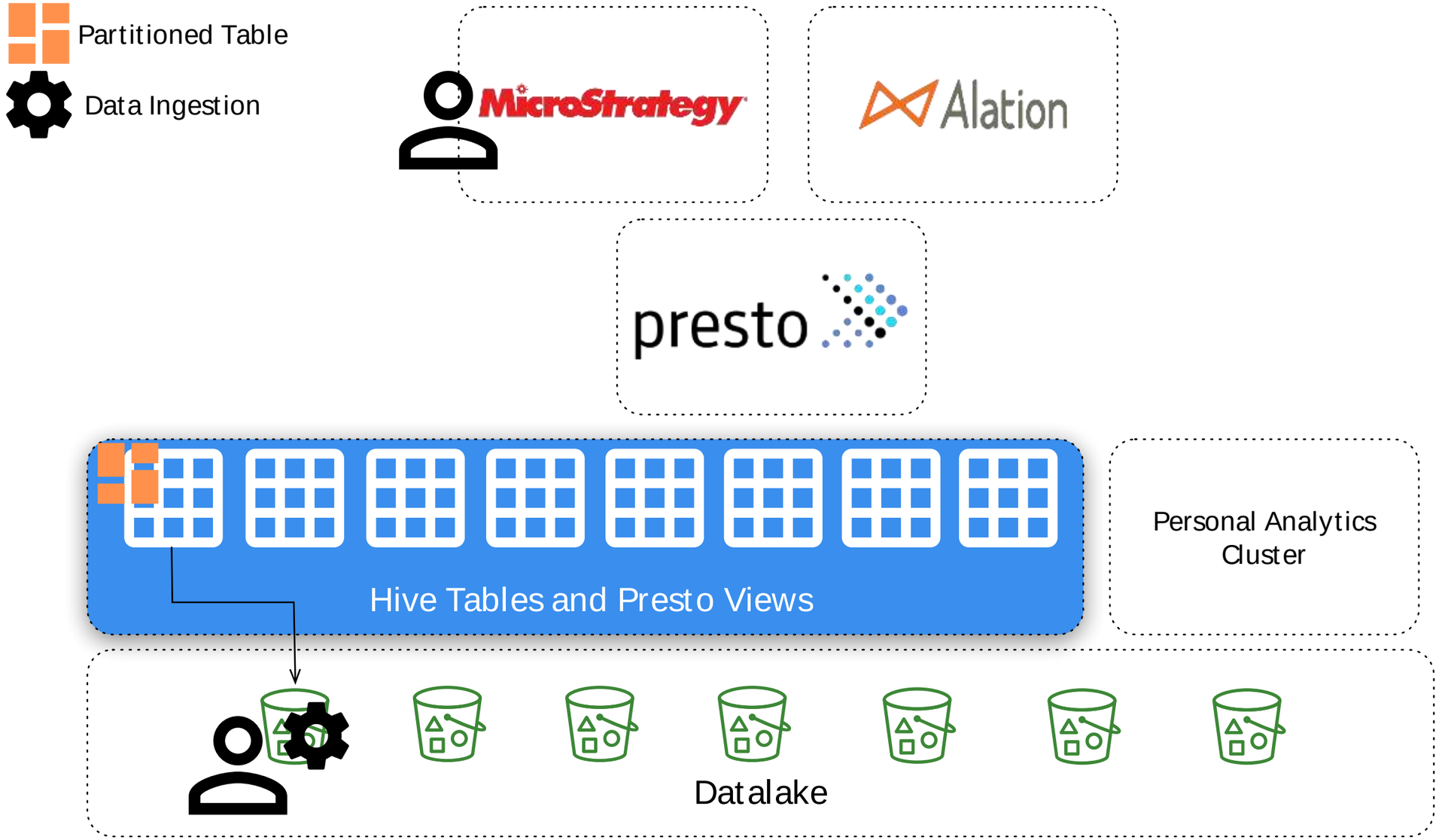
Automated Partition Detection – A Motivation



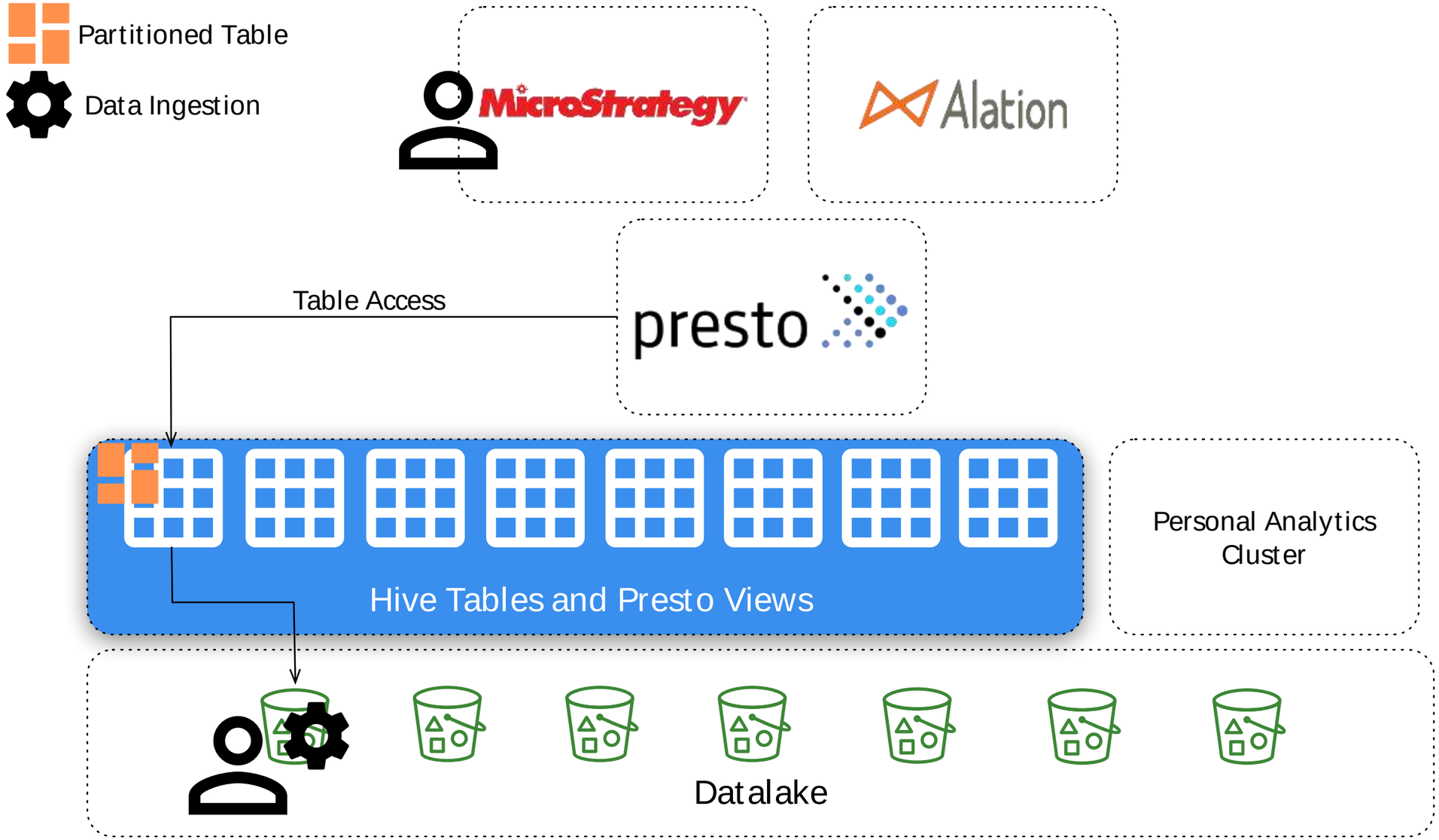
Automated Partition Detection – A Motivation



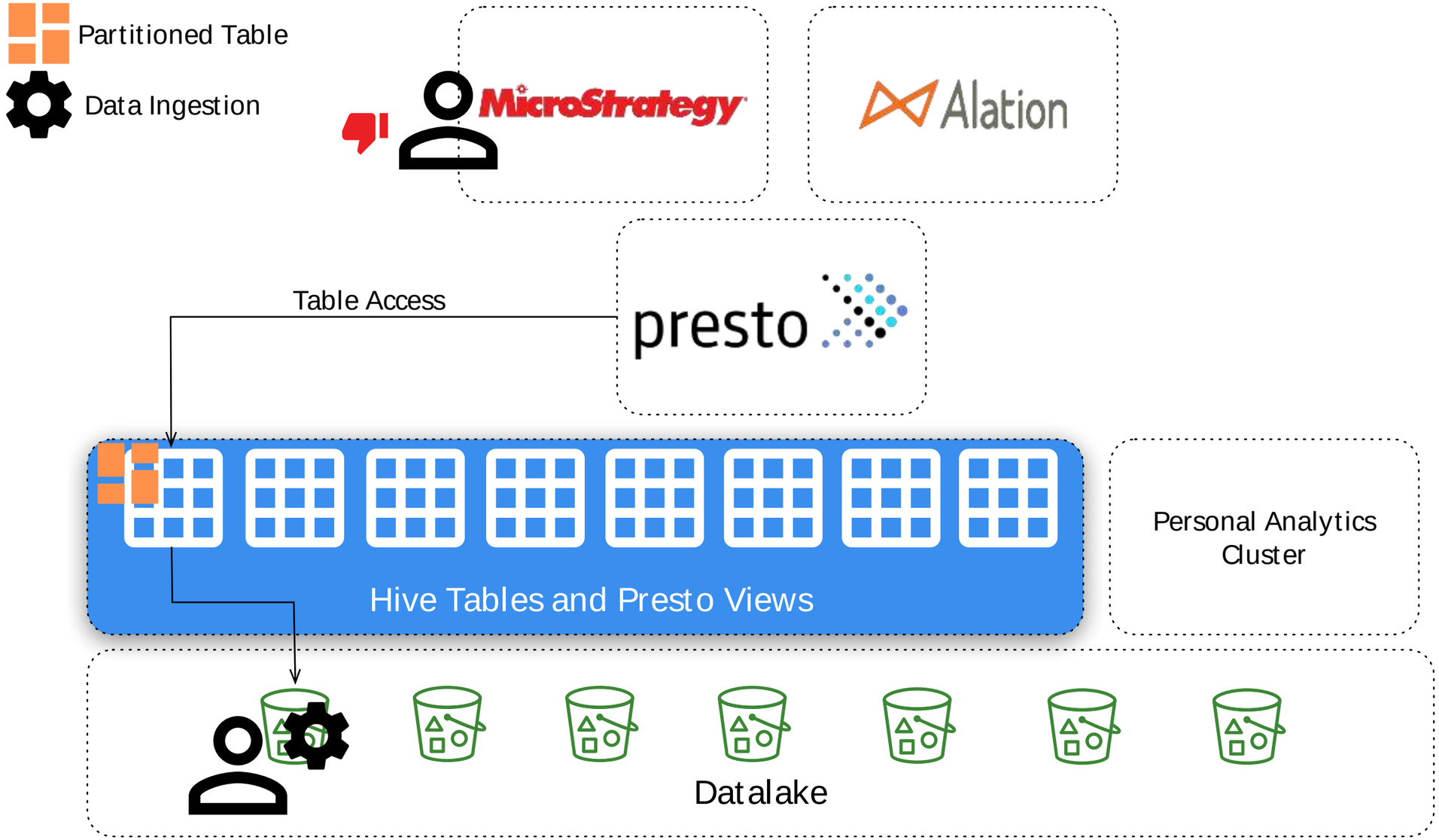
Automated Partition Detection – A Motivation



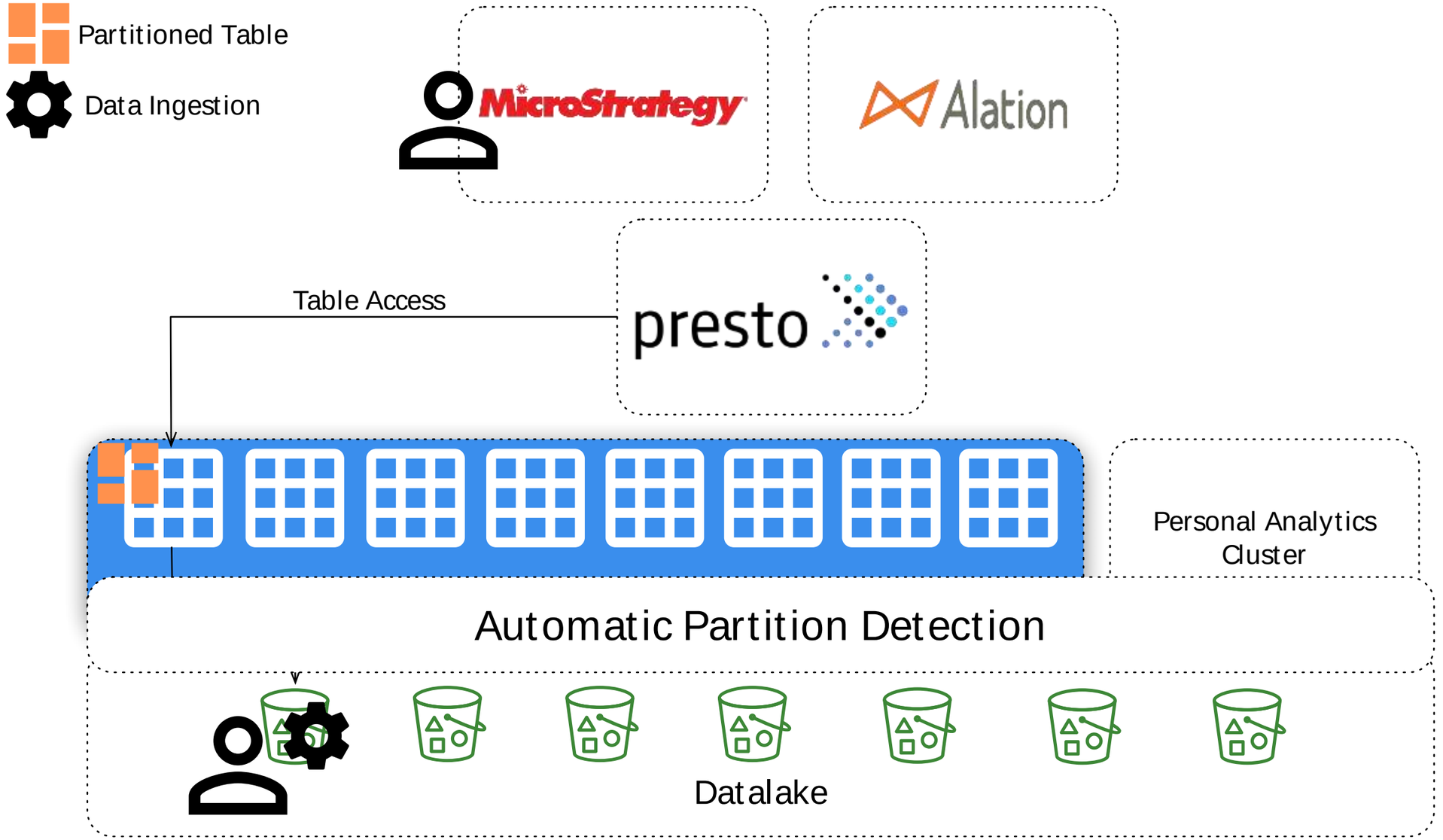
Automated Partition Detection – A Motivation



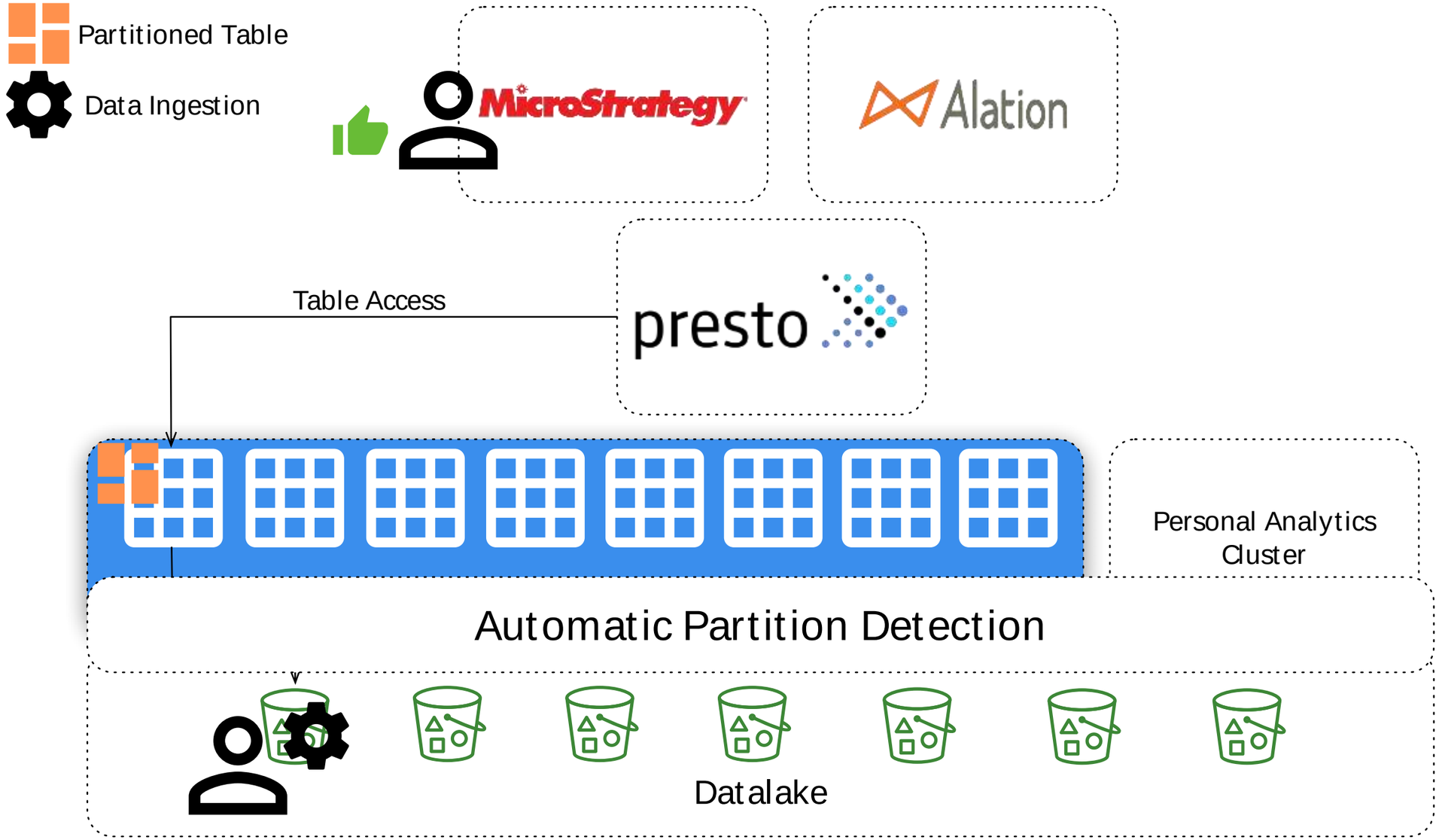
Automated Partition Detection – A Motivation



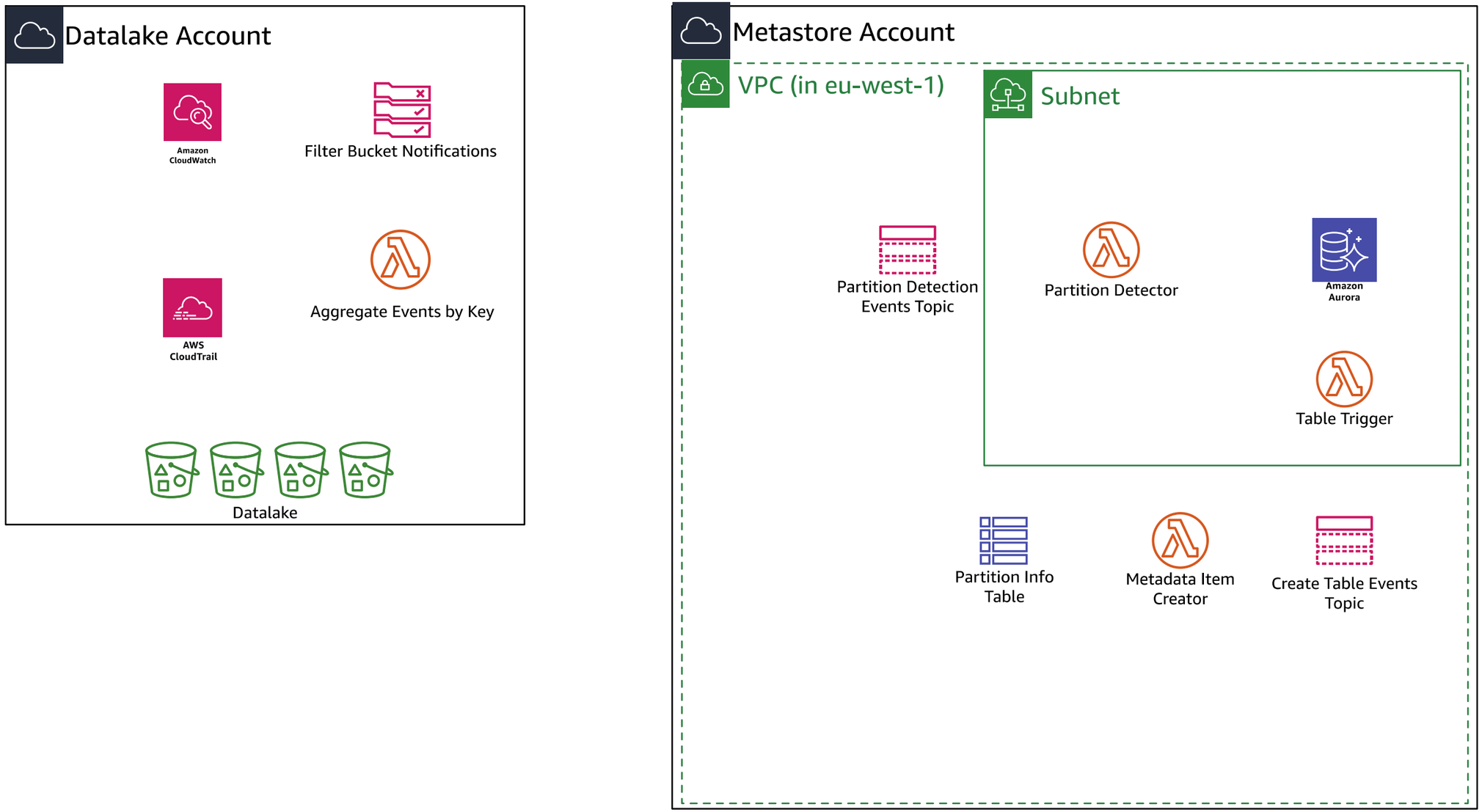
Automated Partition Detection – A Motivation



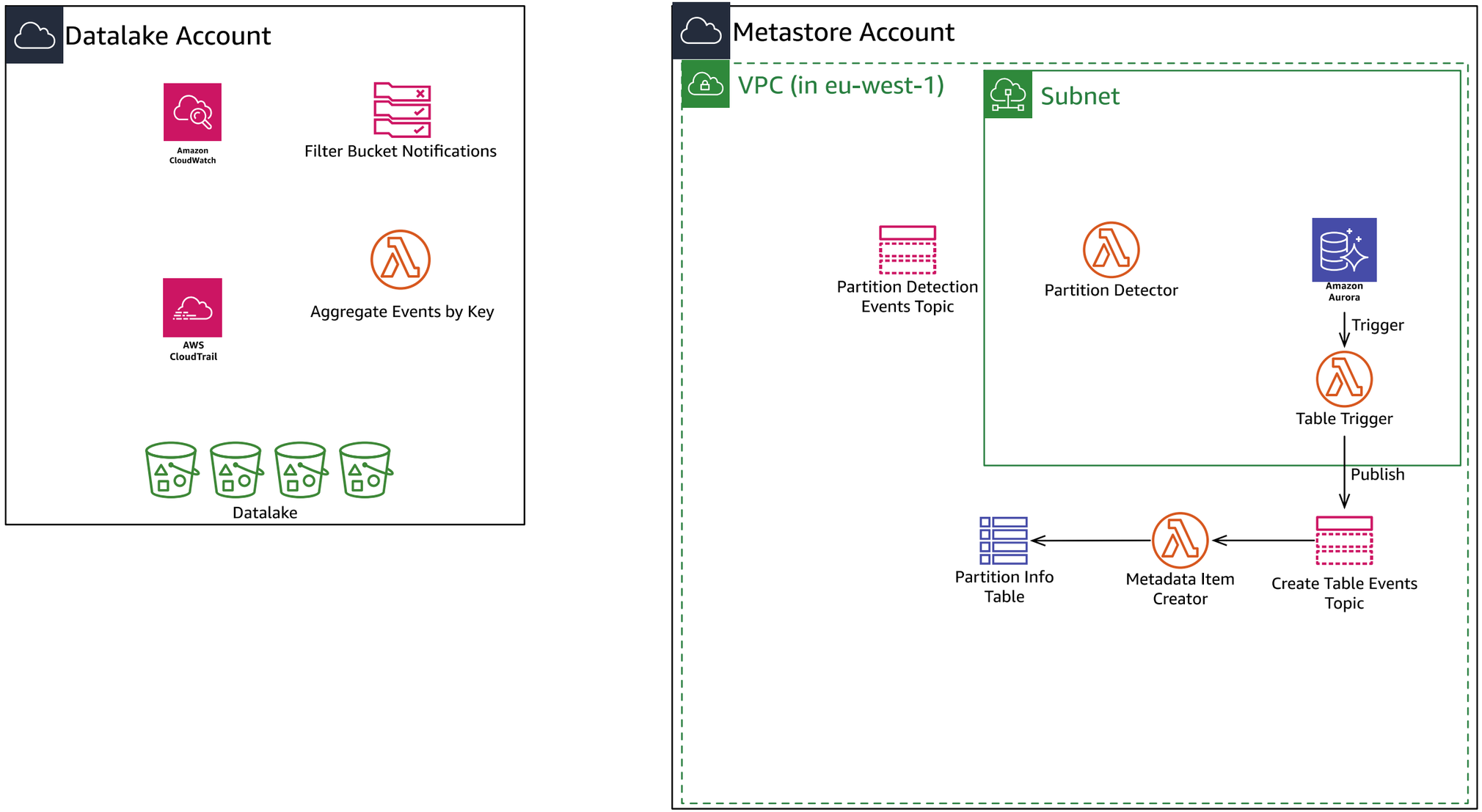
Automated Partition Detection – A Motivation



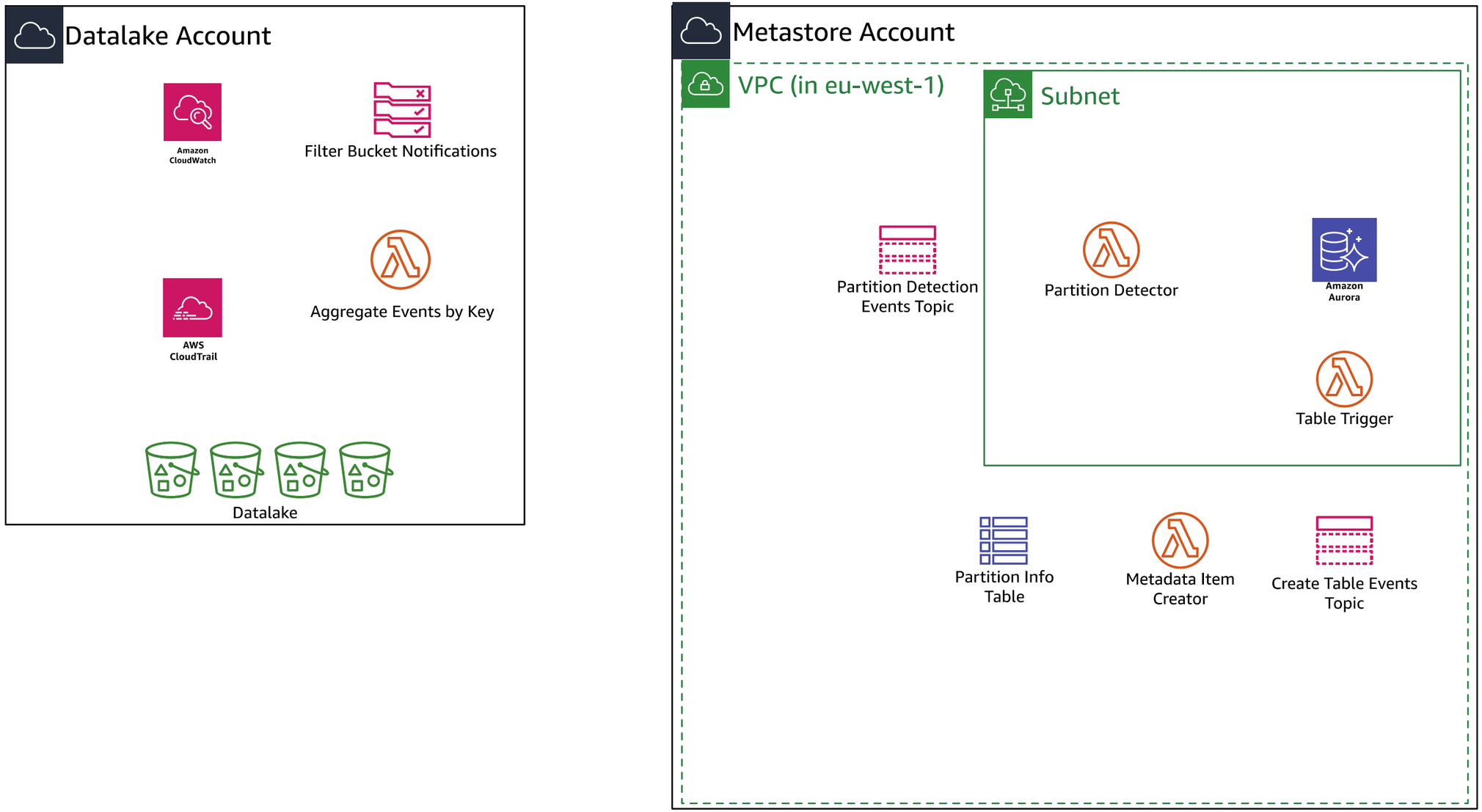
Partition Detection Architecture



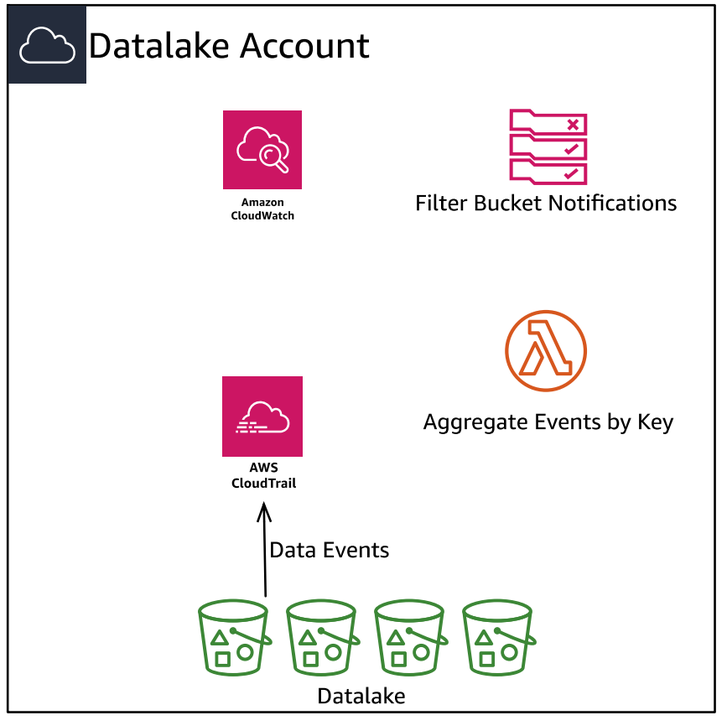
Partition Detection Architecture

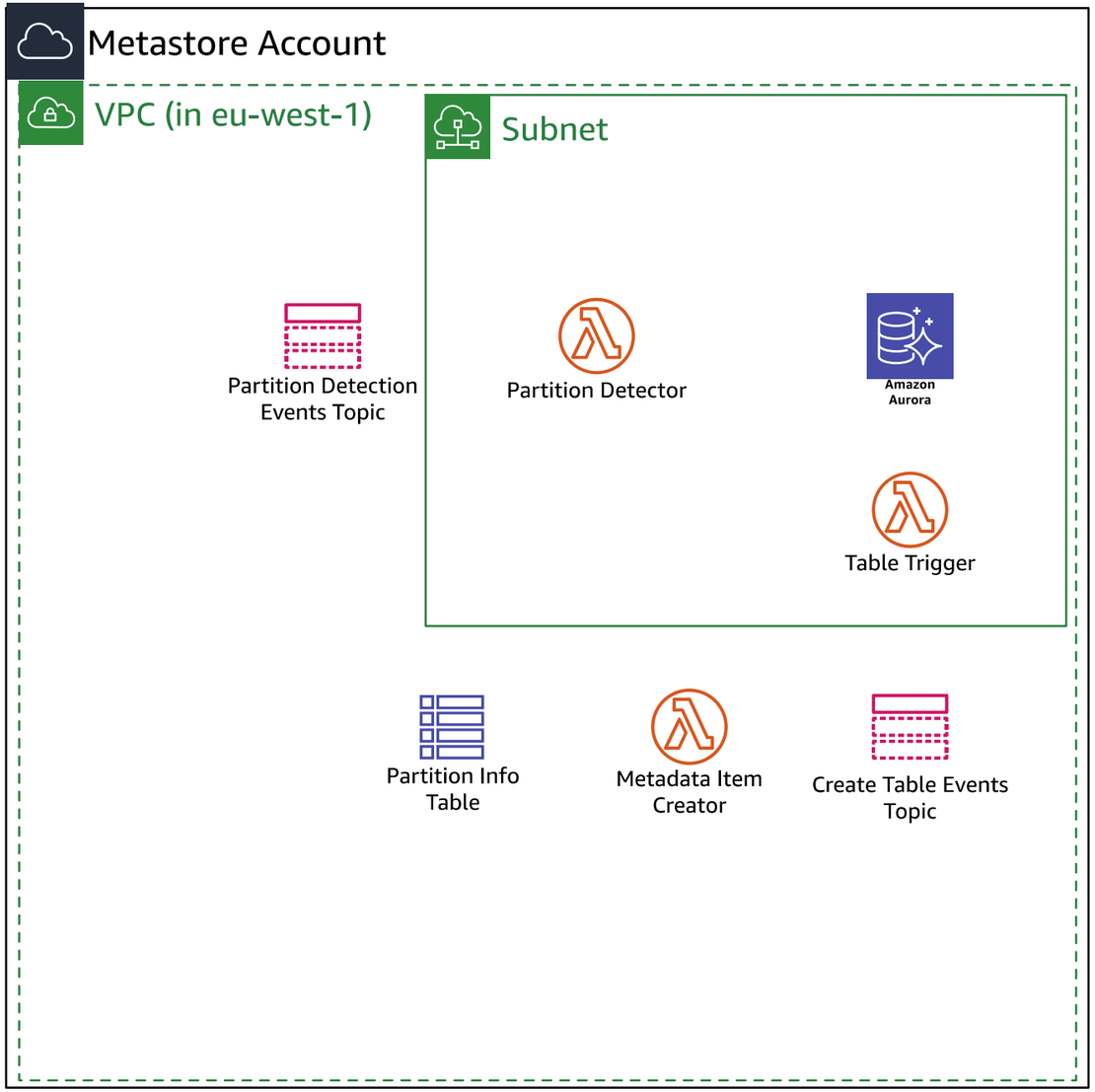


Partition Detection Architecture

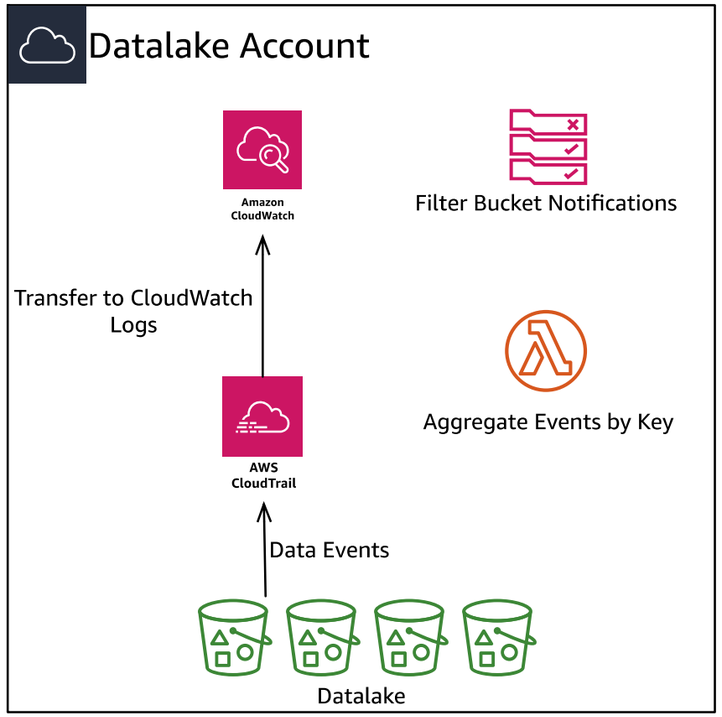


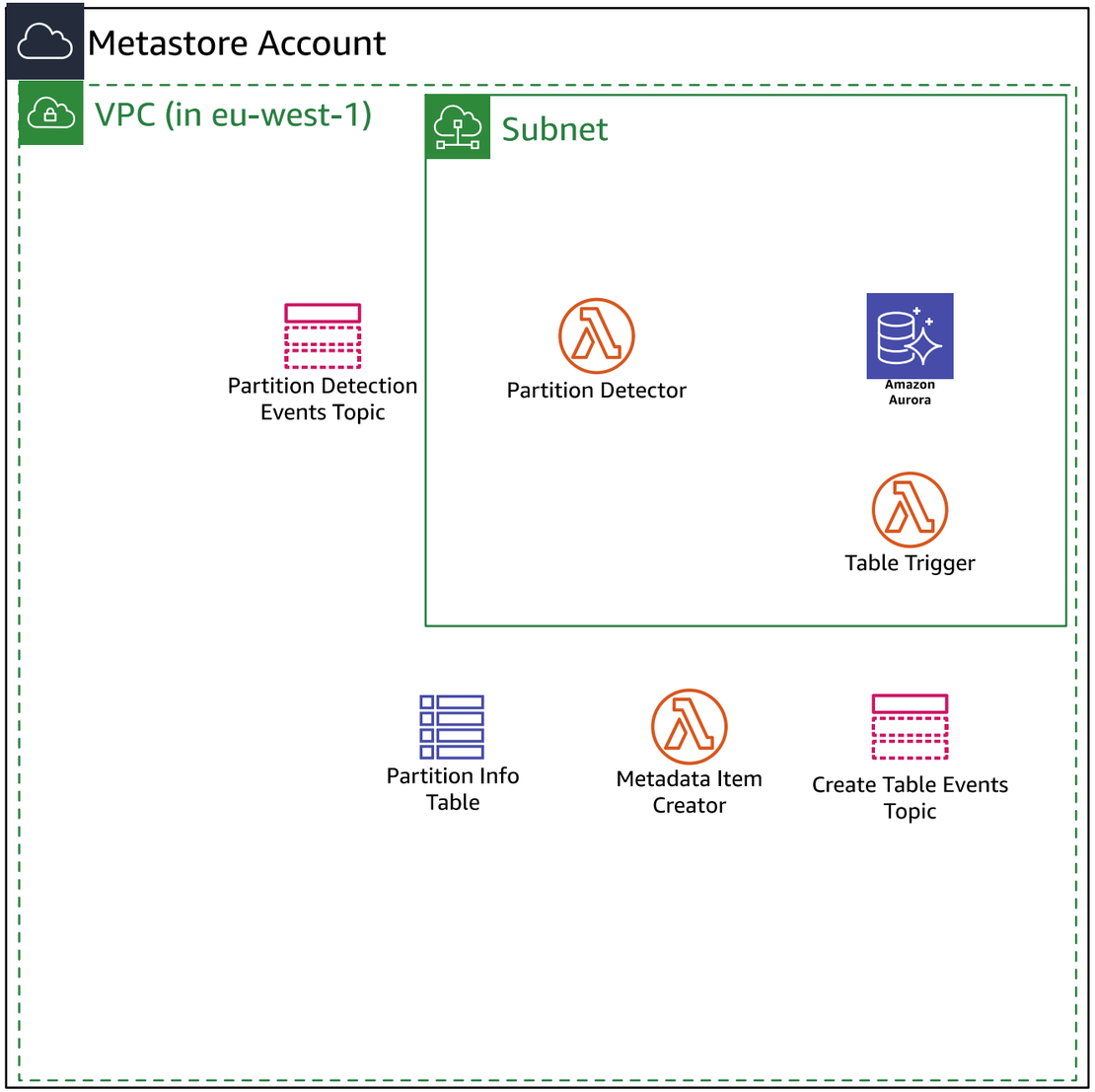
Partition Detection Architecture



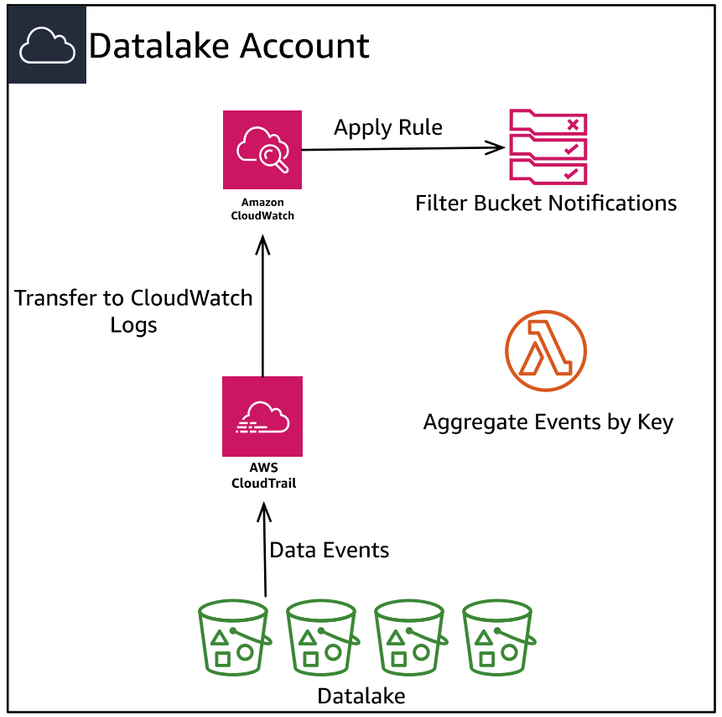


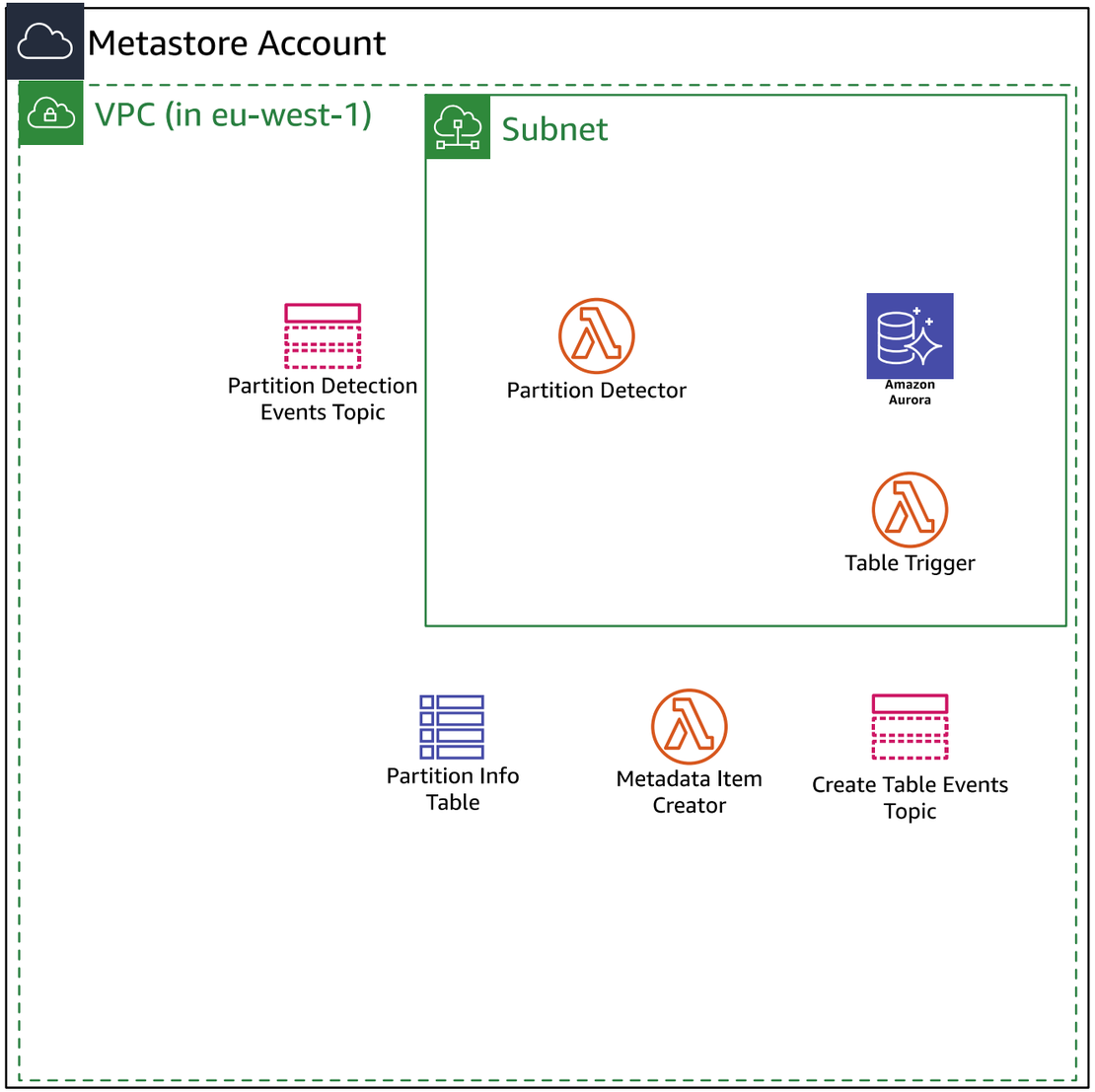
Partition Detection Architecture



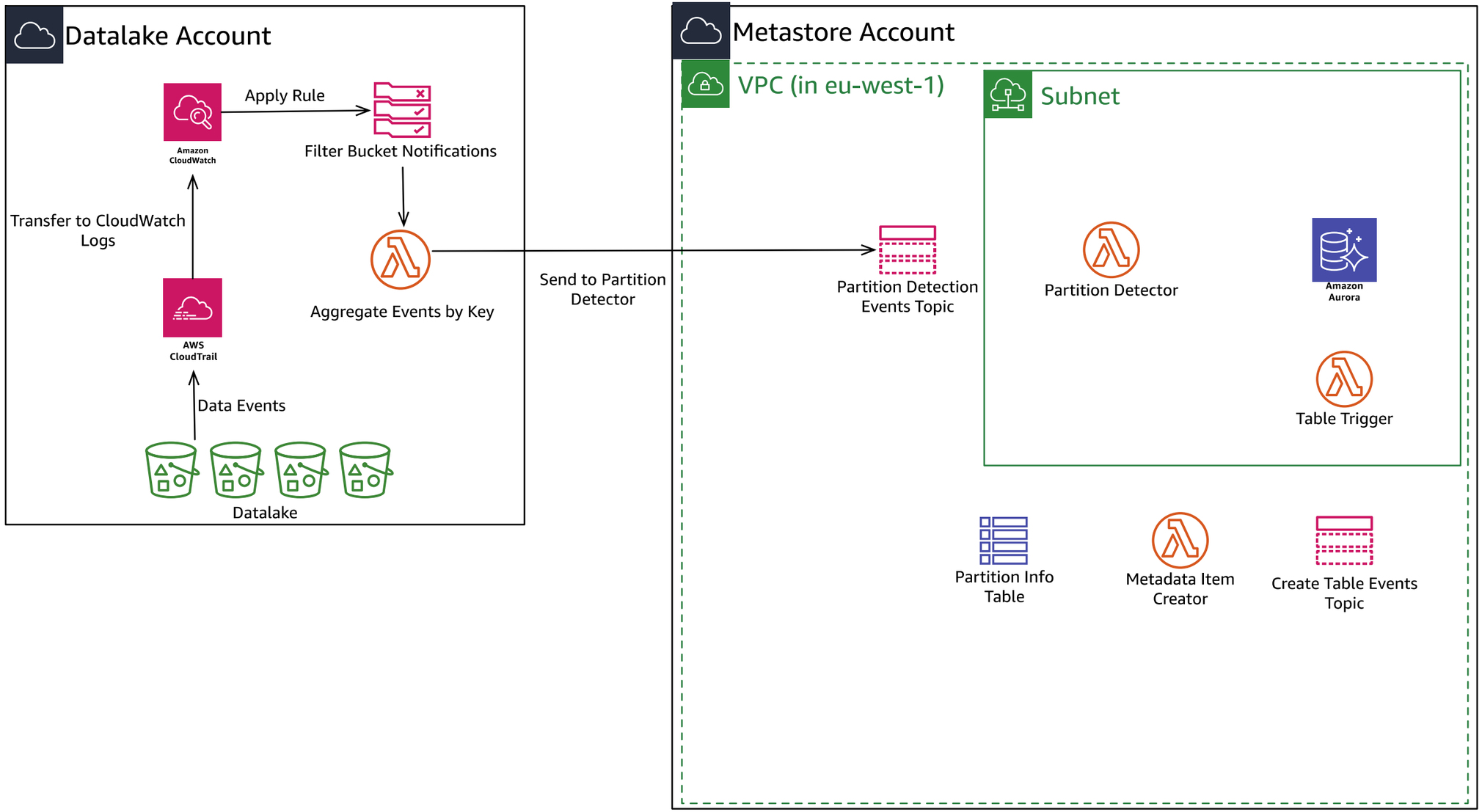


Partition Detection Architecture

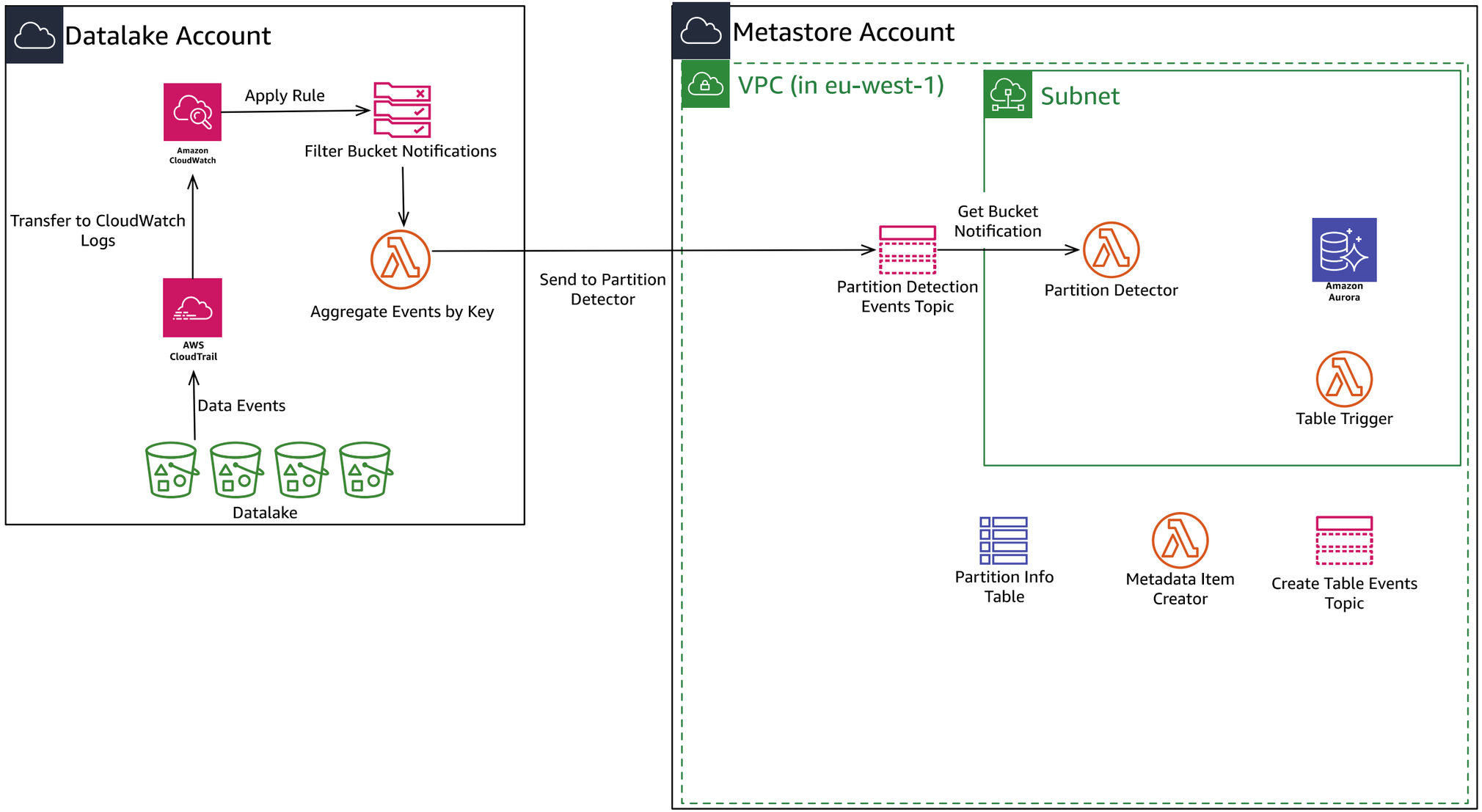




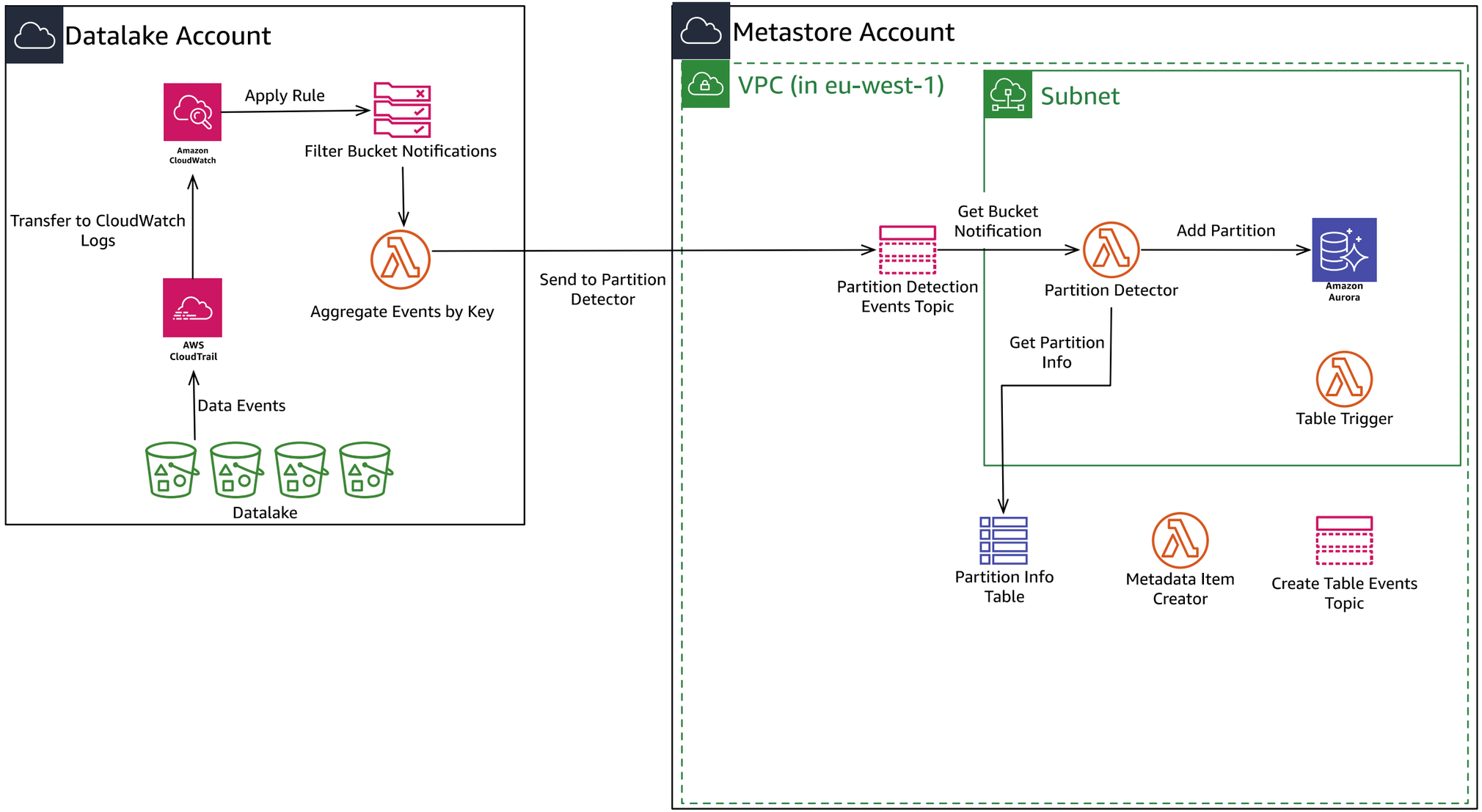
Partition Detection Architecture



Partition Detection Architecture



Partition Detection Architecture



Conclusion

aws SUMMIT

Build our own vs. AWS managed services

Metastore → Glue

Presto → Athena

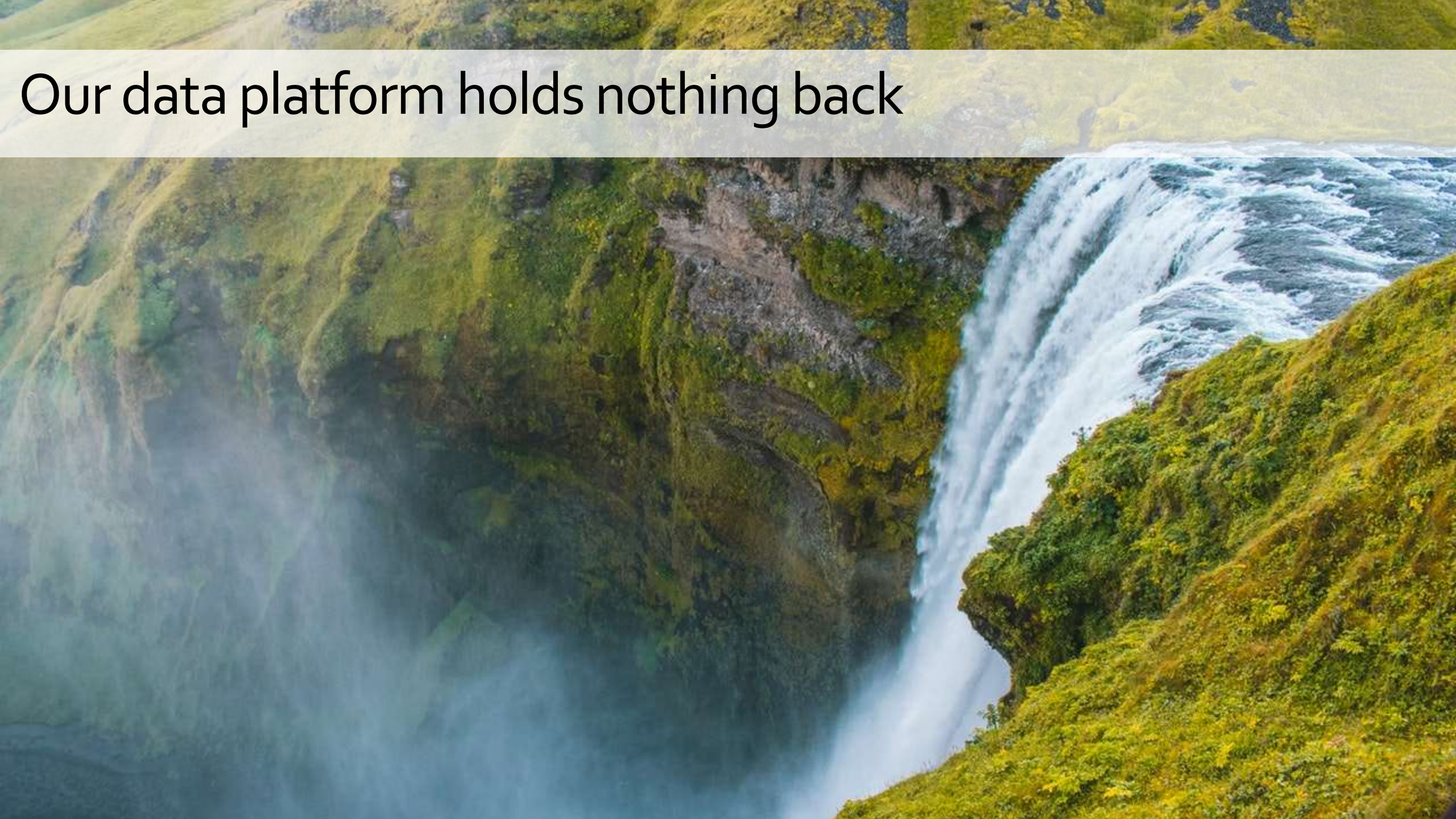
DataWario → Glue, Step function, Lambda, …

Personal Analytics Cluster → Glue notebooks, Sagemaker

We hope to throw out **most** of the custom components we build.

Our data warehouse was a bottle neck

Our data platform holds nothing back



Thank you!

Sean Gustafson
Senior Technical Product Manager
Scout24

Raffael Dzikowski
Senior Data Engineer
Scout24

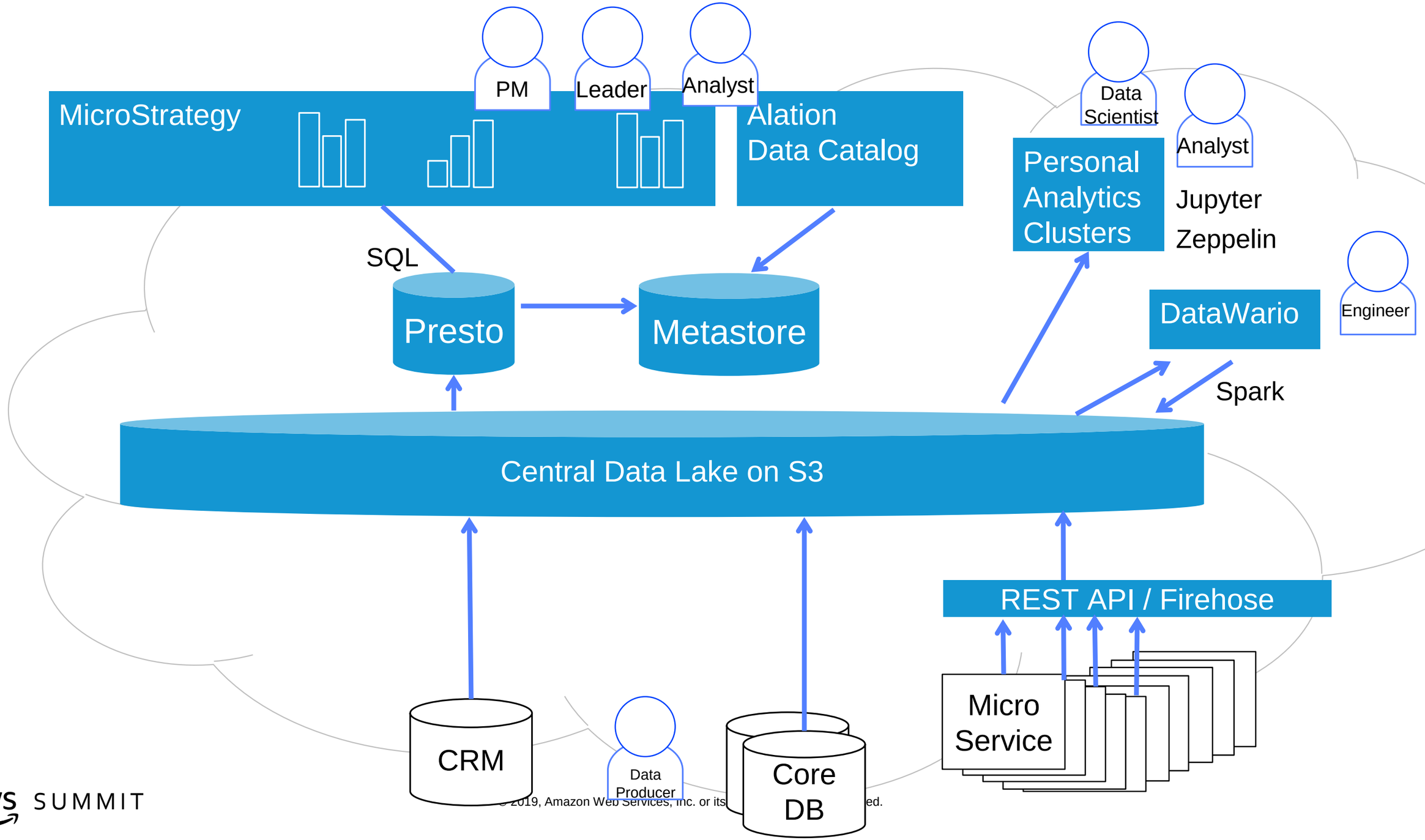
aws SUMMIT

Please complete the session survey.

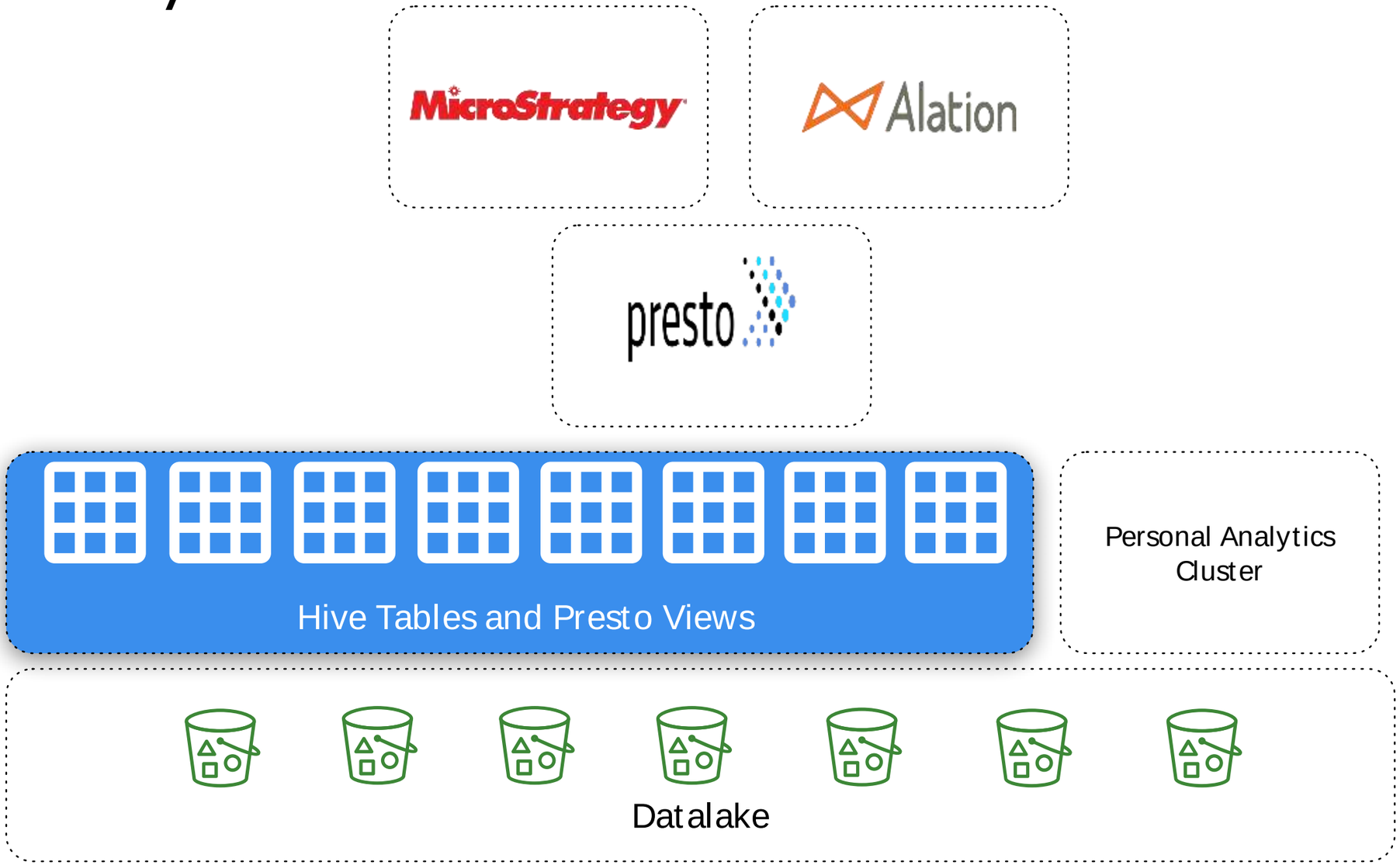
Extra slides

Data Warehouse vs. Data Platform

Centralized	Federated
Control	Autonomy
Perfection	Scale
Pull	Push
Product is Data	Product is Platform
Reporting	Reporting, Advanced Analytics, Machine Learning, etc.



Our Journey to Presto

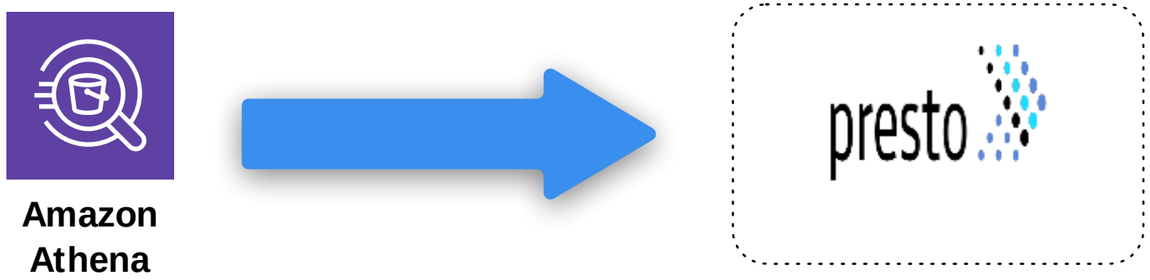


Our Journey to Presto

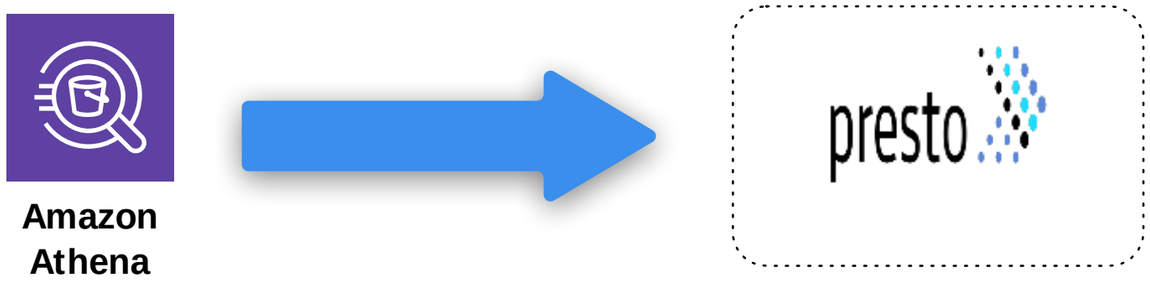
Our Journey to Presto

Amazon Athena

Our Journey to Presto

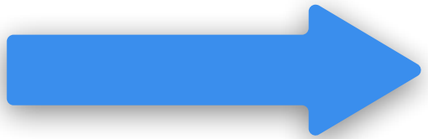


Our Journey to Presto



Our Journey to Presto

Amazon Athena



On EMR

- Cross Account Support (OneScout Hive Metastore)
- Leverages Datalake Access Roles (EMRFS)
- Scheduled Scaling Configurations
- Fits our GDPR Concept (multiple isolated Clusters)

Expectations of Data Platform Users

SCOUT24 DATA LANDSCAPE MANIFESTO

ROLES, RESPONSIBILITIES, AND VALUES FOR A DATA-DRIVEN COMPANY AT SCALE

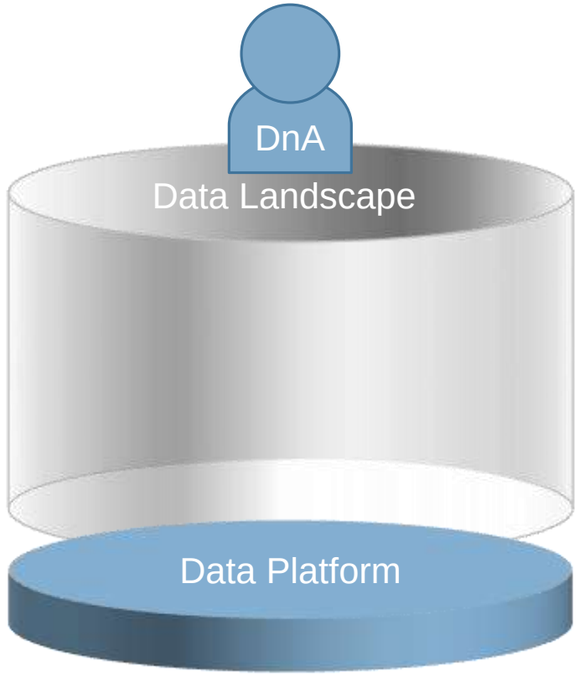
aws SUMMIT

#1 Preamble

Data is a key asset of our company.

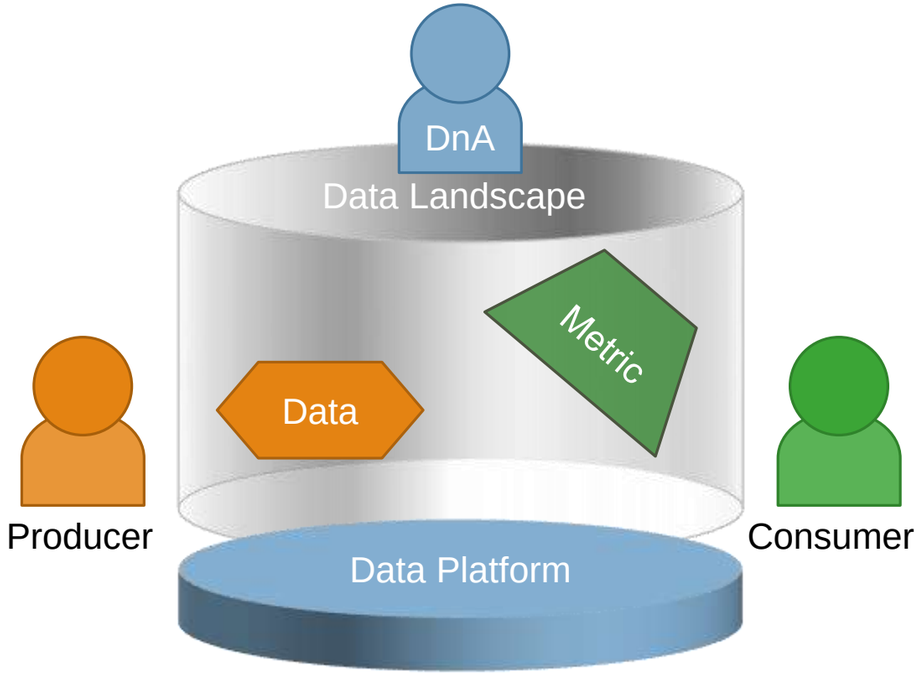
#2 Our Responsibility

We, Data & Analytics, are responsible for providing a solid Data Platform as well as clear guidelines and training how to participate in the Data Landscape.



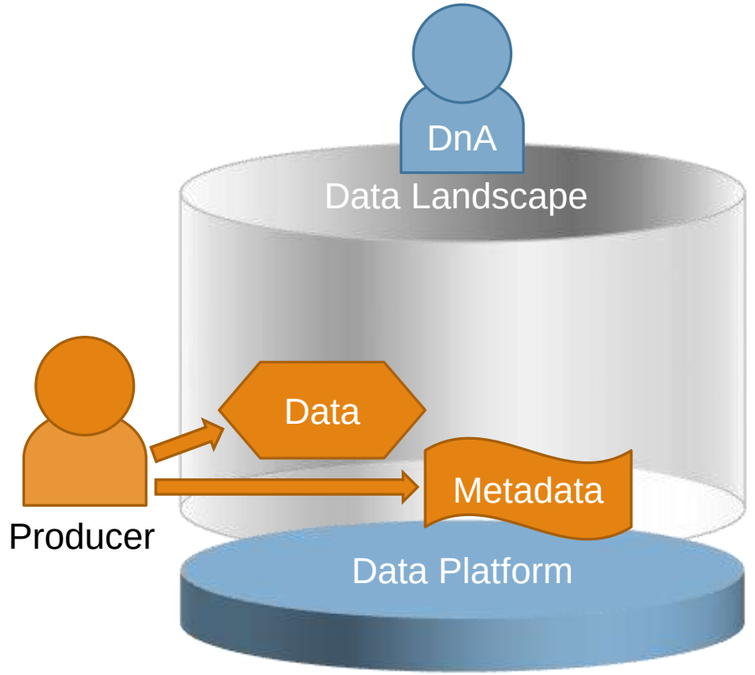
#3 Data Autonomy, Not Anarchy

Data autonomy puts data producers & data consumers in control of their data & of their metrics and thereby allows us to be data-driven at scale, but this comes with responsibility.



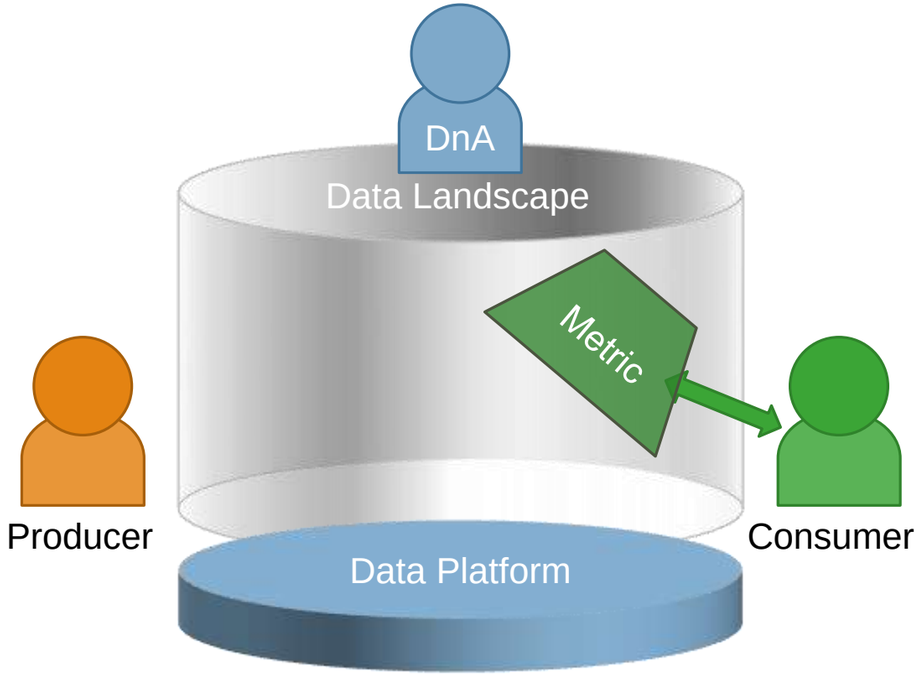
#4 Producer’s Responsibility

Data producers are responsible for publishing data to the central Data Lake, for the data's quality, and for publishing metadata that makes it easy to find and consume the data.



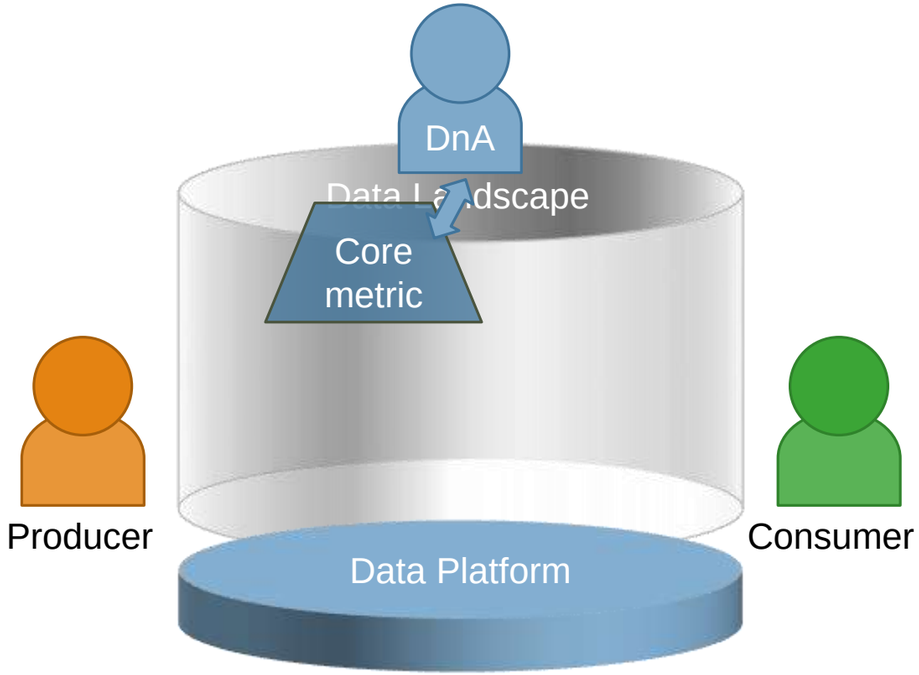
#5 Consumer's Responsibility

Data consumers are responsible for the definition & visualization of metrics and for driving the implementation and maintenance of these metrics.



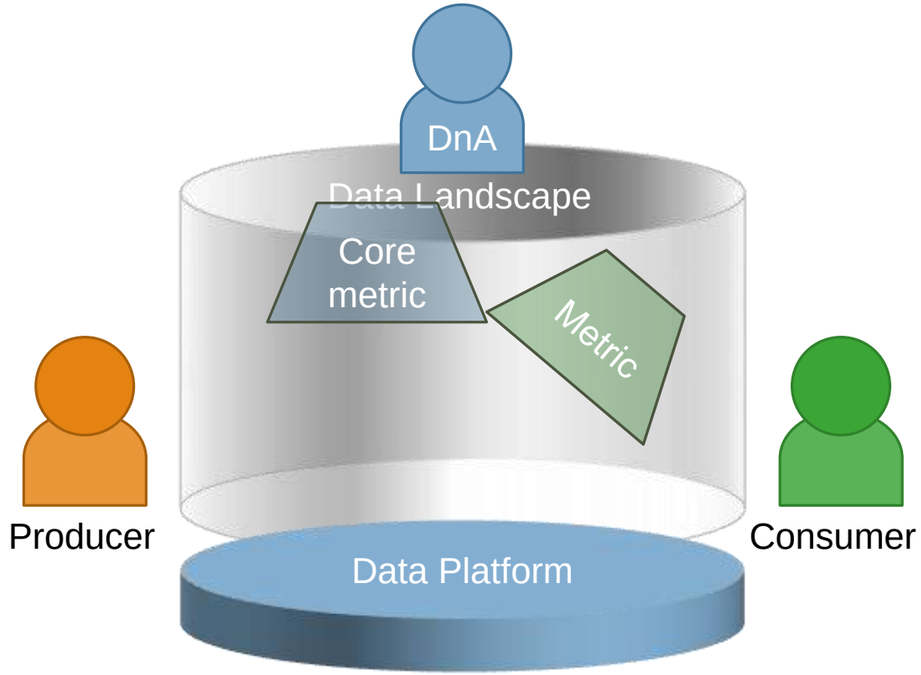
#6 Exception: Core KPIs

We, Data & Analytics, take the full ownership and responsibility of the few top company-wide core KPIs.



#7 Transparency Over Continuity

We value data transparency over data continuity, which means we may break metric comparability if it is for the cause of enabling better insights.



The Ultimate Goal

A federal landscape of data producers and consumers with just enough rules to ensure seamless co-operation without severely impeding autonomy.