T-Mobile
PROCUREMENT & SUPPLY

LESSONS IN LAKEHOUSES AND DELTA SHARING

DAIS, 2024

SYNERGIES BY THE BILLIONS

INTRO

- I’m Geoff Freeman. I’m a solution architect for the T-Mobile finance department. My team has built a lakehouse which is a pretty cool thing and I think after hearing this you will want to have one too.
- I’m Luke Barnes. I was a software engineer that built T-Mobile’s access management platform. Recently, I’m a Data Strategy Manager focused on unifying data architecture and user experience.

T-Mobile
PROCUREMENT & SUPPLY

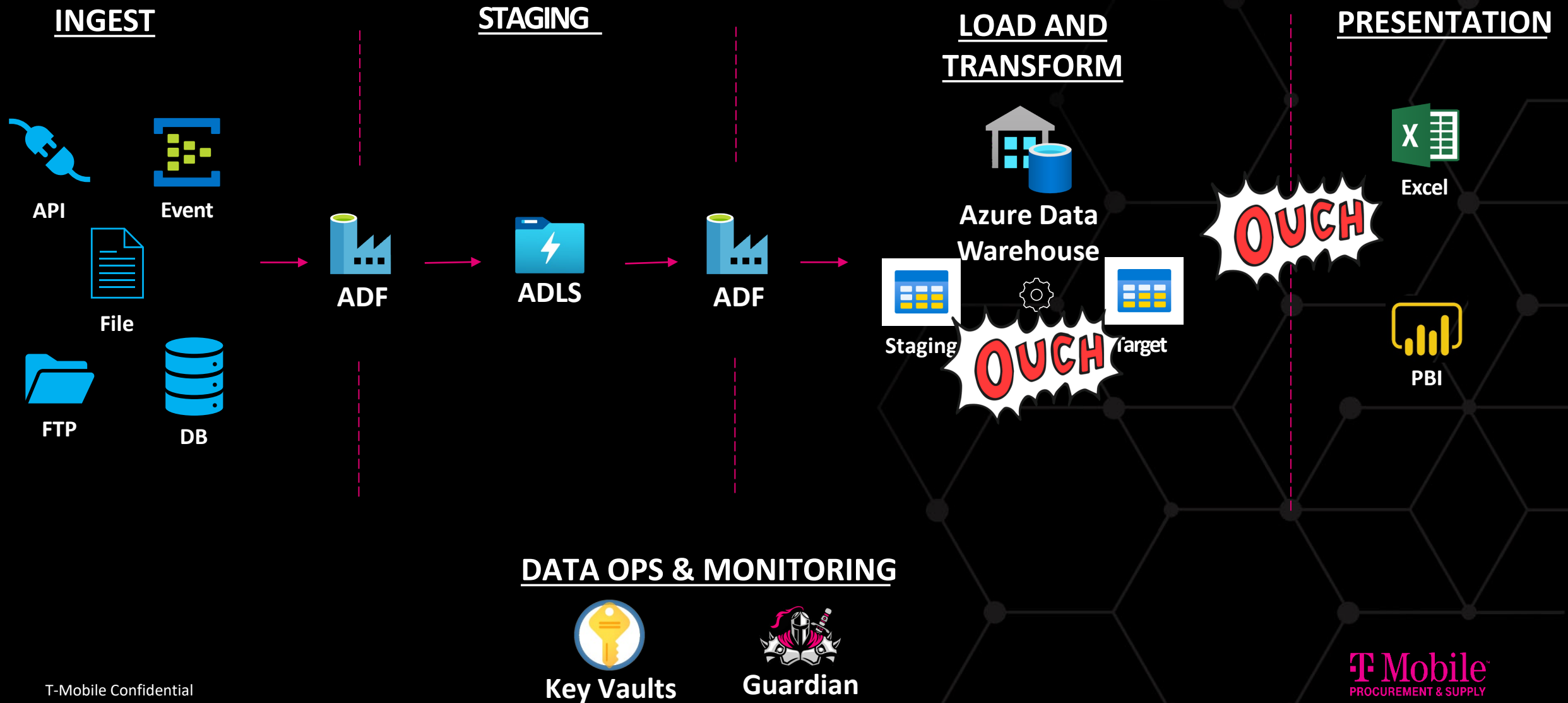
Why You’re Here

You’re interested in building a Lakehouse and/or Data Mesh, and we’ve already built one. We’ll tell you about our learnings, big lessons, and road map for the future.

- Why we built a lakehouse
- Lessons learned
- Why we built a data mesh
- How UC helped
- How Delta Sharing helps
- How we’re going to use Delta Sharing in the future

T-Mobile PROCUREMENT & SUPPLY

Legacy Architecture



The Data Strategy

Platform Vision: Scalable data management that sets the enterprise standard

Low Data Latency
The **most up to date raw and transformed data** is made available for all consumers and use cases.

High Data Mobility
Data is easily **consumed where it resides** and **democratized**. Various organizations and solutions can consume and **share data seamlessly**.

Predictable Performance
Workload isolation through serverless workloads ensures **consistent query experience,** eliminating “noisy neighbor” problem.

Economies of Scale
Ensure a **cost-effective management** of Lakehouse architecture and realize **enterprise data management efficiencies**.

T-Mobile PROCUREMENT & SUPPLY

Elevate Lakehouse Architecture

Separation of **C**hurch and **S**tate

Compute: disconnected from the data

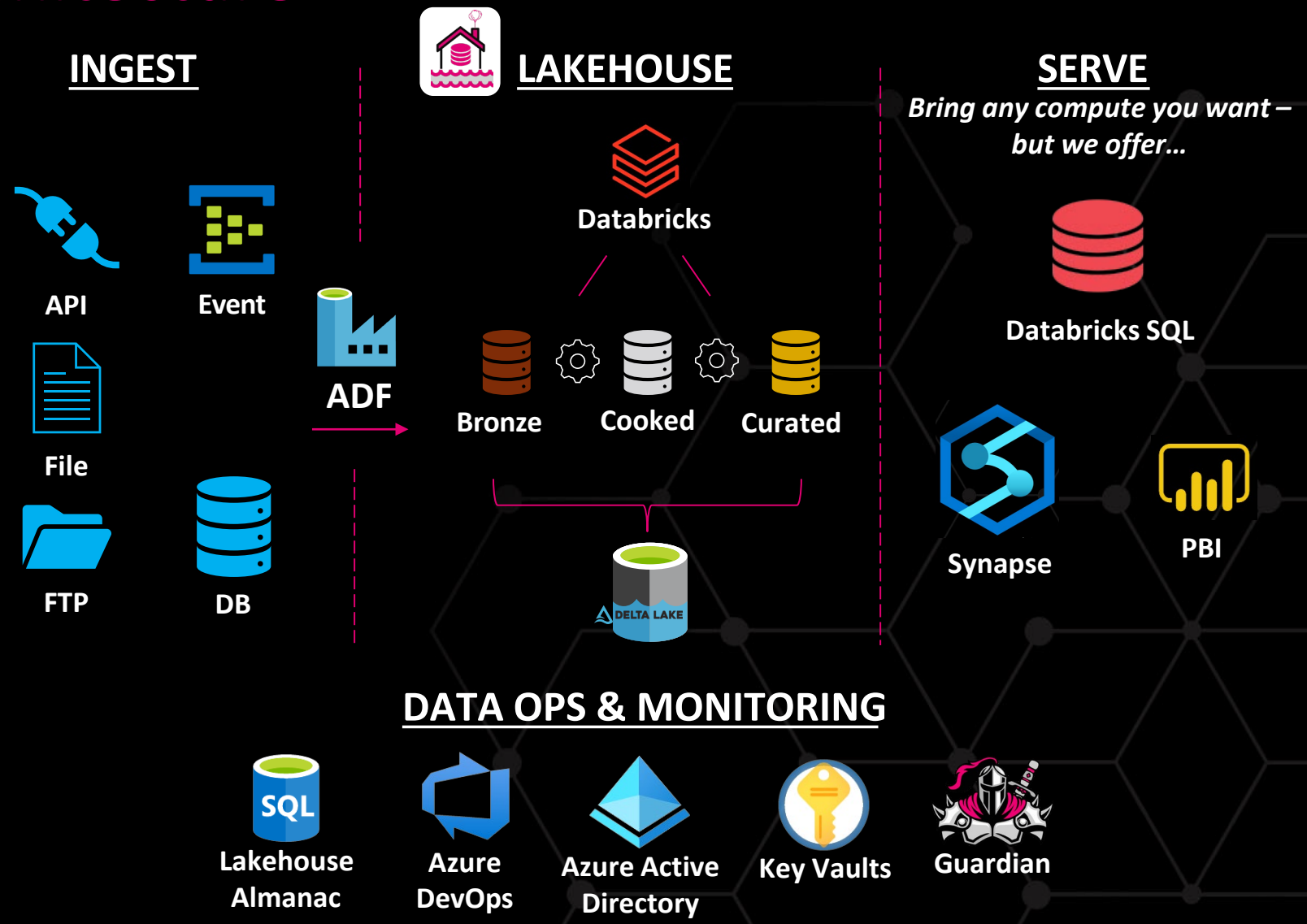
Supported Compute Religions

- Sql Server
- Spark (Databricks)
- Snowflake
- Redshift
- PowerBi/Fabric
- PrestoDb
- Big Query(coming soon)

Storage: open-source format Delta Tables

Supported Storage Provider

- Azure Data Lake Storage(ADLS)
- Google Cloud
- S3
- Oracle Cloud
- IBM Cloud
- HDFS (Hadoop)



The Data Strategy

Platform Vision: Scalable data management that sets the enterprise standard

Low Data Latency

The **most up to date raw and transformed data** is made available for all consumers and use cases.

- Refresh Cadence
  - ⬇ **75%** *Average refresh cadence from 6hrs to 1.5hrs*
- ETL Processing Time
  - ⬇ **88%** *Complex build times reduced from 4hrs to 30mins*

Predictable Performance

Workload isolation through serverless workloads ensures **consistent query experience,** eliminating "noisy neighbor" problem.

- Data Availability
  - ⬆ **1%** Data availability increase from 98% to 99%
- Query Failures
  - ⬇ **60%** Reduction in query failures from 1.5% to 0.6%

High Data Mobility

Data is easily **consumed where it resides** and **democratized**. Various organizations and solutions can consume and **share data seamlessly**.

- # of Data Connections (in and out)
  - ⬆ **30%** *55 connection points (41 in / 14 out)*
- Time to Onboard Data
  - ⬇ **80%** *Data onboarding time from 5 days to 1 day*

Economies of Scale

Ensure a **cost-effective management** of Lakehouse architecture and realize **enterprise data management efficiencies**.

- Infrastructure Spend
  - ⬇ **30%** $120k Azure spend saved per month
- Resource Efficiency
  - ⬆ **30%** 10 sources added with equivalent staff

T-Mobile PROCUREMENT & SUPPLY

Lessons Learned

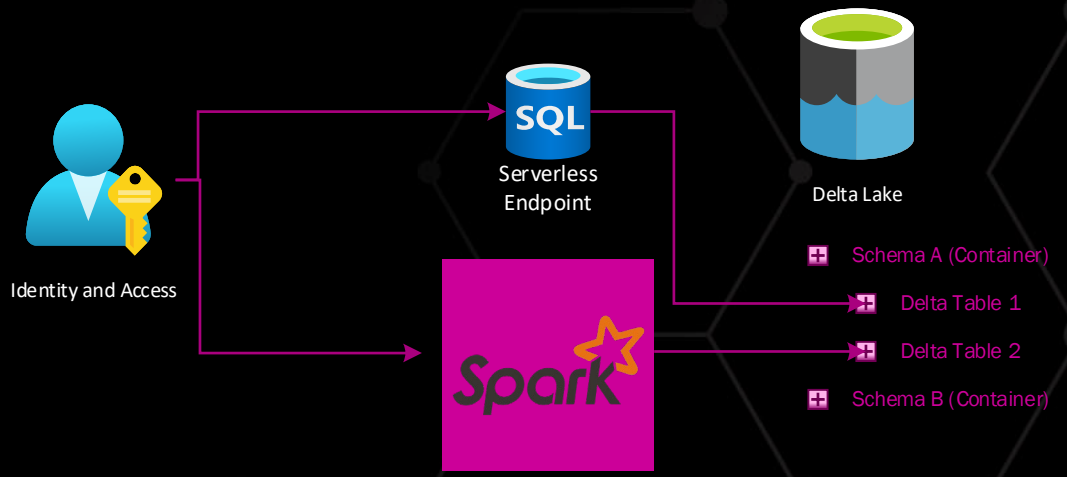
- Have a roadmap for the transition
- Automate everything
- Data engineering is software engineering
- KISS – don’t build for perfection
- Embrace commodity hardware

Compute optimization – don’t do it - at least, not yet

LAKEHOUSE ARCHITECTURE – ACCESS AND SECURITY

Security handled with **Azure Active Directory:**

- One group leveraged per security context
- Security applied at data lake layer, and all other fabrics apply security uniformly
- Requires active directory pass-through on the connection
- We call this **Conformed Security**



Access provisioned through **Provisionator** (via AAD groups and SCIM) for:

- Data access
- Compute access
  - Serverless SQL
  - Synapse Spark compute
  - Databricks SQL
- Storage
- Orchestration tools like ADF
- Code repositories

T-Mobile PROCUREMENT & SUPPLY

Provisionator – Unified Access Management

Conformed security also provides the opportunity for a unified access management experience for your customers

Product Mission

To streamline and democratize access to T-Mobile's data, so everyone can make data-driven decisions

- Automated approval & provisioning process
- Integrated into data catalogs & applications
- Auditable access logs
- Automated revocation when users leave TMO
- Timeboxed access
- One-stop-shop for access to all data at T-Mobile

T-Mobile PROCUREMENT & SUPPLY

Conformed Security

Sensibility

Technology and best practices change, so design with change in mind. Keep it simple. Adapt best practices to fit your needs. Spend time locking down things that matter most, first.

Single Source of Truth

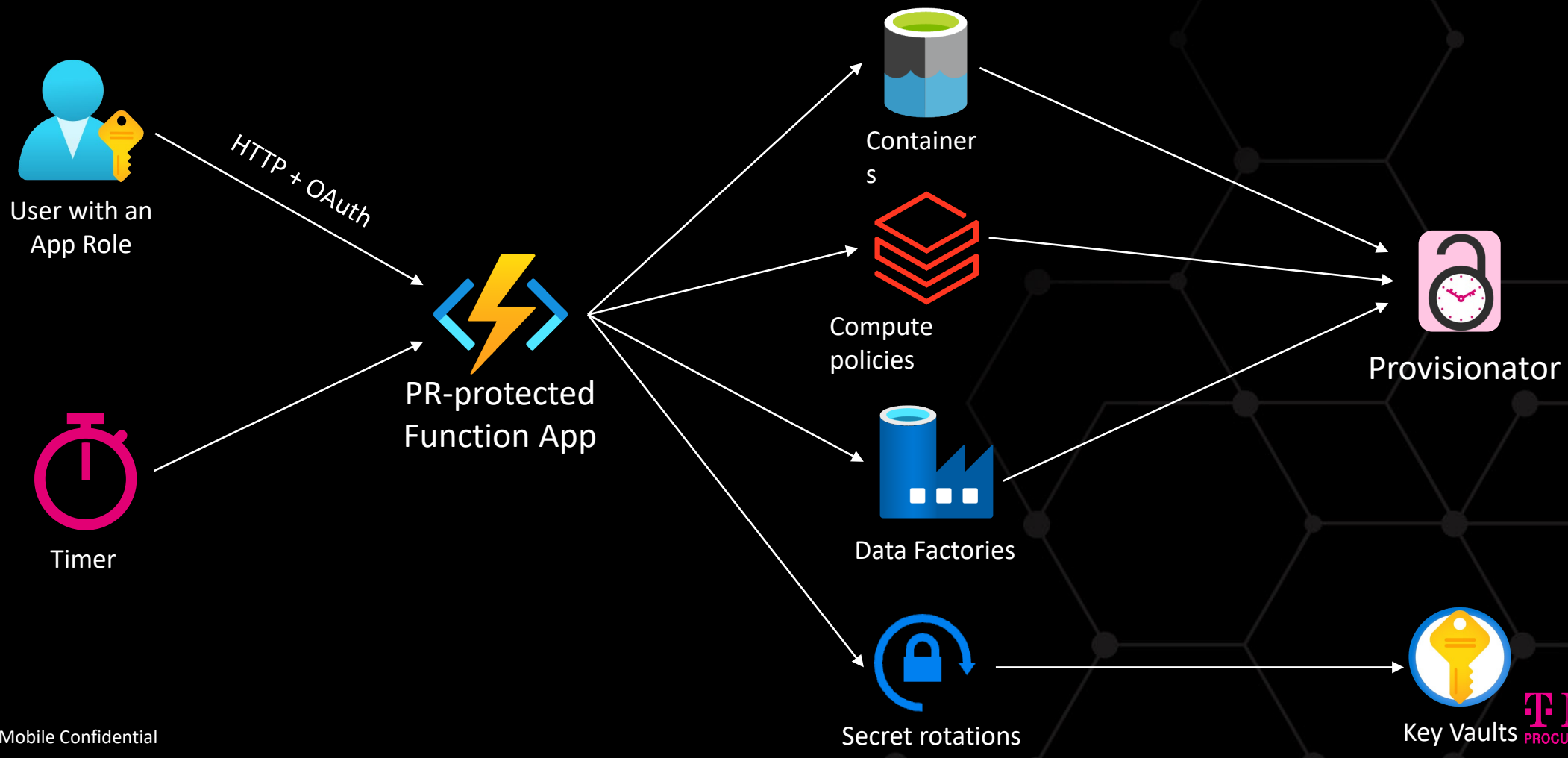
Use a single Identity Provider (IdP). Most products offer SCIM connectors so you can manage permissions in your IdP of choice. Do not nest security groups – inheritance is hard to audit.

Access Management

This is how customers will engage with your security policy, so it should be as seamless as possible. If possible, secure the data in just one place, and let the permissions flow downstream to all compute providers. If not, make sure access is conformed.

Automated Scaffolding

Because we have conformed security, any automation or tooling built for administering our Lakehouse has much more impact because it can be uniformly deployed into every data platform at the company. Some examples of what and how:



Delta Sharing is the Future

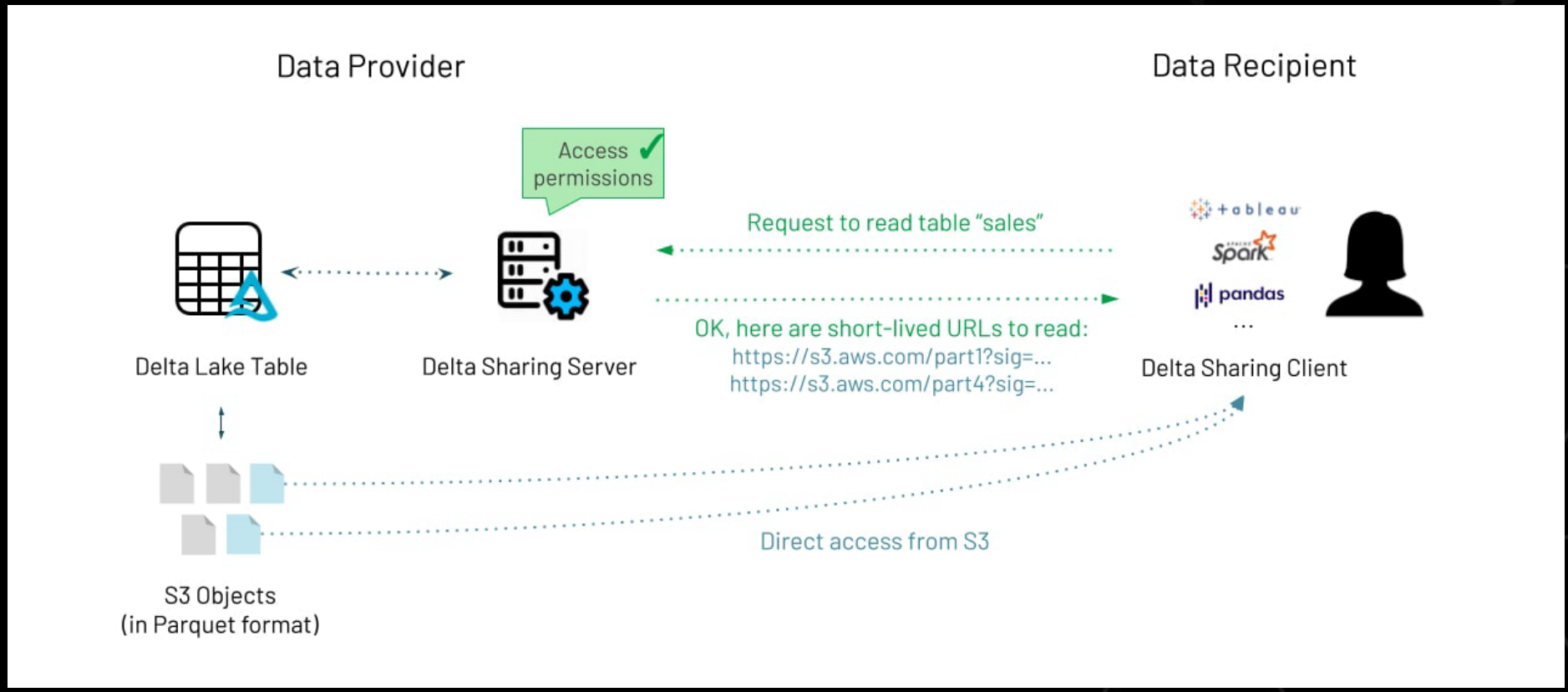
These companies have or are implementing delta sharing
- Salesforce
- SAP
- Oracle
- o9
- Scopeworker

Scopeworker

T-Mobile
PROCUREMENT & SUPPLY

How Delta Sharing Works – Conceptual Model

- Auth used varies by cloud, but uses short lived shared access signatures per file
- A delta share sends a manifest of available tables. No actual data is sent until query time
- At query time, each file gets its own URL

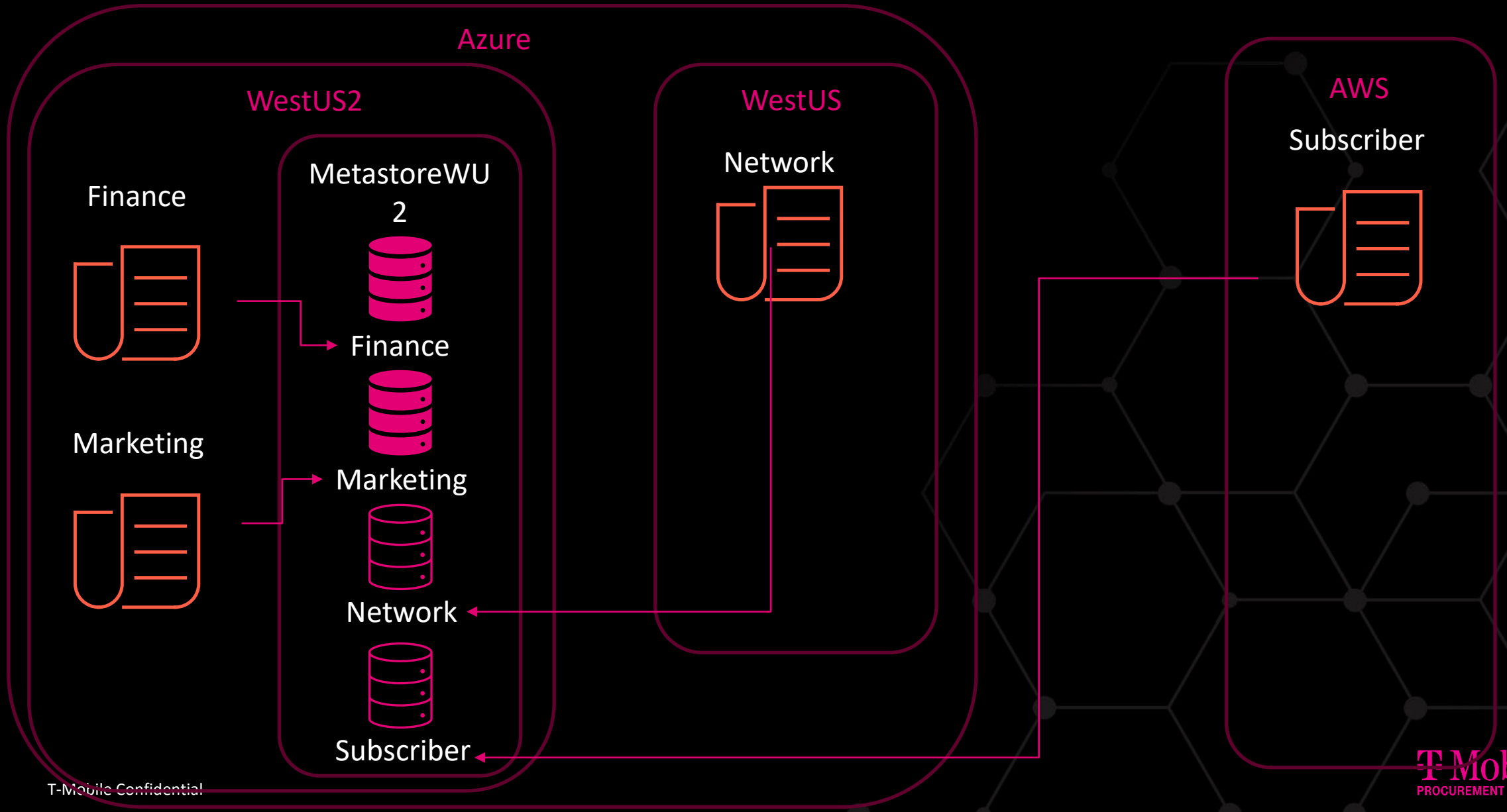


T-Mobile PROCUREMENT & SUPPLY

Benefits – The Pros of Delta Sharing

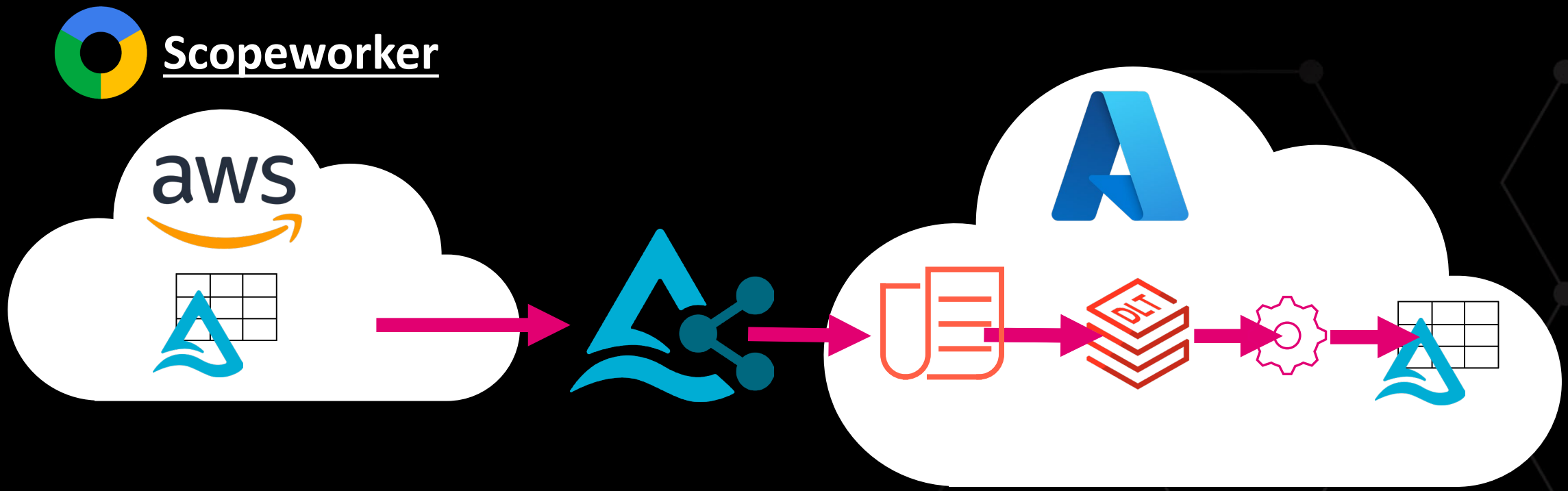
- Reduced latency
- Reduced data engineering to setup/maintain
- Reduced data movement
- Simplification of architecture

Internal Implementation



T-Mobile
PROCUREMENT & SUPPLY

ScopeWorker – Our Implementation

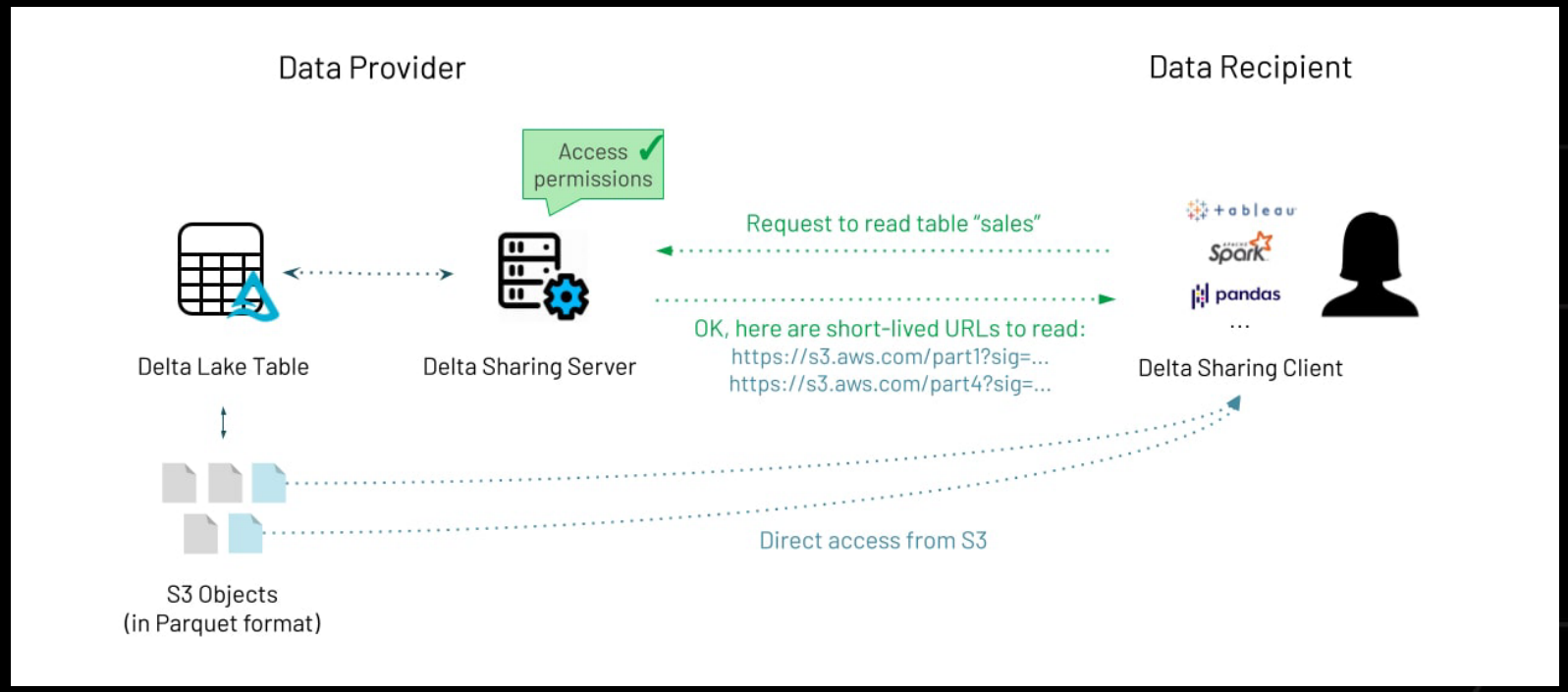


- Scopeworker shares their tables through Delta Sharing
- Delta Shares are mounted to T-Mobile Unity Catalog
- Delta Live Table runs every 5 minutes to copy across to our ADLS
- No consumers are granted direct access to the Delta Share in UC

Delta Sharing Gotchas

What features of Delta Sharing are implemented, both on the producers and consumers?

- Predicate pushdown?
- ChangeDataFeed?
- What happens if you scan the entire table?
  - Even if fully implemented, consider **`SELECT DateKey, COUNT(*) FROM {Table}`**. It'll have to read every file.



- Every file needs to be copied. Stores can generally handle it, but **will the authorization service that's creating short lived signed URLs**?
- What will egress fees look like?

Final thoughts

Data engineering is software engineering

Automate yourself out of a job