

Power BI on Databricks Best Practice



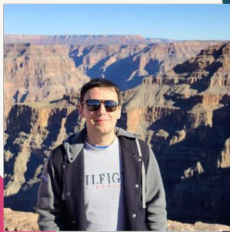
Meet the speakers



Chris Webb

Principal Program Manager @ Microsoft

<https://www.linkedin.com/in/chriswebb6/>



Marius Panga

Senior Solutions Architect @ Databricks

<https://www.linkedin.com/in/mariuspanga/>



Denny Lee

Sr. Staff Developer Advocate @ Databricks

<https://www.linkedin.com/in/dennyglee/>



Liping Huang

Senior Solutions Architect @ Databricks

<https://www.linkedin.com/in/lipinght/>



Performance



Performance–Power BI

- Think Import vs Direct Query
- Star Schema
- Composite Model/Aggregations/Hybrid Tables
- Referential Integrity
- Query Folding/Push down transformation logic
- Use Efficient DAX calculations
- Evaluation configuration settings
 - Maximum connections per data source
 - Maximum number of simultaneous evaluations
 - Maximum number of concurrent jobs
 - MaxParallismPerQuery
- Reduce number of visuals on a page
- Apply/Clear All Slicers button
- Query reduction setting



Performance–Delta Tables

- Use Delta format
- Set NOT NULL where possible
- Use Predictive Optimization (Public Preview)
- Use AUTO OPTIMIZE
- Run ANALYZE TABLE regularly
- Use Z-ordering on selective join keys or common selective query predicates
- Avoid partitioning for data < 1 TB
- Use Liquid Clustering (Public Preview)
- Avoid using wide data types to reduce model size of Power BI Dataset



Key Differentiators

Performance, Community, Reliability

1.7+
Exabytes
processed /
day

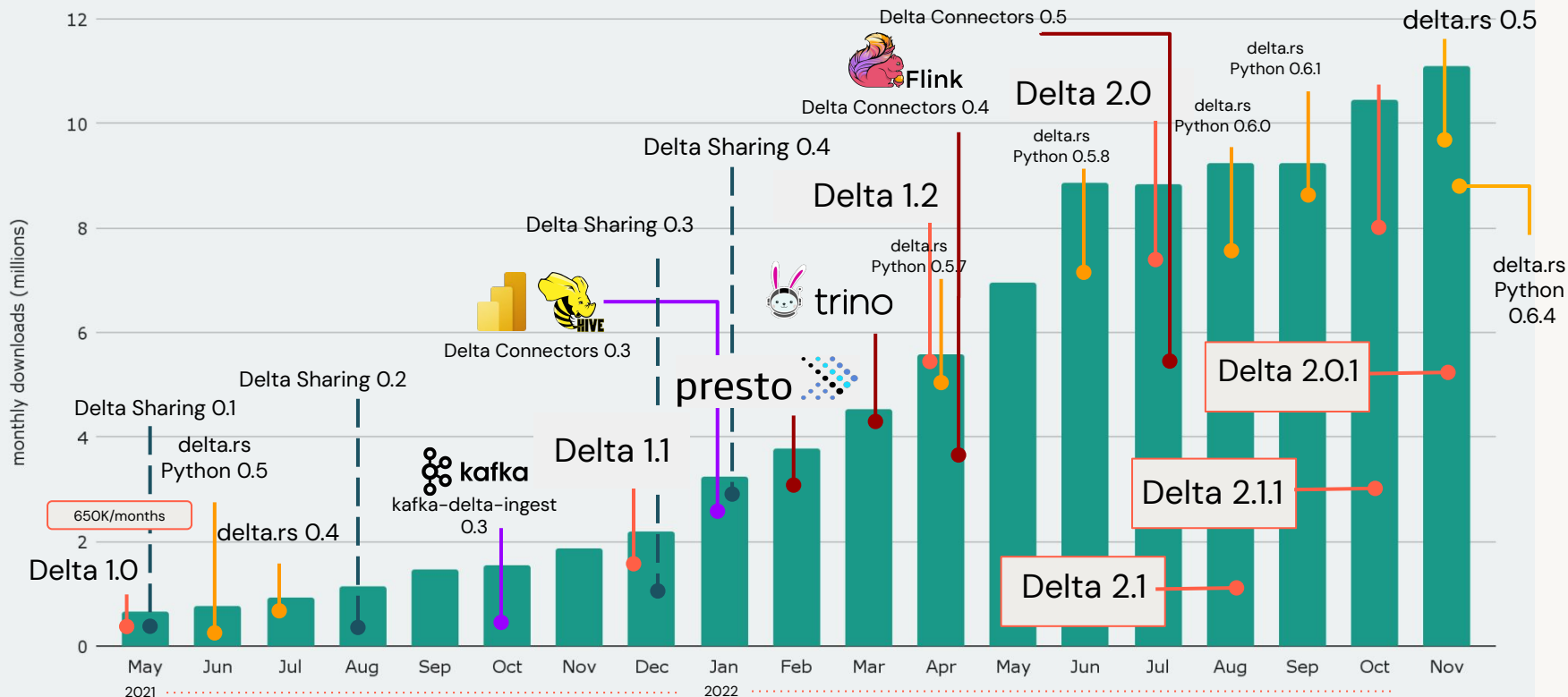
663%
Increase in
contributor
strength over last
three years

7K+
Companies in
Production



The most widely used lakehouse format in the world

11.1M downloads/month



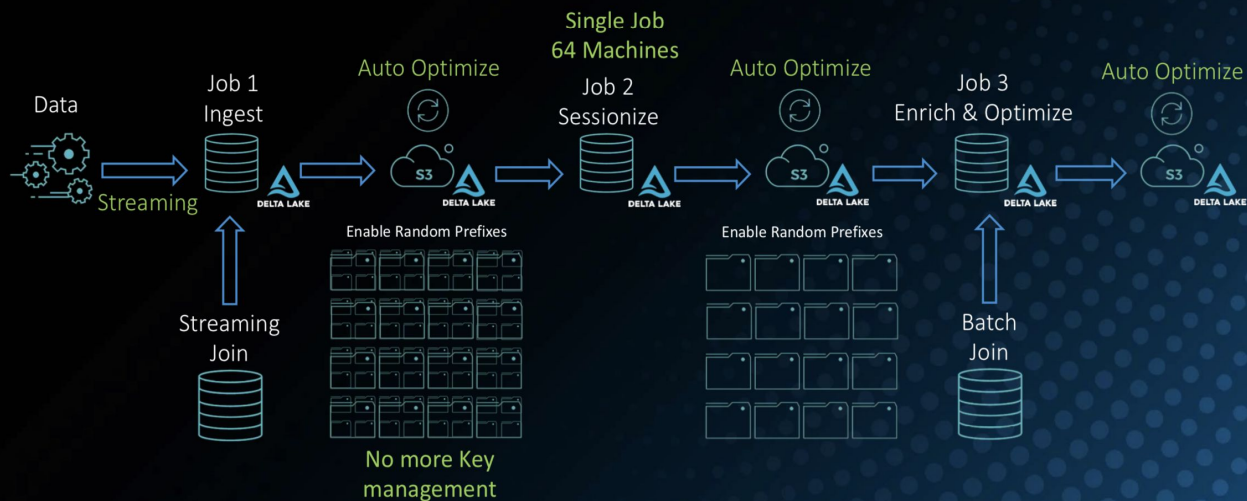


Improved reliability:
Petabyte-scale jobs

10x lower compute:
640 instances to 64!

Simpler, faster ETL:
84 jobs → 3 jobs
halved data latency

SESSIONIZATION WITH DELTA LAKE



FASTER QUERIES, RELIABLE PIPELINES, 10X REDUCTION IN COMPUTE!

14





ABN·AMRO

10x
faster TTM

100+
use cases

500+
empowered
users

Delta Lake allows ABN AMRO to create data pipelines that are not only fast but highly reliable — critical for analytics and the data science teams who rely on complete and accurate data for decision-making, analytics and model training.





Threat Detection and Response at Scale Dominique Brezinski (Apple) and Michael Armbrust (Databricks)

Databricks

Streaming ETL Telemetry and Logs

DELTA Streaming Table

DELTA Streaming Refinement

DELTA Alerts Table

Analysts

Pager Alerts

SPARK+AI SUMMIT 2018

16:57

vimeo

Spark + AI Summit 2018, Day 1 Keynote

3.6M rec/s
input rate

3.4M rec/s
processing rate

> 100TB new data / day
>300B events/day

Most queried table
1,149,012,553,409 rows

“We have been running our entire system writing out hundreds of TB of data a day on Delta Lake since the very beginning”

Dominique Brezinski, Distinguished Engineer/Director, Apple



Performance–Databricks SQL

- Use SQL Warehouse instead of All-purpose compute
- Use SQL Serverless
- Scale out for concurrency. Consider auto-scaling
- Scale up for larger datasets. Start with Medium for most BI workloads
- Set Auto stop
- Utilize the query history in the UI and the System table
- Pushdown calculations to Databricks SQL
- Can use multiple SQL Warehouses for isolation of workload or BU
- Use Materialized Views
- Use Lakehouse Federation

Databricks SQL Serverless

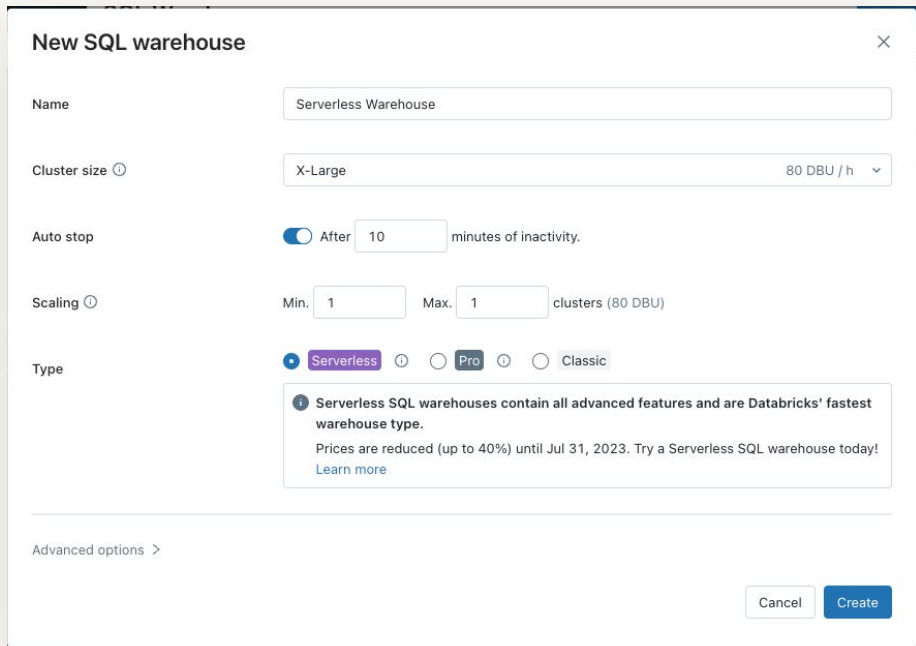
Instant, scalable compute for all DW/BI workloads

Get best performance, lower costs,
and focus on delivering value rather
than managing infrastructure.

Instant, elastic compute decoupled from storage

Eliminate management overhead

Lower TCO with AI powered optimizations



The screenshot shows the 'New SQL warehouse' configuration interface. It includes fields for 'Name' (Serverless Warehouse), 'Cluster size' (X-Large, 80 DBU/h), 'Auto stop' (After 10 minutes of inactivity), and 'Scaling' (Min. 1, Max. 1 clusters). The 'Type' section shows 'Serverless' as the selected option, with 'Pro' and 'Classic' as alternatives. A note states: 'Serverless SQL warehouses contain all advanced features and are Databricks' fastest warehouse type. Prices are reduced (up to 40%) until Jul 31, 2023. Try a Serverless SQL warehouse today! Learn more'. At the bottom, there is an 'Advanced options' link and 'Cancel' and 'Create' buttons.

Materialized View

```
CREATE MATERIALIZED VIEW customer_orders
AS
SELECT
  customers.name,
  sum(orders.amount),
  orders.orderdate
FROM orders
  LEFT JOIN customers ON
    orders.custkey = customers.c_custkey
GROUP BY
  name,
  orderdate;
```



Results are
pre-computed and
incrementally
refreshed

customers
(Table)



orders
(Table)



A type of view that pre-computes and stores the results of a SQL query and keeps them fresh over time.

Benefits:

1. **Accelerate BI dashboards.** Much faster to query data that is pre-computed vs querying base tables.
2. **Reduce data processing costs.** MV results are refreshed incrementally avoiding the need to completely rebuild the view when new data arrives.
3. **Improve data access control.** More tightly govern what data can be seen by consumers by controlling access to base tables.

Lakehouse Federation

Unify your entire data estate with lakehouse

Discover, query, and govern all your data,
no matter where it lives

- **Unified view** into all your data
- **Unified engine** for all your data and use cases
- **Unified governance** across all data sources



Security



Authentication

Authentication Methods

Username/Password

Personal Access Token

OAuth

Power BI Desktop vs. Service

Username/Password vs. Basic

Personal Access Token vs. Key

AAD(OIDC) vs. OAuth2

Tips

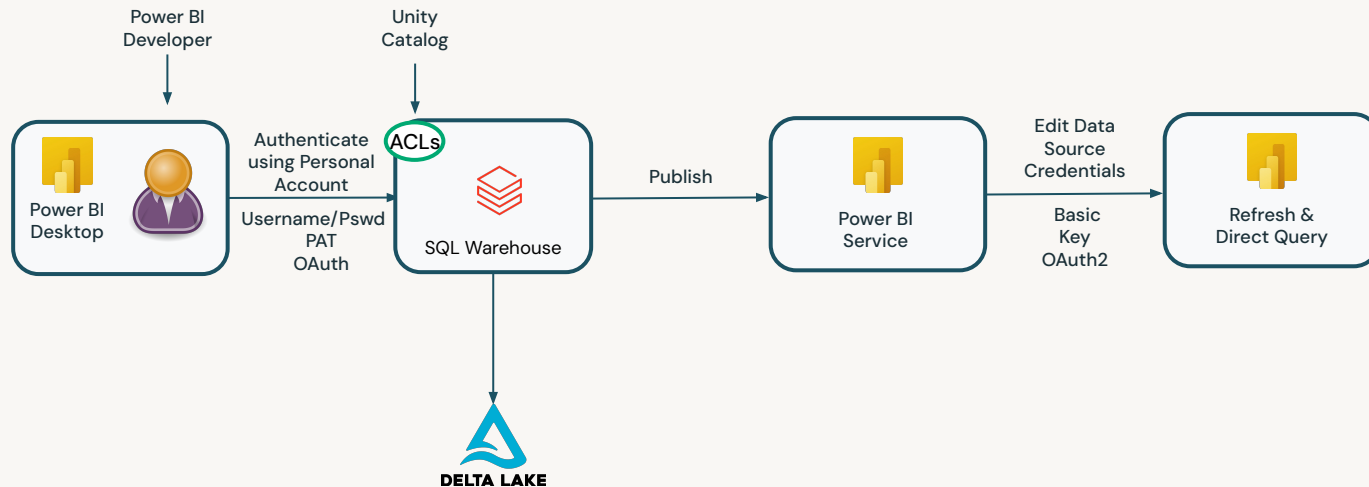
Does not support MFA

Requires key rotation

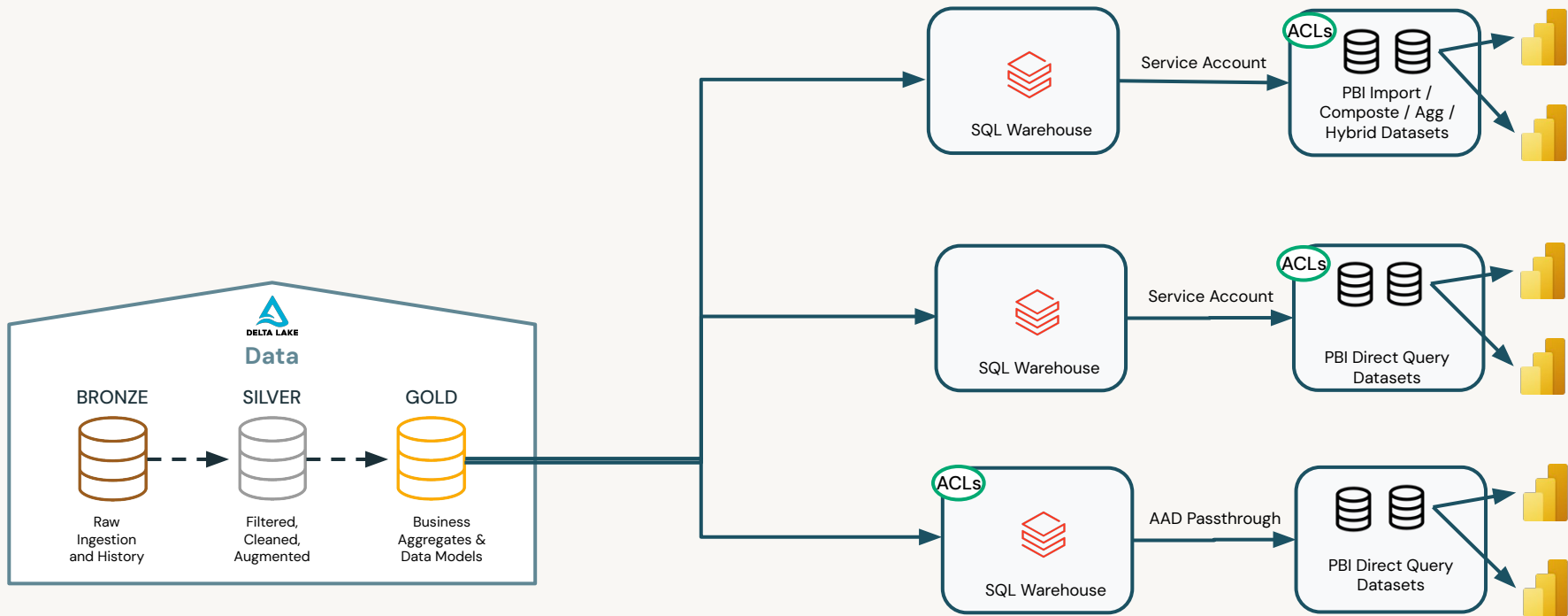
Databricks on AWS requires
Account SSO for AAD passthrough



Authentication in Authoring Process



Access Control Patterns



Unity Catalog Security to Power BI

on Direct Query with AAD Passthrough

Tables Access

Column Masks

Row Filters

Dynamic Views



Networking Security

If you have enabled **Front End Private Link** or **IP Access List**, you will need a gateway

Vnet Gateway

Newly GA

Managed solution

Premium-only feature

On-premise Data Gateway

Needs to be set-up and maintained

Hosted on VMs

Controlled by you



Thank you!

Please give us feedback



https://bit.ly/db_pbi_webinar



Resources

- [Use DirectQuery in Power BI Desktop](#)
- [DirectQuery model guidance](#)
- [About using DirectQuery in Power BI](#)
- [Connect Power BI to Azure Databricks](#)
- [DirectQuery optimisation scenarios with the Optimize ribbon](#)
- [Query Catching](#)
- [Z-Order](#)
- [Analyze Table](#)
- [Auto Optimize](#)
- [Feature Comparison Databricks SQL Warehouse Types](#)