STRATEGIC PARTNER

Microsoft

GOLD SPONSOR

elitmind
think bright

Future Processing

SILVER SPONSOR

dbWatch
DATABASE CONTROL

nationale nederlanden

BRONZE SPONSOR

predica.

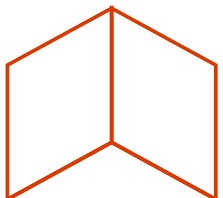
Data Community

Azure Databricks 101

Microsoft® MVP Most Valuable Professional

Kamil Nowiński

Principal Microsoft Consultant



SQL Player
Play with data & have fun!

SQL Day

Kamil Nowiński

Microsoft Data Platform **MVP**
Speaker, blogger, data enthusiast
Principal Microsoft Consultant at Altius (www.altiusdata.com)
Almost 20 yrs experience as DEV/BI/(DBA)
Member of the Data Community PL
Project member of „SCD Merge Wizard"
Founder of blog SQLPlayer (www.SQLplayer.net)

SQL Server Certificates:
MCITP, MCP, MCTS, MCSA, MCSE Data Platform,
MCSE Data Management & Analytics
Moreover: Bicycle, Running, Digital photography
@NowinskiK, @SQLPlayer

Blog

SQL Player
Play with data & have fun!

www.SQLPlayer.net

- Technical posts
- Various skill level
- Cheet sheets
- Recommended books
- Many useful other links
- Interviews (Podcast)
- YouTube Channel

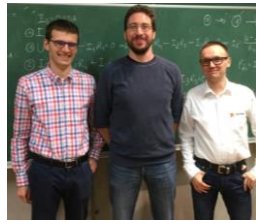
Scan me

“Ask SQL Family” #podcast

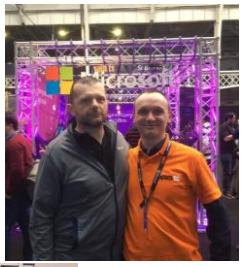


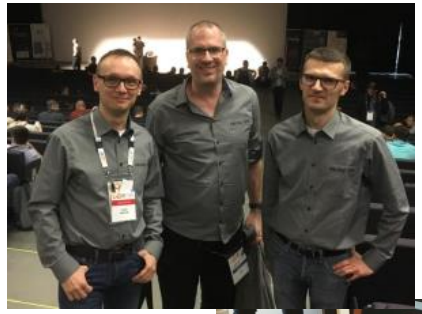
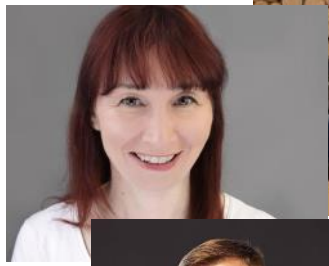


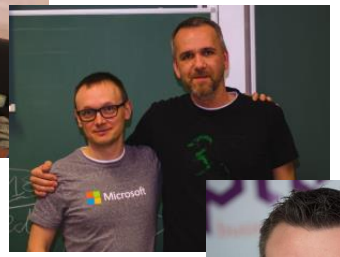




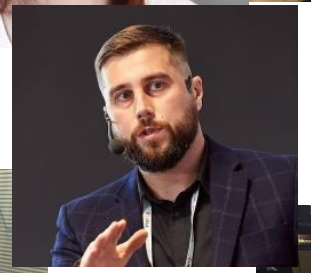


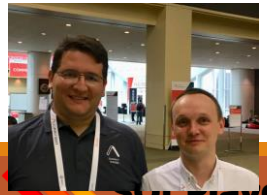










Scan me

YouTube

www.SQLPlayer.net/YouTube

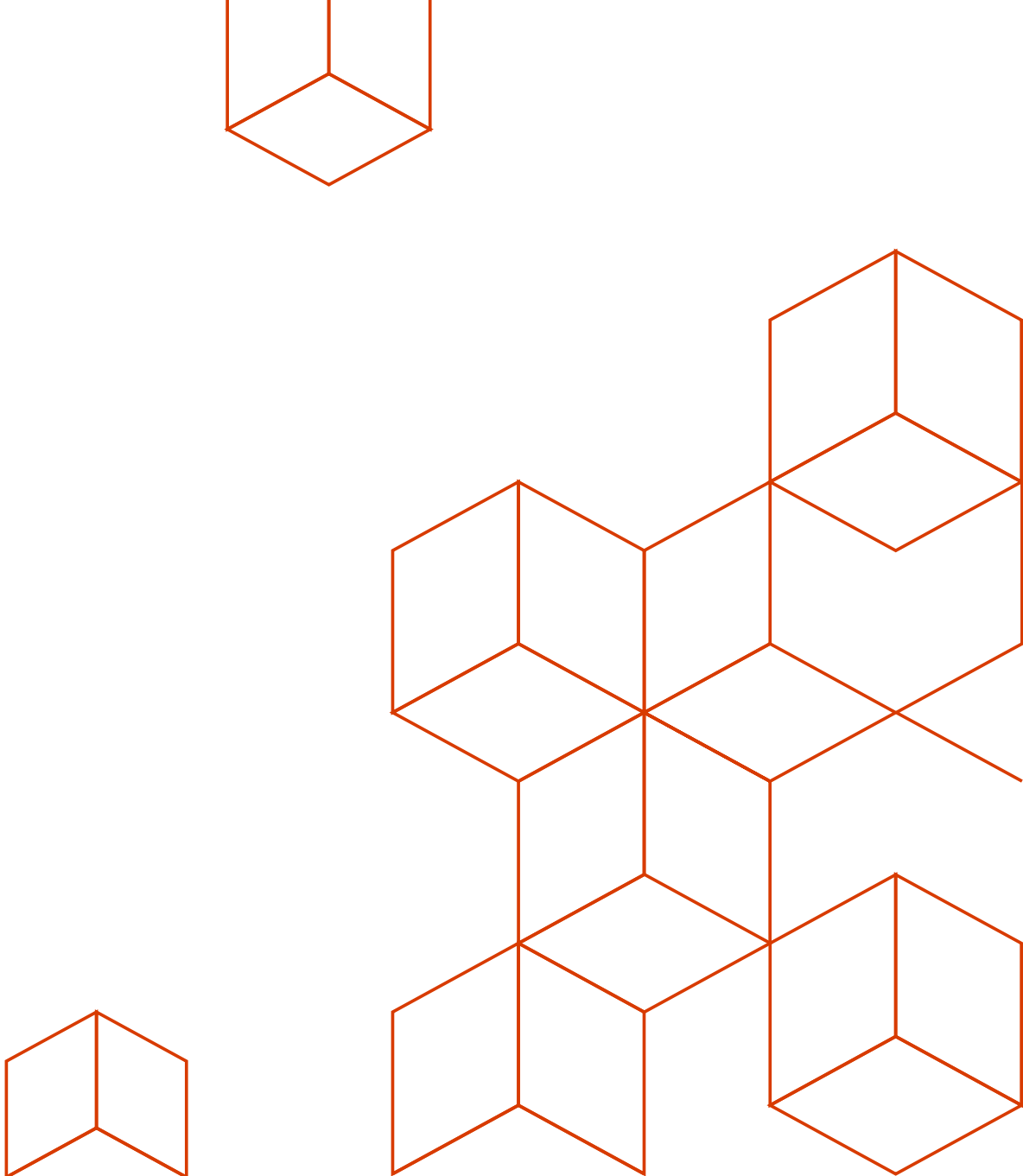
Slides available

github.com
/NowinskiK/CommunityEvents

GitHub

- github.com/sqlplayer/
  - SCD-Merge-Wizard
  - DataScriptWriter
  - azure.datafactory.tools
  - azure.datafactory.devops

Azure Databricks



What is Azure Databricks?

Microsoft

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure

databricks

Best of Databricks

+

Microsoft

Best of Microsoft

Designed in collaboration with the founders of Apache Spark

One-click set up; streamlined workflows

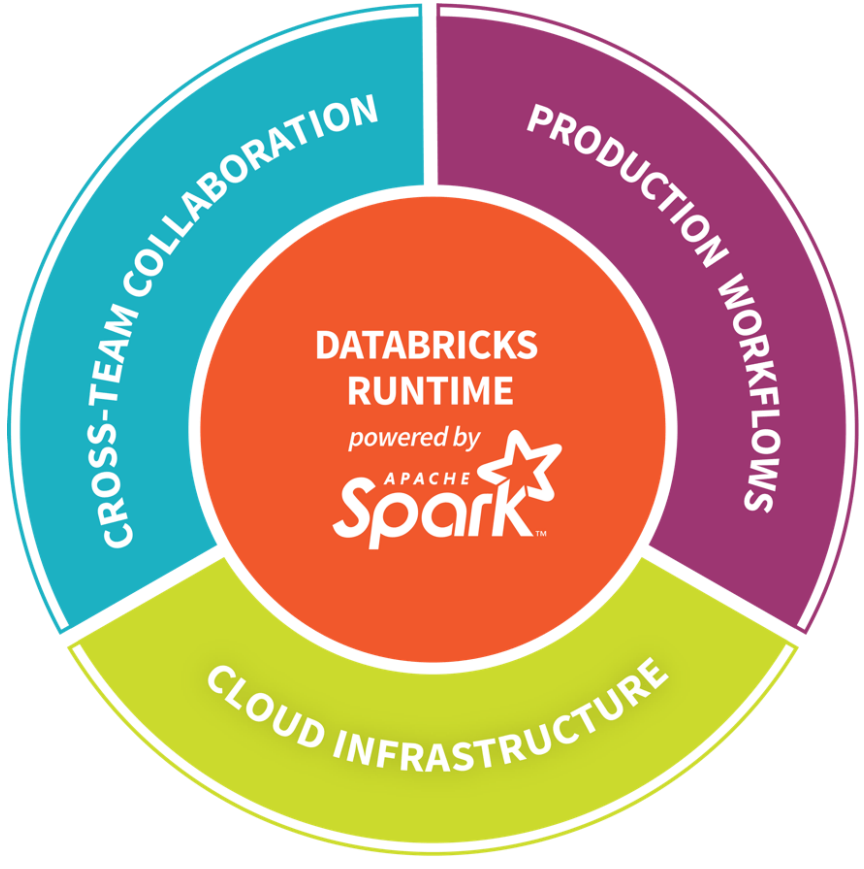
Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)

Enterprise-grade Azure security (Active Directory integration, compliance, enterprise -grade SLAs)

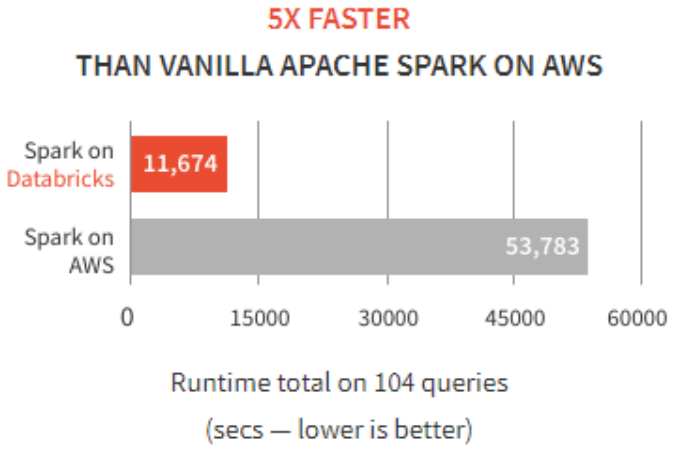
DATABRICKS - COMPANY OVERVIEW

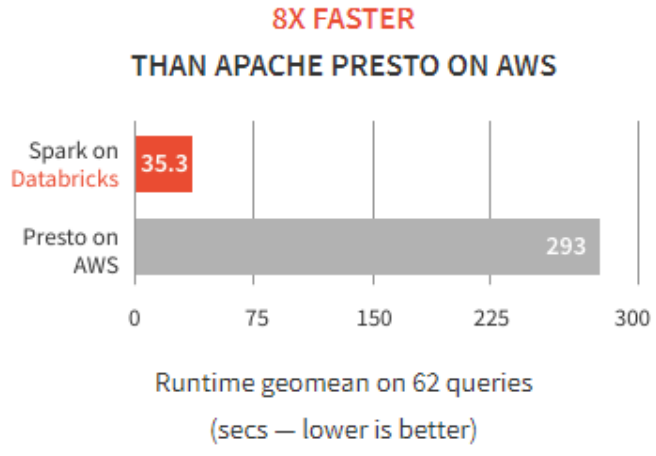
- Founded in late 2013
- By the creators of Apache Spark, original team from UC Berkeley
- Largest code contributor code to Apache Spark
- Main Product: The Unified Analytics Platform

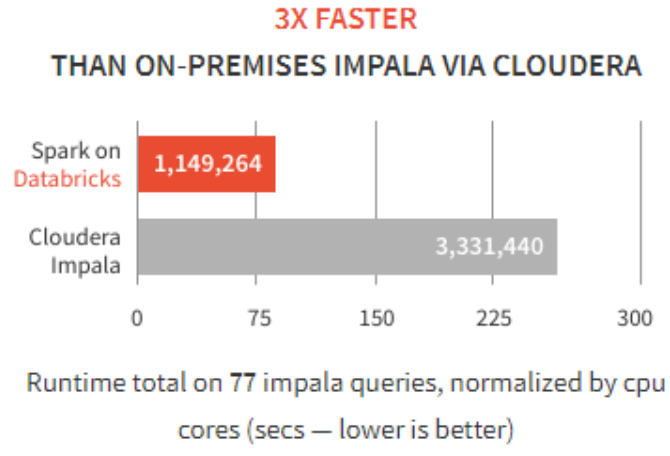


DATABRICKS SPARK IS FAST

Benchmarks have shown Databricks to often have better performance than alternatives







SOURCE: Benchmarking Big Data SQL Platforms in the Cloud

Why Spark?

- Open-source data processing engine built around speed, ease of use, and sophisticated analytics
- In memory engine that is up to 100 times faster than Hadoop
- Largest open-source data project with 1000+ contributors
- Highly extensible with support for Scala, Java and Python alongside Spark SQL, GraphX, Streaming and Machine Learning Library

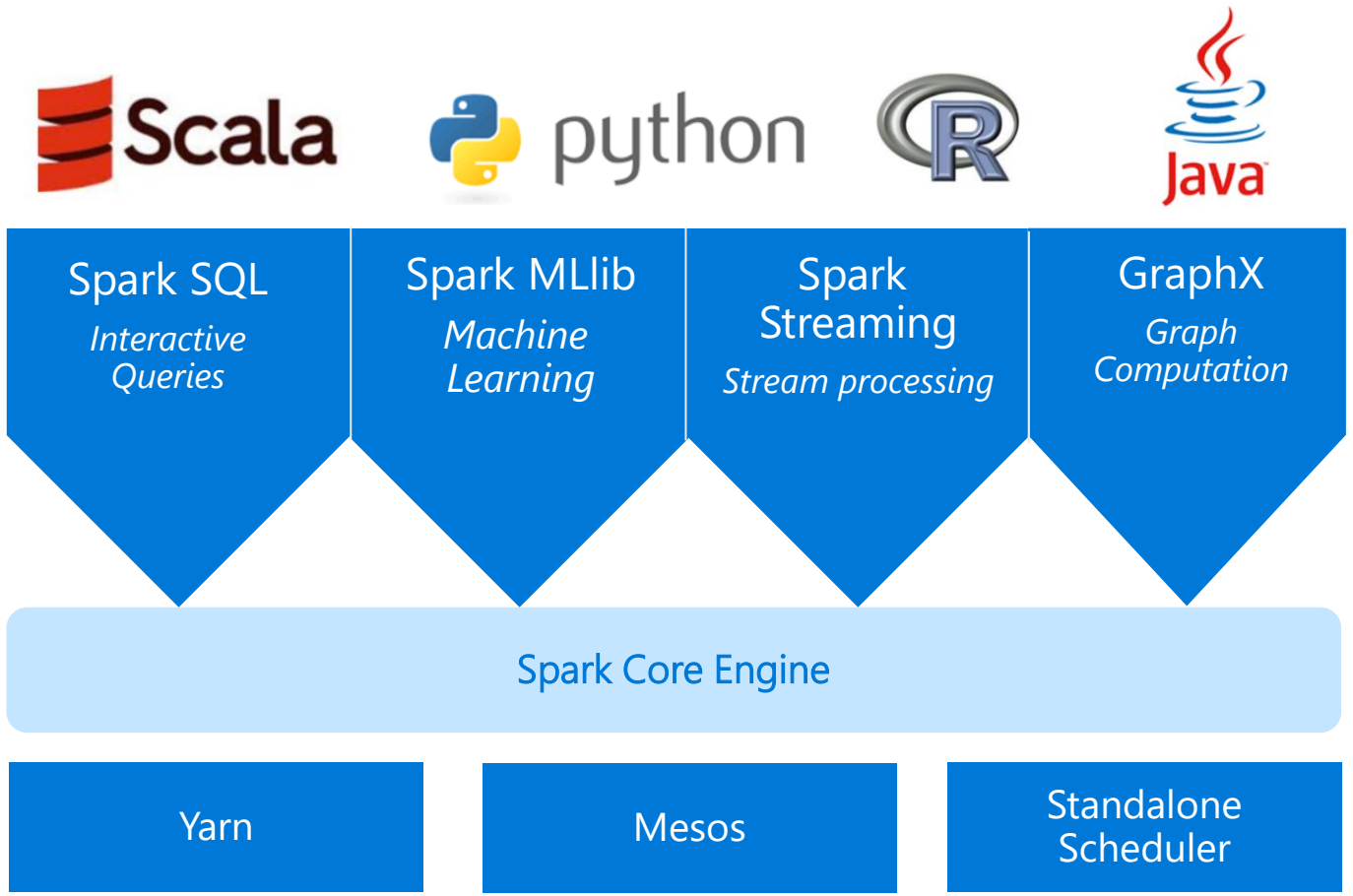
Microsoft

APACHE SPARK

A **unified**, **distributed, open source** engine for large-scale data processing

Spark Unifies:

- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing



AZURE DATABRICKS

Enhance Productivity

IoT / streaming data

Cloud storage

Data warehouses

Hadoop storage

Build on secure & trusted cloud

Azure Databricks

Collaborative Workspace

DATA ENGINEER ⟷ DATA SCIENTIST ⟷ BUSINESS ANALYST

Deploy Production Jobs & Workflows

MULTI-STAGE PIPELINES

JOB SCHEDULER

NOTIFICATION & LOGS

Optimized Databricks Runtime Engine

DATABRICKS I/O

APACHE SPARK

SERVERLESS

Rest APIs

Scale without limits

Machine learning models

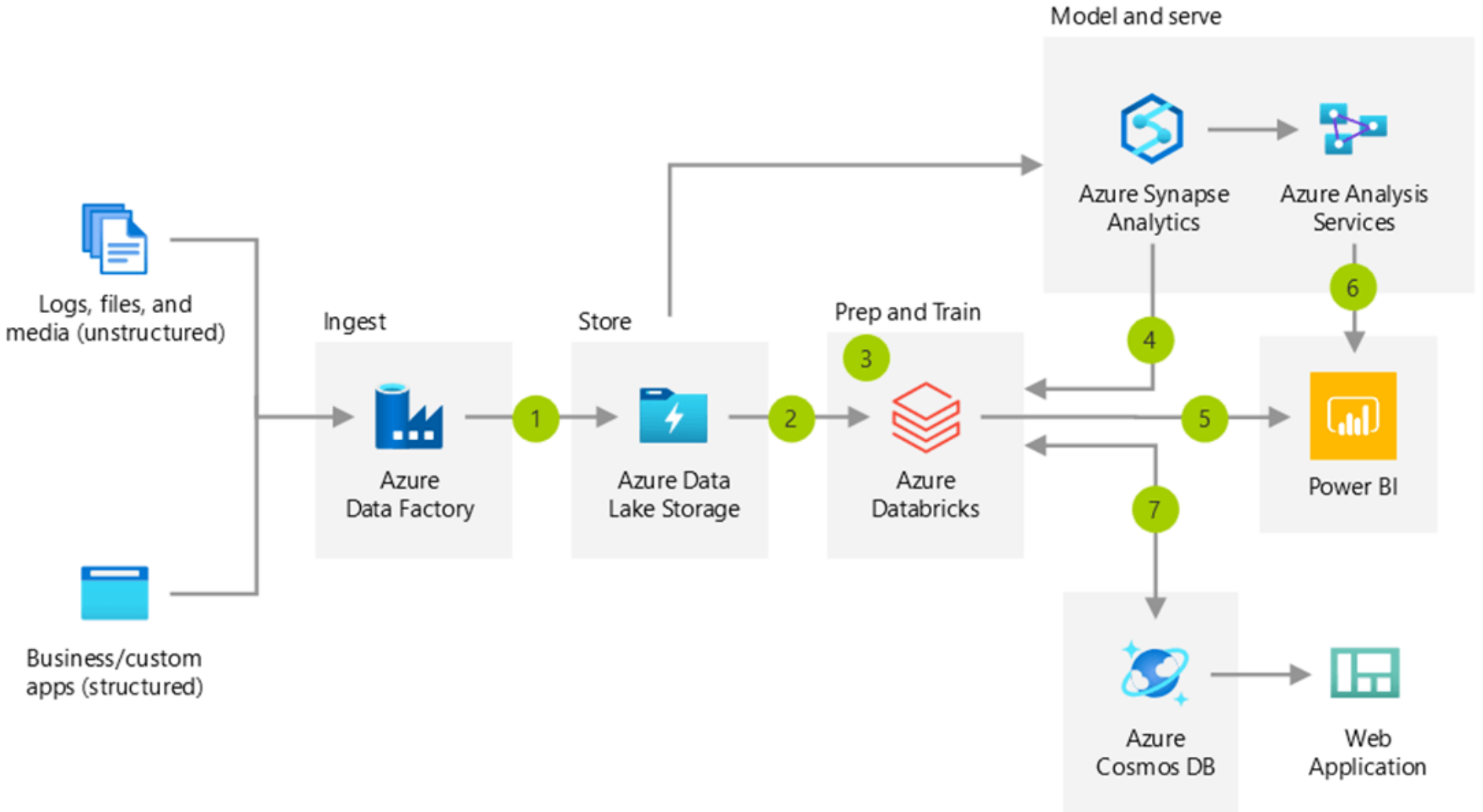
BI tools

Data exports

Data warehouses

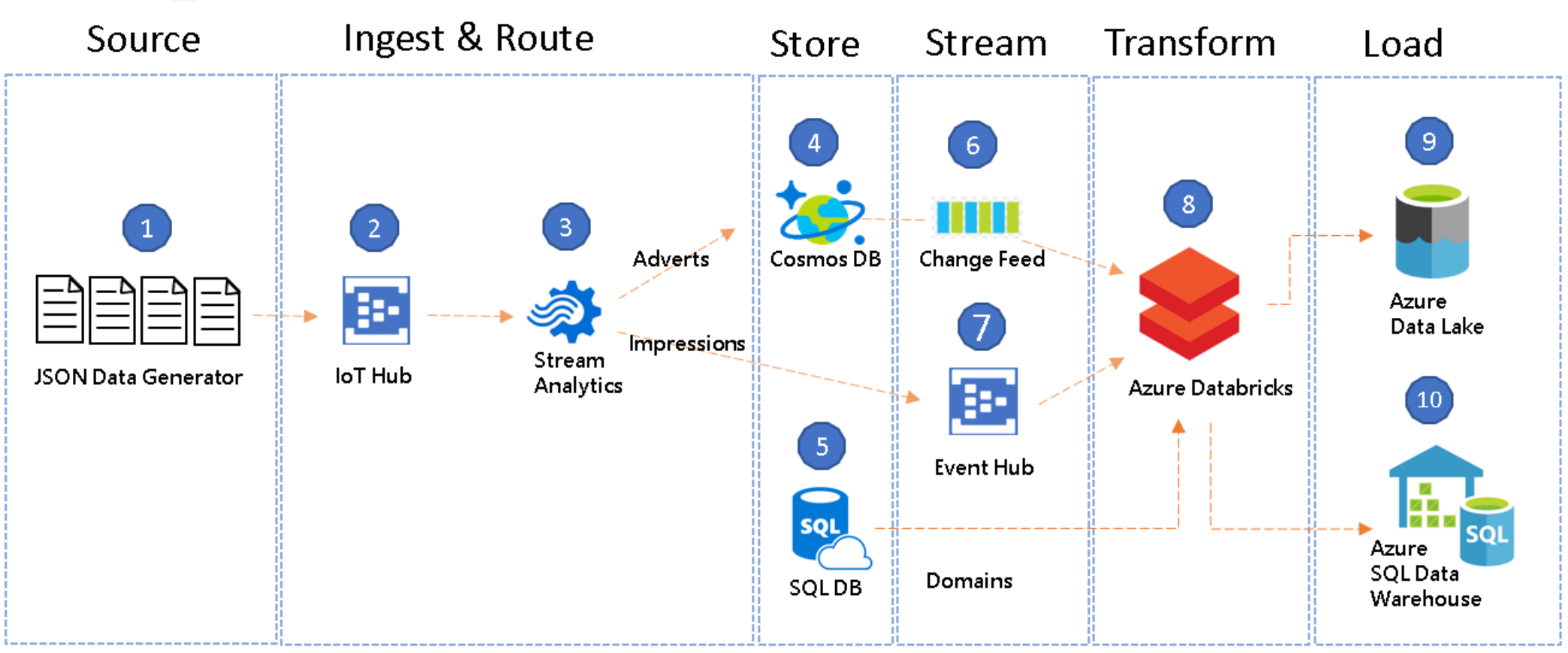
Azure

ARCHITECTURE EXAMPLE



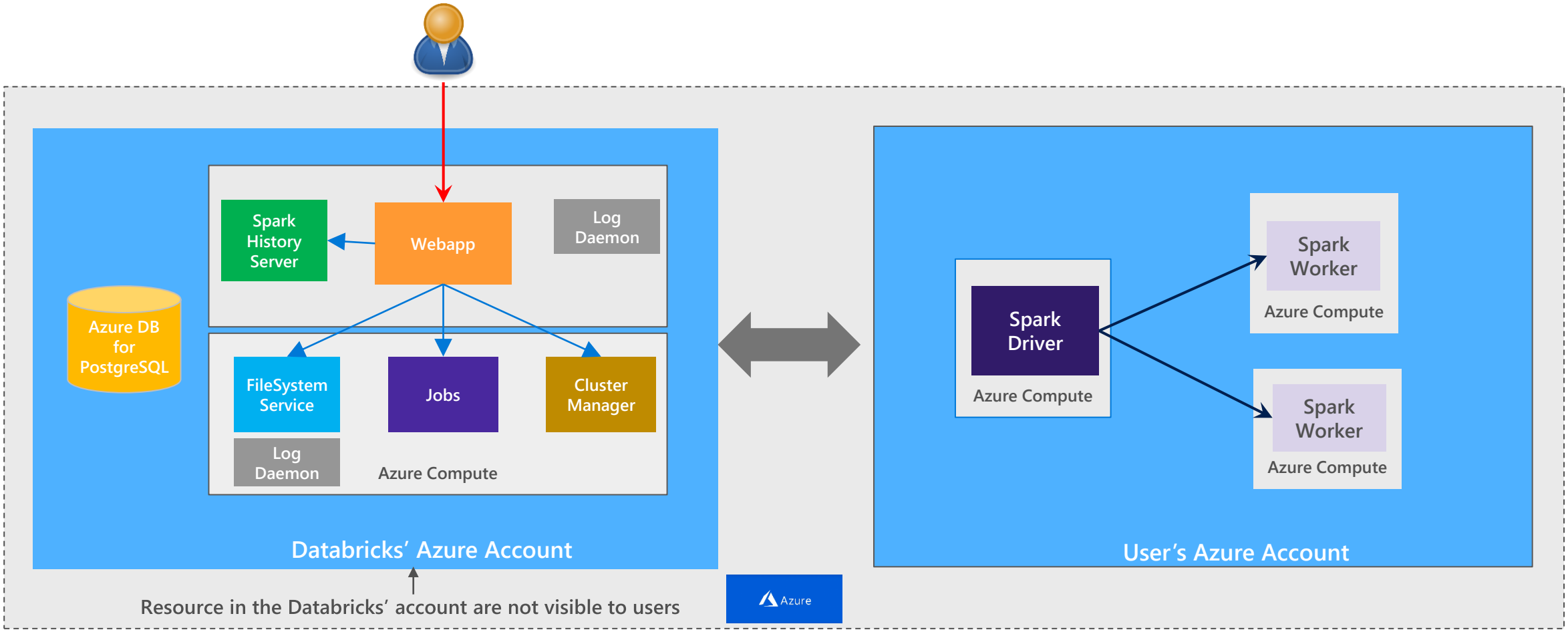
https://databricks.com/blog/2020/08/18/how-to-accelerate-your-etl-pipelines-from-18-hours-to-as-fast-as-5-minutes-with-azure-databricks.html

ARCHITECTURE EXAMPLE

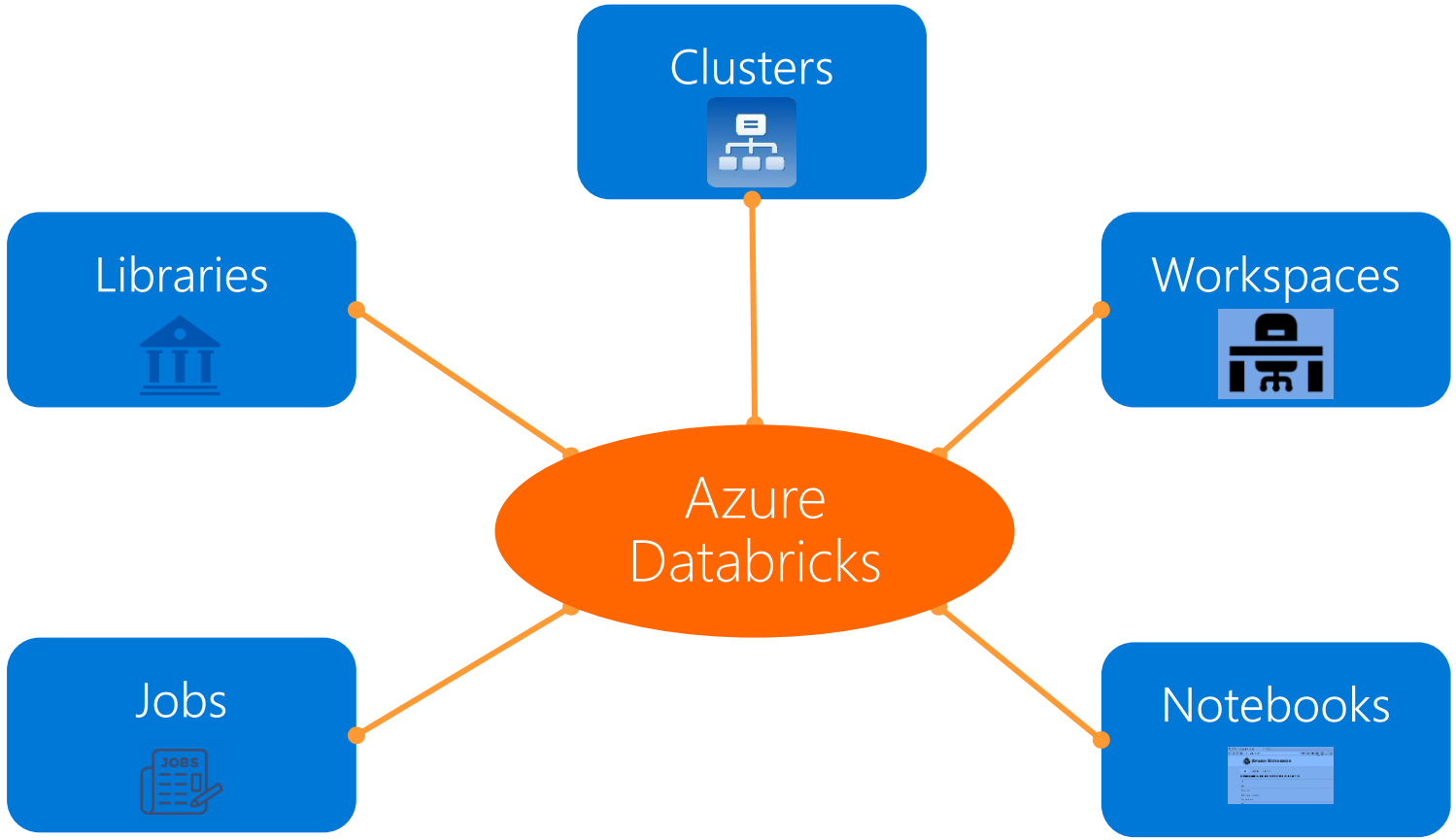


https://medium.com/microsoftazure/an-introduction-to-streaming-etl-on-azure-databricks-using-structured-streaming-databricks-16b369d77e34

AZURE DATABRICKS CLUSTER ARCHITECTURE



AZURE DATABRICKS CORE ARTIFACTS



Azure Databricks

Demo

Azure Databricks

Delta Lake

DELTA LAKE - Reliable Data Lakes at Scale on Azure


Data Versioning


ACID Transactions


Optimized Layouts


Fast Streaming


Efficient Upserts


Schema Enforcement

ACID Transaction Guarantees
- Atomic, Consistent, Isolated, Durable

Versioned parquet files
- Delta transaction log keeps track of all operations

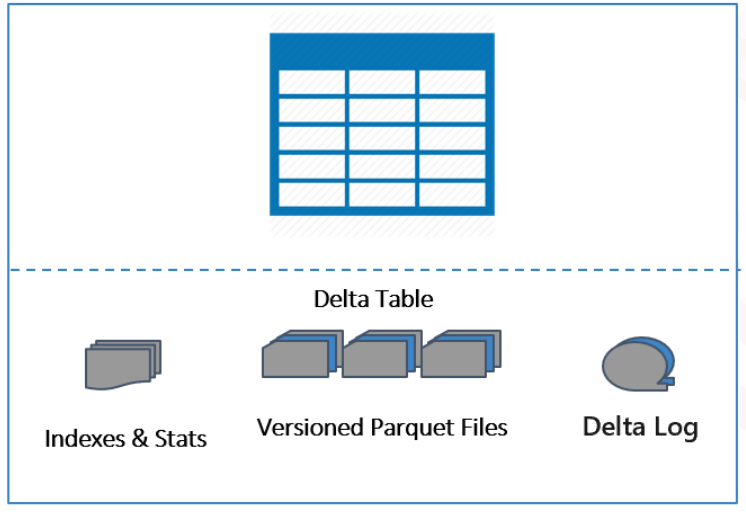
Efficient Upserts
- *MERGE, DELETE, UPDATE*

Small file compaction w/ no interrupt to availability
- *OPTIMIZE* and *VACUUM*

Z-Order partitioning w/ up to 100x perf
- New multidimensional partitioning enables data skipping

Time Travel
- Audit history, Pipeline Debugging, Data Reproducibility

Delta Table =
Parquet + Transaction Log + Indexes/Stats



Open Source

THE APACHE SOFTWARE FOUNDATION + THE LINUX FOUNDATION

Resources

[Azure Databricks – strona główna]

[Apache Spark]

[Databricks Community Edition]

[Microsoft Learn: Databricks]

[PowerShell module: azure.databricks.cicd.tools]

[Databricks CLI]

[Spark by {Examples}]

Questions?

Thank you!

kamil@nowinski.net

@NowinskiK @SQLPlayer

SQLPlayer.net

https://github.com/NowinskiK/CommunityEvents

Kamil Nowinski
Microsoft Data Platform MVP
MCSE Data Platform & MCSE Data Management and Analytics