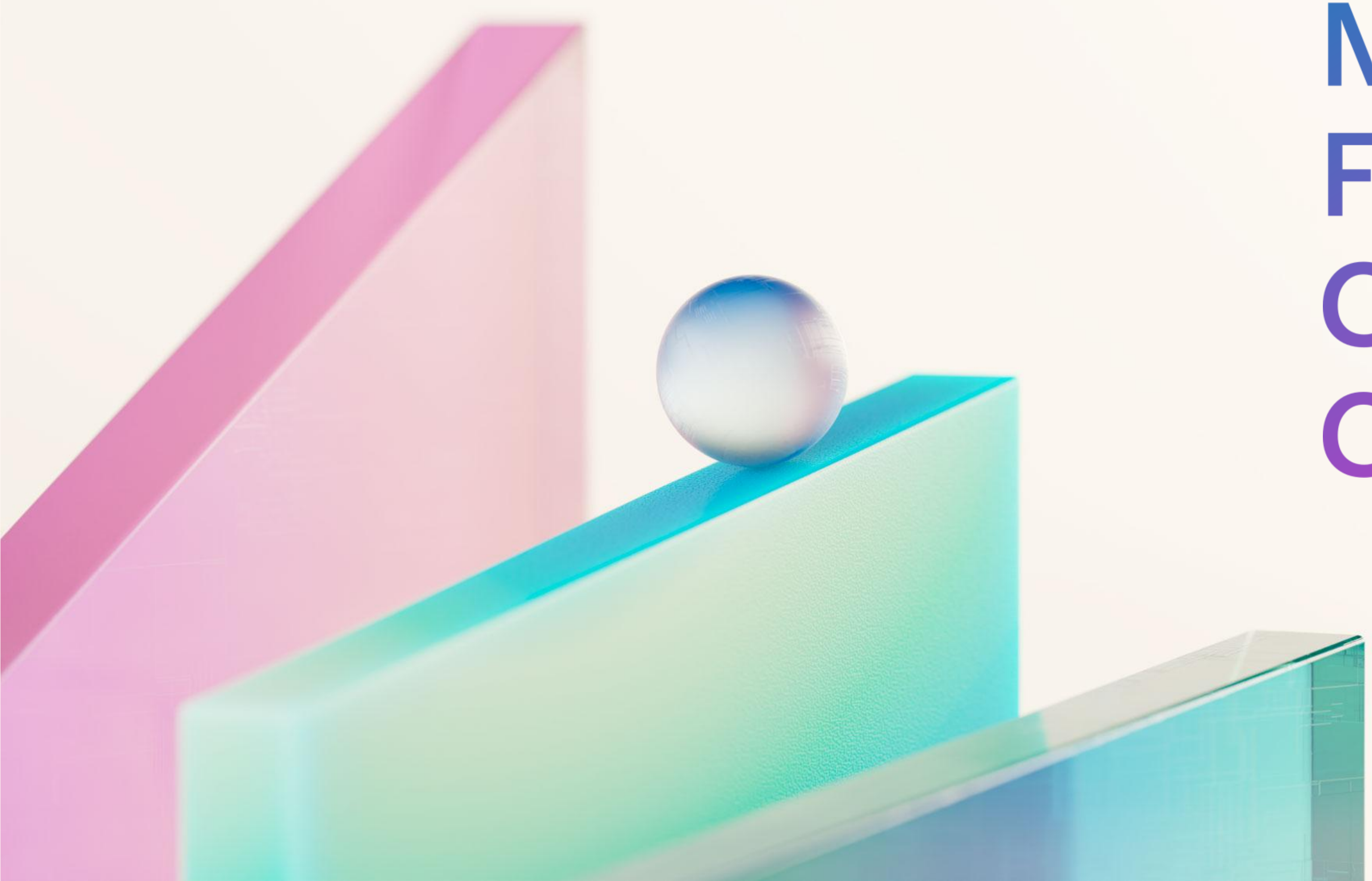
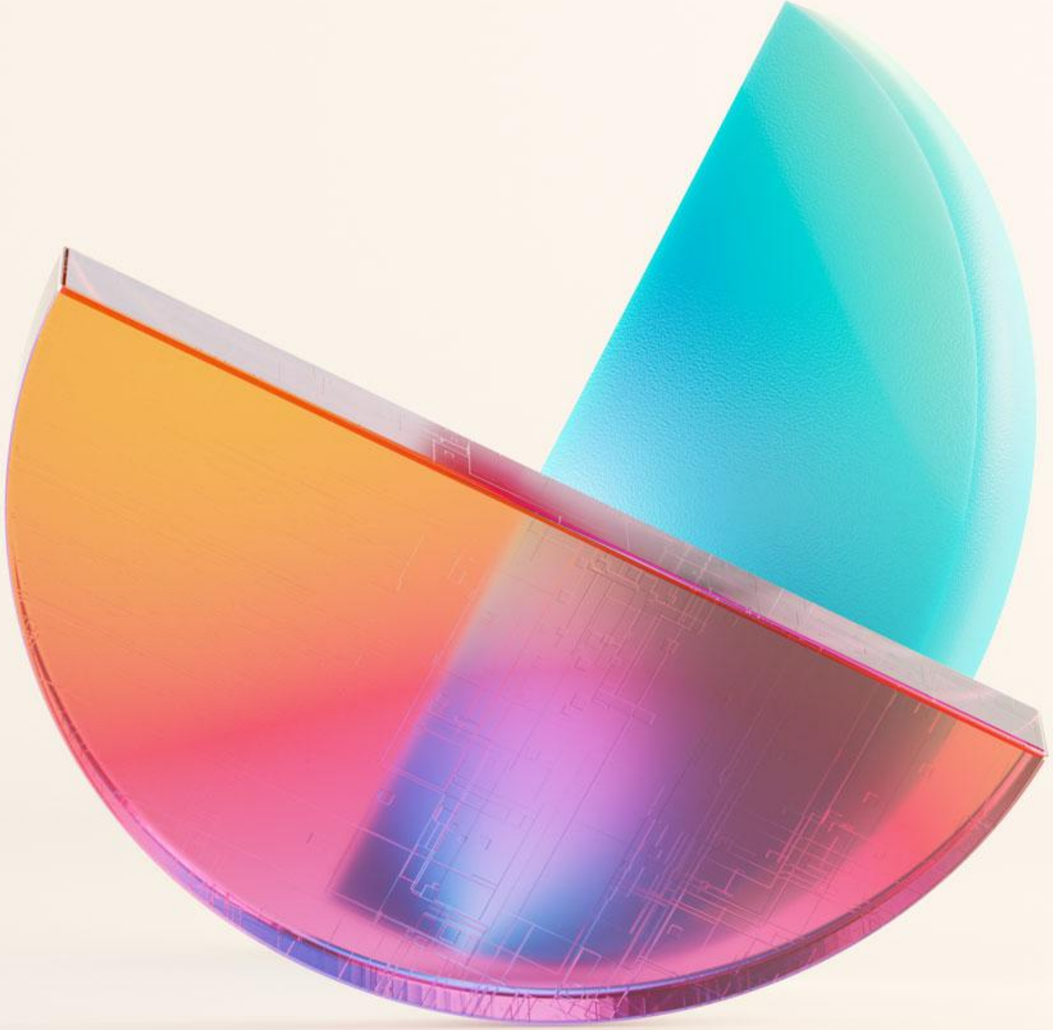
Microsoft



Microsoft Fabric Community Conference

Microsoft

Microsoft Fabric
Community Conference



Harnessing Apache Spark for Next-Gen Analytics in Microsoft Fabric

Anu Venkataraman, Miles Cole, Long Tian

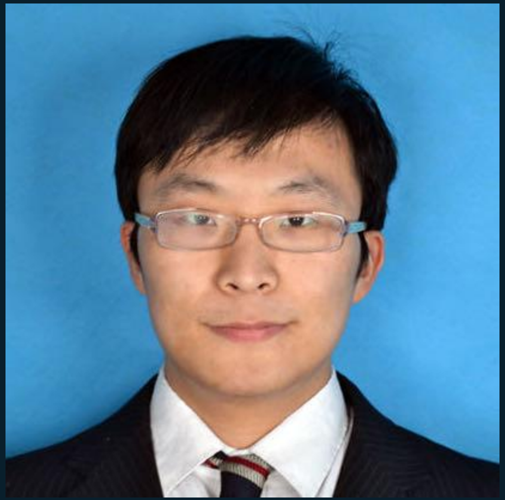
Microsoft

Anu Venkataraman

Senior Program Manager
Microsoft

Miles Cole

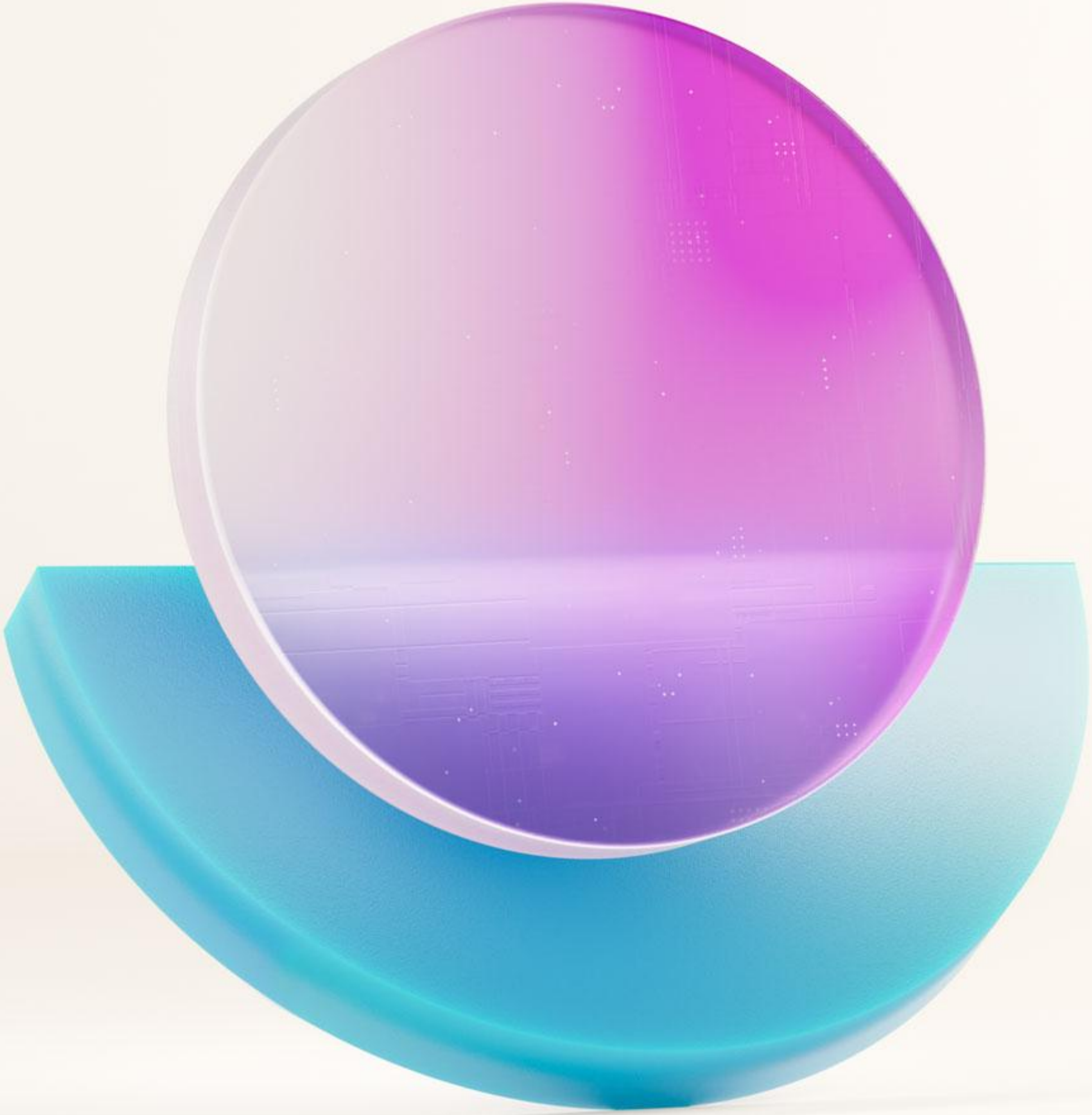
Principal Program Manager
Microsoft



Long Tian

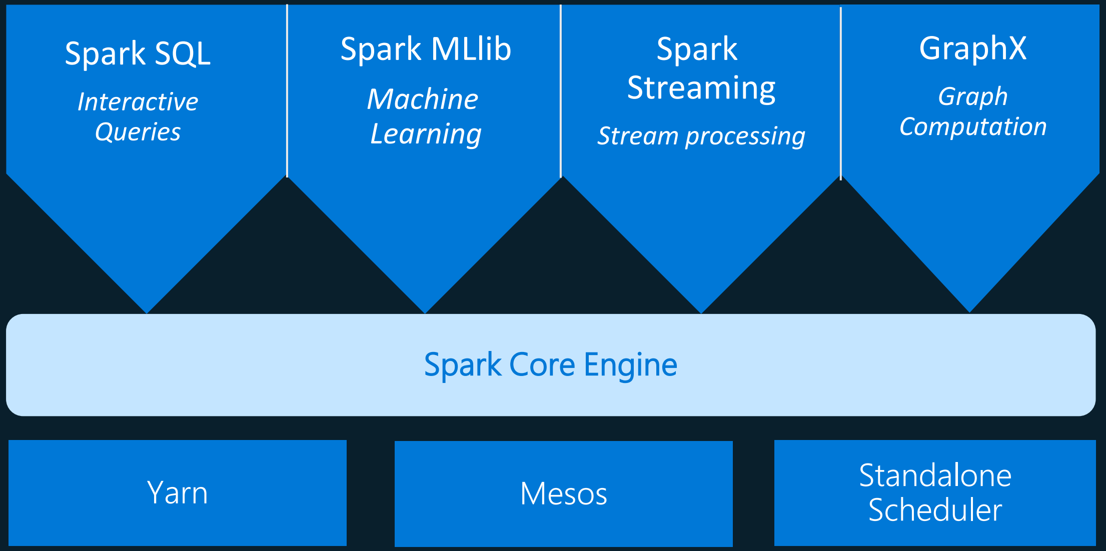
Principal Software Eng. Manager
Microsoft

Microsoft Fabric
Community Conference



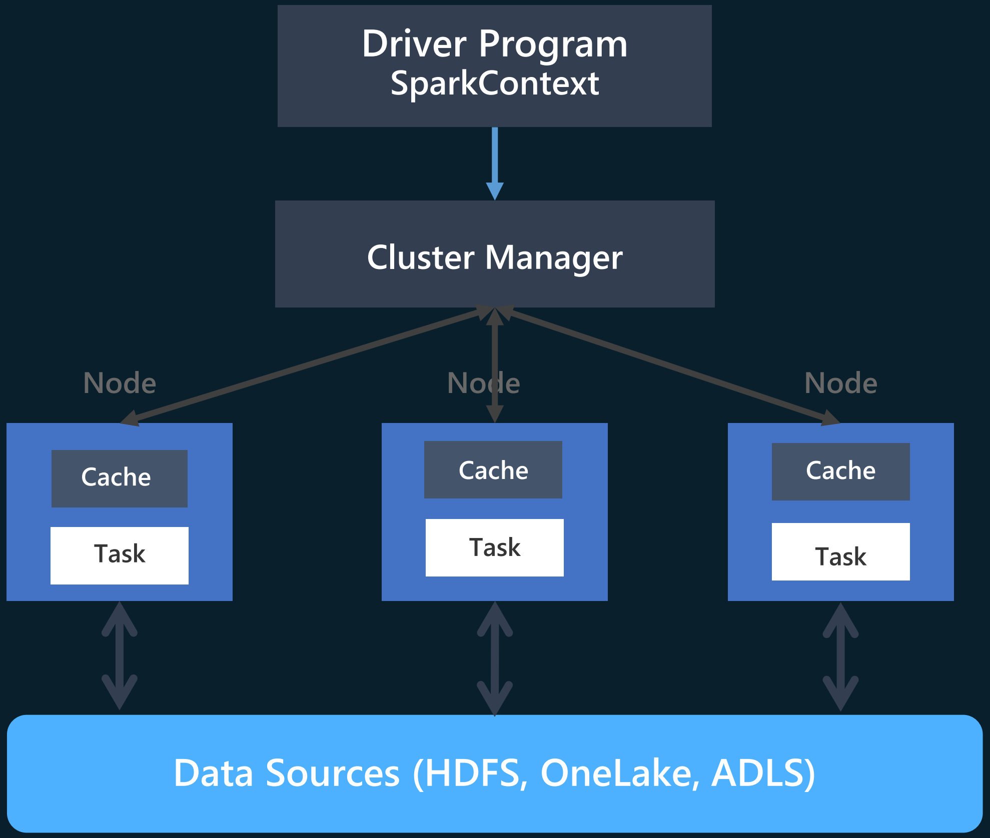
Apache Spark
An overview

Apache Spark

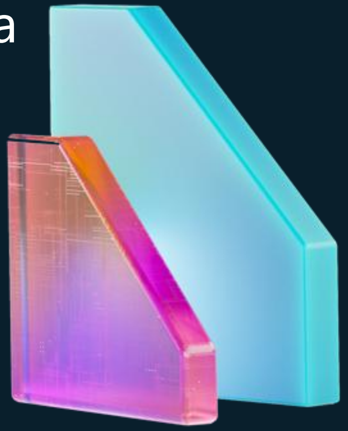


- An open-source unified analytics engine for large-scale data processing
- Supports batch, interactive and streaming data processing
- Massive in-memory distributed and parallel processing capabilities
- Allows writing code in Python, Scala, Java, R and SQL
- Commonly used for complex analytics, data transformation, machine learning and AI tasks on big data

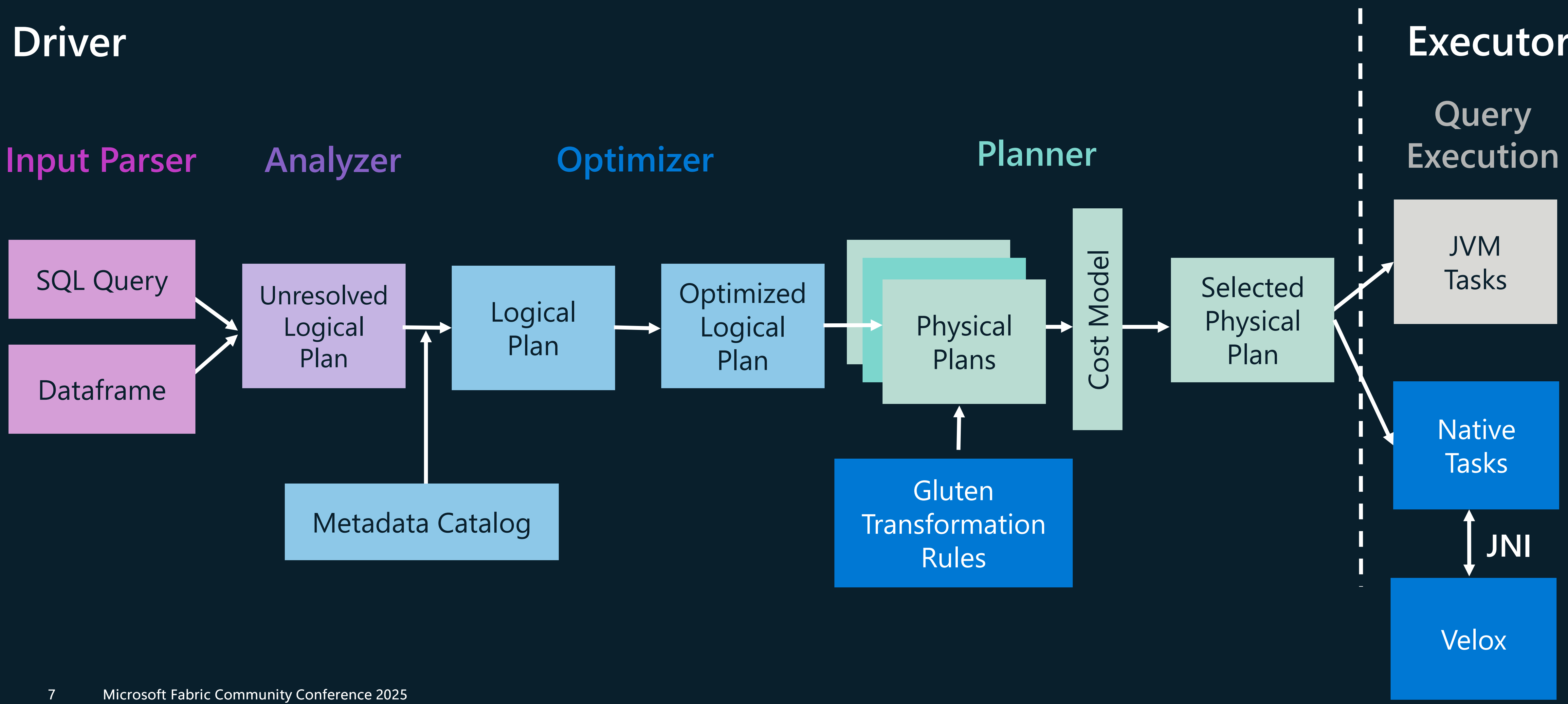
Spark design and job execution

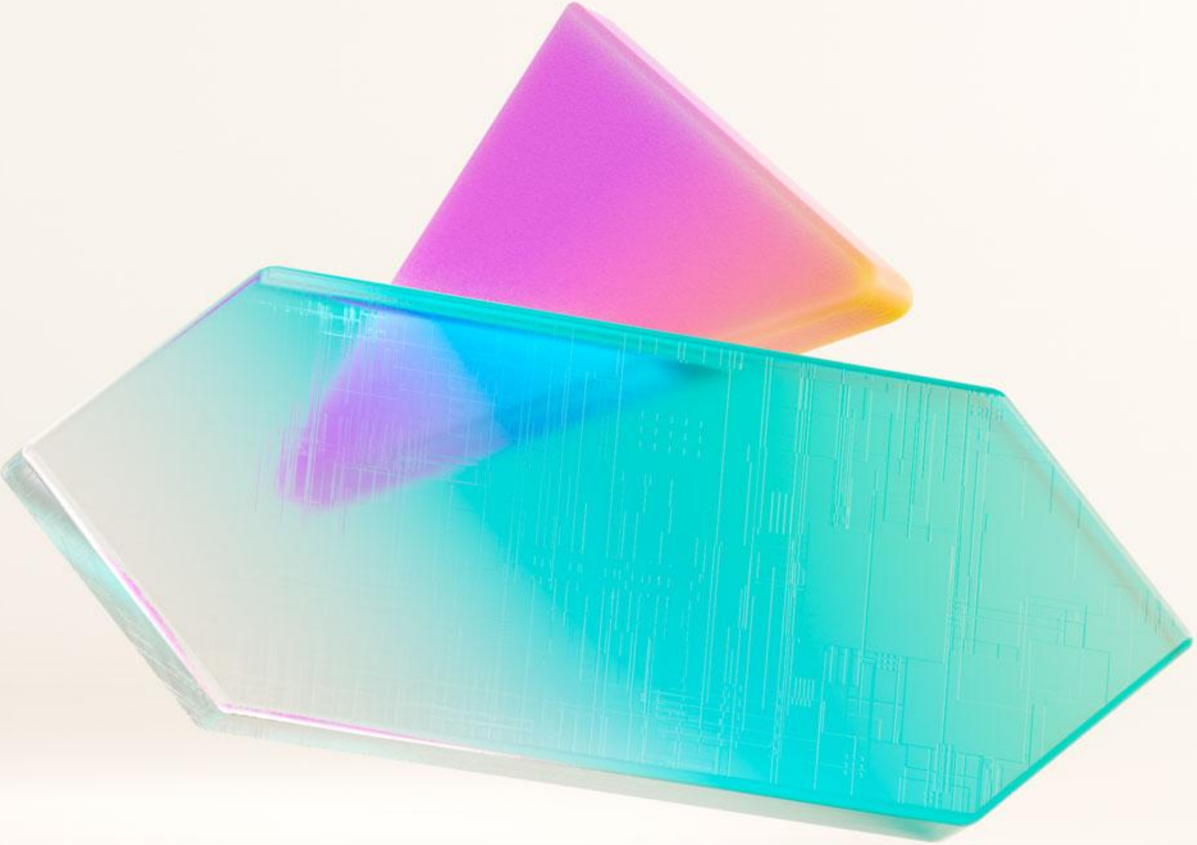


- The *Driver* runs the user's *main* function and executes the various parallel operations on the worker nodes.
- The worker nodes read and write data from/to data source.
- Spreads the processing and data onto different Worker modes
- The results of the operations are collected by the driver
- Worker nodes process data in memory using Resilient Distributed Data Sets (RDDs) and DataFrames



Spark Execution Architecture

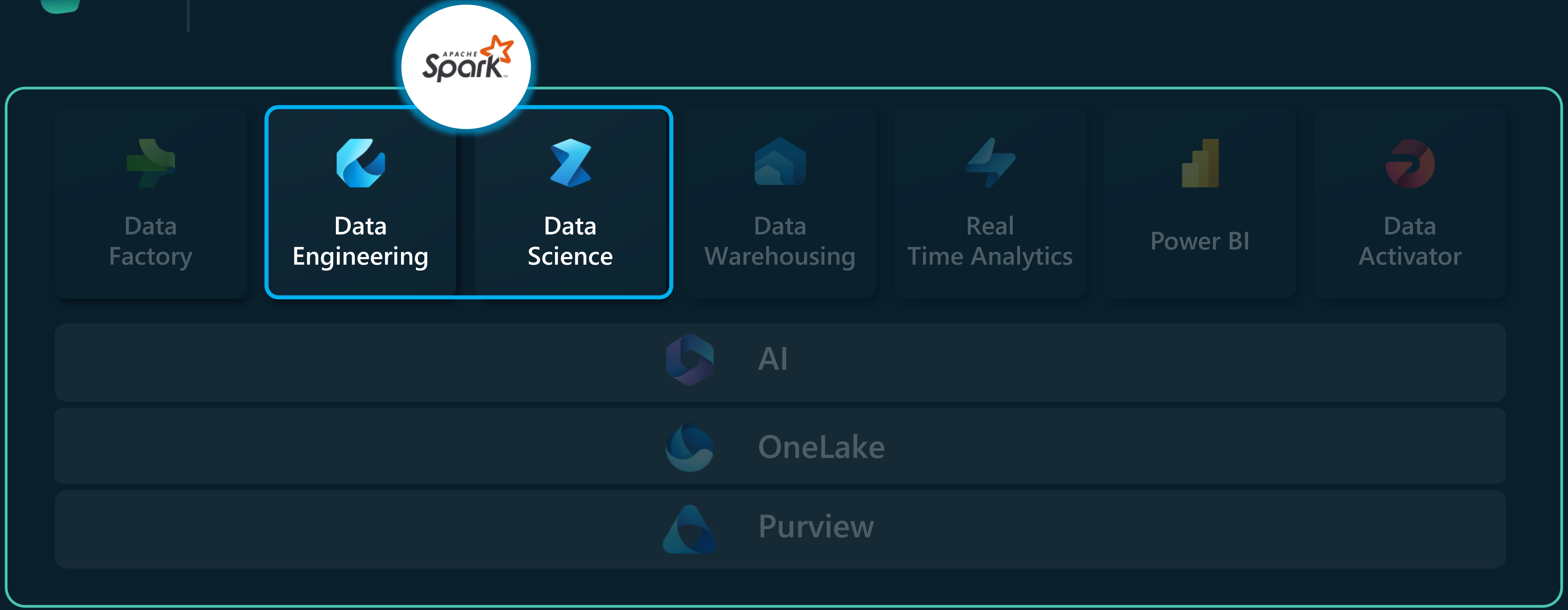




Apache Spark

Leverage the power of
Spark in Fabric

The unified data platform for the era of AI



Spark

Data Factory

Data Engineering

Data Science

Data Warehousing

Real Time Analytics

Power BI

Data Activator

AI

OneLake

Purview

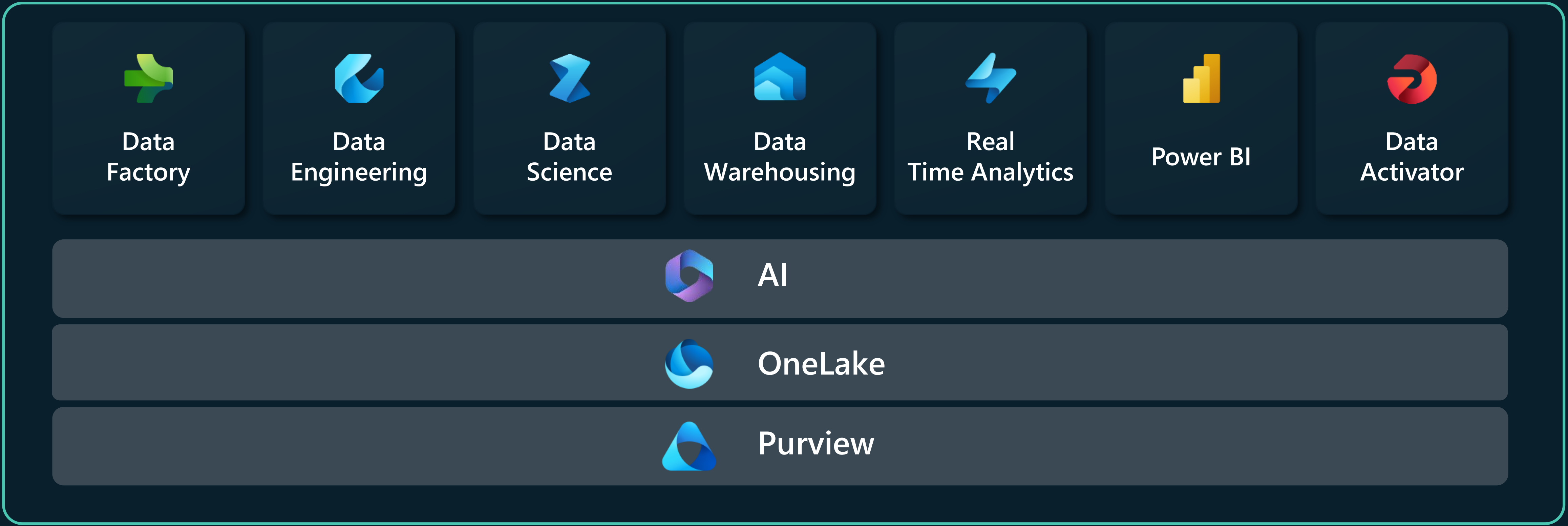
Unified architecture

Unified experience

Unified governance

Unified business model

The unified data platform for the era of AI



Data Factory

Data Engineering

Data Science

Data Warehousing

Real Time Analytics

Power BI

Data Activator

AI

OneLake

Purview

Unified architecture

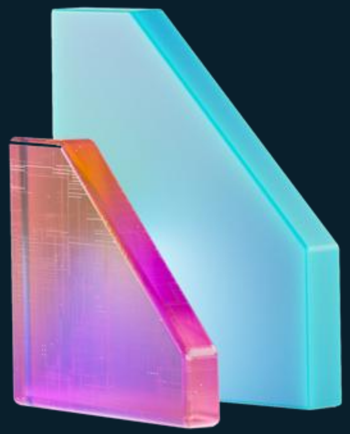
Unified experience

Unified governance

Unified business model

Spark in Microsoft Fabric

- Powers Data Engineering and Data Science experiences
- World-class Spark serverless compute
- Spark sessions start in sub 10 seconds and dynamically scale in/out, pause, and resume.
- Fully managed - no need to create or manage compute
- Provides flexibility to write code in notebooks with
  - Multiple supported languages
  - Co-editing and co-authoring
- Integrates with your Lakehouse
- Orchestrate and schedule data transformations in notebooks and Spark jobs with pipelines



Learning Agenda

Module 1

Developing Spark Applications

Module 2

Orchestrating Spark

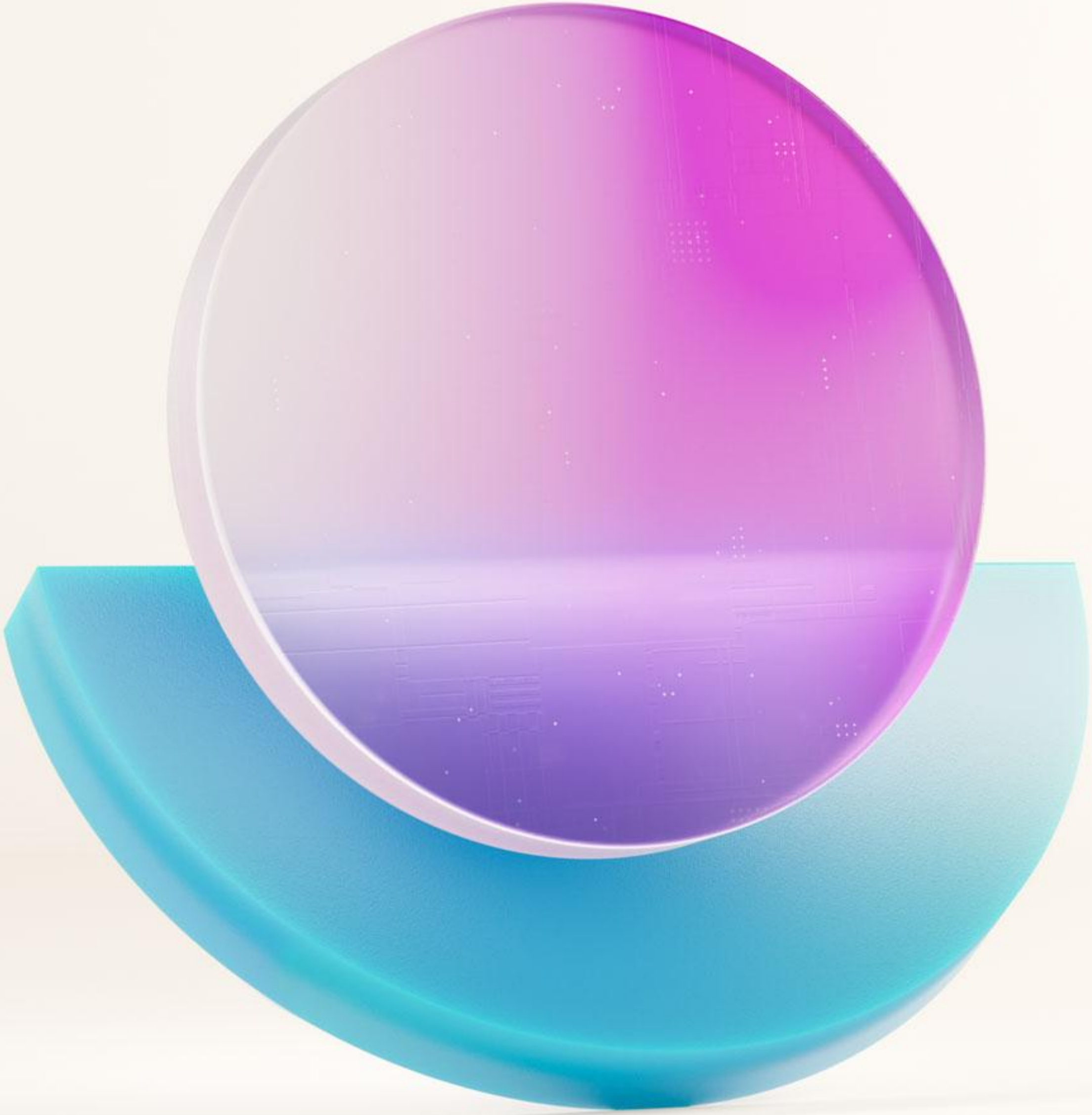
Module 3

Job Scheduling, Monitoring, and Debugging

Module 4

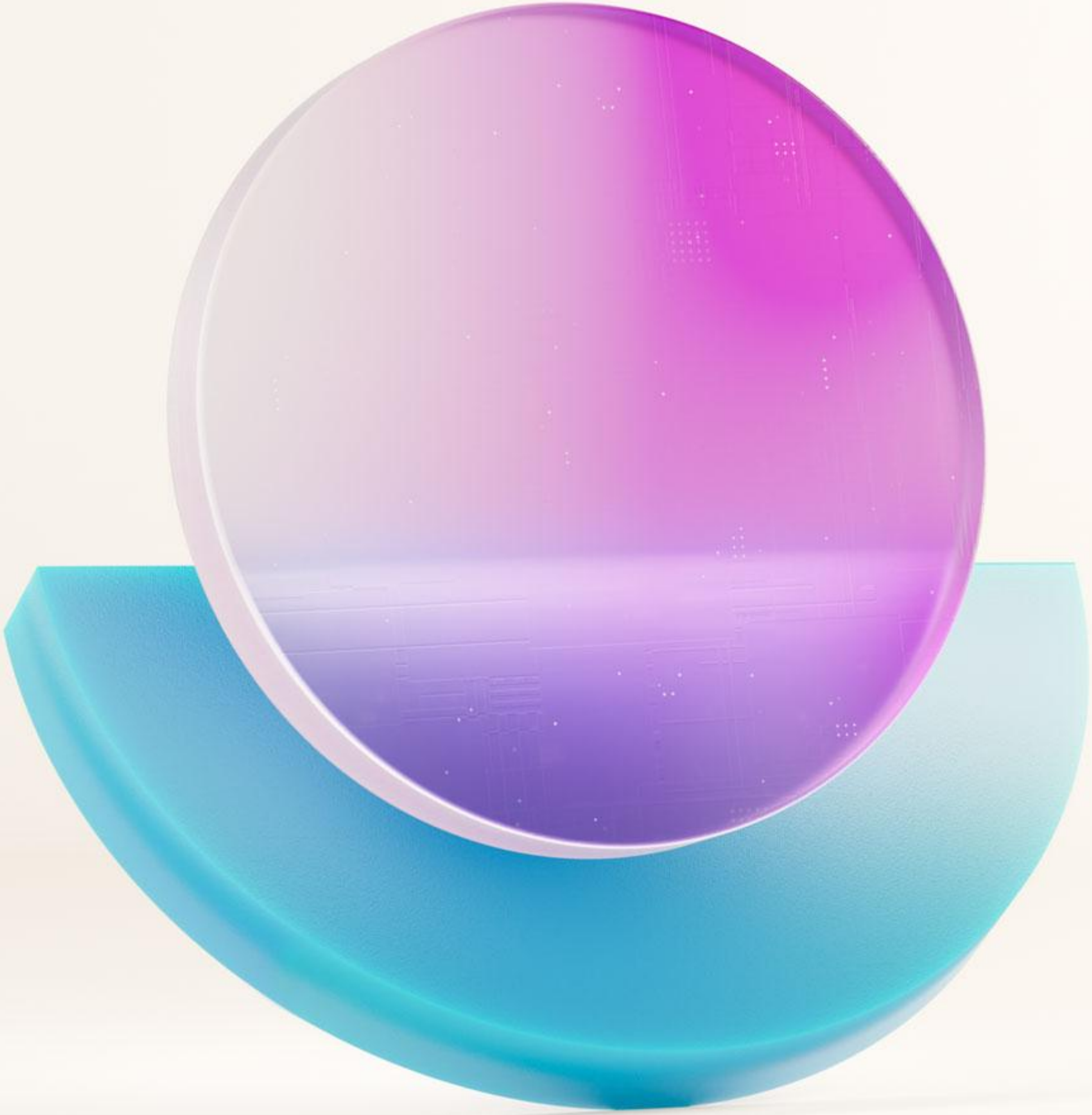
Performance Tuning, Optimizing, and Scaling

Microsoft Fabric
Community Conference



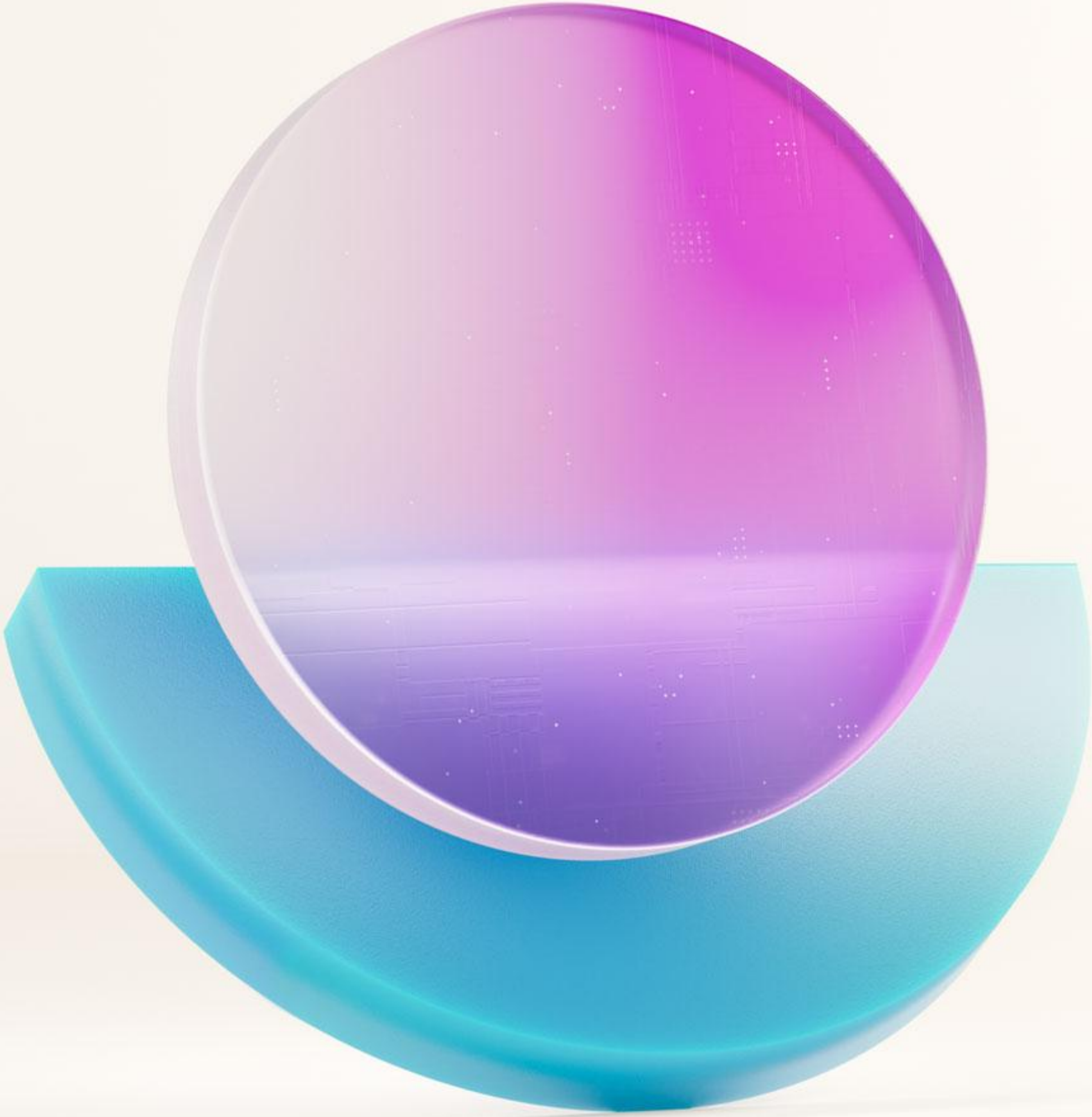
Module 1
Developing Spark Applications

Microsoft Fabric
Community Conference



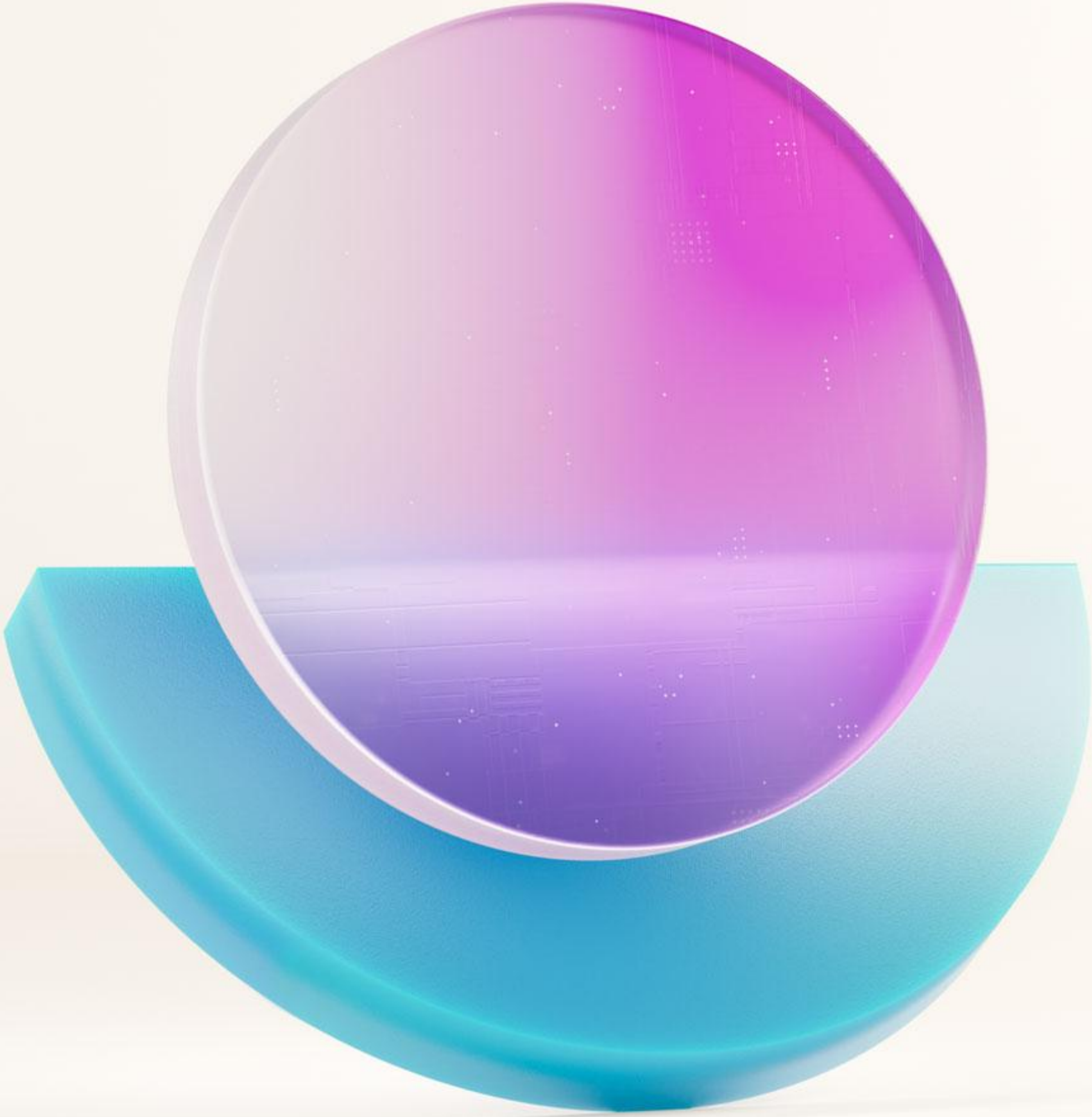
Module 2
Orchestrating Spark

Microsoft Fabric
Community Conference



Module 3
Job Scheduling, Monitoring, and Debugging

Microsoft Fabric
Community Conference



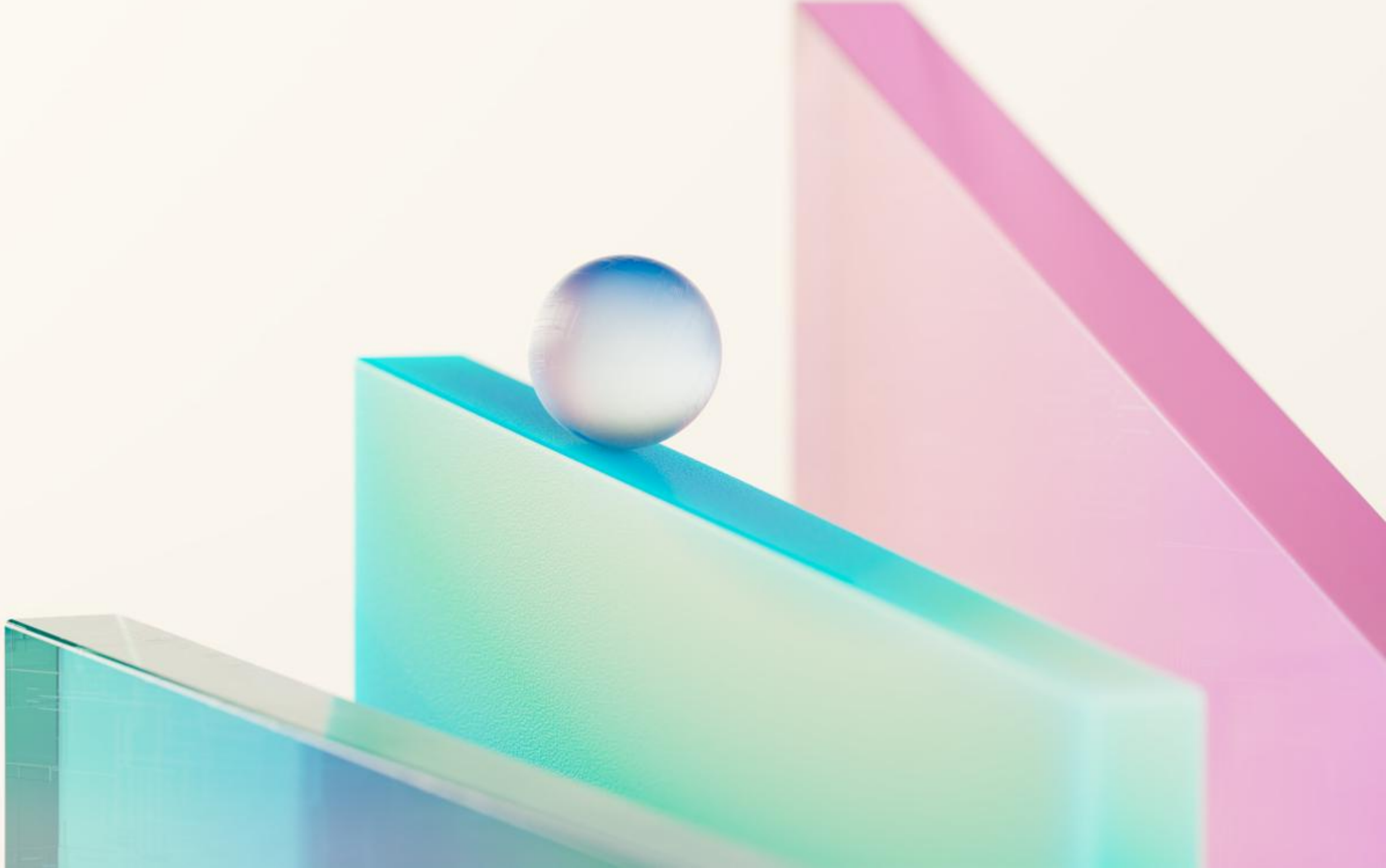
Module 4

Tuning, Optimizing, and Scaling

Microsoft

Microsoft Fabric
Community Conference

Q&A



Continued Learning Resources

MFCC Sessions

Taking Large-Scale Data Engineering Projects to Production

Tuesday, 9:15 AM to 10:15 AM – Grand Ballroom 115

Data Engineering with Spark for SQL Server Professionals

Wednesday, 11:15 PM to 12:15 PM – Boulevard Ballroom 160

Fabric Data Engineering Roadmap

Tuesday, 8:00 AM to 9:00 AM – Premier Ballroom 313

Optimizing Fabric Spark and Best Practices for Production-Ready Workload

Tuesday, 1:45 PM to 2:45 PM – Grand Ballroom 121

Spark Workload Acceleration with the Native Execution Engine in Fabric

Wednesday, 3:15 PM to 4:15 PM – Grand Ballroom 119

Mastering Fabric Data Engineering Admin and Capacity Management

Tuesday, 11:15 AM to 12:15 PM – Boulevard Ballroom 157

The Future Unfolded: Microsoft Fabric for AI and Data Science

Monday, 3:00 PM to 4:00 PM – Boulevard Ballroom 163

Microsoft Fabric - The Command Line Way

Wednesday, 2:00 PM to 3:00 PM – Grand Ballroom 119

MFCC Sessions

Fabric Git/ALM End-to-End Journey with Lakehouse, Shortcuts and Variable Library

Wednesday, 2:00 PM to 3:00 PM – Grand Ballroom 119

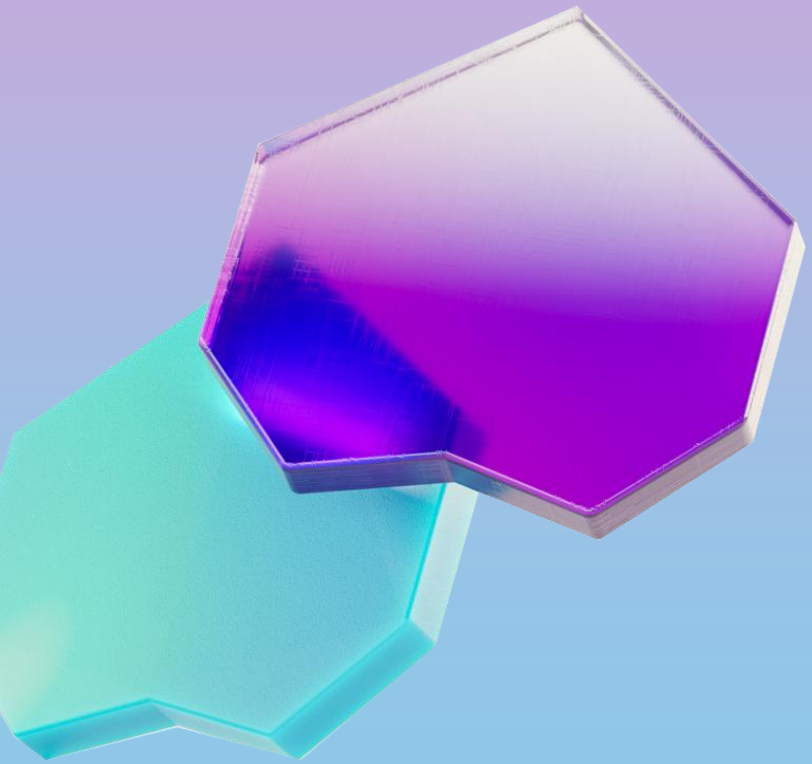
Adopting Microsoft Fabric: Medallion Architecture and Data Mesh Made Easy

Monday, 3:00 PM to 4:00 PM – Boulevard Ballroom 120

Intro to PySpark in Microsoft Fabric

Tuesday, 8:00 AM to 9:00 AM – Grand Ballroom 122

Documentation



Microsoft Fabric Documentation

https://learn.microsoft.com/fabric/

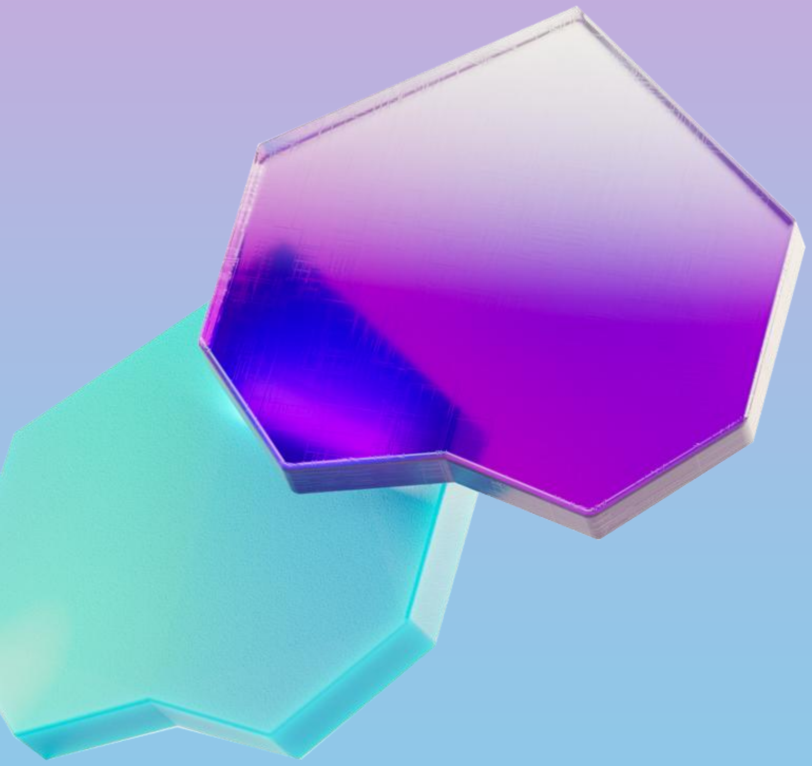
Apache Spark Documentation

https://spark.apache.org/docs/

Delta Lake Documentation

https://docs.delta.io/

Blogs



blog.fabric.microsoft.com

Microsoft Fabric Blog

milescole.dev

Miles Cole | Microsoft Fabric CAT

murggu.medium.com

Aitor Murguzur | Microsoft Fabric CAT

fabric.guru

Sandeep Pawar | Hitachi Solutions | MVP

More References

Best Practices for Fabric Spark - Quick Tips

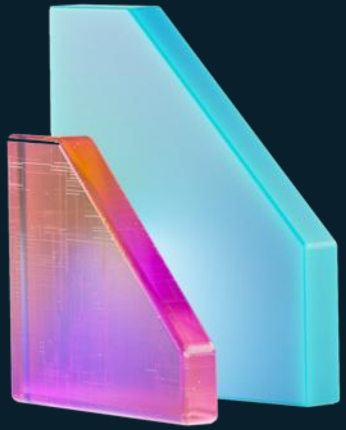
Legacy Timestamp Support in Native Execution Engine for Fabric Runtime 1.3 | Microsoft Fabric Blog | Microsoft Fabric

To V-Order or Not: Making the Case for Selective Use of V-Order in Fabric Spark | Miles Cole

Profiling Microsoft Fabric Spark Notebooks with Sparklens | Microsoft Fabric Blog | Microsoft Fabric

How to use notebooks - Microsoft Fabric | Microsoft Learn

Develop, execute, and manage notebooks - Microsoft Fabric | Microsoft Learn



Microsoft

Get Involved in the Fabric Community

aka.ms/FabricCommunity
Connect with community members, ask questions, and learn more about Fabric

aka.ms/FabricUserGroups
Find a user group that matches your interests in your area or online

aka.ms/SuperUsers
Spread your Fabric knowledge, insights, and best practices with others

aka.ms/MVP
Technology experts that share their knowledge and passion with the community