Spark SQL

databricks

Bryan Cafferky
Data Solutions Enabler

Where Are We Heading?

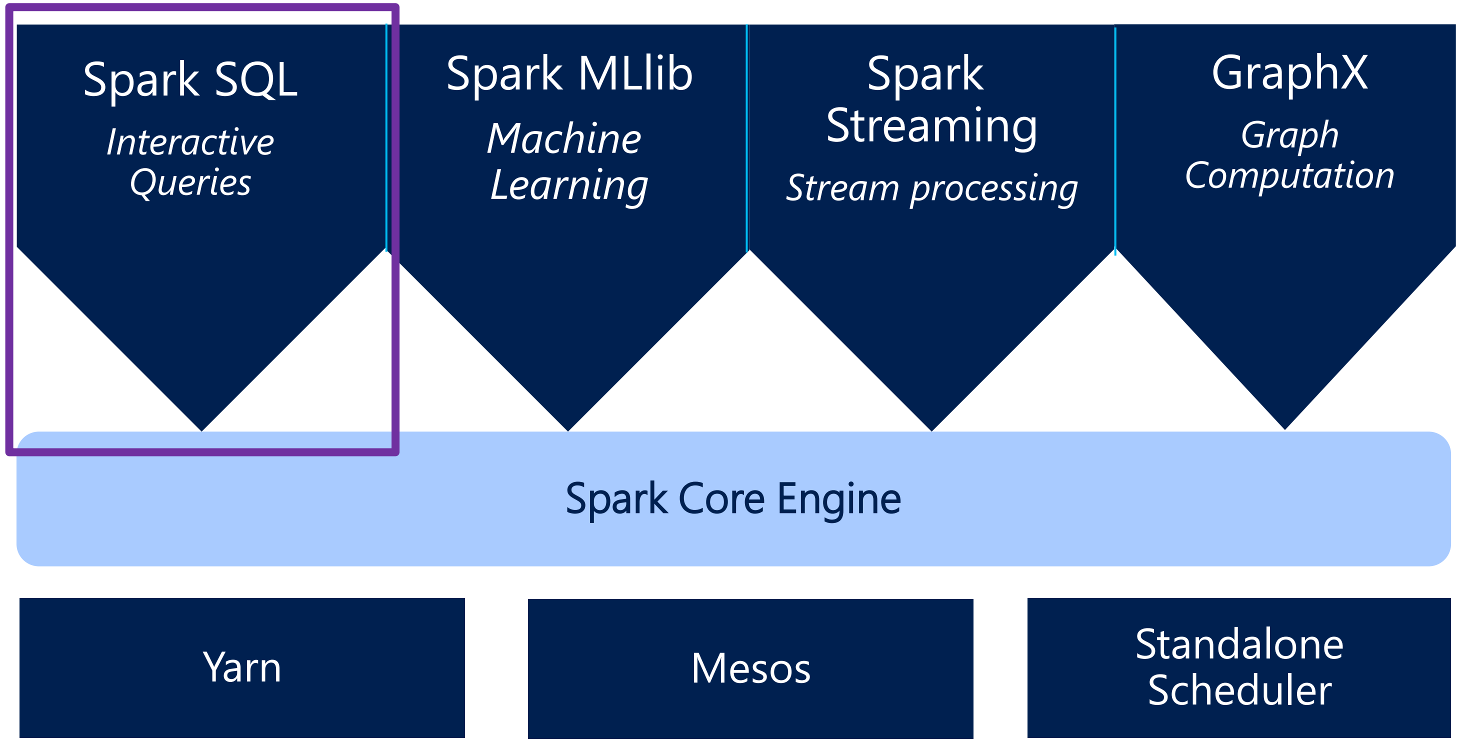
- Why SQL on Spark?
- Schema on Read
- Performance Tuning and Why RDMS had the Right Idea

APACHE SPARK

An unified, open source, parallel, data processing framework for Big Data Analytics

Spark Unifies:

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

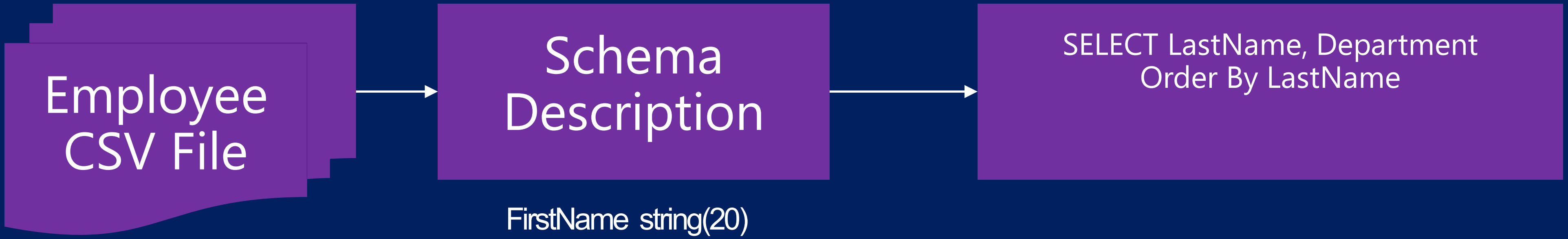


Why Structured Query Language (SQL)?

- ✓ It’s An Awesome Querying Language
- ✓ People Already Know It
- ✓ Can Be Used from Other Spark Languages
- ✓ Supports Performance Tuning and Optimization

Schema On Read

- Data Is Not in an RDMS
- External File Is Described Structurally



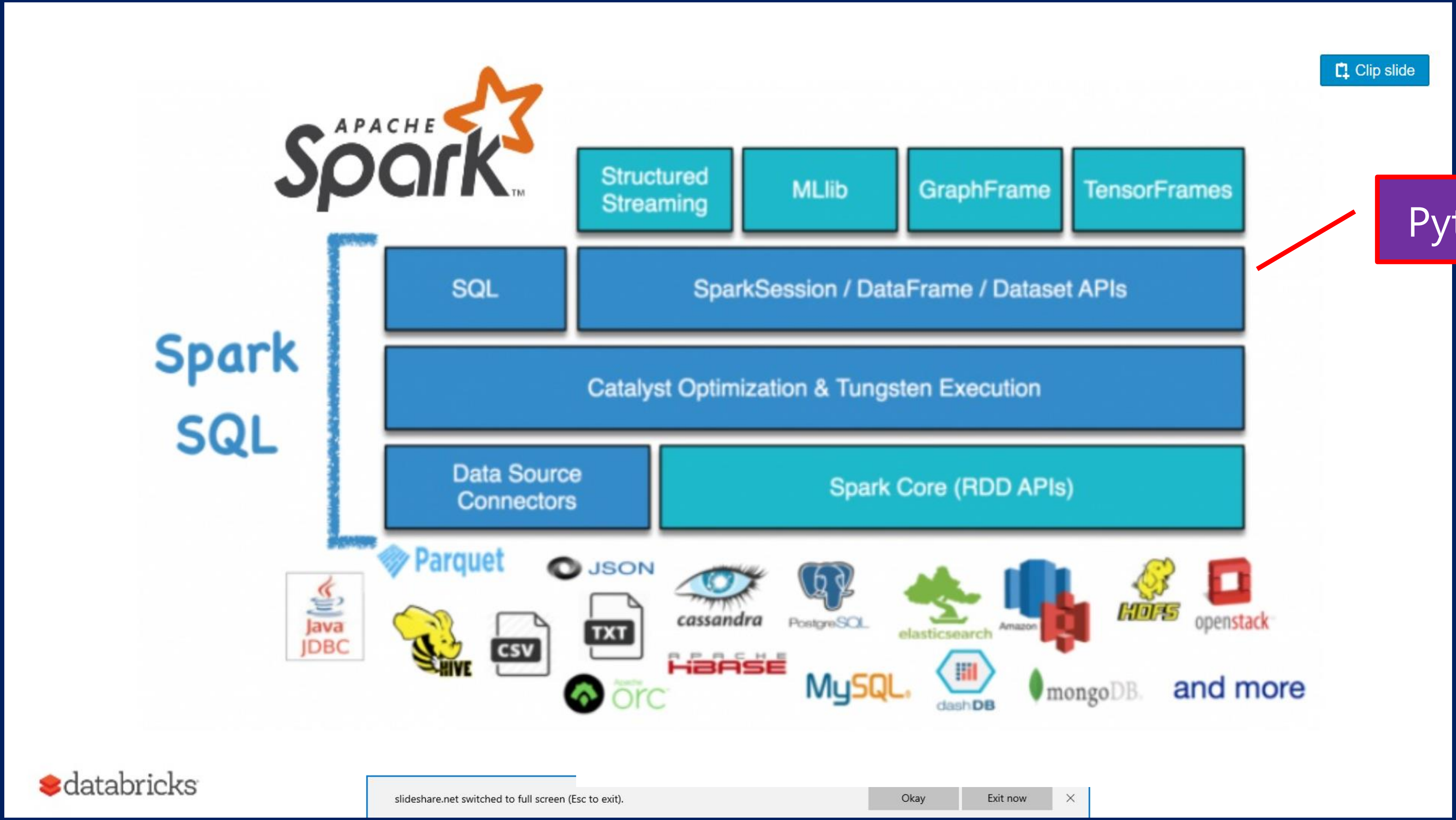
Employee CSV File

Schema Description

SELECT LastName, Department
Order By LastName

FirstName string(20)
Lastname string(20)
HireDate Date
Department String(20)

Spark SQL and Performance Optimization

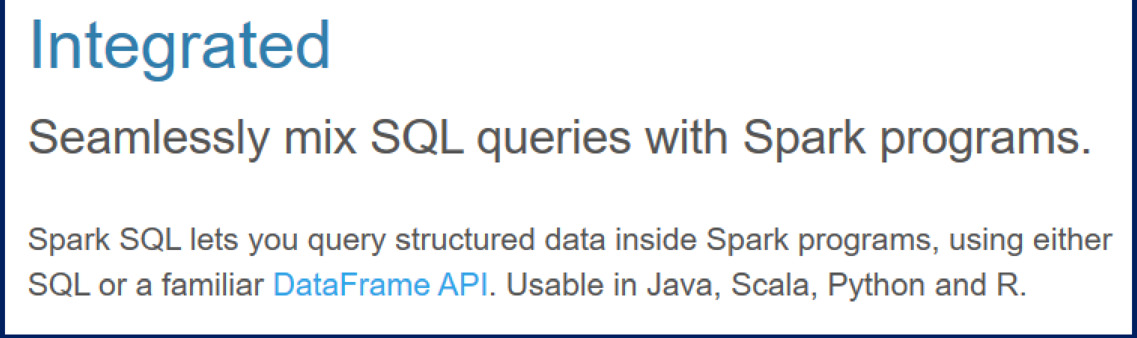


slideshare.net switched to full screen (Esc to exit). Okay Exit now

Python, R & Scala

https://www.slideshare.net/databricks/deep-dive-into-spark-sql-with-advanced-performance-tuning-with-xiao-li-wenchen-fan

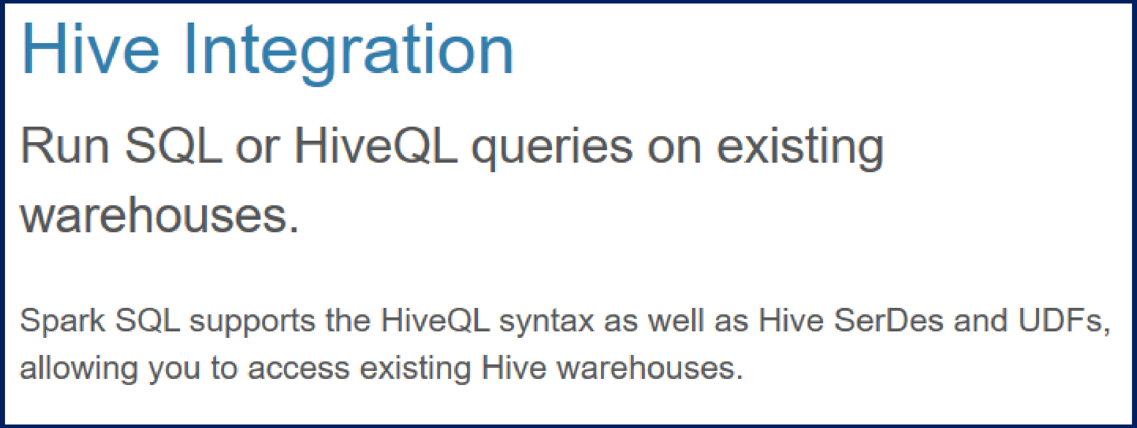
Azure Databricks SQL Benefits



Integrated

Seamlessly mix SQL queries with Spark programs.

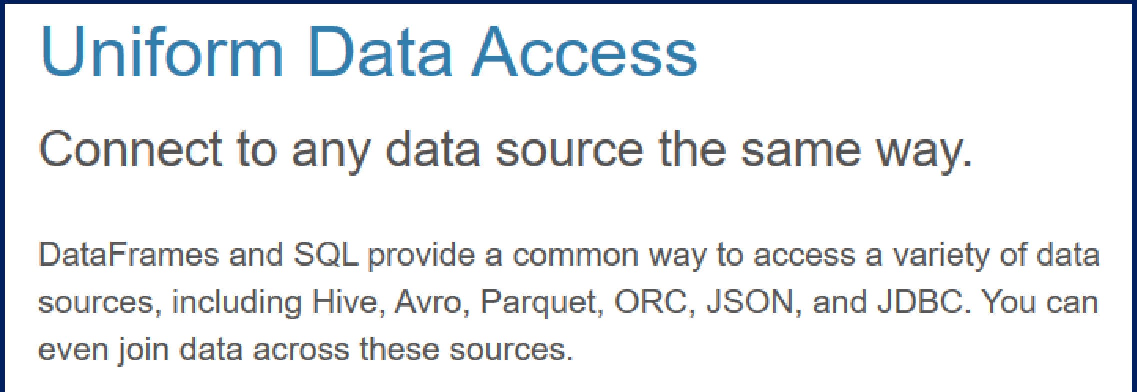
Spark SQL lets you query structured data inside Spark programs, using either SQL or a familiar DataFrame API. Usable in Java, Scala, Python and R.



Hive Integration

Run SQL or HiveQL queries on existing warehouses.

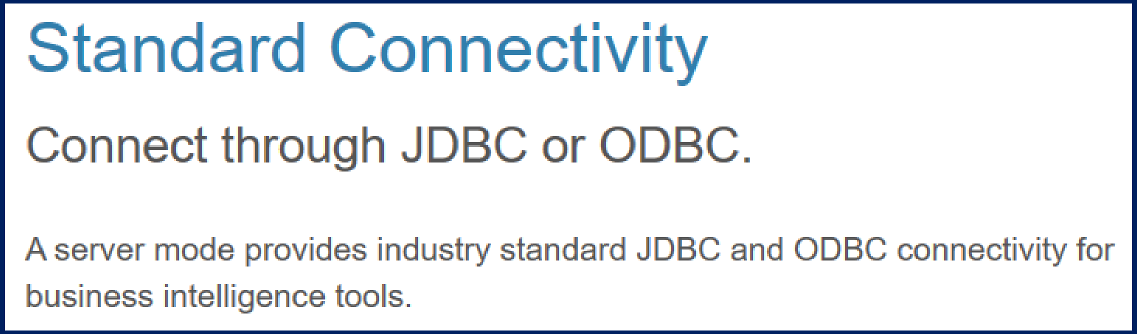
Spark SQL supports the HiveQL syntax as well as Hive SerDes and UDFs, allowing you to access existing Hive warehouses.



Uniform Data Access

Connect to any data source the same way.

DataFrames and SQL provide a common way to access a variety of data sources, including Hive, Avro, Parquet, ORC, JSON, and JDBC. You can even join data across these sources.



Standard Connectivity

Connect through JDBC or ODBC.

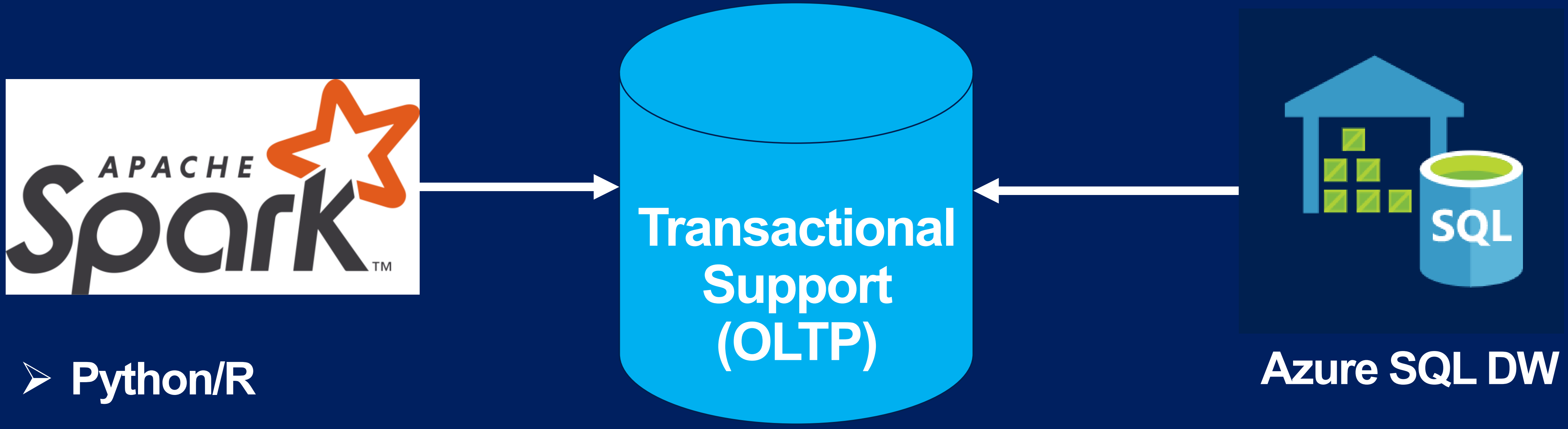
A server mode provides industry standard JDBC and ODBC connectivity for business intelligence tools.

https://spark.apache.org/sql/

Spark SQL

- ✓ Supports Most Standard SELECT syntax.
- ✓ Does NOT have a database catalog.
- ✓ Does not support stored procedures or functions.
- ✓ Does not support referential integrity.
- ✓ Limited Security Support, i.e. Grant and Revoke.

Non-Relational vs. DBMS Converging



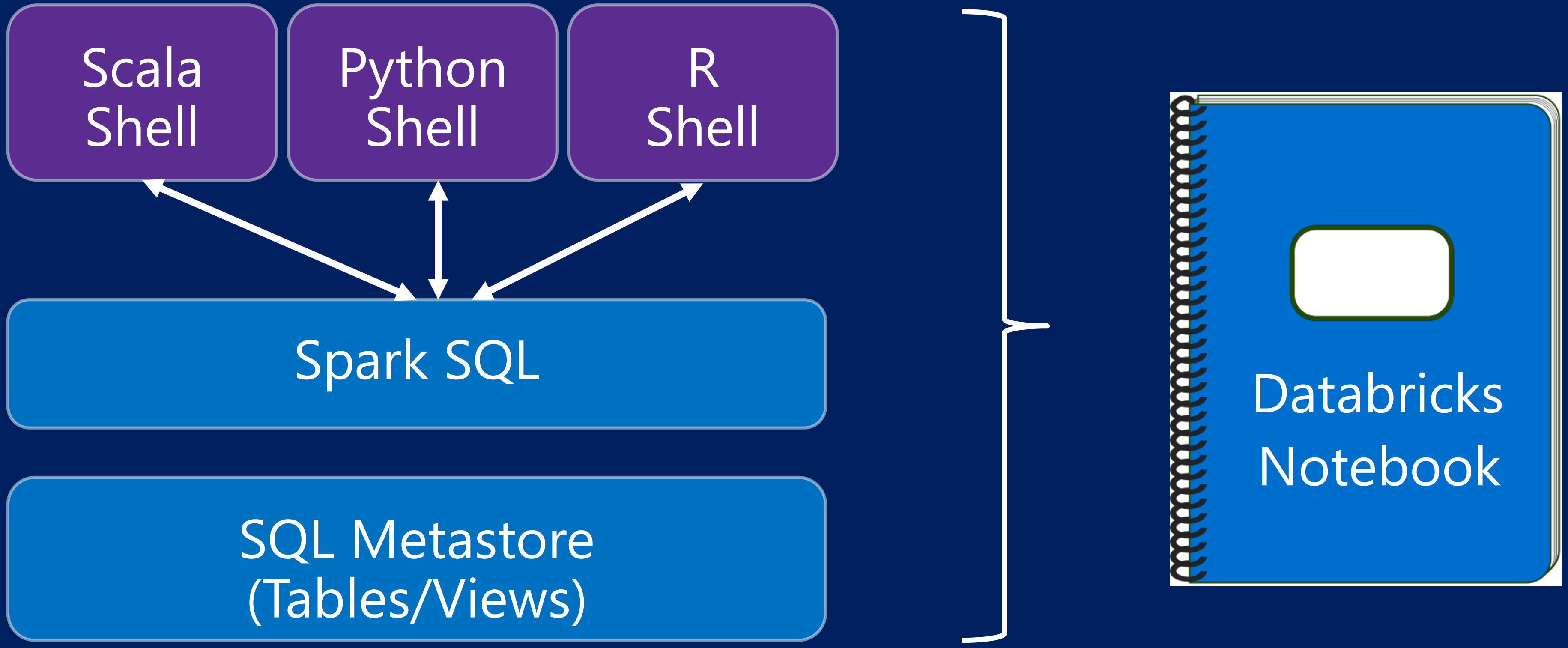
Transactional Support (OLTP)

Azure SQL DW

- Python/R
- Machine Learning
- Non-Structured Data

- Snowflake
- BigQuery
- RedShift

Spark Shells



Scala Shell

Python Shell

R Shell

Spark SQL

SQL Metastore (Tables/Views)

Databricks Notebook

Azure Databricks SQL Deep Dive

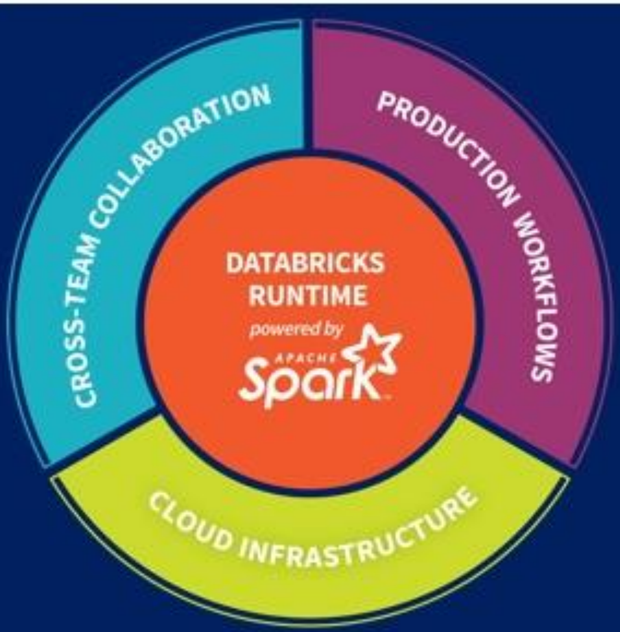
- Using Notebooks
- Using Spark SQL

Review

- Why SQL on Spark?
- Schema on Read
- Performance Tuning and Why RDMS had the Right Idea

Azure Databricks SQL Deep Dive

Microsoft Azure



Bryan Cafferky

Technical Solutions Professional