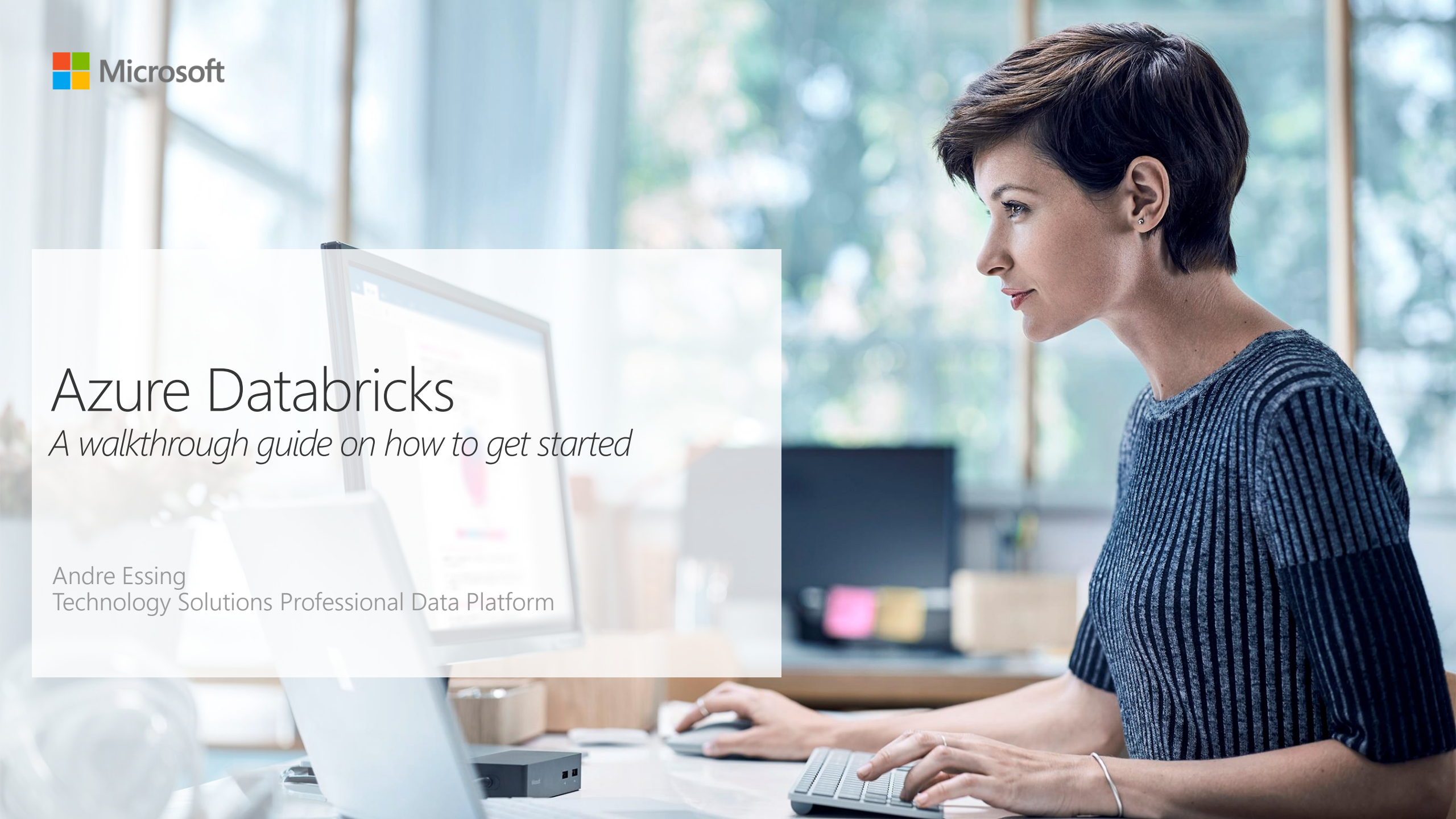




# Azure Databricks

*A walkthrough guide on how to get started*

Andre Essing  
Technology Solutions Professional Data Platform





# Andre Essing

## Technology Solutions Professional

### Microsoft Deutschland GmbH

Andre advises customers in topics all around the Microsoft Data Platform. Since version 7.0, Andre gathering experience with the SQL Server product family. Today Andre concentrates on working with data in the cloud, like Modern Data Warehouse architectures, Artificial Intelligence and new scalable database systems like Azure Cosmos DB.

 [andre.essing@microsoft.com](mailto:andre.essing@microsoft.com)

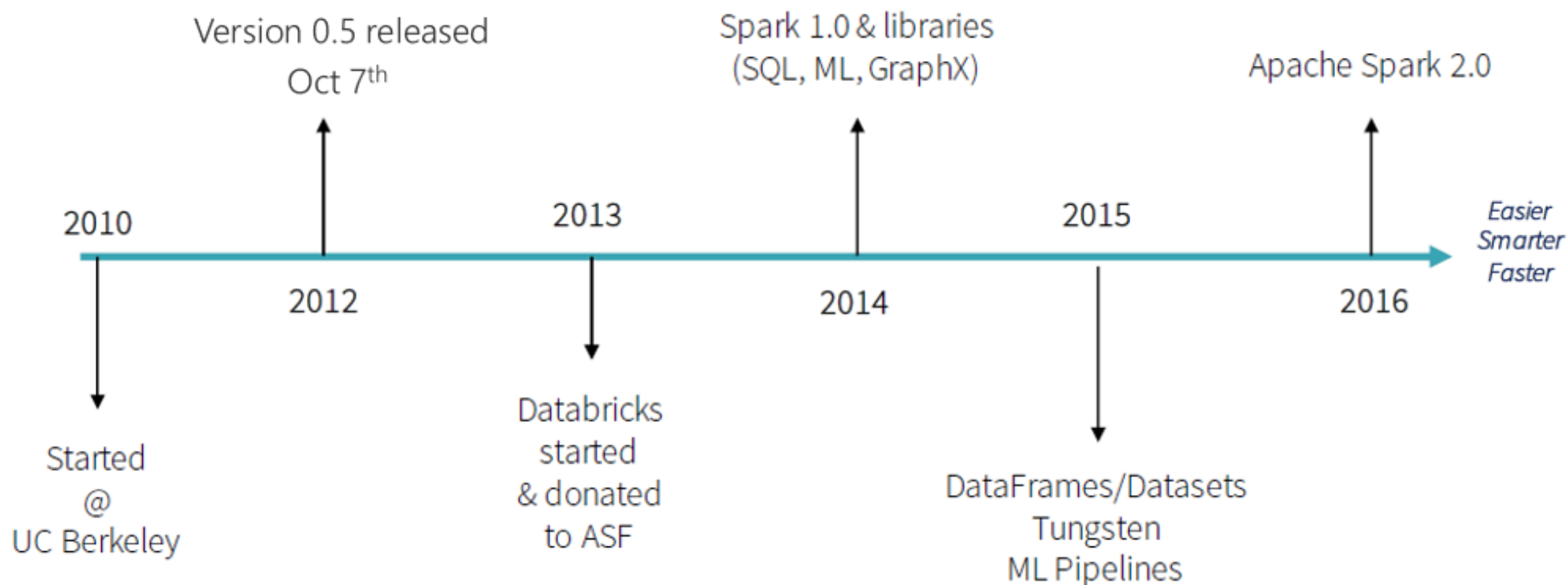
 [andreessing.de](http://andreessing.de)

 [aessing](https://www.linkedin.com/in/aessing)

 [@aessing](https://twitter.com/aessing)

 [aessing](https://github.com/aessing)

# SPARK: A BRIEF HISTORY

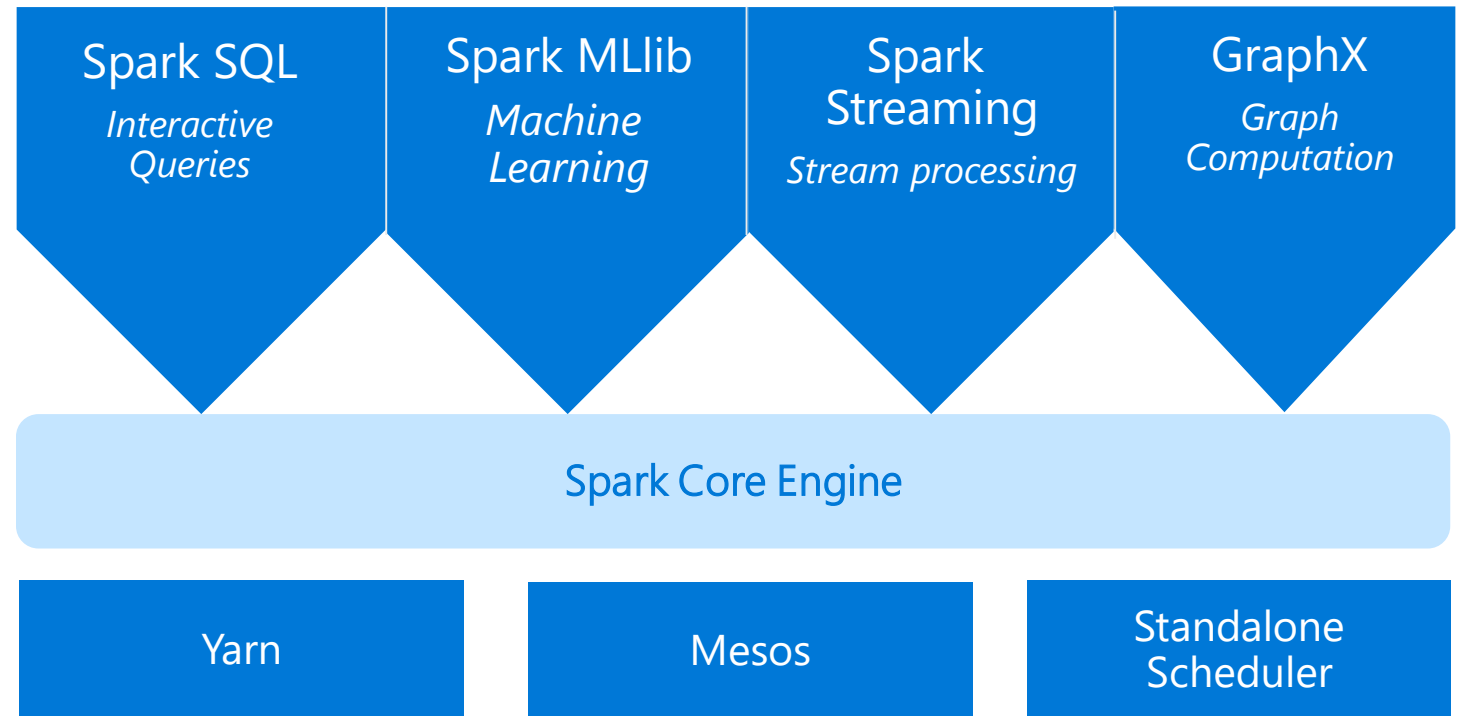


# A P A C H E   S P A R K

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

Spark Unifies:

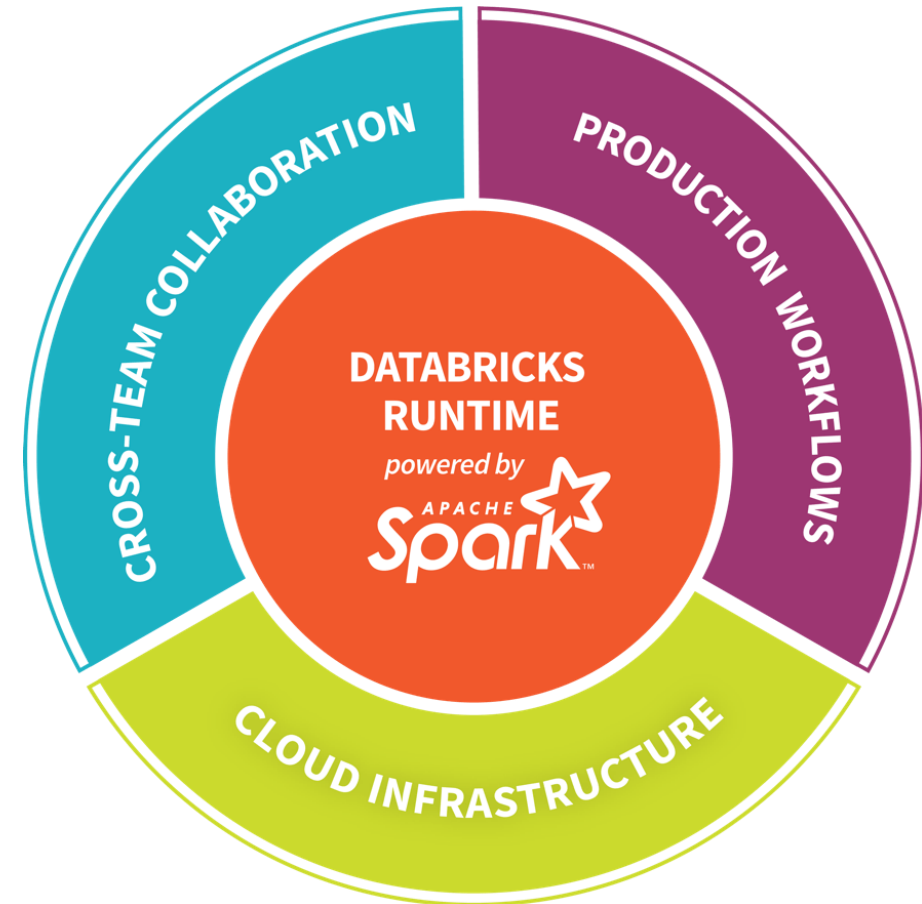
- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing





# DATABRICKS - COMPANY OVERVIEW

- Founded in late 2013
- By the creators of Apache Spark, original team from UC Berkeley AMPLab
- Largest code contributor code to Apache Spark
- Level 2/3 support partnership with
  - Hortonworks
  - MapR
  - DataStax
- Provides [certifications](#) such as Databricks Certified Application, Databricks Certified Distribution and Databricks Certified Developer
- Main Product: The [Unified Analytics Platform](#)
- In Oct 2017, introduced [Databricks Delta](#) (currently in private preview).



# A Z U R E   D A T A B R I C K S

- Azure Databricks is a **first party** service on Azure.
  - Unlike with other clouds, it is not an Azure Marketplace or a 3<sup>rd</sup> party hosted service.
- Azure Databricks is integrated seamlessly with Azure services:
  - [Azure Portal](#): Service can be launched directly from Azure Portal
  - [Azure Storage Services](#): Directly access data in Azure Blob Storage and Azure Data Lake Store
  - [Azure Active Directory](#): For user authentication, eliminating the need to maintain two separate sets of users in Databricks and Azure.
  - [Azure SQL DW and Azure Cosmos DB](#): Enables you to combine structured and unstructured data for analytics
  - [Apache Kafka for HDInsight](#): Enables you to use Kafka as a streaming data source or sink
  - [Azure Billing](#): You get a single bill from Azure
  - [Azure Power BI](#): For rich data visualization
- Eliminates need to create a separate account with Databricks.

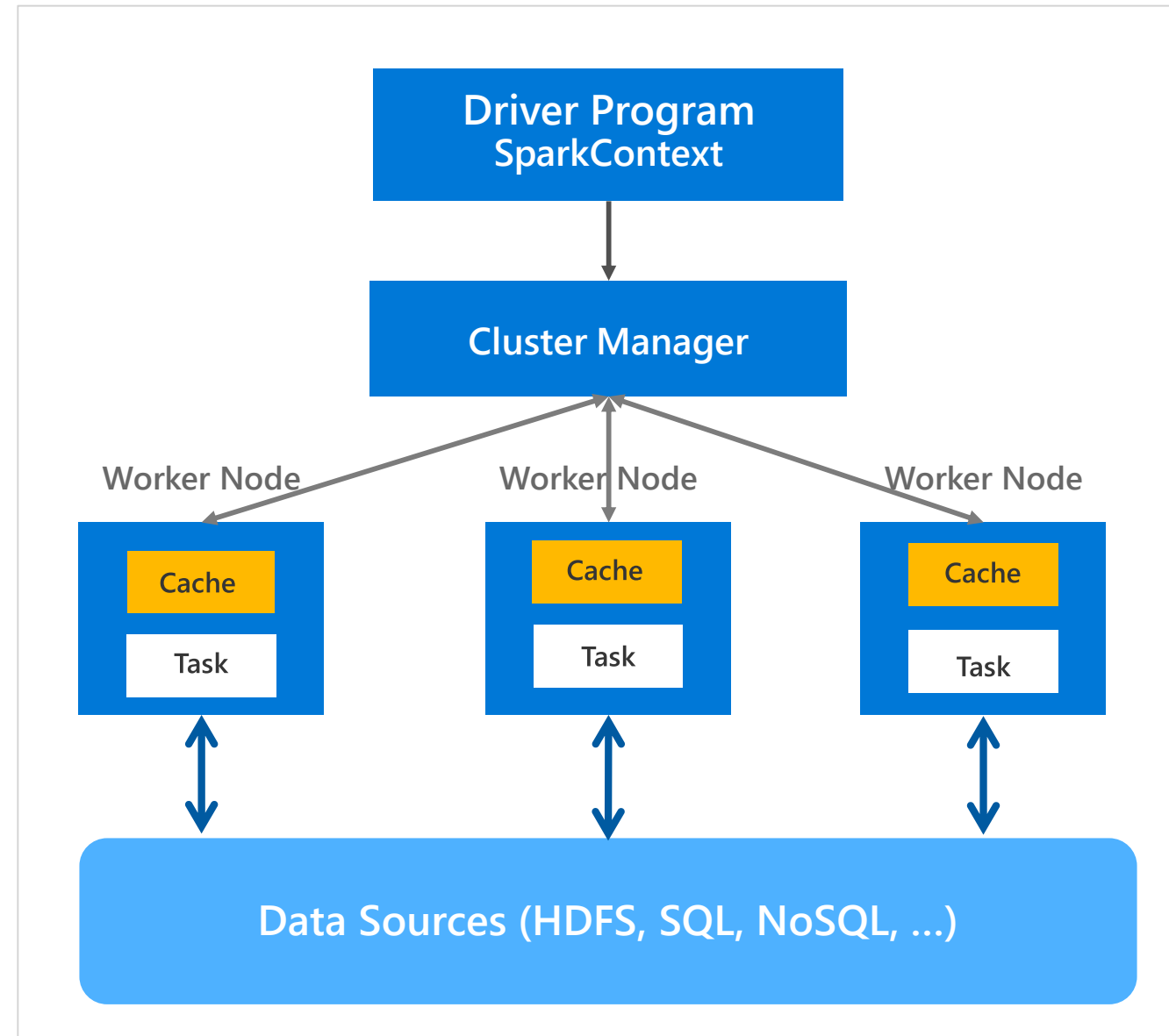


# Azure Databricks

## Core Concepts

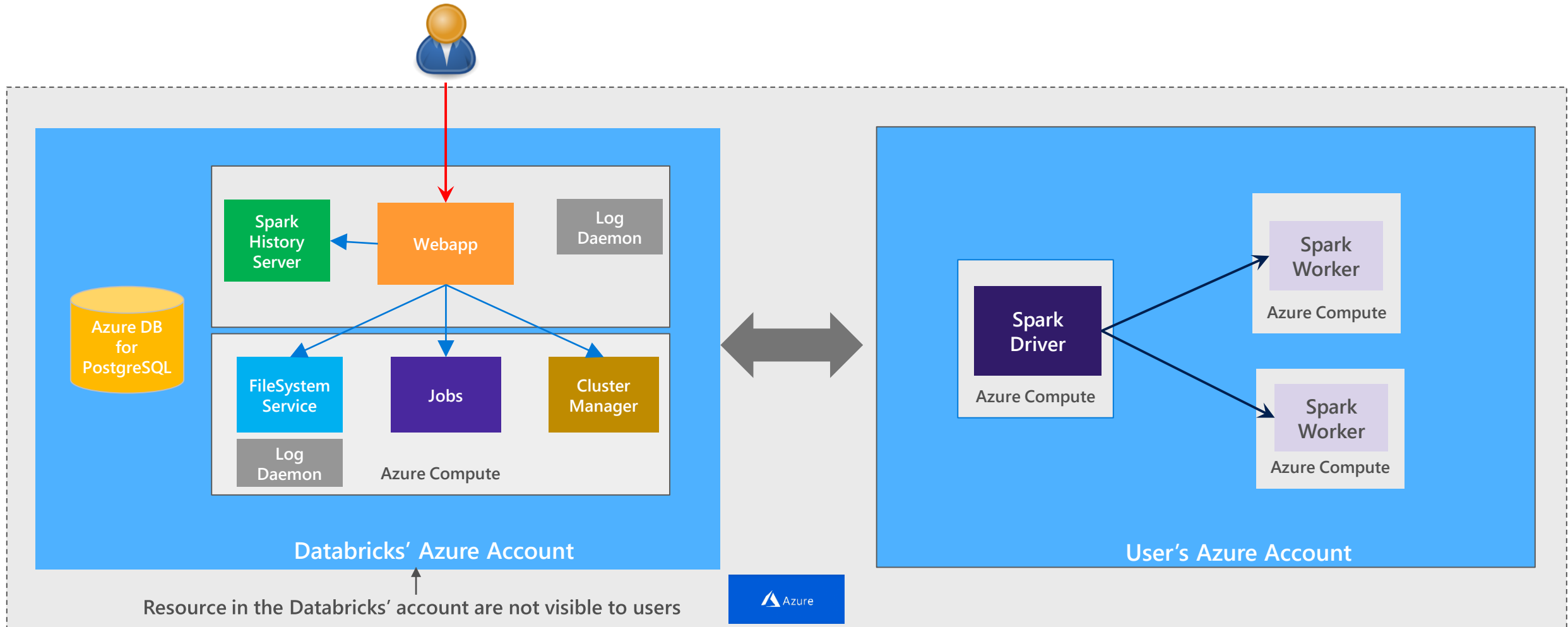
# GENERAL SPARK CLUSTER ARCHITECTURE

- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).





# AZURE DATABRICKS CLUSTER ARCHITECTURE



# SECURE COLLABORATION

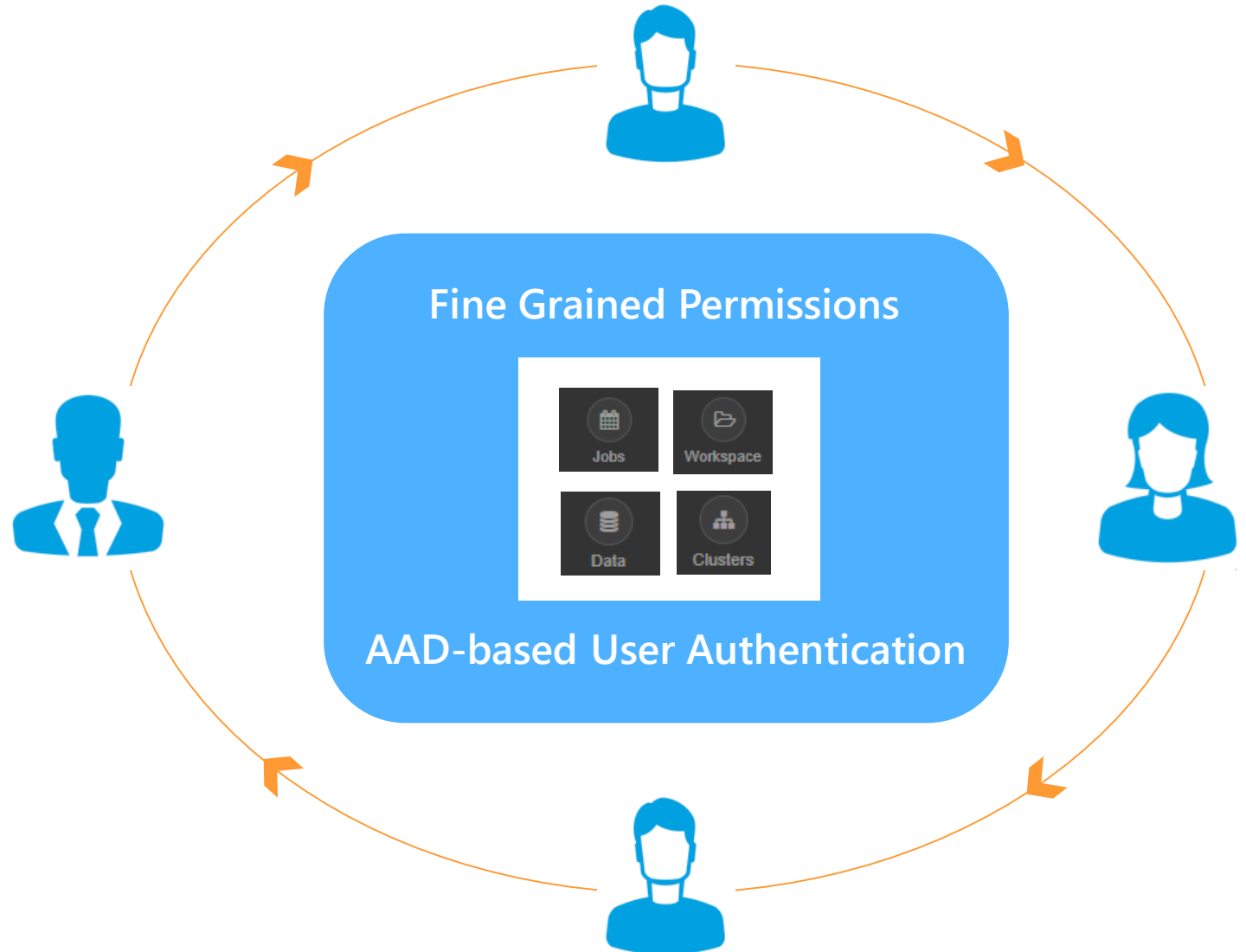
Azure Databricks enables *secure* collaboration between colleagues

- With Azure Databricks colleagues can *securely share* key artifacts such as Clusters, Notebooks, Jobs and Workspaces
- Secure collaboration is enabled through a combination of:

**Fine grained permissions:** Defines who can do what on which artifacts (access control)



**AAD-based authentication:** Ensures that users are actually who they claim to be



# DATABRICKS ACCESS CONTROL


Access control can be defined at the user level via the Admin Console

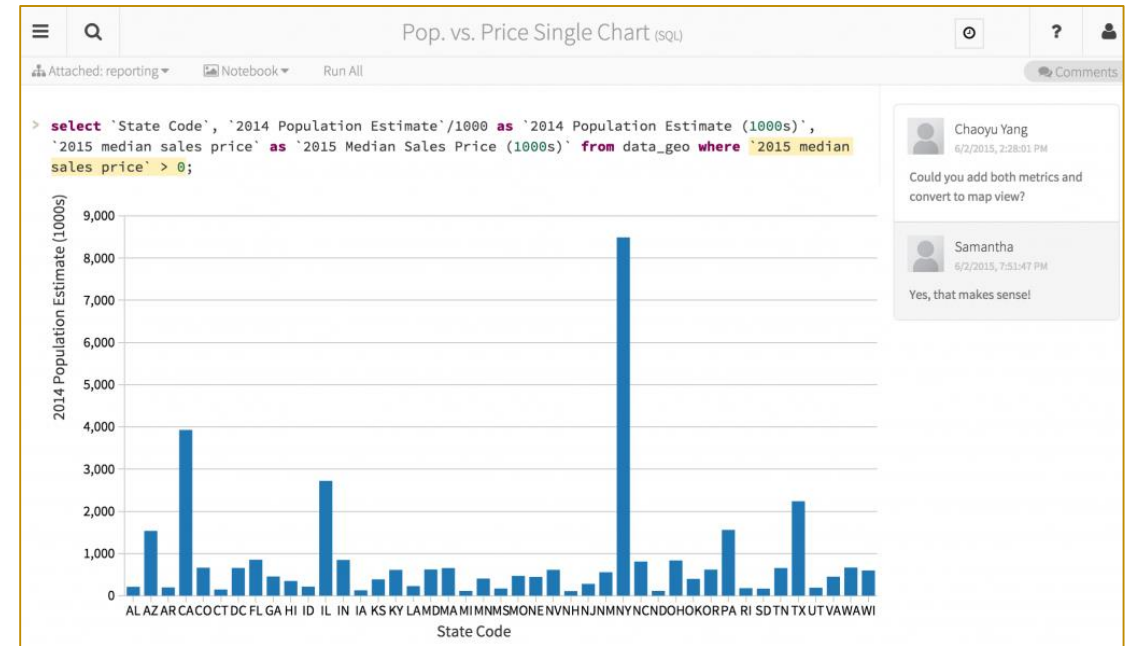
Access Control can be defined for Workspaces, Clusters, Jobs and REST APIs

Databricks Access Control	Workspace Access Control	Defines who can who can view, edit, and run notebooks in their workspace
	Cluster Access Control	Allows users to who can attach to, restart, and manage (resize/delete) clusters.  Allows Admins to specify which users have permissions to create clusters
	Jobs Access Control	Allows owners of a job to control who can view job results or manage runs of a job (run now/cancel)
	REST API Tokens	Allows users to use personal access tokens instead of passwords to access the Databricks REST API

# A Z U R E   D A T A B R I C K S   N O T E B O O K S   O V E R V I E W

Notebooks are a popular way to develop, and run, Spark Applications

- Notebooks are not only for authoring Spark applications but can be *run/executed directly* on clusters
  - **Shift+Enter**
  - click the  at the top right of the cell in a notebook
  - Submit via Job
- Notebooks support fine grained permissions—so they can be *securely shared* with colleagues for collaboration (see following slide for details on permissions and abilities)
- Notebooks are well-suited for prototyping, rapid development, exploration, discovery and iterative development



Notebooks typically consist of code, data, visualization, comments and notes

# MIXING LANGUAGES IN NOTEBOOKS

You can mix multiple languages in the same notebook

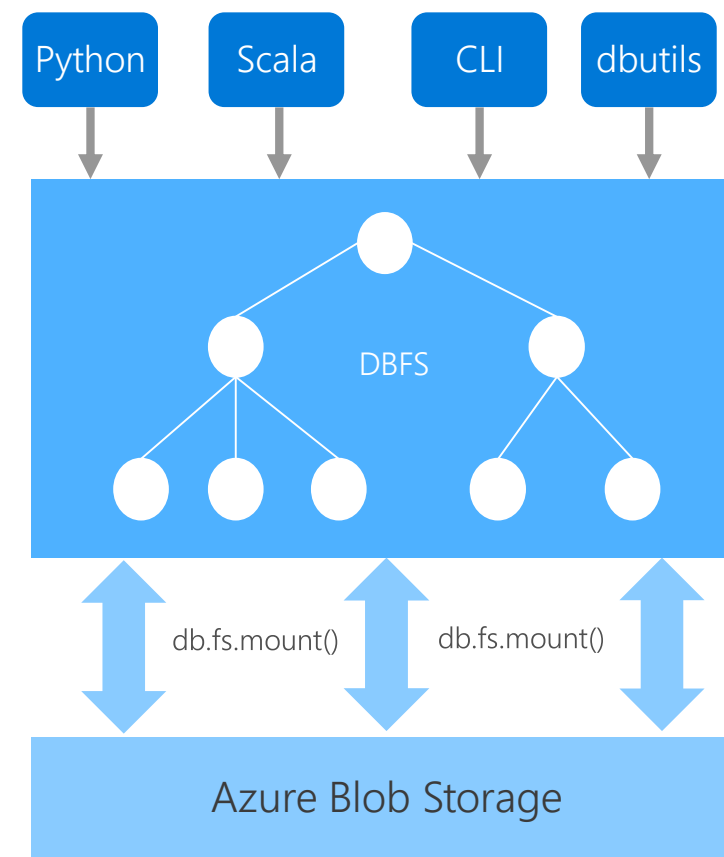
Normally a notebook is associated with a specific language. However, with Azure Databricks notebooks, you can mix multiple languages in the same notebook. This is done using the language magic command:

- `%python` Allows you to execute python code in a notebook (even if that notebook is not python)
- `%sql` Allows you to execute sql code in a notebook (even if that notebook is not sql).
- `%r` Allows you to execute r code in a notebook (even if that notebook is not r).
- `%scala` Allows you to execute scala code in a notebook (even if that notebook is not scala).
- `%sh` Allows you to execute shell code in your notebook.
- `%fs` Allows you to use Databricks Utilities - dbutils filesystem commands.
- `%md` To include rendered markdown
- `%run` Allows you to run another notebook from within another

# DATABRICKS FILE SYSTEM (DBFS)

Is a distributed File System (DBFS) that is a layer over Azure Blob Storage

- Azure Storage buckets can be mounted in DBFS so that users can directly access them without specifying the storage keys
- DBFS mounts are created using `dbutils.fs.mount()`
- Azure Storage data can be cached locally on the SSD of the worker nodes
- Available in both Python and Scala and accessible via a DBFS CLI
- Data persist in Azure Blob Storage – is not lost even after cluster termination
- Comes pre-installed on Spark clusters in Databricks







© 2017 Microsoft Corporation. All rights reserved. Microsoft, Windows, and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.