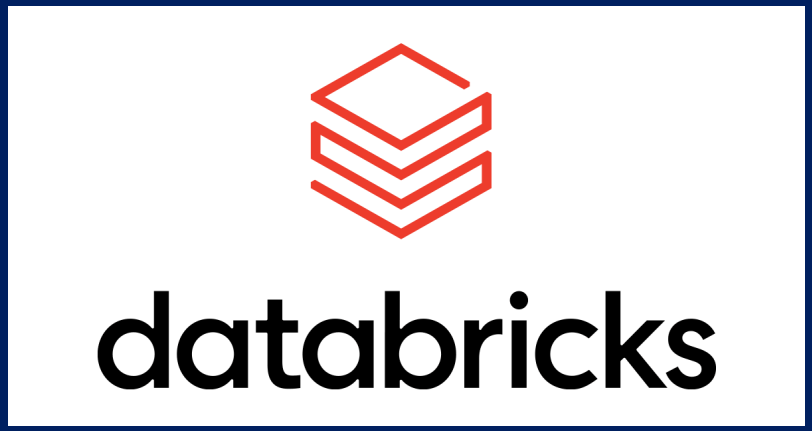
Databricks
Overview



Bryan Cafferky
Big Data and AI Consultant

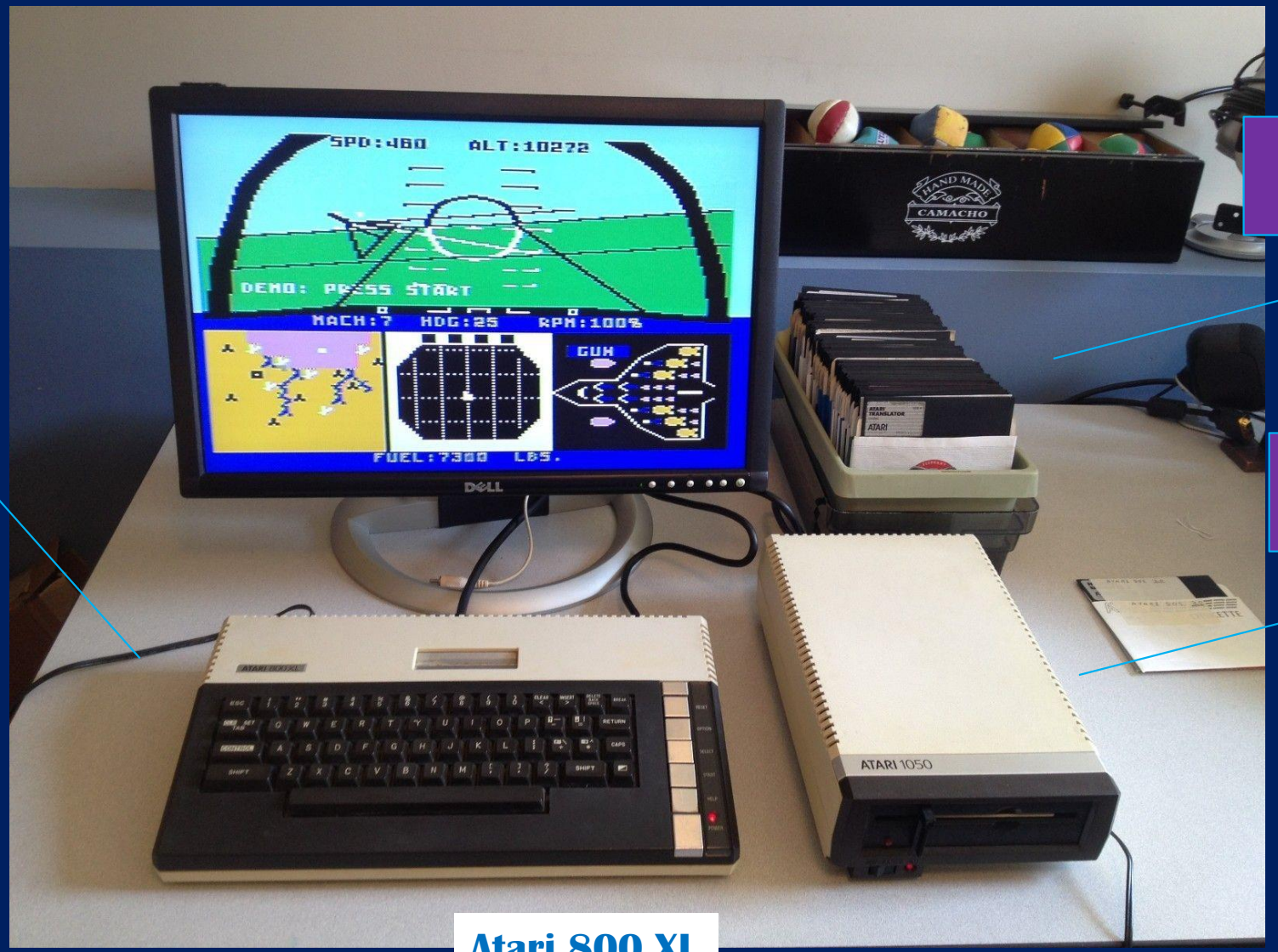
https://github.com/bcafferky/shared

Remember how you started?

Proprietary Sound and Graphics Chips

- 6502
- ANTIC
- GTIA

Sound
- POKEY

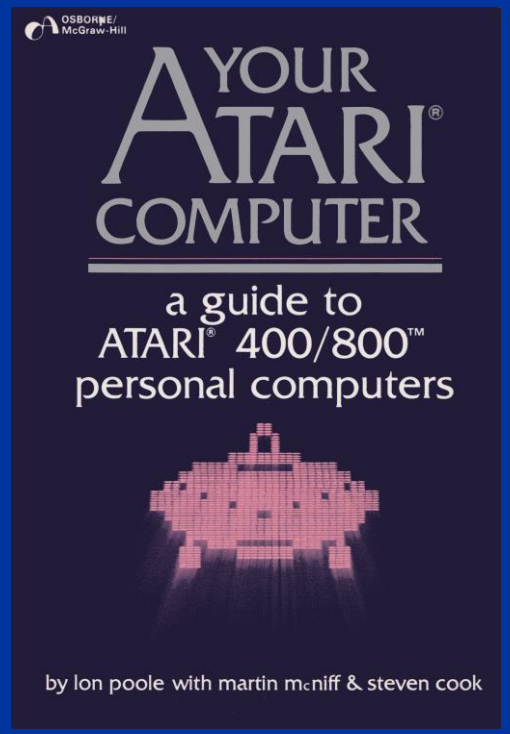


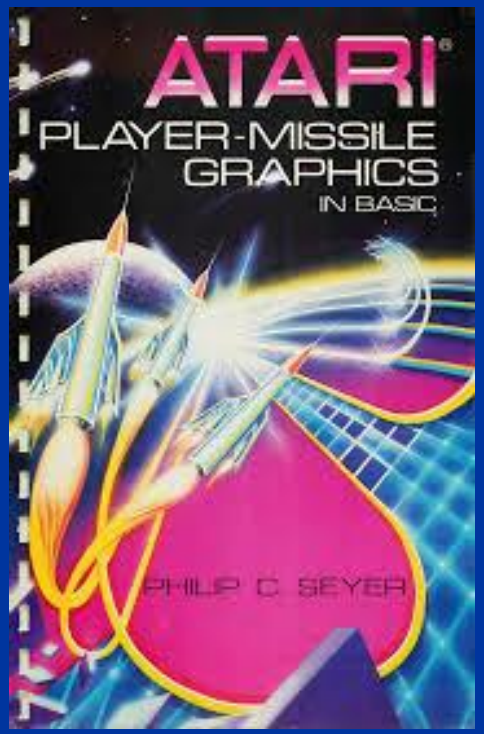
Atari 800 XL

My Programs

170k 5 ¼" Floppy

The Excitement and Fun!





Time to Find Your Inner Child!

Where Are We Heading?

Your Mission Should You Choose to Accept It

Scale Up vs. Scale Out and Barry the Weightlifter

The Apache Hadoop Project & Spark

What is Spark?

What is Databricks?

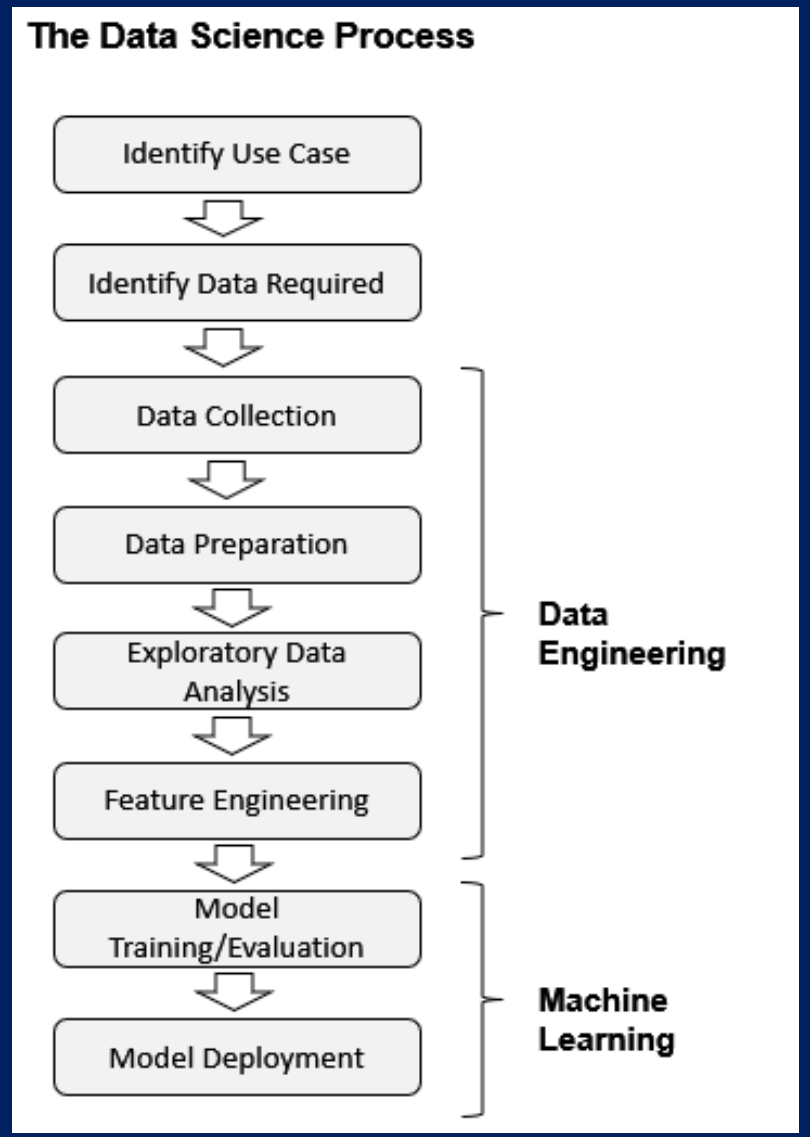
Your mission…

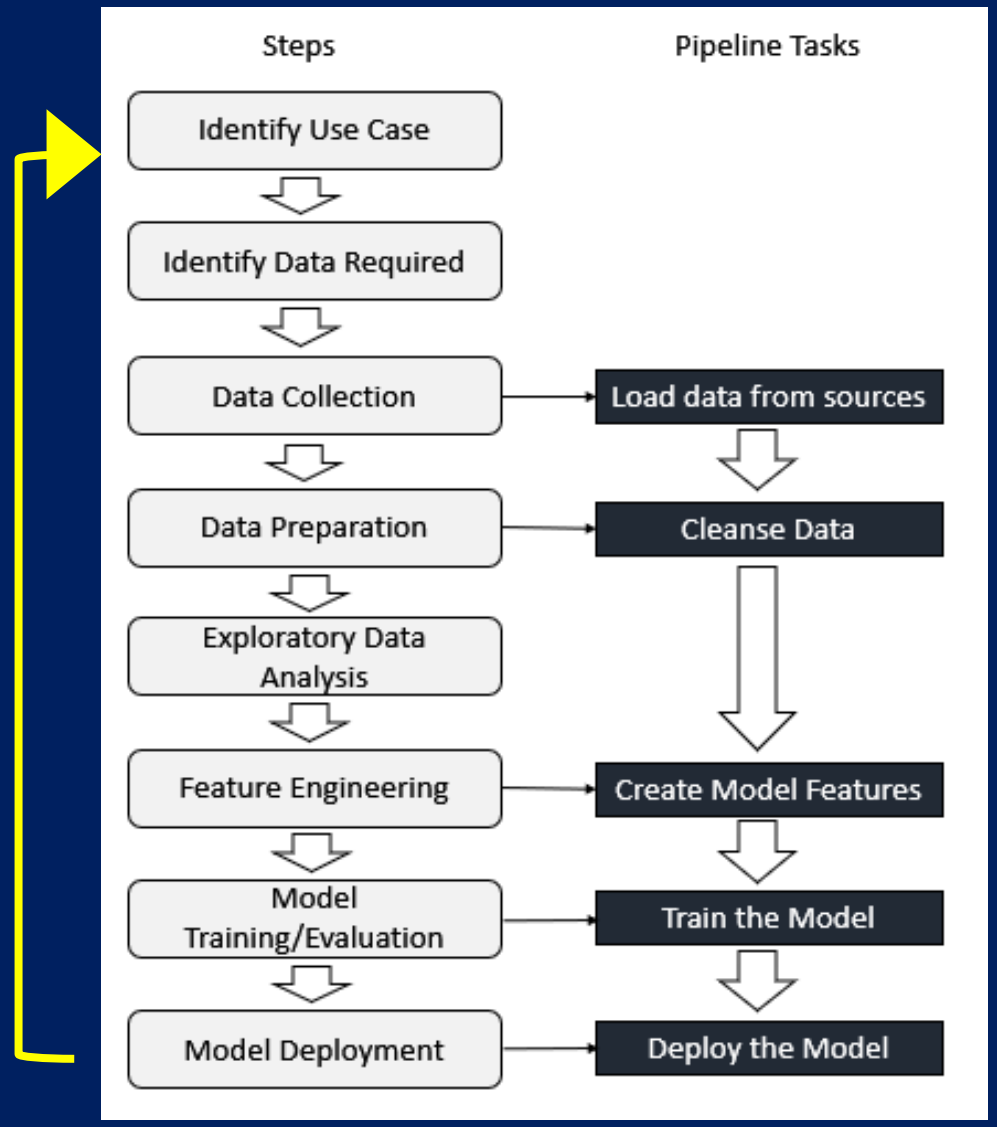
AdventureWorks is new to using data analysis and AI tools.
Management adopted Databricks for these tasks and assigned you to:

- Analyze sales data to identify trends and patterns like what products are growing and in which demographics.
- Provide these insights via dashboards with compelling visualizations.
- Develop a machine learning model that will predict the bike a customer is likely to purchase.
- Create a repeatable process to maintain the above assets.

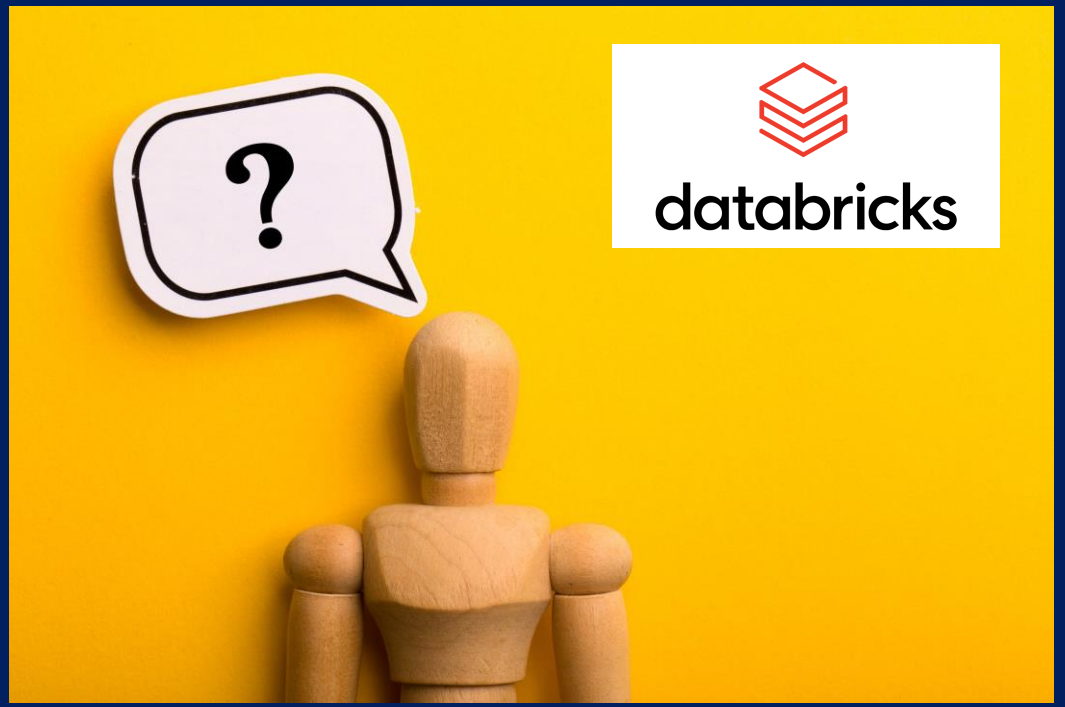
* Historical data is stored in the company's SQL Server data warehouse.

The Data Science Process





Can Databricks Help Us?



Clarifying Confusing Terms

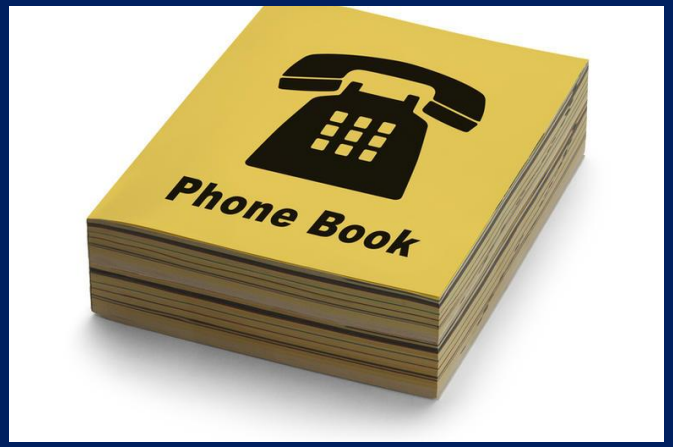
Analytics	Visualizations and/or Machine Learning
~~Big~~ Data Challenging	Non-traditional: Movies, Images, massive, streaming
Data Lake	Storage where you place data files probably.
Artificial Intelligence	Includes AI, Machine Learning, Deep Learning, Predictive Modeling

Scale Up vs. Scale Out

Scale Up

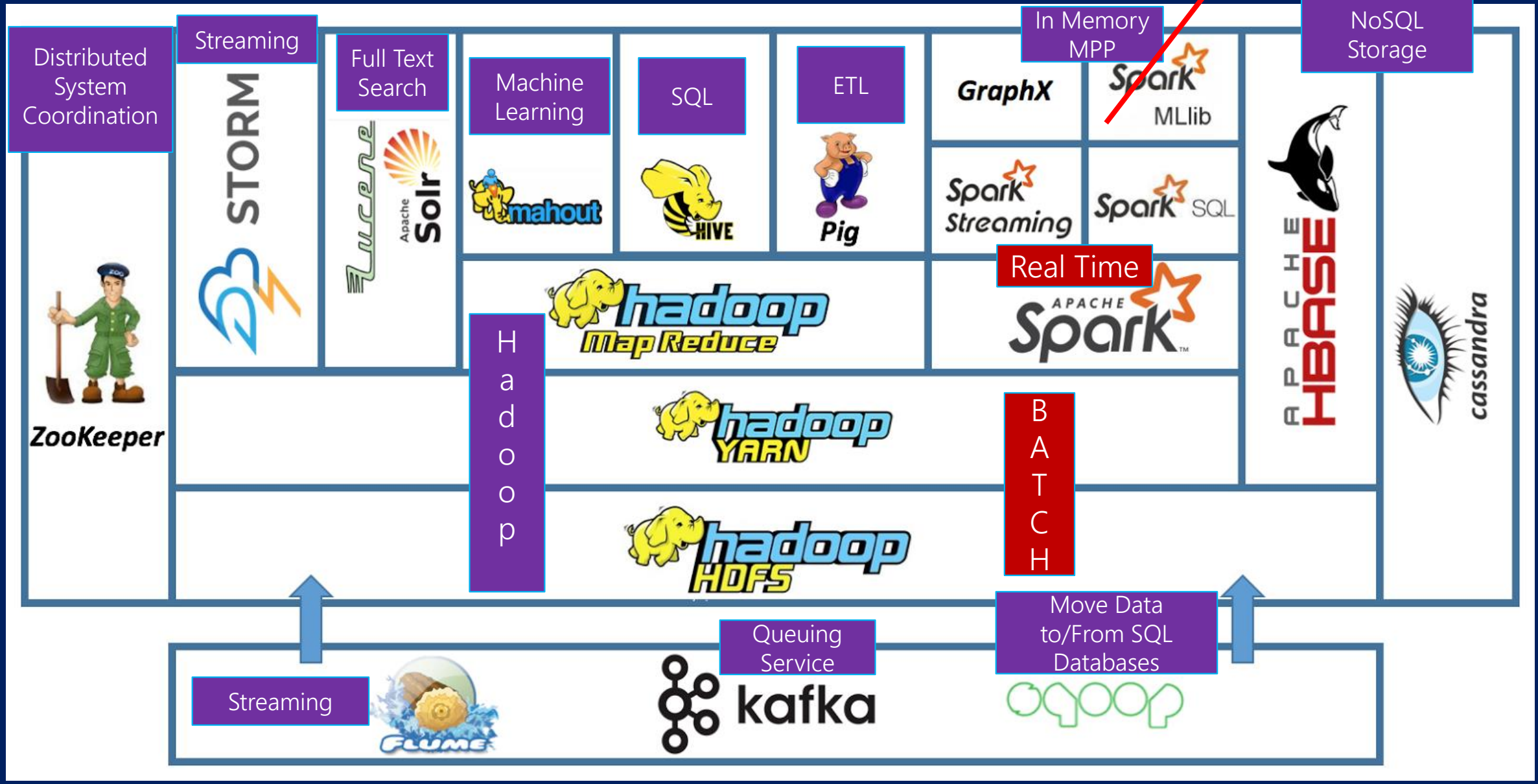
Scale Out

Scale Up vs. Scale Out



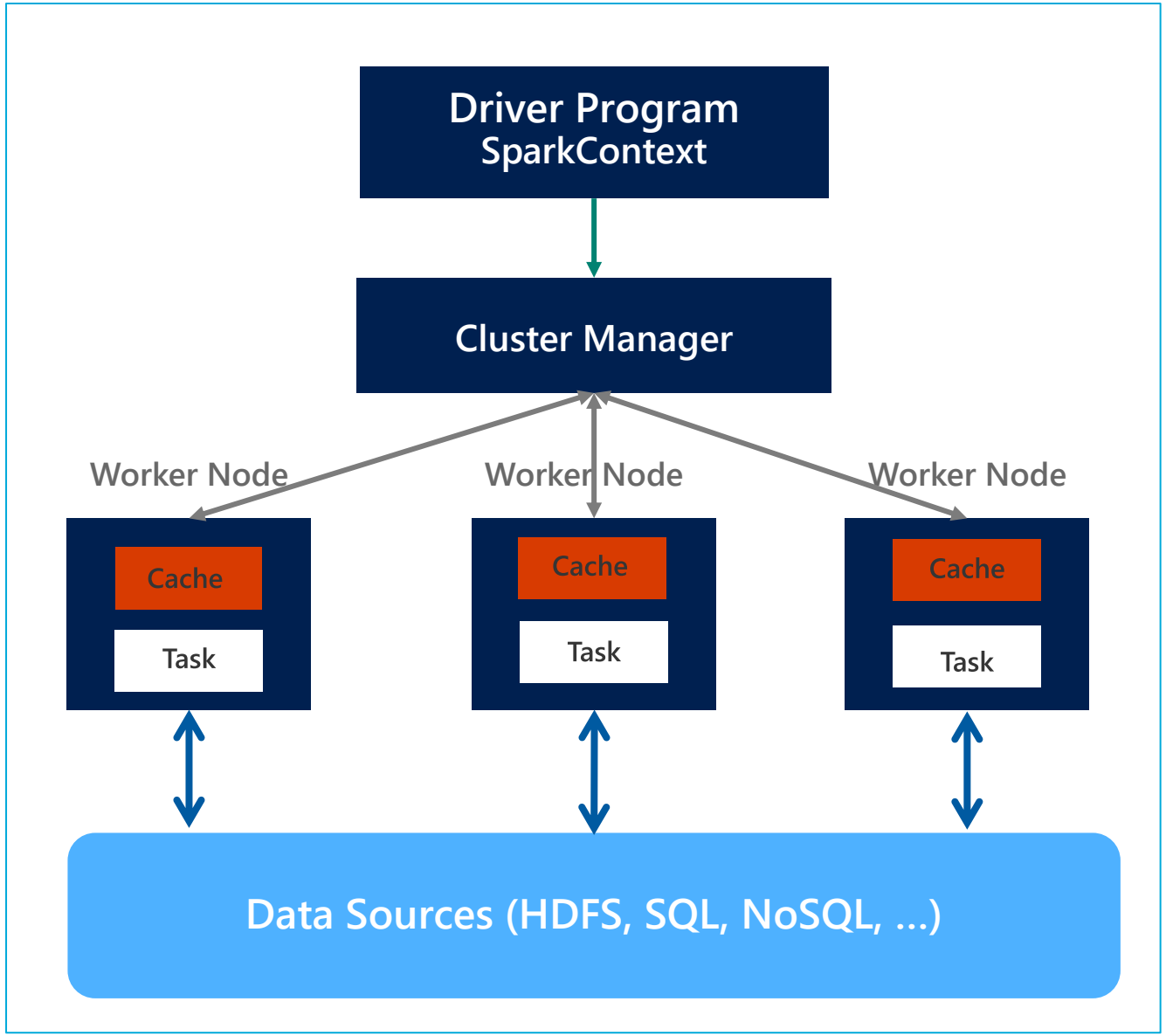
Scale Out

The Apache Hadoop Ecosystem



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).

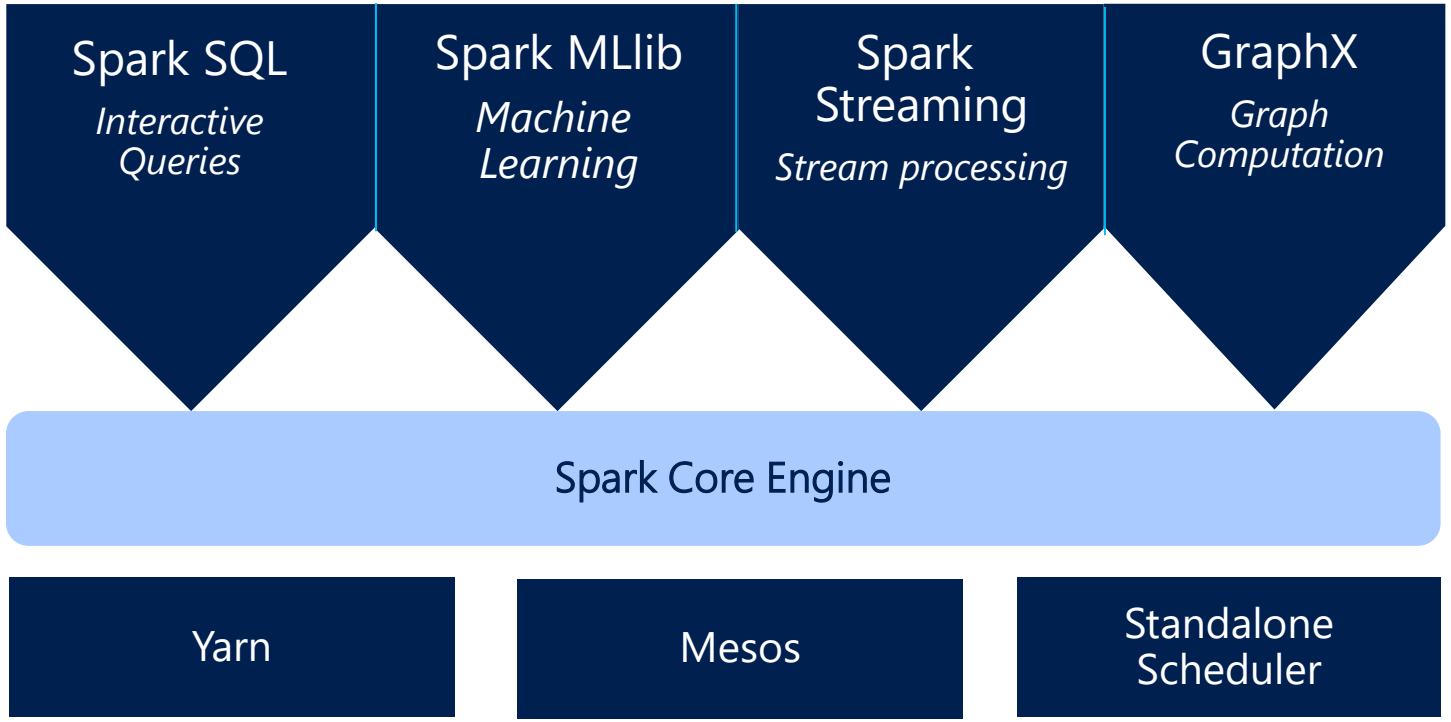


APACHE SPARK

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

Spark Unifies:

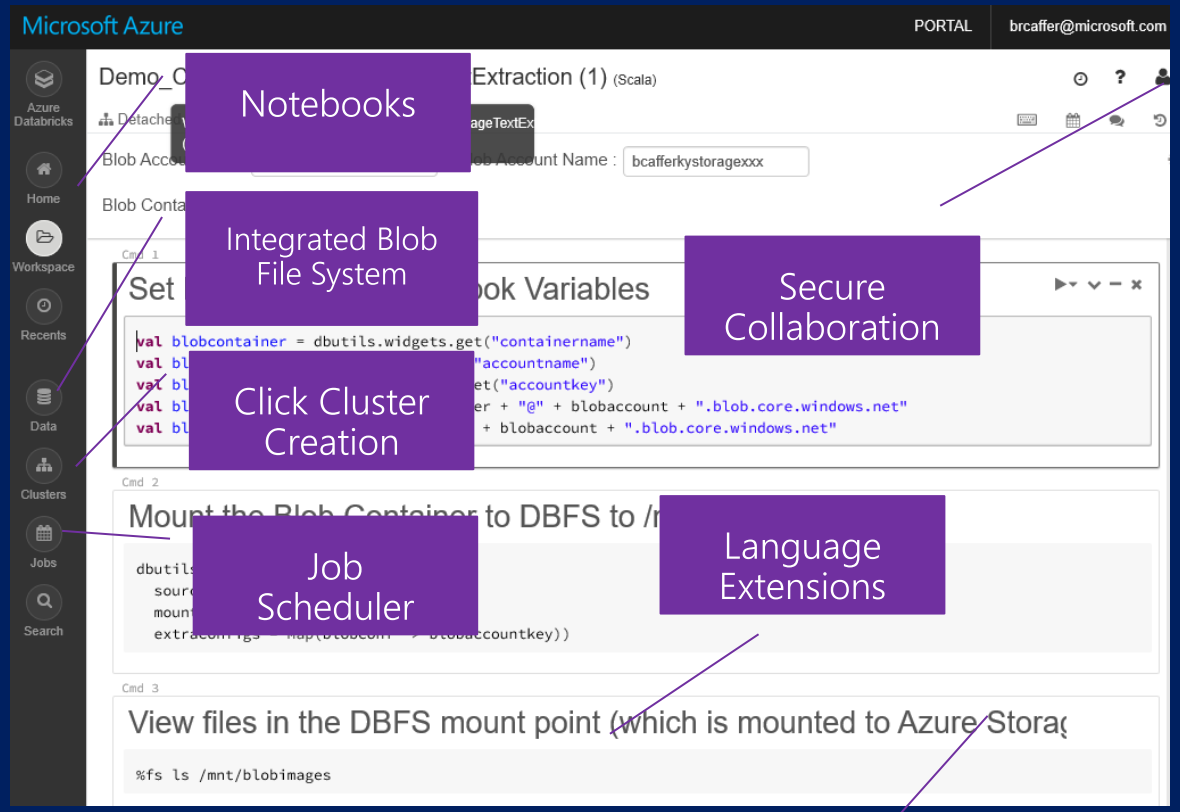
- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing



Scala Java python R

Azure Databricks

It all runs on Spark



Notebooks

Integrated Blob File System

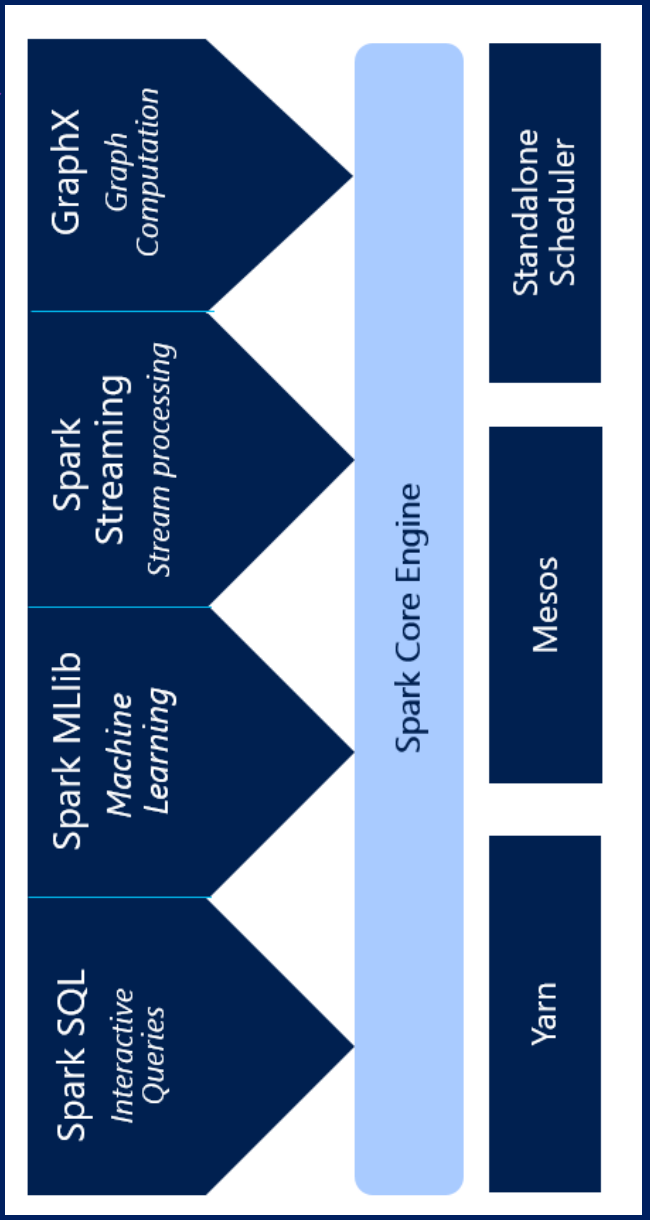
Click Cluster Creation

Job Scheduler

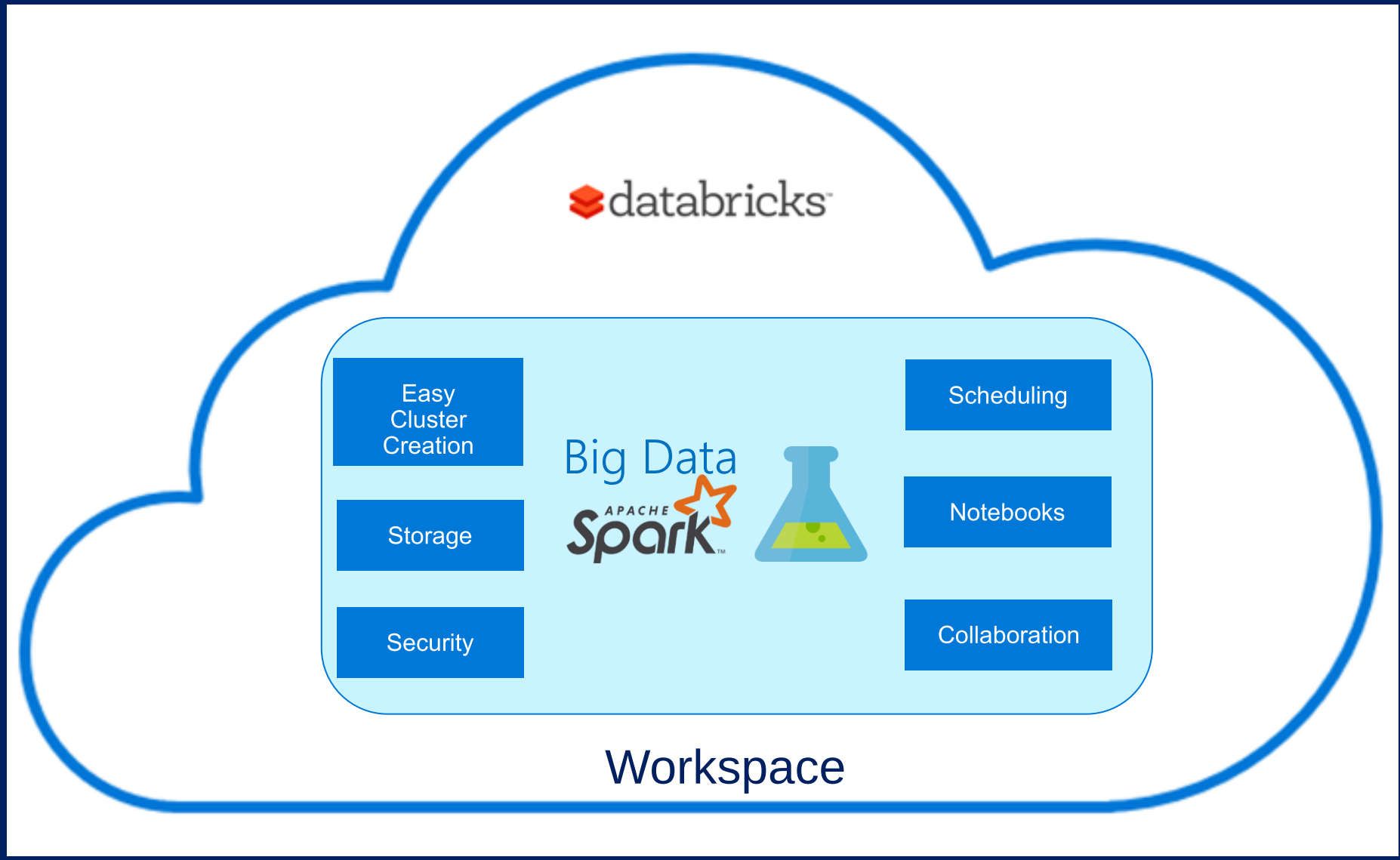
Secure Collaboration

Language Extensions

Spark Optimizations



Databricks Services



Demonstration

- Creating the Databricks Workspace
- Creating a Cluster
- Uploading Data
- Workspace
- Using Notebooks
  - Dashboards
  - Security
  - Cell Actions
  - Schedule
  - Revision History
  - Comments
- Libraries
- Jobs

https://docs.databricks.com/notebooks/notebooks-use.html

https://docs.databricks.com/notebooks/index.html

Demonstration Notebooks

Databricks Demo Notebooks: https://databricks.com/discover/notebook-gallery

https://github.com/dennyglee/databricks

https://databricks.com/resources?_sft_resource_type=example-notebook#databricks-jump-start

Databricks Demos: https://databricks.com/discover/demos

https://docs.databricks.com/notebooks/index.html

Where We've Been

Your Mission Should You Choose to Accept It

Scale Up vs. Scale Out and Barry the Weightlifter

The Apache Hadoop Project & Spark

What is Spark?

What is Databricks?

Lab Tasks

- Create a Cluster
- Create a SQL Table Using the GUI
- Upload the Remaining Files Using the GUI