Developing Batch Processing Solutions with Azure Databricks

Tim Warner

AUTHOR EVANGELIST, PLURALSIGHT

@TechTrainerTim TechTrainerTim.com

Overview

Manage data sources: Azure blog storage; Azure Data Lake Storage; Azure SQL Data Warehouse

Deploy Azure Databricks service and Spark cluster

Perform ETL batch processing operations from various data sources

Process data streamed from Event Hubs

Understand Azure Databricks

Azure Databricks

Microsoft's hosted Databricks environment

- Databricks is a hosted Apache Spark environment

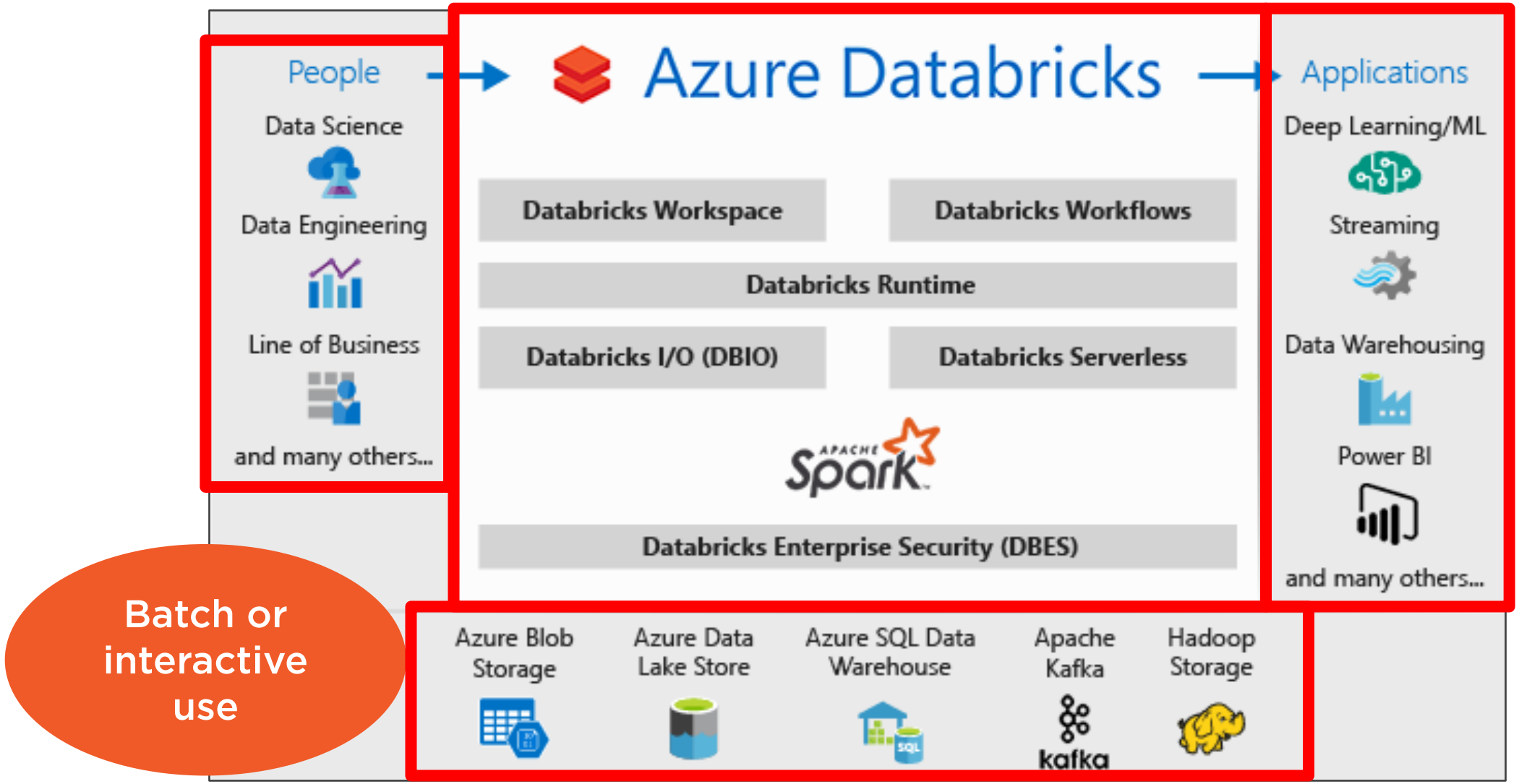
Fast, in-memory distributed data processing and analysis

Azure Databricks

Azure Databricks added value:

- Auto scale
- Multi-modal notebook model
- Built-in machine learning libraries
- Full Azure integration
  - RBAC/Azure AD
  - Tie-in with other Azure resources

Azure Databricks Ecosystem



https://timw.info/95cee

MapReduce vs. Spark

Batch processing

Disk data processing

Java API

Generally non-interactive

HiveQL

External ML support (Mahout)

Batch and real-time (streaming) processing

In-memory data processing

Scala, R, Python, SQL, Java APIs

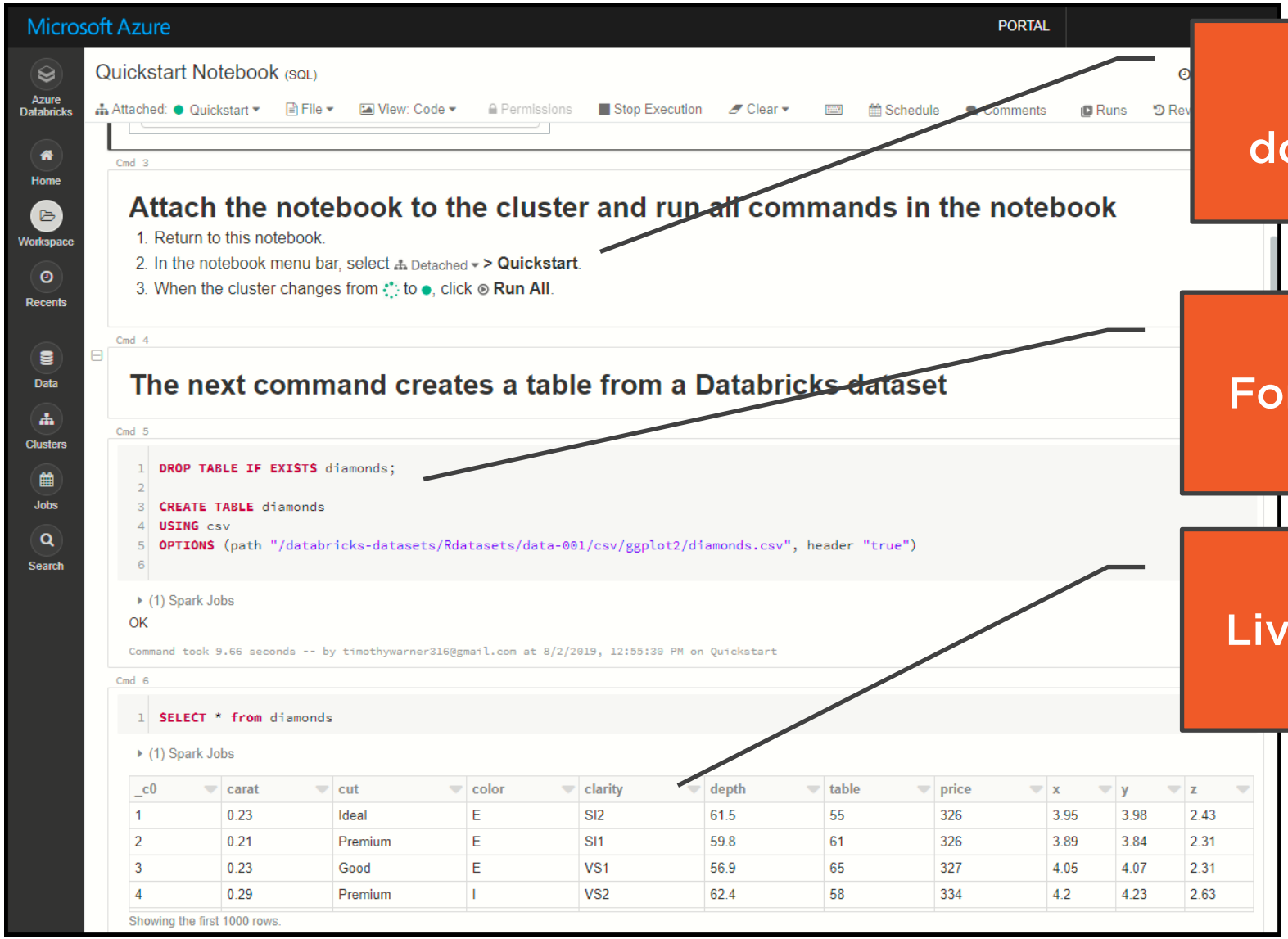
Interactive emphasis

Spark SQL

Native MLlib support

ETL with Azure Databricks

Notebook Paradigm for Data Analysis



Markdown documentation

Formatted code

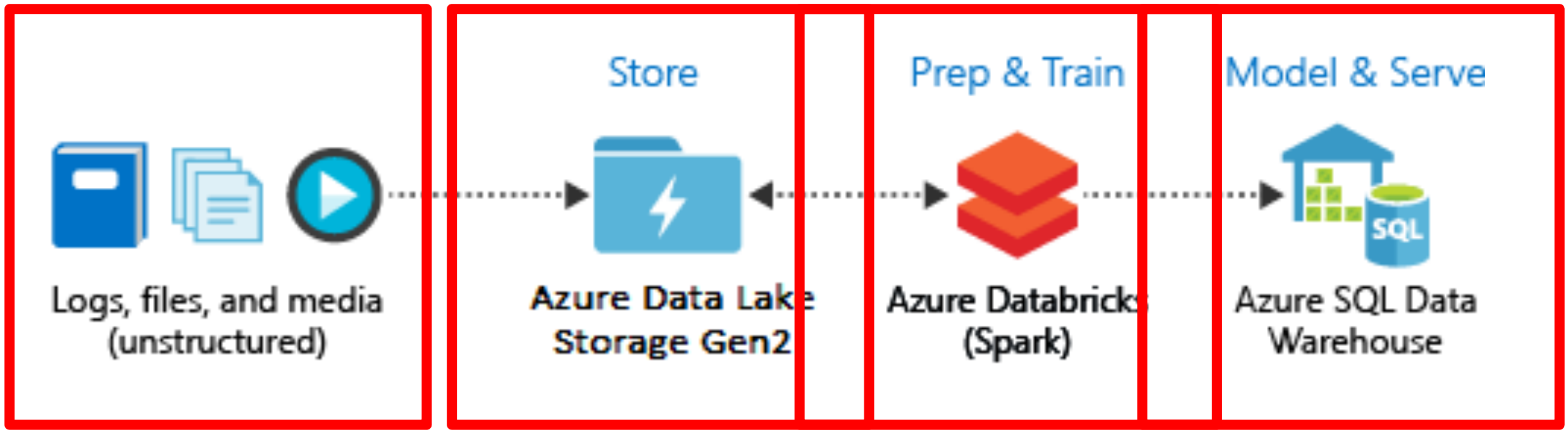
Live data results

Notebook Paradigm for Data Analysis

Collaboration

Visualizations

Our Databricks ETL Job



https://timw.info/11f4b

Demo

1

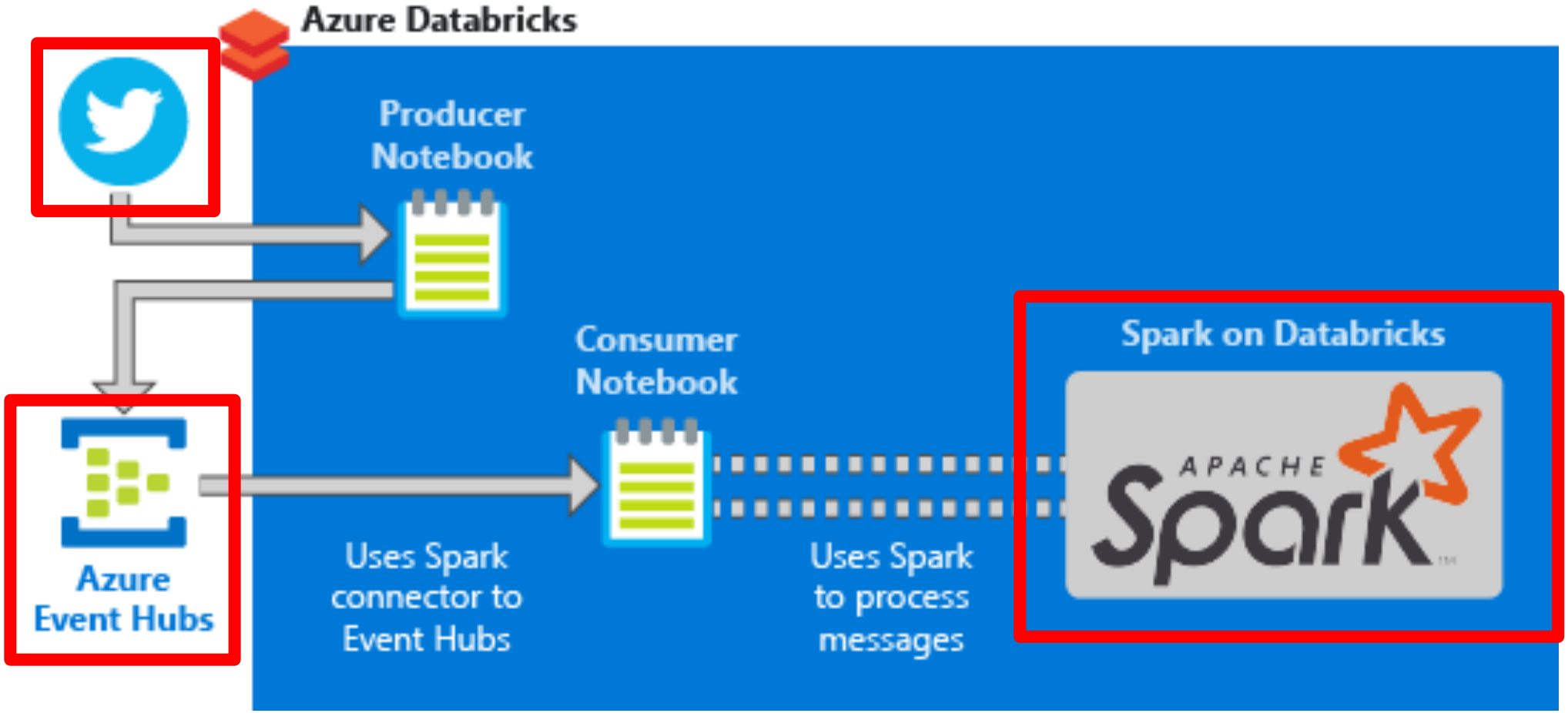
ETL with Databricks:
https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

Stream Processing with Azure Databricks

Azure Event Hub

- PaaS solution for high-volume data streaming and event ingestion
- Can receive and process millions of events per second
- Ingest telemetry/event streams to Data Lake Storage Gen2
- Forms the "front door" to Azure event pipelines

Our Databricks Event Streaming Job



https://timw.info/e170c

Demo

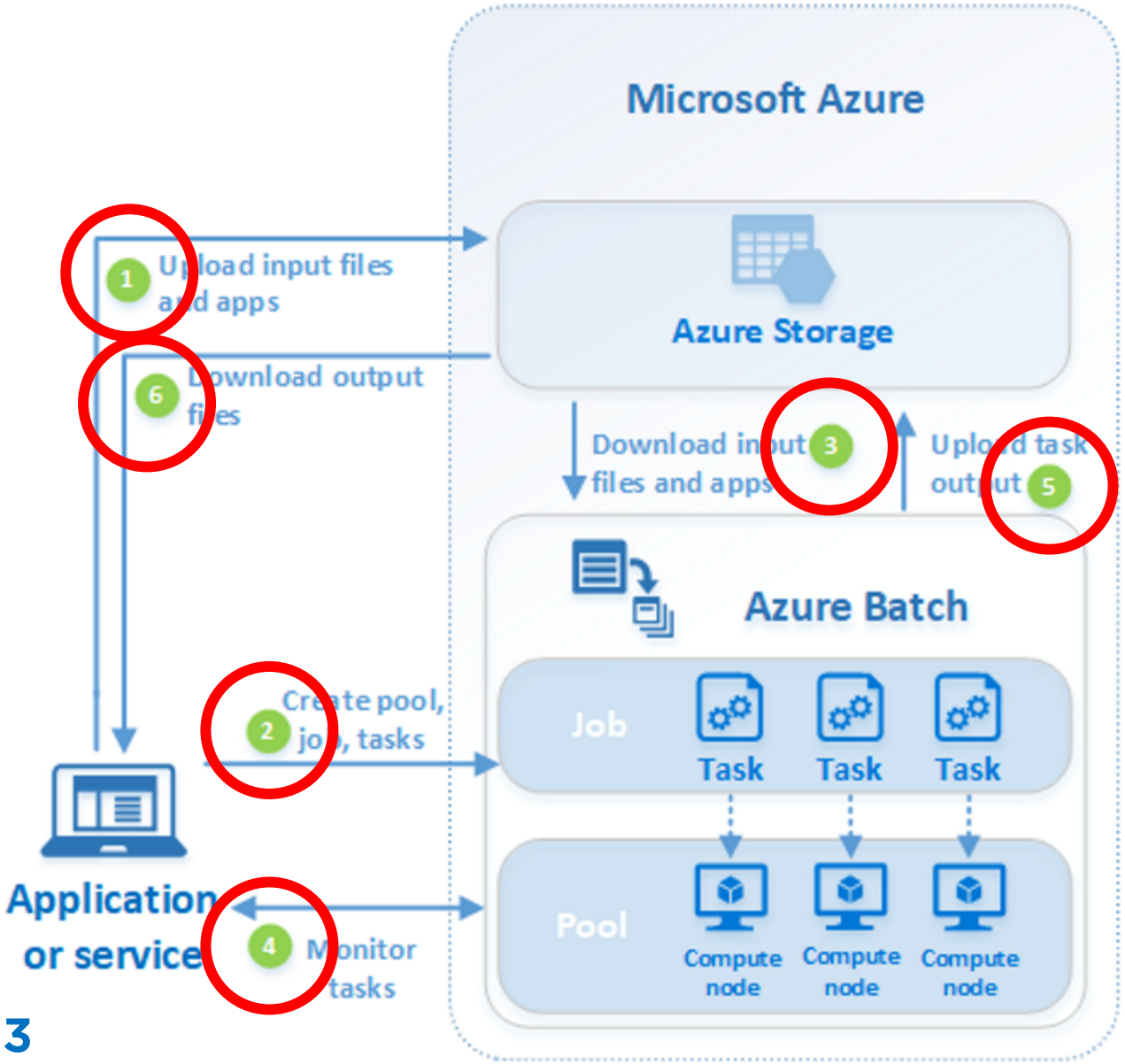
2

Stream Event Hub to Databricks: https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-stream-from-eventhubs

Azure Distributed Data Engineering Toolkit (aztk)

- Open-source Python CLI application
- Provision on-demand Spark clusters
- Programatically submit Spark jobs
- Built atop Azure Batch
- Employs (BYO) Docker containers
- Low-priority VMs offer 80% discount

Azure Batch



https://timw.info/3b733

For Further Learning

Getting Started with Spark 2
(Janani Ravi)
Remember that Apache Spark is the underlying technology

Design and Document Data Flows with Microsoft Azure
(John Savill)
See the module "Providing a Data Flow Solution" for Databricks coverage

Summary

Thank you!

Email: **tim-warner@pluralsight.com**

Twitter: **@TechTrainerTim**

Web: **TechTrainerTim.com**