

Developing Batch Processing Solutions with Azure Databricks



Tim Warner

AUTHOR EVANGELIST, PLURALSIGHT

@TechTrainerTim TechTrainerTim.com



Overview



Manage data sources: Azure blob 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

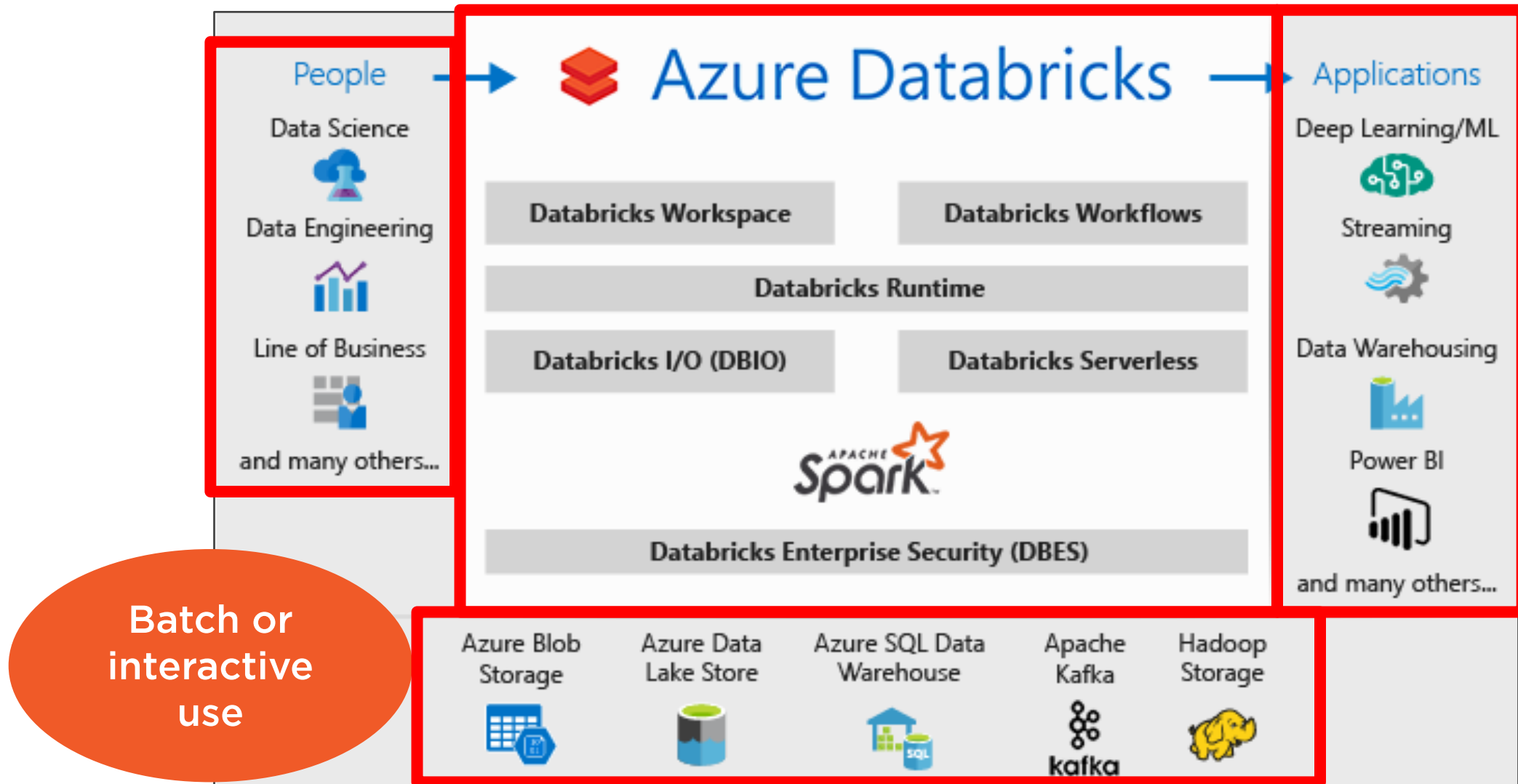


Azure Databricks added value:

- 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



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



Microsoft Azure

Quickstart Notebook (SQL)

Attached: Quickstart File View: Code Permissions Stop Execution Clear Schedule Comments Runs Rev

Cmd 3

Attach the notebook to the cluster and run all commands in the notebook

1. Return to this notebook.
2. In the notebook menu bar, select **Detached** > **Quickstart**.
3. When the cluster changes from to , click **Run All**.

Cmd 4

The next command creates a table from a Databricks dataset

Cmd 5

```
1 DROP TABLE IF EXISTS diamonds;
2
3 CREATE TABLE diamonds
4 USING csv
5 OPTIONS (path "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv", header "true")
6
```

► (1) Spark Jobs
OK
Command took 9.66 seconds -- by timothywarner316@gmail.com at 8/2/2019, 12:55:30 PM on Quickstart

Cmd 6

```
1 SELECT * from diamonds
```

► (1) Spark Jobs

_c0	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63

Showing the first 1000 rows.

Markdown
documentation

Formatted code

Live data results



Notebook Paradigm for Data Analysis

The screenshot displays a Databricks notebook interface. It features a main code editor with two command blocks. The first block, labeled 'Cmd 6', contains SQL code to drop and create a table named 'diamonds' from a CSV file. The second block, labeled 'Cmd 7', contains a SQL query to select all data from the 'diamonds' table. To the right of the code editor is a sidebar showing the user profile of 'Tim Warner' and a message about sample data. Above the main editor, a partial view of 'Cmd 11' is visible.

```
Cmd 6
1 DROP TABLE IF EXISTS diamonds;
2
3 CREATE TABLE diamonds
4 USING csv
5 OPTIONS (path "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv", header "true")
6
```

Cmd 7

```
1 SELECT * from diamonds
```

Tim Warner
8/2/2019, 1:03:50 PM

This is part of the sample data that Azure loads for you

Collaboration

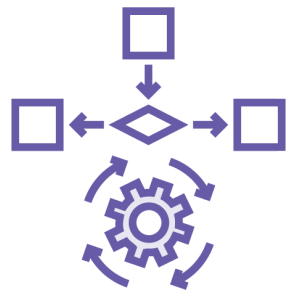
This block shows the visualization toolbar and command execution details. The toolbar includes icons for a table, a bar chart, a dropdown menu, a 'Plot Options...' button, and a download icon. Below the toolbar, a status bar indicates that the command took 2.11 seconds to execute, was run by 'timothywarner316@gmail.com' on 8/2/2019 at 12:55:30 PM, and was executed on the 'Quickstart' cluster.

Plot Options...

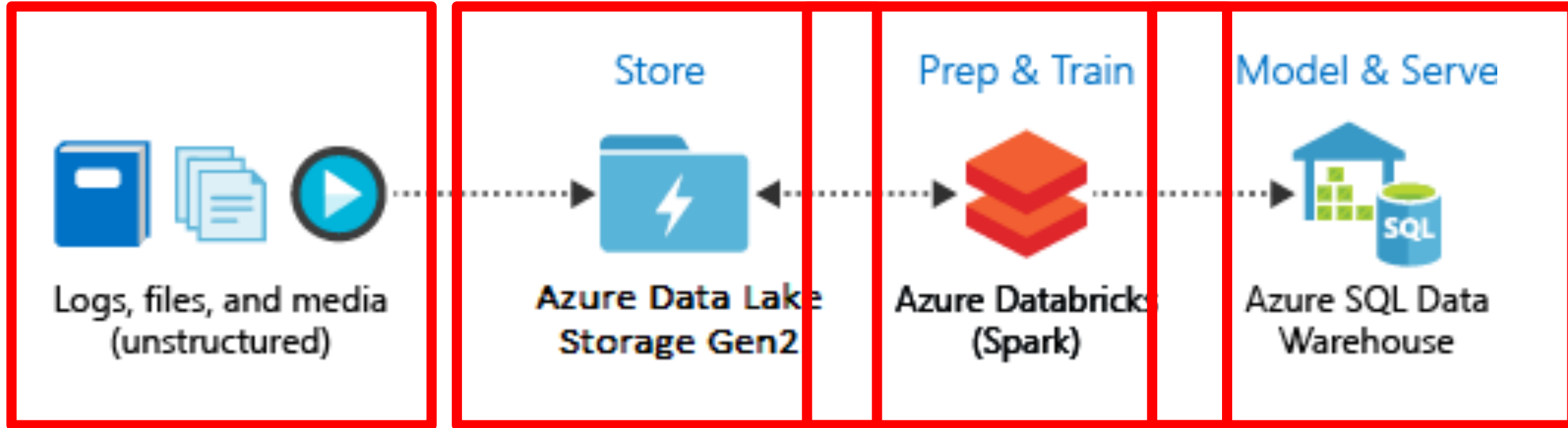
Command took 2.11 seconds -- by timothywarner316@gmail.com at 8/2/2019, 12:55:30 PM on Quickstart

Visualizations





Our Databricks ETL Job



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



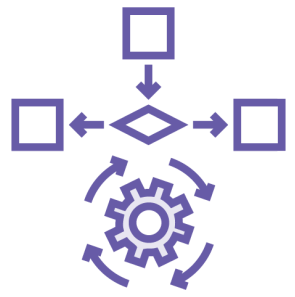
PaaS solution for high-volume data streaming and event ingestion

Can receive and process millions of events per second

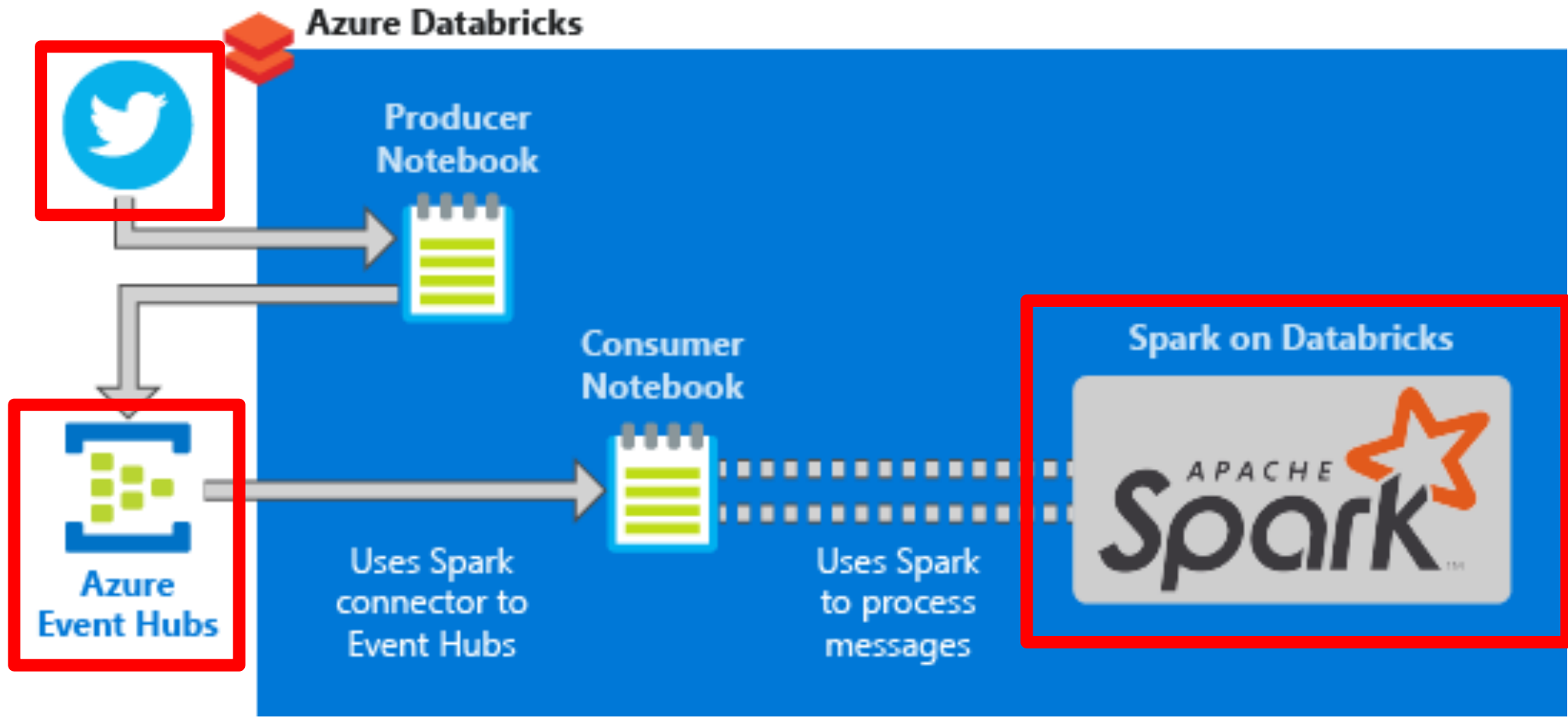
Ingest telemetry/event streams to Data Lake Storage Gen2

Forms the "front door" to Azure event pipelines





Our Databricks Event Streaming Job



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**

**Programmatically
submit Spark jobs**

**Built atop Azure
Batch**

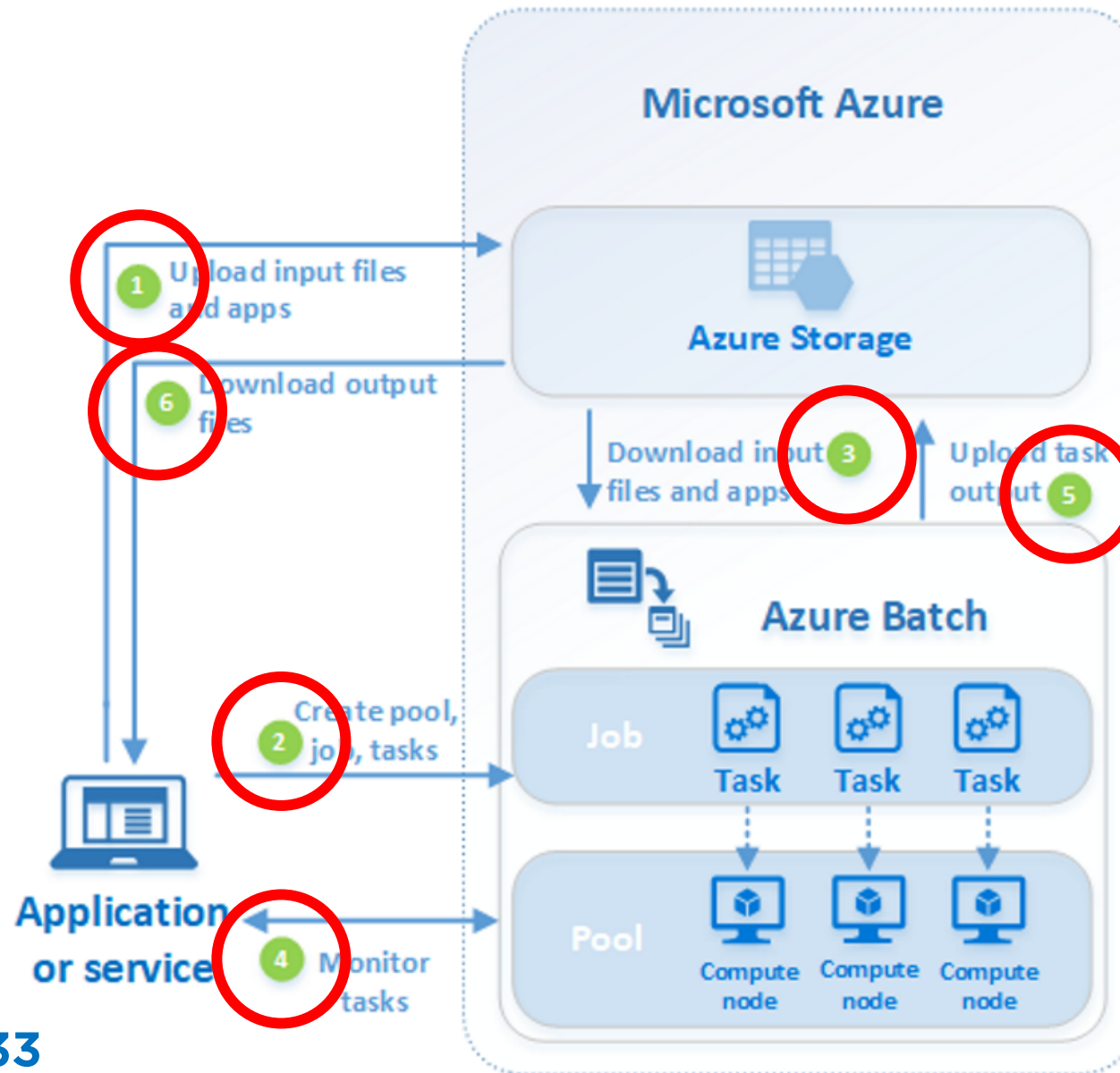
**Employs (BYO)
Docker containers**

**Low-priority VMs
offer 80%
discount**





Azure Batch





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](https://twitter.com/TechTrainerTim)

Web: TechTrainerTim.com

