

A Tour of Azure Databricks

Jonathan Wood
Software Consultant, Wintellect
@JWood



Wintellect Core Services



Consulting

Custom software application development and architecture



Instructor Led Training

Microsoft's #1 training vendor for over 14 years having trained more than 50,000 Microsoft developers



On-Demand Training

World class, subscription based online training

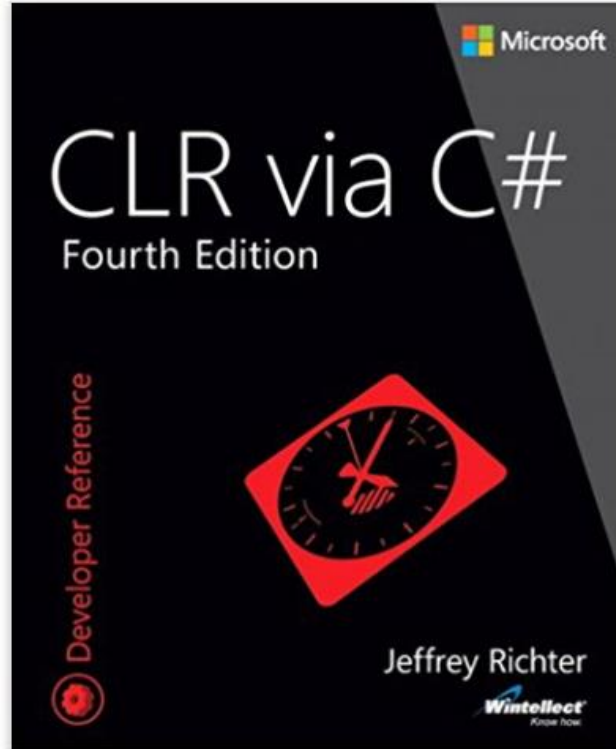


Gold Cloud Platform
Silver DevOps
Silver Application Development



Industry Influencers

We wrote the book (over 30 of them)



We help
companies
build better
software,
faster.

Some Highlights

- Gold Cloud Platform Partner, Gold DevOps Partner, Gold Data Platform Partner
 - Multiple ALM Rangers
- 2016 IAMCP Gold Partner of the Year for the U.S. announced at WPC
- CEO is Microsoft Regional Director (RD) for Atlanta
- Software Development competency partner
- Xamarin Premier Consulting Partner
 - Multiple Xamarin Certified Engineers
 - Chosen to teach the 2-day Xamarin University pre-con at Evolve 2016
- Other: Visual Studio Integration Partner, Azure Circle Partner, ALM Inner Circle Partner, MVP of the Year, and more...

Agenda

- What is Databricks?
- Why use Spark?
- Why Azure Databricks?
- Components of Databricks
 - Clusters
 - Notebooks
 - Advanced features
 - Jobs
- Demos

Poll

What is Databricks?

An easy to use, collaborative, Apache Spark-based analytics platform.



What is Databricks?



Data Engineer

Loads data



Data Scientist

Analyzes data



Business Analyst

Review and
make decisions
from data

What is Databricks?

- Bring teams together in an interactive workspace.
- Fully managed environments with one-click setup.
- Integrate with a wide variety of data stores and services
- Add advanced AI capabilities instantly and share insights



What is Databricks?

Built around Spark

- Optimized for performance
- Support for Python, R, Scala, and machine learning APIs



Why use Spark?

- In-memory engine
 - Up to 100x faster than MapReduce
- DataFrame API
 - Optimized for performance
 - Easy to use for analysis
- Machine Learning library
- Streaming API

Why Azure Databricks?

Microsoft Azure

Integrations

- Power BI
- Azure Storage

Security

- Active Directory
- Compliance

Auto scaling

- Scales by load
- Reduces cost

Clusters

Easy setup

Fully managed

Configure Spark engine

Where notebooks run

Auto scales based on load

Terminate after a set time of inactivity

demo Edit Start Clone

Configuration Notebooks (0) Libraries (0) Event Log Spark UI Driver Logs Spark Cluster UI - Master ▾

Cluster Type

Serverless Pool (beta, R/Python/SQL) Standard [Learn more about Serverless Pools ?](#)

Databricks Runtime Version

4.0 (includes Apache Spark 2.3.0, Scala 2.11)

Python Version ?

3

Driver Type

Standard_DS3_v2 14.0 GB Memory, 4 Cores, 0.75 DBU

Worker Type

Standard_DS3_v2 14.0 GB Memory, 4 Cores, 0.75 DBU

Min Workers

2

Max Workers

8

☒ Enable Autoscaling ?

Auto Termination ?

☒ Terminate after

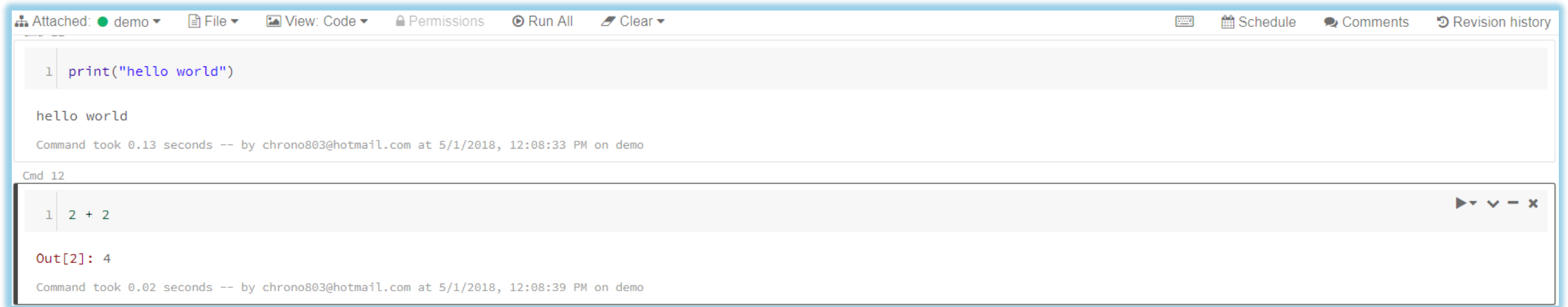
30

 minutes of inactivity

Notebooks

Interface to interact with Databricks

Can be in Python, R, Scala, or SQL for code, and markdown for text



The screenshot displays the Databricks notebook interface. At the top, there is a toolbar with options: Attached: demo, File, View: Code, Permissions, Run All, and Clear. On the right side of the toolbar are icons for keyboard shortcuts, Schedule, Comments, and Revision history.

The first code block contains the following Python code:

```
1 print("hello world")
```

The output of this code is:

```
hello world
```

Below the output, a status message reads: "Command took 0.13 seconds -- by chrono803@hotmail.com at 5/1/2018, 12:08:33 PM on demo".

The second code block is labeled "Cmd 12" and contains the following code:

```
1 2 + 2
```

The output of this code is:

```
Out[2]: 4
```

Below the output, a status message reads: "Command took 0.02 seconds -- by chrono803@hotmail.com at 5/1/2018, 12:08:39 PM on demo".

Advanced Notebook Features

Widgets

- Execute notebook with different parameters
- Dropdown, text, combo box, multiselect



hello this is a widget : 7

```
Cmd 1
1 dbutils.widgets.dropdown("1", "1", [str(x) for x in range(1, 10)], "This is a widget")

Command took 0.11 seconds -- by chrono803@hotmail.com at 5/1/2018, 1:38:16 PM on demo
```

```
Cmd 2
1 dbutils.widgets.get("1")

Out[2]: '7'

Command took 0.02 seconds -- by chrono803@hotmail.com at 5/1/2018, 1:44:04 PM on demo
```

Workflows

- Run multiple notebooks in a pipeline
- Can pass parameters

```
1 result = 2 + 2
2 dbutils.notebook.exit(int(result))
```

```
1 dbutils.notebook.run("./run", 60)
```

Notebook job #2

Out[1]: '4'

Jobs

Run notebooks on a schedule

Set alerts for job failures or timeouts

Set number of retries if job fails

Can also set wait time between retries

Can use existing cluster or create new cluster

Create Schedule

Schedule

Every ▾ hour ▾ starting at 00 ▾ : 00 ▾ US/Pacific ▾

☐ Show Cron Syntax

A new cluster will be created each time this schedule runs. You can modify settings from the [Jobs page](#) once the job is created.

Cancel

Ok

Run 1 of ETL

✕ Delete

Started: 2018-05-01 15:04:52 EDT

Duration: 27s

Status: Succeeded

Job ID: 8

Run ID: 7

Task: Notebook at /Users/chrono803@hotmail.com/workflow_example

▸ Parameters:

Cluster: demo (42 GB, Running, 4.0 (includes Apache Spark 2.3.0, Scala 2.11)) - [View Spark UI](#) / [Logs](#)

Output

```
dbutils.notebook.run("./run", 60)
```

Notebook job #9

Out[1]: '4'

Command took 16.98 seconds

Demo

Get Started with Azure Databricks

Demo

Sample ETL Pipeline

Takeaways

Clusters

- Very easy to create
- Reduce costs

Notebooks

- Version control
- Comments
- Widgets

Jobs

- Create schedules
- Run on new cluster
- Alerts

Integration

- Blob Storage
- SQL Data Warehouse

Questions?

Jonathan Wood
Software Consultant
@JWood