A Tour of Azure Databricks

Jonathan Wood
Software Consultant, Wintellect
@JWood

Wintellect Core Services

Consulting

Custom software application development and architecture

Instructor Led Training

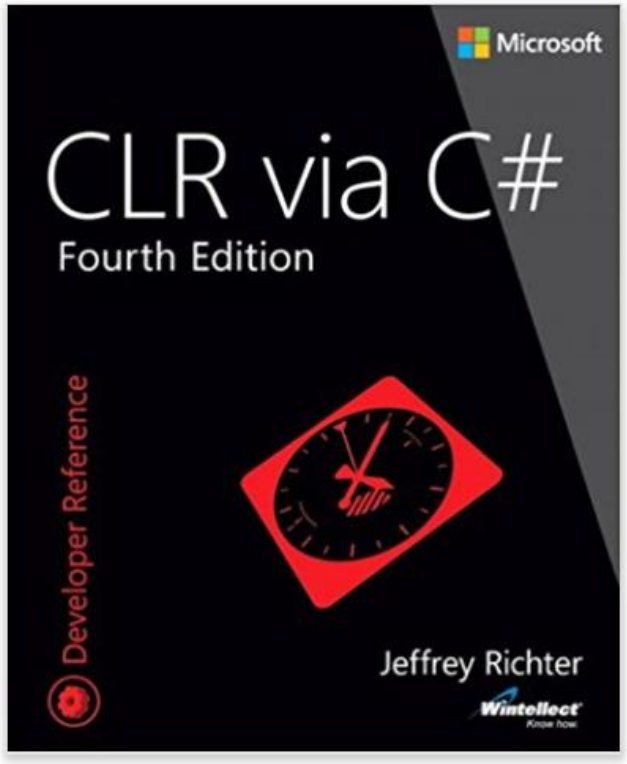
Microsoft's #1 training vendor for over 14 years having trained more than 50,000 Microsoft developers

On-Demand Training

World class, subscription based online training

Industry Influencers

We wrote the book (over 30 of them)



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	Security	Auto scaling
• Power BI • Azure Storage	• Active Directory • Compliance	• Scales by load • Reduces cost

Clusters

Easy setup

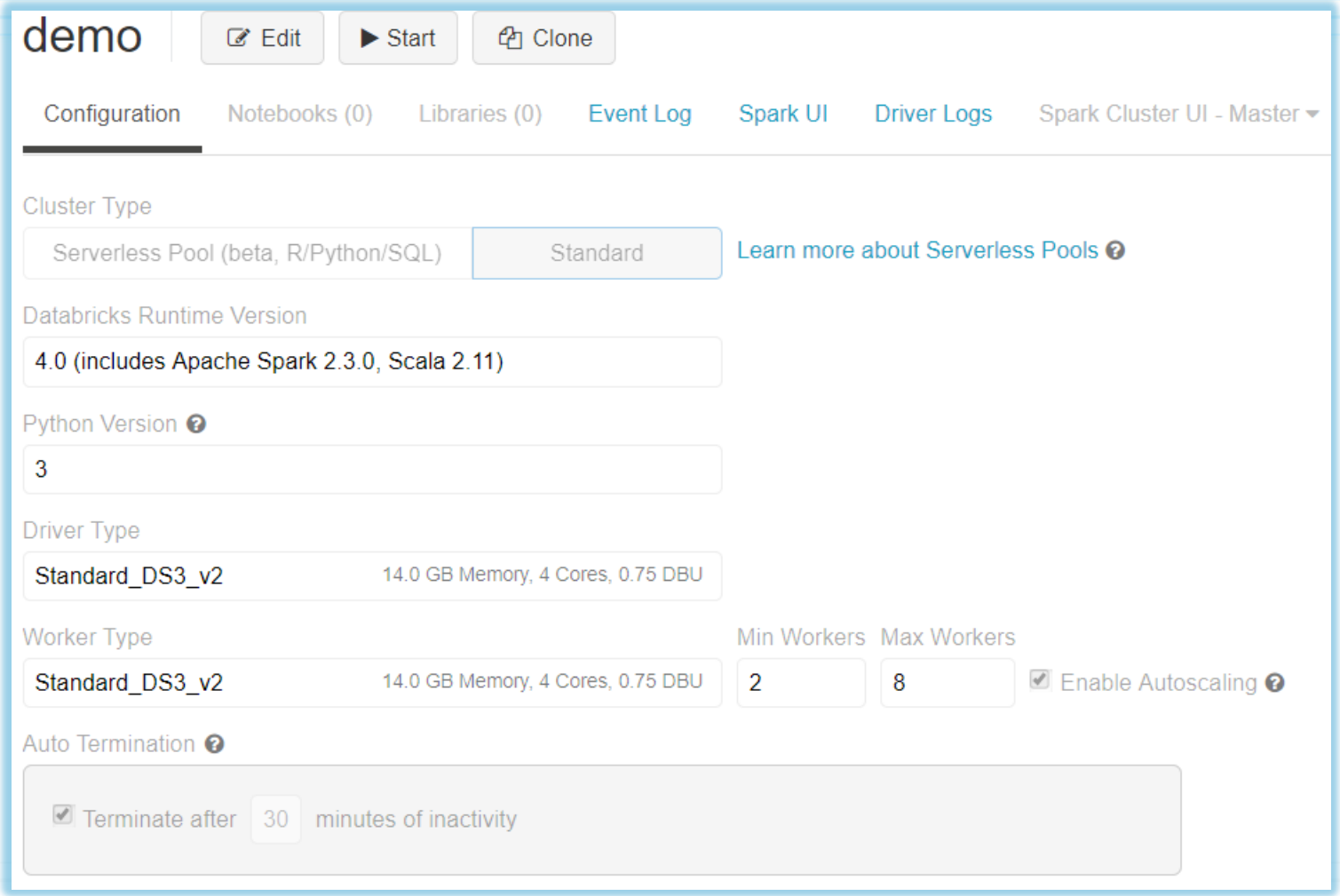
Fully managed

Configure Spark engine

Where notebooks run

Auto scales based on load

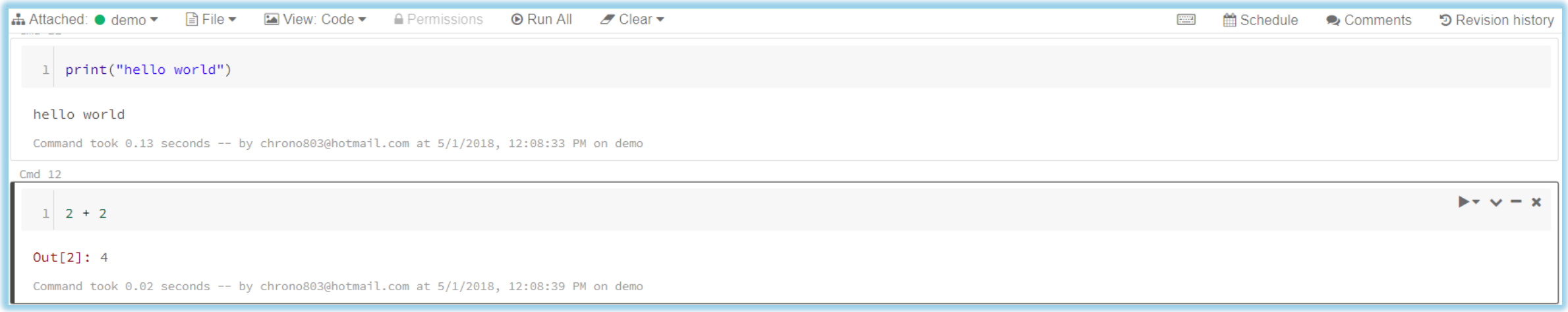
Terminate after a set time of inactivity



Notebooks

Interface to interact with Databricks

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



Advanced Notebook Features

Widgets

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

hello this is a widget : 7

Cmd 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

```
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

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

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

Notebook job #2

Out[1]: '4'

Jobs

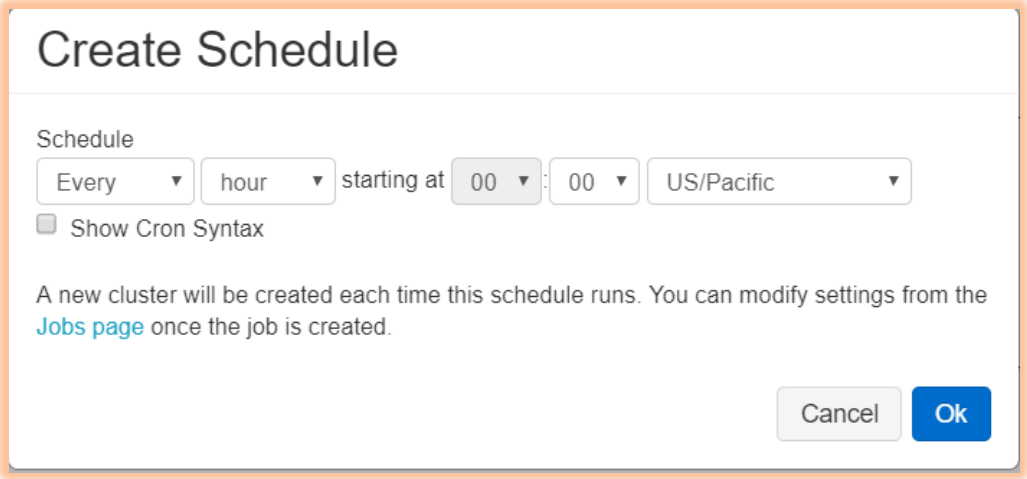
Run notebooks on a schedule

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



Demo

Get Started with Azure Databricks

Demo

Sample ETL Pipeline

Takeaways

Clusters

- Very easy to create
- Reduce costs

Jobs

- Create schedules
- Run on new cluster
- Alerts

Notebooks

- Version control
- Comments
- Widgets

Integration

- Blob Storage
- SQL Data Warehouse

Questions?

Jonathan Wood
Software Consultant
@JWood