

R on the cloud



Omri Mendels
omri.mendels@Microsoft.com

About me

- Part of a team that work with partners on novel projects
- Write open source code and blogs with the solutions
- *#RealLifeCode*

Unsupervised driver safety estimation at scale, a collaboration with Pointer Telocation

May 24, 2018



Categories

Azure App Services
Big Data
Blockchain
Bots
Cognitive Services
Containers
DevOps
Frameworks
Internet Of Things
Machine Learning
Uncategorized

Follow us on Social Media



Authors



omri374 / driver_behavior_analysis

Unwatch

Star

Fork

Code

Issues

Pull requests

Projects

Wiki

Insights

Settings

No description, website, or topics provided.

6 commits

1 branch

0 releases

1 contributor

MIT

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download

omri374 Update README.md

Latest commit 93051c7 23 days ago

| | | |
|--|----------------------------|--------------|
| .gitignore | gitignore | 23 days ago |
| Driver safety estimation - pandas.ipynb | initial code + data commit | 23 days ago |
| Driver safety estimation - pyspark.ipynb | initial code + data commit | 23 days ago |
| LICENSE | Initial commit | 2 months ago |
| README.md | Update README.md | 23 days ago |
| dataset.csv | initial code + data commit | 23 days ago |

README.md

Driver behavior analysis

This code repository holds the jupyter notebooks for estimating driver safety on Pointer's dataset. It contains a sample of the dataset as well as Python (pandas) and PySpark implementations of the process.

It's best first to go over the python notebook as it contains more details, and then to the pyspark notebook to see the same implementation on pyspark.

We'll cover
today:

Using the Data Science Virtual Machine on Ubuntu

R Studio Server

Jupyter Notebooks and Azure Notebooks

R on Spark

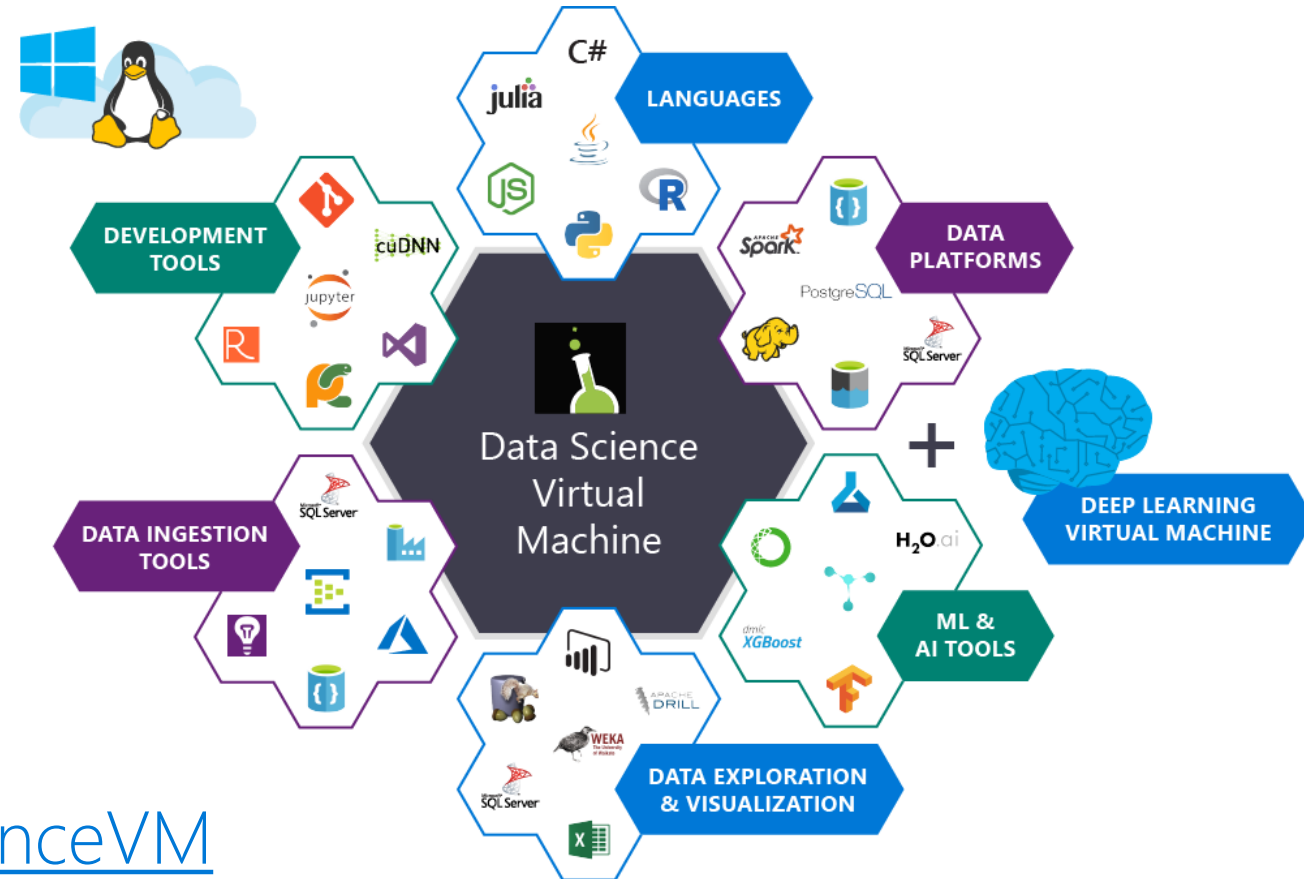
Links and Resources

* Many of the slides are courtesy of [David Smith](#)

Data Science Virtual Machine

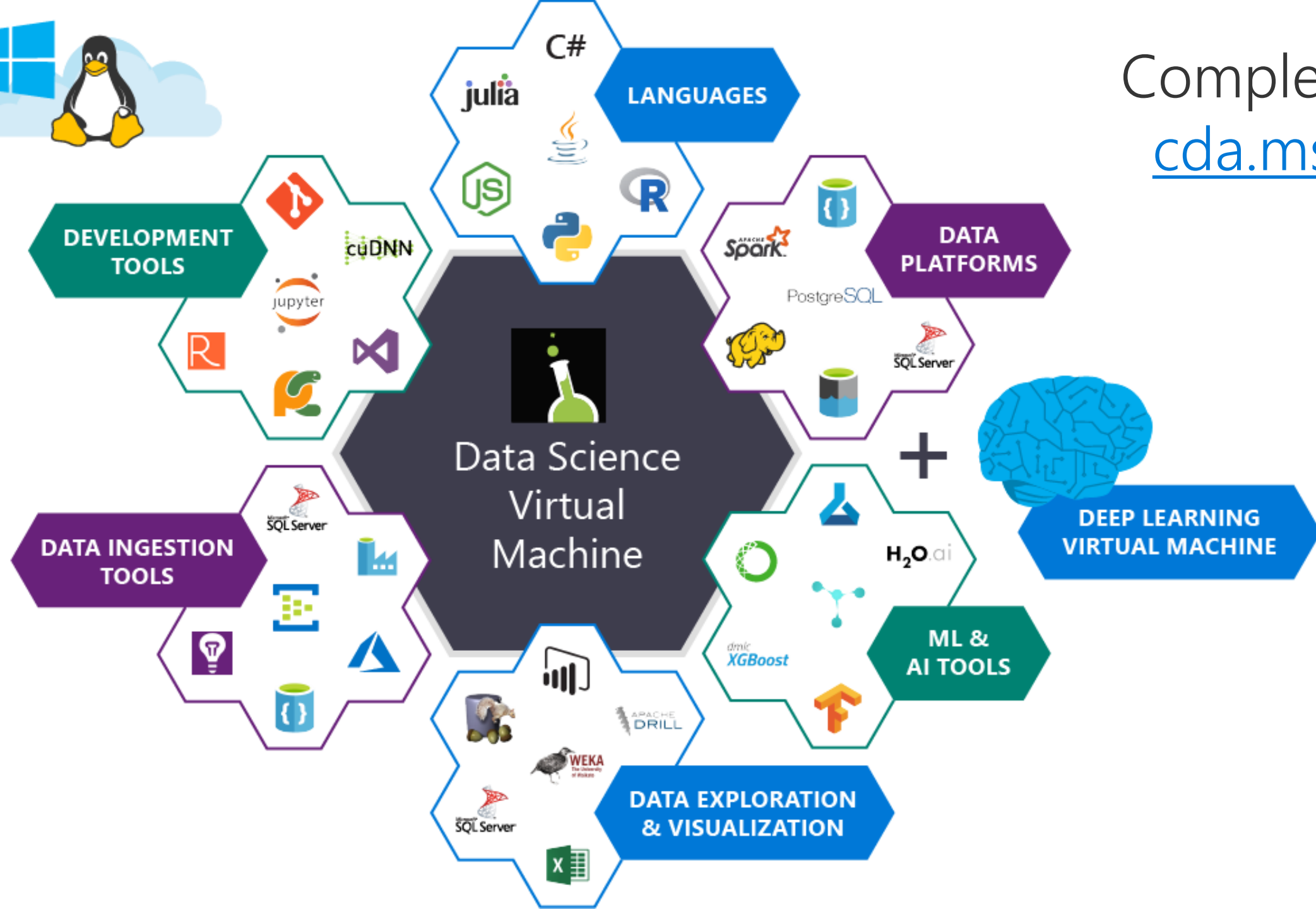
What is the Data Science VM?

- VM image with many Data Science & ML tools pre-installed
- Regularly updated and tested for compatibility
- Windows and Ubuntu
- Pay only standard VM rates



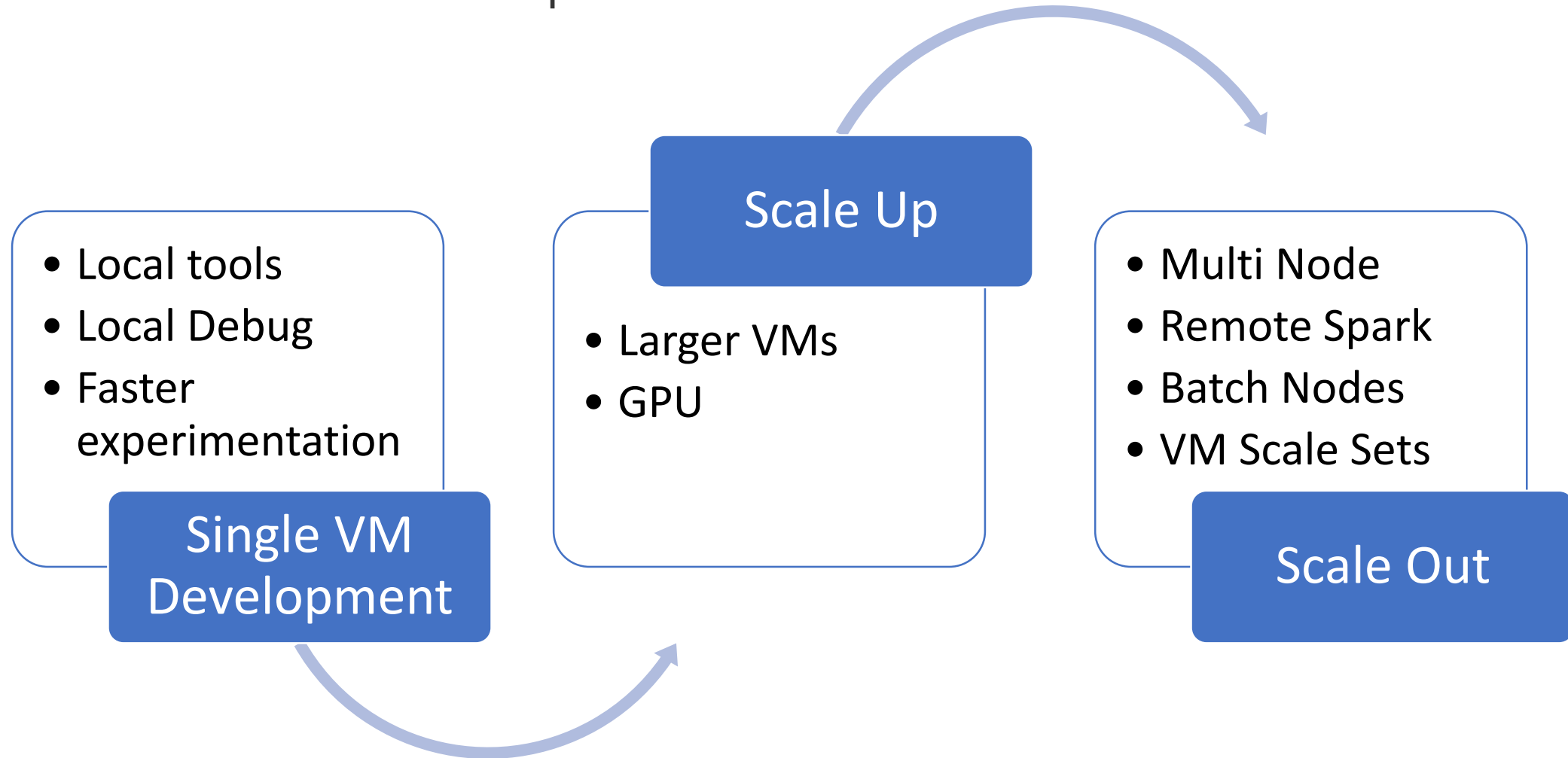
<https://aka.ms/dsvm/overview>

<https://github.com/Azure/DataScienceVM>



Complete list:
cda.ms/sN

Iterative Development with DSVM



Connecting to the DSVM

- Windows: Remote Desktop
 - Tip: download RDP file for quick connections
- Ubuntu:
 - `ssh`
 - X windows (X2Go client)
- App-Specific Interfaces
 - JupyterHub: <https://xxx.xxx.xxx.xxx:8000/>
 - RStudio Server: <http://xxx.xxx.xxx.xxx:8787/>
 - how to launch on DSVM: <https://cda.ms/s0>

Azure Notebooks: notebooks.azure.com

The screenshot displays the Microsoft Azure Notebooks web interface. At the top, the header shows 'Microsoft Azure Notebooks Preview' and the user 'davidsmi'. Below the header is a navigation bar with links for 'Libraries', 'What's New', 'Status', and 'Help'. A large blue banner features the text 'Qcon Code Lab: AI for R Users'. Below this, a breadcrumb trail indicates the current location: 'davidsmi > Libraries > qcon'. The main content area shows a toolbar with icons for 'Run', 'Share', 'Clone', 'Preview', 'Edit File', and other actions. A search bar is present with the text 'Search' and a 'Show hidden items' button. Below the search bar is a table listing files and notebooks. The table has three columns: 'FILE NAME', 'FILE TYPE', and 'MODIFIED'. The files listed are 'Custom Vision API.ipynb', 'hotdogs-good.txt', 'keys.txt', 'nothotdog-find-data.R', 'nothotdogs-good.txt', 'README.md', and 'Vision API.ipynb'. The bottom of the interface shows a pagination control indicating 'Showing 7 search results (1 hidden)' and a page number '1'.

Microsoft Azure Notebooks Preview davidsmi

Libraries What's New Status Help

Qcon Code Lab: AI for R Users

davidsmi > Libraries > qcon

Run + Share Clone 14 Clones Star (0) Preview Edit File

Search Show hidden items

| FILE NAME | FILE TYPE | MODIFIED |
|-------------------------|-----------|--------------|
| Custom Vision API.ipynb | Notebook | Apr 12, 2018 |
| hotdogs-good.txt | Text | Apr 2, 2018 |
| keys.txt | Text | Apr 12, 2018 |
| nothotdog-find-data.R | R | Apr 2, 2018 |
| nothotdogs-good.txt | Text | Apr 2, 2018 |
| README.md | Markdown | Apr 12, 2018 |
| Vision API.ipynb | Notebook | Apr 11, 2018 |

Showing 7 search results (1 hidden) < 1 >

README.md

Distributed ML with Spark and R

What is Spark?

- Distributed data processing engine
 - Store and analyze massive volumes in a robust, scalable cluster
- Successor to Hadoop
 - in-memory engine 100x faster than map-reduce
- Highly extensible, with machine-learning capabilities
 - Supports Scala, Java, Python, R, SQL ...
- Largest open-source data project
 - Apache project with 1000+ contributors
- Managed cloud services available
 - Azure Databricks (completely managed analytics environment) & HDInsight (managed cluster)



Spark + R

- Two main packages
 - SparkR (official package)
 - Write R code that translates to
 - sparklyr (unofficial but integrates better with dplyr)
- How to use Spark with R
 - Run Spark locally
 - Provision a cluster in the cloud (either yourself or a managed solution)
 - Use DataBricks – a managed cluster that provides additional cluster/jobs/code management capabilities

Demo: R on spark with sparklyr

- sparklyr: R interface to Spark
 - open-source R package from RStudio
- Move data between R and Spark
- “References” to Spark Data Frames
 - Familiar R operations, including dplyr syntax
 - Computations offloaded to Spark cluster, and deferred until needed
 - CPU/RAM/Disk consumed in cluster, not by R
- Interfaces to Spark ML algorithms
- Integrated with RStudio Server

Demo: Machine Learning with sparklyr

Links and references

Provision new DSVM: <https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/dsvm-ubuntu-intro>

Azure Notebooks: <https://notebooks.azure.com>

Sparklyr: <http://spark.rstudio.com/>

DataBricks: <https://databricks.com/>

Sparklyr Demo:

<https://gist.github.com/omri374/90c51349294298bda1161eab8220495d>

Thank you!