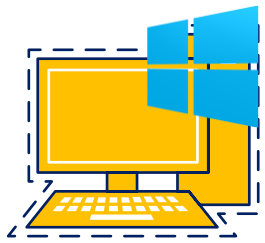
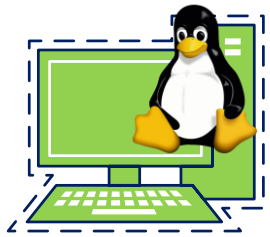
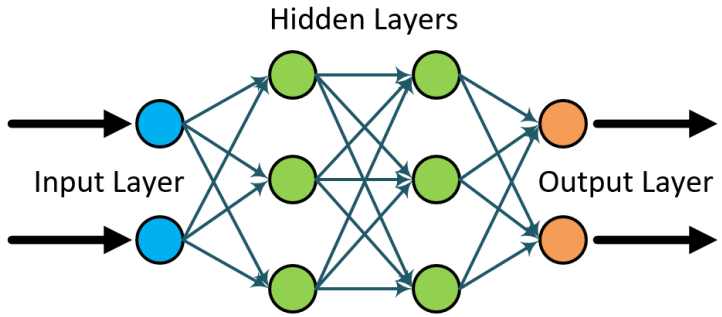
Data Science Virtual Machine – A Walkthrough of end-to-end Analytics Scenarios

Barnam Bora
Program Manager - Engineering







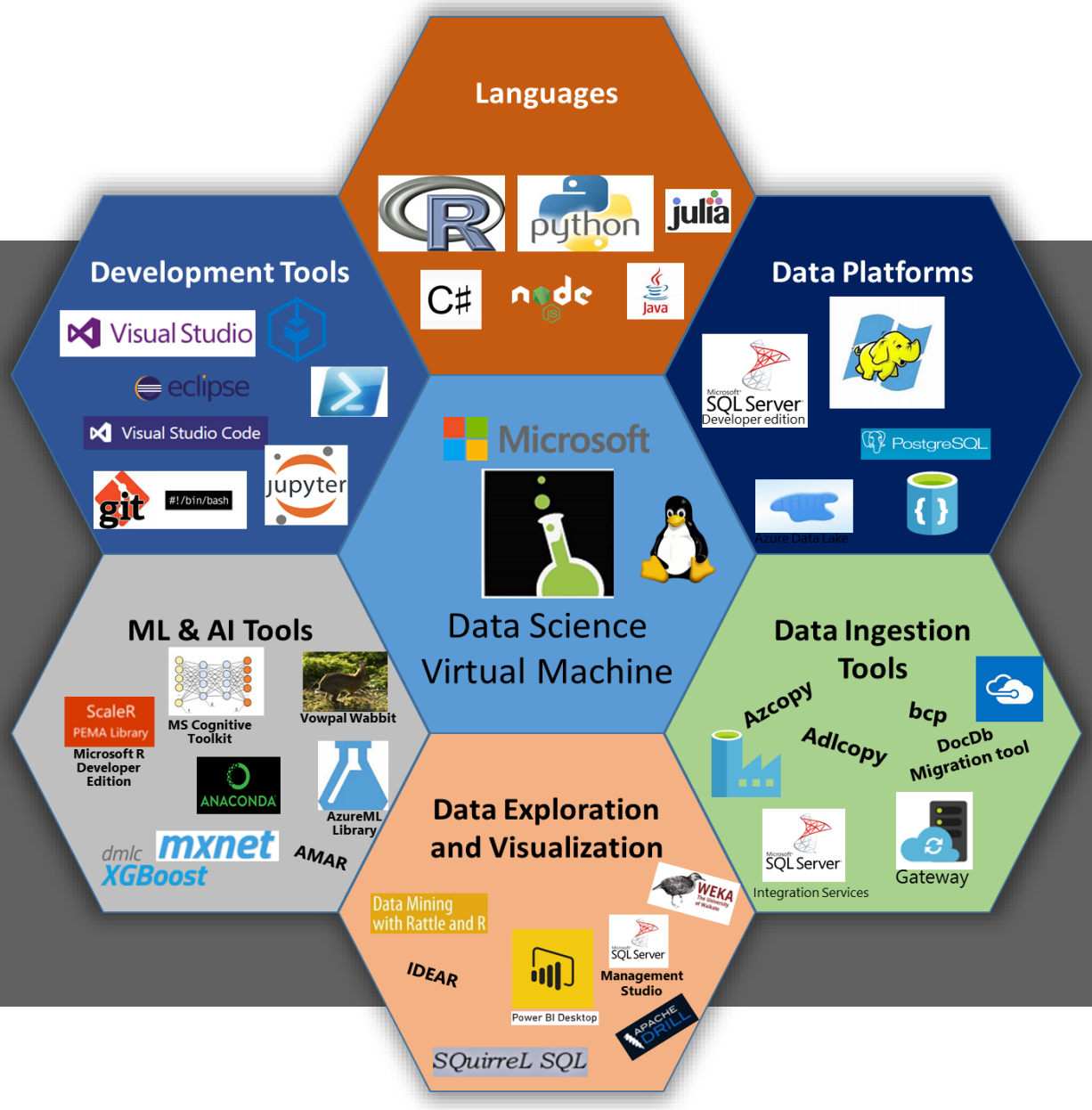
Agenda:

- Brief introduction to the Data Science Virtual Machines in Azure
- Scenario Walkthroughs:
  - ✓ SQL Server R Services: - Dev>Train>Test>Deploy>Score
  - ✓ Using the Local Spark instance on the DSVM for Dev & Test
  - ✓ Training and Deploying Deep Learning Models Using the 'Deep Learning Toolkit for the DSVM' on GPU based Azure VMs
  - ✓ Briefly Querying and wrangling across platforms
- Roadmap
- Q and A + Conclusion

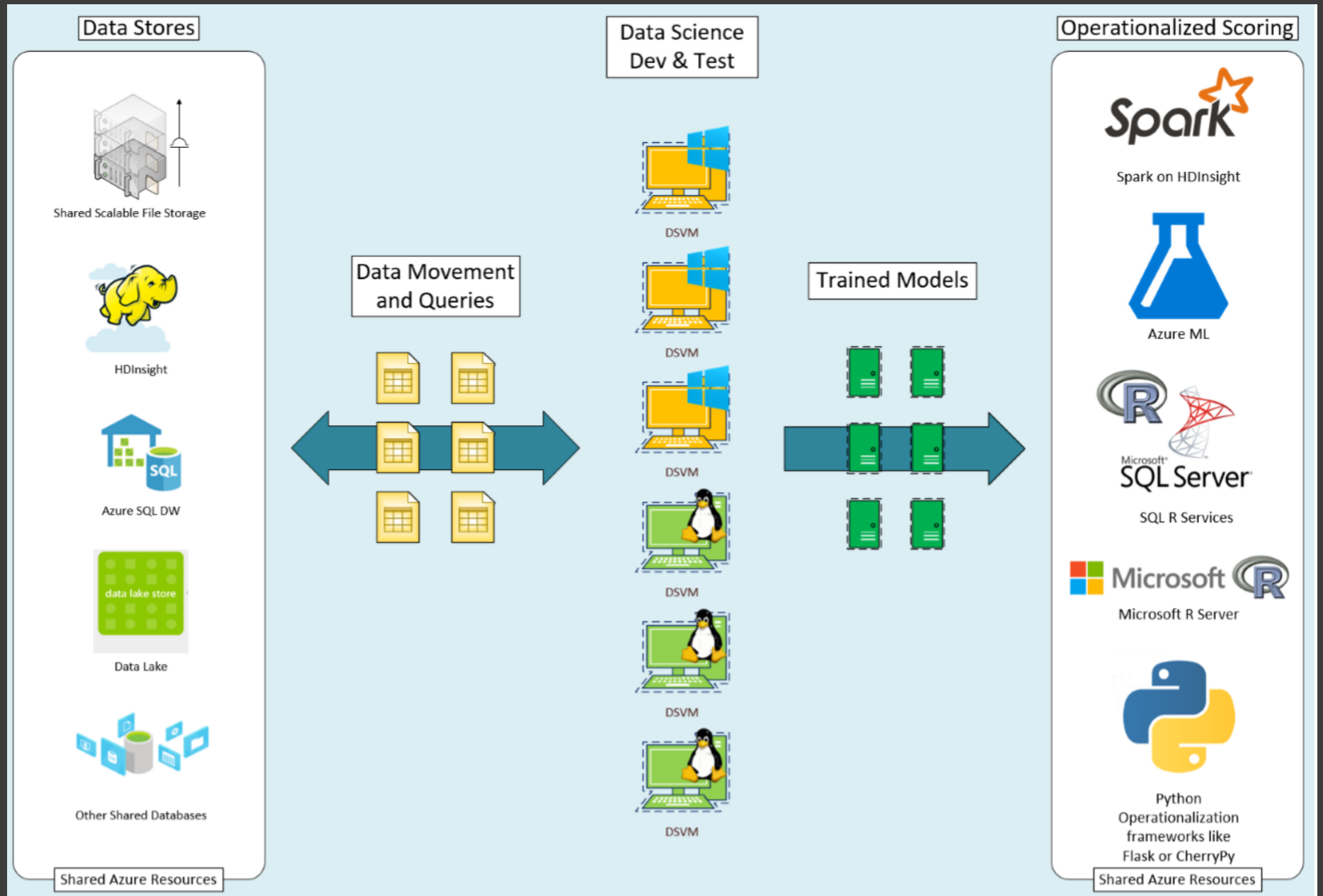
This session aims to familiarize attendees to some popular scenarios enabled by the DSVM and the included tools. This is not a general training module for Data Science. Please visit http://learnanalytics.microsoft.com

Data Science VM ?

Comprehensive cloud based Data Science Environment to empower Data Scientists



Development flow with DSVM



VM Versions comparison – Quick Reference

Windows Edition

- ✓ Microsoft R Open with popular packages pre-installed
- ✓ Microsoft R Server Developer Edition
- ✓ Anaconda Python 2.7, 3.5
- ✓ JuliaPro with popular packages pre-installed
- ✓ Jupyter Notebook Server (R, Python, Julia)
- ✓ SQL Server 2016 Developer Edition: Scalable in-database analytics with R services
- ✓ IDEs and Editors
  - ➥ Visual Studio Community Edition 2015 (IDE)
  - ➥ Azure HDInsight (Hadoop), Data Lake, SQL Server Data tools
  - ➥ Node.js, Python, and R tools for Visual Studio
  - ➥ RStudio Desktop

- ✓ Power BI desktop - (BI Dashboard Design & Analysis)
- ✓ Machine Learning Tools
  - ➥ Integration with Azure Machine Learning
  - ➥ Microsoft Cognitive toolkit (CNTK) - (deep Learning/AI)
  - ➥ Xgboost (popular ML tool in data science competitions)
  - ➥ Vowpal Wabbit (fast online learner)
  - ➥ Rattle (visual quick-start data and analytics tool)
  - ➥ Mxnet (deep learning/AI)
  - ➥ Tensorflow
- ✓ SDKs to access Azure and Cortana Intelligence Suite of services
- ✓ Tools for data movement and management of Azure and Big Data resources: Azure Storage Explorer, CLI, PowerShell, AdlCopy (Azure Data Lake), AzCopy, dtui (for DocumentDB), Microsoft Data Management Gateway
- ✓ Git, Visual Studio Team Services plugin
- ✓ Windows port of most popular Linux/Unix command-line utilities accessible through GitBash/command prompt
- ✓ Weka
- ✓ Apache Drill

Linux Edition

- ✓ Microsoft R Open with popular packages pre-installed
- ✓ Microsoft R Server Developer Edition
- ✓ Anaconda Python 2.7, 3.5 with popular packages pre-installed
- ✓ Julia with popular packages pre-installed
- ✓ JupyterHub: Multi-user Jupyter notebooks (R, Python, Julia, PySpark)
- ✓ PostgreSQL, SQuirreL SQL (database tool), SQL Server drivers, and command line (bcp, sqlcmd)
- ✓ IDEs and editors
  - ➥ Eclipse with Azure toolkit plugin
  - ➥ Emacs (with ESS, auctex) gedit
  - ➥ IntelliJ IDEA
  - ➥ PyCharm
  - ➥ Atom
  - ➥ Visual Studio Code
  - --
- ✓ Machine Learning Tools
  - ➥ Integrations with Azure Machine Learning
  - ➥ Microsoft Cognitive toolkit (CNTK)-(deep Learning/AI)
  - ➥ Xgboost (popular ML tool in data science competitions)
  - ➥ Vowpal Wabbit (fast online learner)
  - ➥ Rattle (visual quick-start data and analytics tool)
  - ➥ Mxnet (deep learning/AI)

- ✓ SDKs to access Azure and Cortana Intelligence Suite of services
- ✓ Tools for data movement and management of Azure and Big Data resources: Azure Storage Explorer, CLI

- ✓ Git
  --

- ✓ Weka
- ✓ Apache Drill
- ✓ Apache Spark - local instance

Most of Today's Examples are worked out as jupyter Notebooks – Included on the DSVMs

- Popular open-source application
- A browser based **Read–Eval–Print Loop** (**REPL**) environment
- Used to create and share documents that contain:
  - ✓ Live code
  - ✓ Equations
  - ✓ Visualizations
  - ✓ Explanatory text/documentation
  - ✓ Stored Outputs etc.

The focus of this session is skewed predominantly towards demonstrating the scenarios as opposed to discussing the Data Science algorithms and methods used in the examples.

Please visit http://learnanalytics.microsoft.com for dedicated Data Science Training

Dataset Refresher - The 2013 NYCTaxi Data:
Data wrangling, manipulations, modeling, and evaluation

```
##                          medallion                      hack_license
## 1 D7D598CD99978BD012A87A76A7C891B7 82F90D5EFE52FDFD2FDEC3EAD6D5771D
## 2 5455D5FF2BD94D10B304A15D4B7F2735 177B80B867CEC990DA166BA1D0FCAF82
## 3 93D6821F86A12B537C5EADBDFB432CA7 28B0AA10202F83FEB0F4E69340CA8F86
##   vendor_id rate_code store_and_fwd_flag     pickup_datetime
## 1       VTS         1                NA 2013-12-01 00:13:00
## 2       VTS         1                NA 2013-12-01 00:40:00
## 3       VTS         1                NA 2013-12-01 02:21:00
##     dropoff_datetime passenger_count trip_time_in_secs trip_distance
## 1 2013-12-01 00:31:00               1              1080          3.90
## 2 2013-12-01 00:48:00               6               480          3.20
## 3 2013-12-01 02:30:00               5               540          3.28
##   pickup_longitude pickup_latitude dropoff_longitude dropoff_latitude
## 1        -73.97934         40.77665         -73.98186         40.73428
## 2        -73.93967         40.72615         -73.98558         40.71807
## 3        -73.95875         40.76808         -73.95875         40.76808
```

The data used for this exercise is the public NYC Taxi Trip and Fare data-set (2013, December, ~4 Gb, ~13 million rows) available from: http://www.andresmh.com/nyctaxitrips

Demo Scenario

SQL Server R Services: -

- ✓ Dev
- ✓ Train
- ✓ Test
- ✓ Deploy
- ✓ Score

```
CREATE LOGIN [<YourDSVMNameHere>\SQLRUserGroup] FROM WINDOWS
```

Demo Scenario

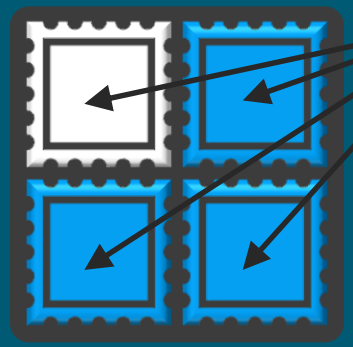
Using the Local Spark instance on the DSVM for Dev & Test

Using the Local Spark instance on the DSVM with 2013 NYCTaxi Data:

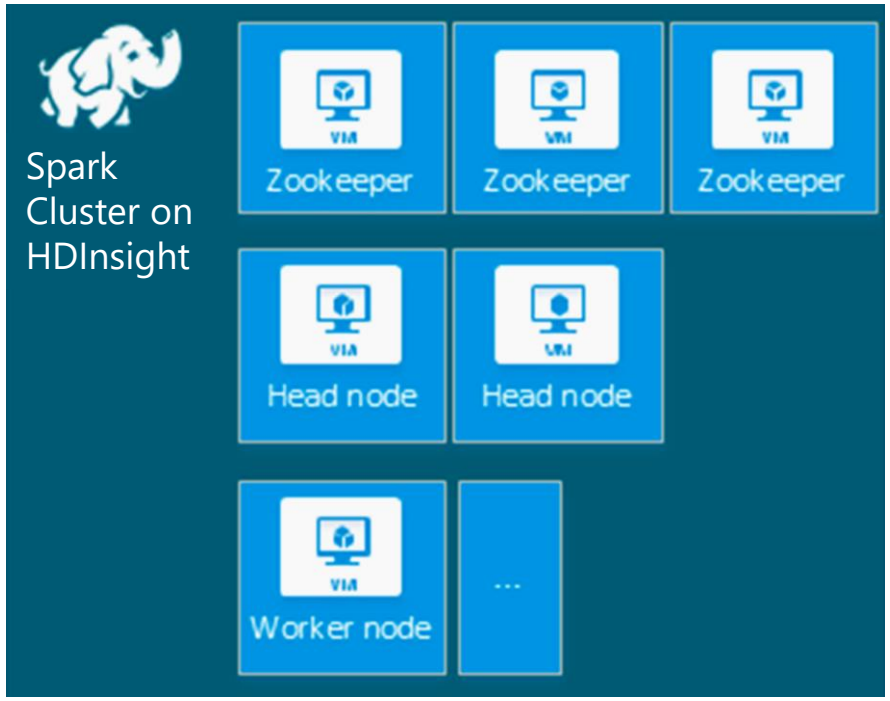
Data wrangling, manipulations, modeling, and evaluation

Easily deployed/scaled interchangeably via YARN

Local Spark on the DSVM



Head and **Worker** Roles handled and optimized on the box by the **Spark Local Process**



Using SparkR on a Local Spark instance with 2013 NYCTaxi Data:

Data wrangling, manipulations, modeling, and evaluation

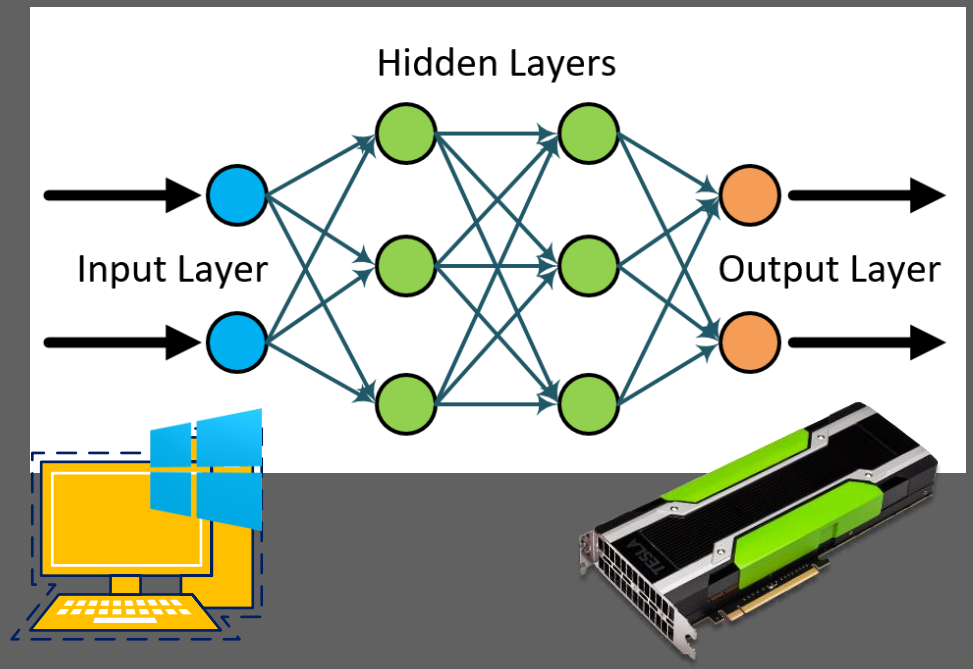
```
##                          medallion                      hack_license
## 1 D7D598CD99978BD012A87A76A7C891B7 82F90D5EFE52FDFD2FDEC3EAD6D5771D
## 2 5455D5FF2BD94D10B304A15D4B7F2735 177B80B867CEC990DA166BA1D0FCAF82
## 3 93D6821F86A12B537C5EADBDFB432CA7 28B0AA10202F83FEB0F4E69340CA8F86
##   vendor_id rate_code store_and_fwd_flag     pickup_datetime
## 1       VTS         1                 NA 2013-12-01 00:13:00
## 2       VTS         1                 NA 2013-12-01 00:40:00
## 3       VTS         1                 NA 2013-12-01 02:21:00
##     dropoff_datetime passenger_count trip_time_in_secs trip_distance
## 1 2013-12-01 00:31:00               1              1080          3.90
## 2 2013-12-01 00:48:00               6               480          3.20
## 3 2013-12-01 02:30:00               5               540          3.28
##   pickup_longitude pickup_latitude dropoff_longitude dropoff_latitude
## 1        -73.97934         40.77665         -73.98186         40.73428
## 2        -73.93967         40.72615         -73.98558         40.71807
## 3        -73.95875         40.76808         -73.95875         40.76808
```

Two sets of files

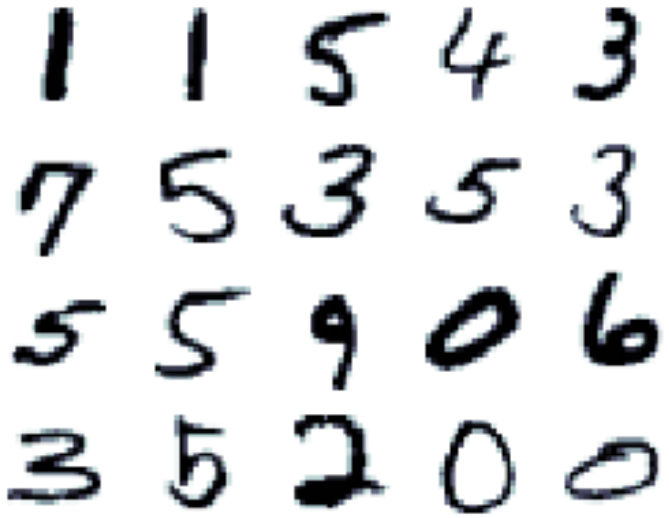
- trip_data CSVs contain trip details
- trip_fare CSVs contain details of fare paid
- Unique key to join trip_data and trip_fare: medallion, hack_licence, and pickup_datetime

Demo Scenario

Training and Deploying Deep Learning Models Using the 'Deep Learning Toolkit for the DSVM' on GPU based Azure VMs



MNIST Feed Forward Network
- Using CNTK (Microsoft Cognitive Toolkit)



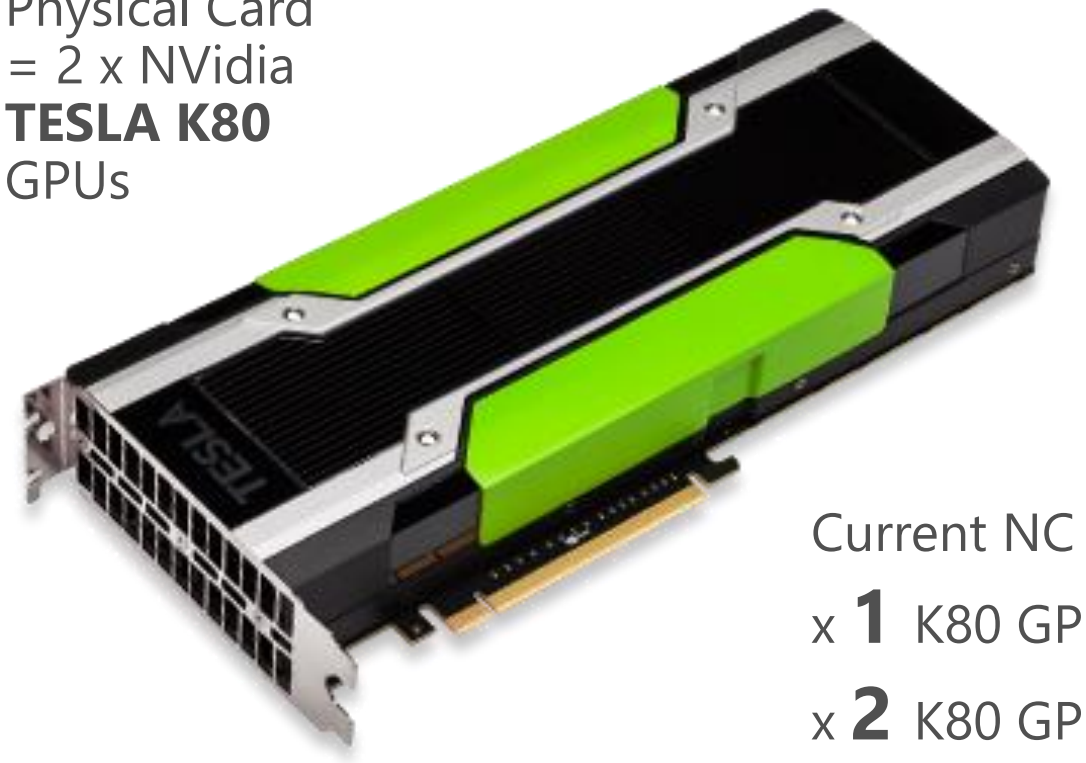
The MNIST database (Mixed National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems.

http://yann.lecun.com/exdb/mnist/

The Deep Learning toolkit for DSVM

[Click Here]

Physical Card
= 2 x NVidia
TESLA K80
GPUs



The Toolkit deploys on **NC** class Azure VMs with GPUs

Great for **Deep Learning** workloads

Current NC Class VM SKU Configurations:

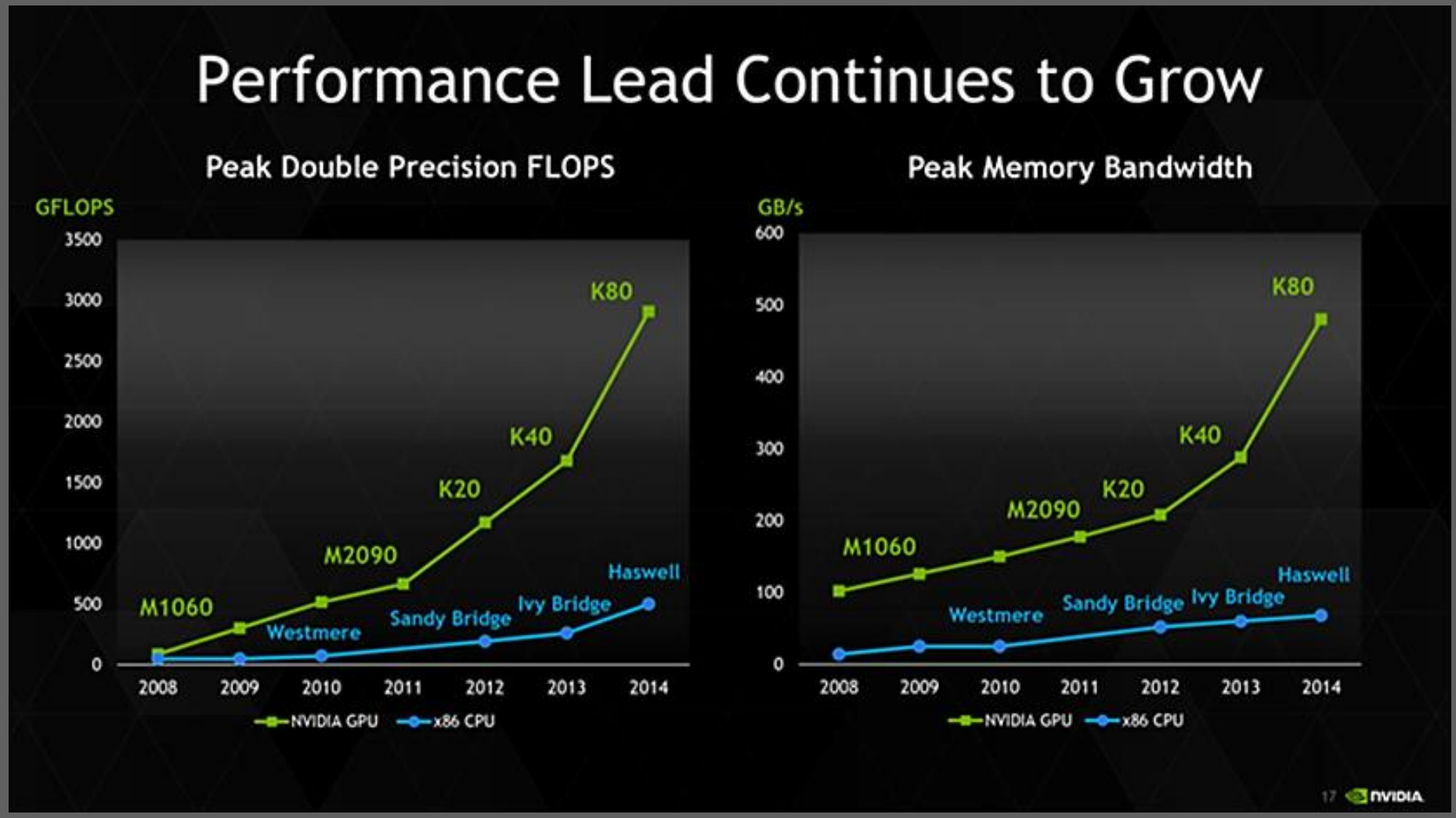
x **1** K80 GPU	- 1/2 Physical Card	- **12** GB GDDR5 VRAM
x **2** K80 GPUs	- 1 Physical Card	- **24** GB GDDR5 VRAM
x **4** K80 GPUs	- 2 Physical Cards	- **48** GB GDDR5 VRAM

Deep Learning Toolkit

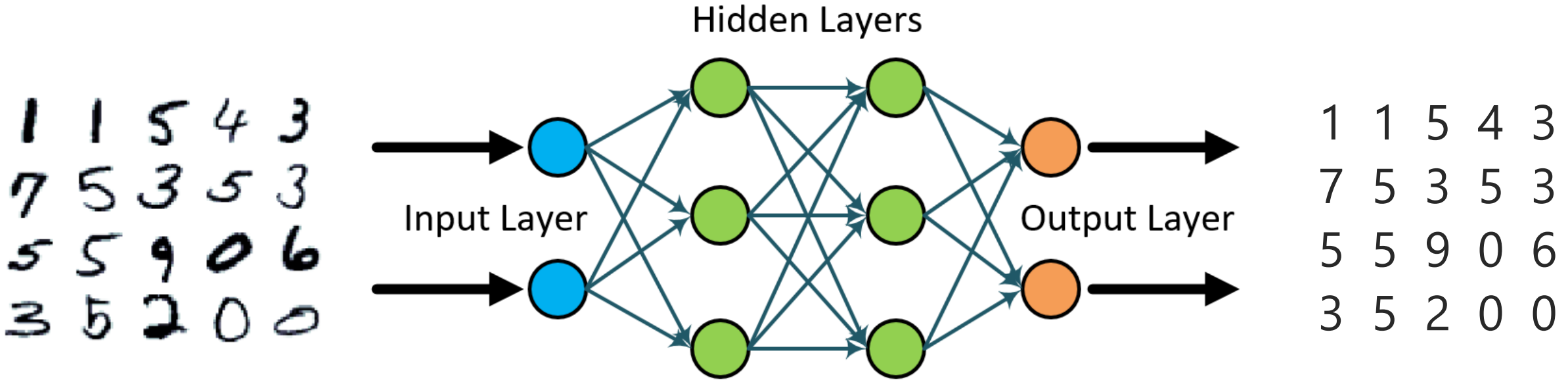
for the Data Science Virtual Machine (DSVM)

Microsoft

GPU based Parallelization provides orders of magnitude increase in Performance when it comes to Deep Learning



MNIST Feed Forward Network
- Using CNTK (Microsoft Cognitive Toolkit)



Demo Scenario

Querying and wrangling across platforms using Apache Drill

Special Appearance by Power BI

Exploring the 2013 NYCTaxi Data using Apache DRILL:

Data wrangling, manipulations, modeling, and evaluation

```
##                          medallion                      hack_license
## 1 D7D598CD99978BD012A87A76A7C891B7 82F90D5EFE52FDFD2FDEC3EAD6D5771D
## 2 5455D5FF2BD94D10B304A15D4B7F2735 177B80B867CEC990DA166BA1D0FCAF82
## 3 93D6821F86A12B537C5EADBDFB432CA7 28B0AA10202F83FEB0F4E69340CA8F86
##   vendor_id rate_code store_and_fwd_flag     pickup_datetime
## 1       VTS         1                NA 2013-12-01 00:13:00
## 2       VTS         1                NA 2013-12-01 00:40:00
## 3       VTS         1                NA 2013-12-01 02:21:00
##     dropoff_datetime passenger_count trip_time_in_secs trip_distance
## 1 2013-12-01 00:31:00               1              1080          3.90
## 2 2013-12-01 00:48:00               6               480          3.20
## 3 2013-12-01 02:30:00               5               540          3.28
##   pickup_longitude pickup_latitude dropoff_longitude dropoff_latitude
## 1        -73.97934         40.77665         -73.98186         40.73428
## 2        -73.93967         40.72615         -73.98558         40.71807
## 3        -73.95875         40.76808         -73.95875         40.76808
```

Two sets of files

- trip_data CSVs contain trip details
- trip_fare CSVs contain details of fare paid
- Unique key to join trip_data and trip_fare: medallion, hack_licence, and pickup_datetime

Roadmap

- Windows Server 2016 based DSVM Offering
  - Containerized Workloads
  - Windows 10 style Desktop Experience
  - Excel (Office 365) – Pre-installed – Needs BYO ProPlus Subscription
- Ubuntu Based Deep Learning DSVM
- Investment into other categories
  - IOT
  - Cognitive Computing

Summary:

- Introduction to the Data Science Virtual Machines in Azure
- Scenario Walkthroughs:
  - ✓ SQL Server R Services: - Dev>Train>Test>Deploy>Score
  - ✓ Using the Local Spark instance on the DSVM for Dev & Test
  - ✓ Training and Deploying Deep Learning Models Using the 'Deep Learning Toolkit for the DSVM' on GPU based Azure VMs
  - ✓ Briefly Querying and wrangling across platforms
- Roadmap
- Q and A

This session familiarizes attendees to some popular scenarios enabled by the DSVM and the included tools. This is not a general training module for Data Science. Please visit http://learnanalytics.microsoft.com

Useful Links:

- DSVM Forum - send questions, feedback and feature requests on the forum - http://aka.ms/dsvm/forum
- DSVM Product Page – http://aka.ms/dsvm
- DSVM Introductory DIY workshop – http://aka.ms/dsvm/workshop
- 2 Page Handout – http://aka.ms/dsvm/handout
- Learn Analytics @ Microsoft – http://learnanalytics.microsoft.com