Microsoft

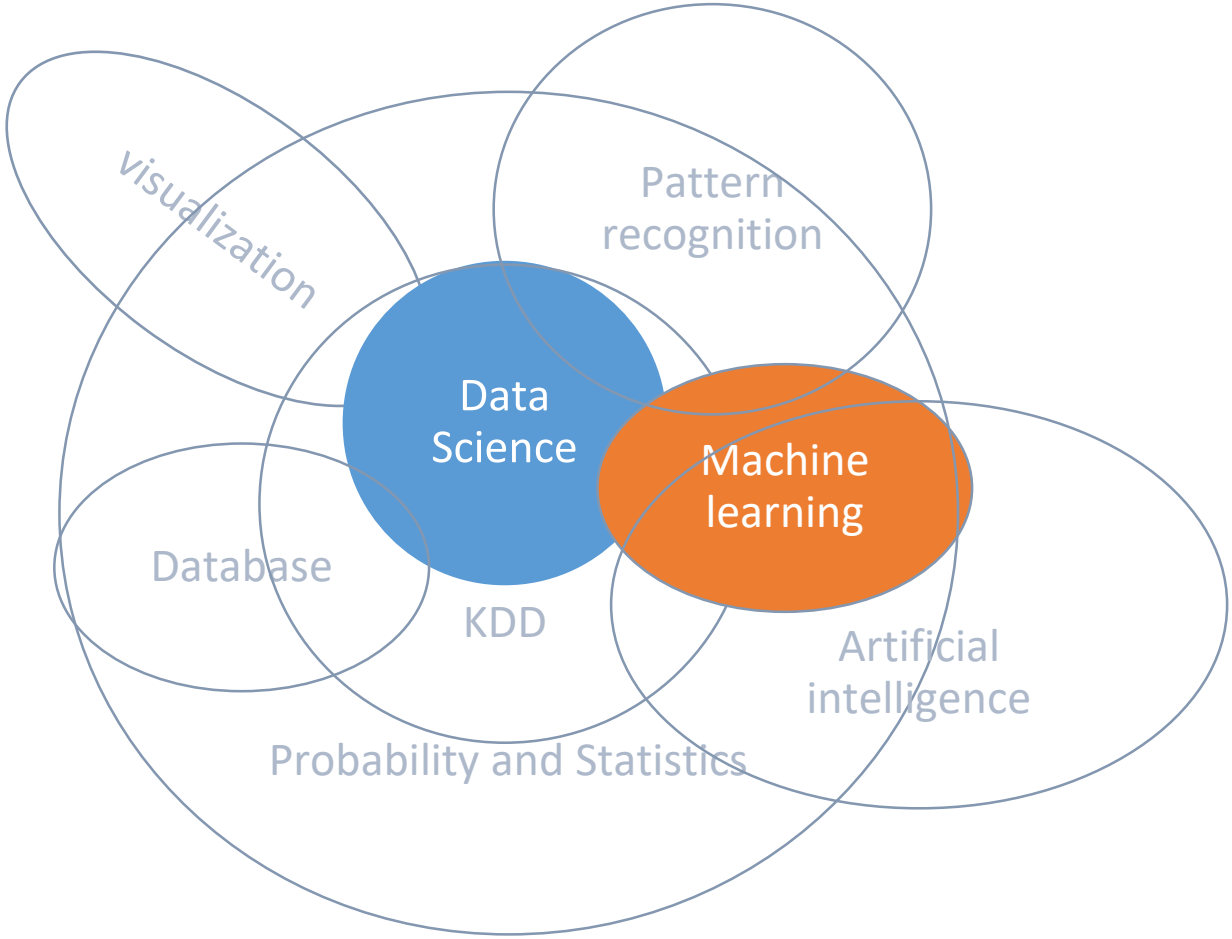
APAC Machine Learning & Data Science Community Summit

2017년 5월 20일(토)
상암동 누리꿈스퀘어 비즈니스타워 3층

Elastic machine learning with Azure Data Science Virtual Machine

Le Zhang, Ph. D.
Data Scientist, Microsoft

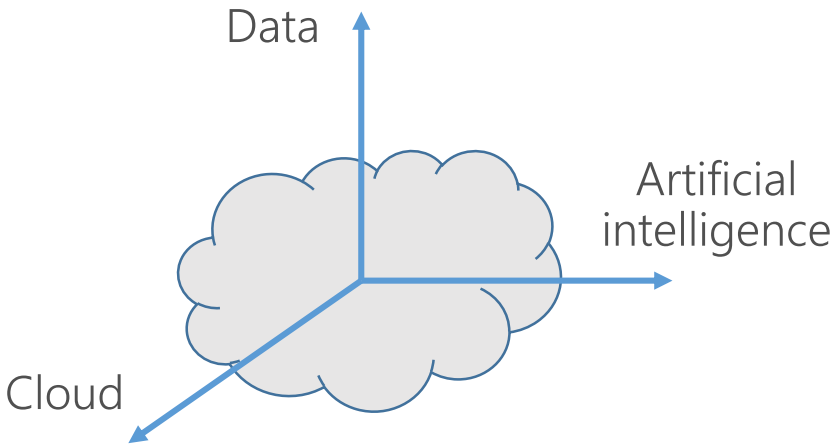
What is DS and ML?



http://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-looking-forwards-sas-data-mining-and-machine-learning/

Data science on cloud

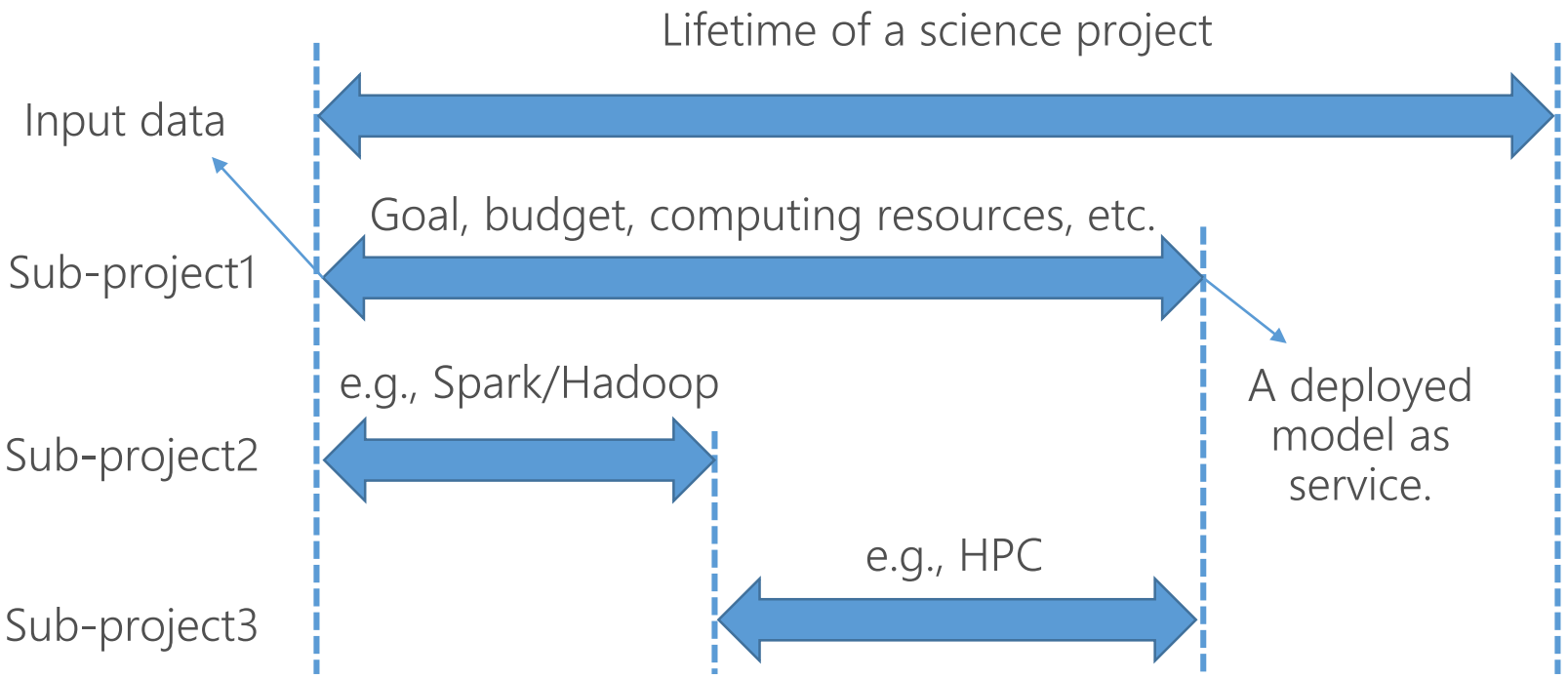
- Need to run scalable data science.
- Cost-effectiveness.
- Collaborative working environment.
- Share of codes, scripts, data, etc.
- Efficiency in prototyping ideas.
- Ecosystem.



https://www.analyticsvidhya.com/blog/2015/09/scalable-data-science-cloud-python-r/

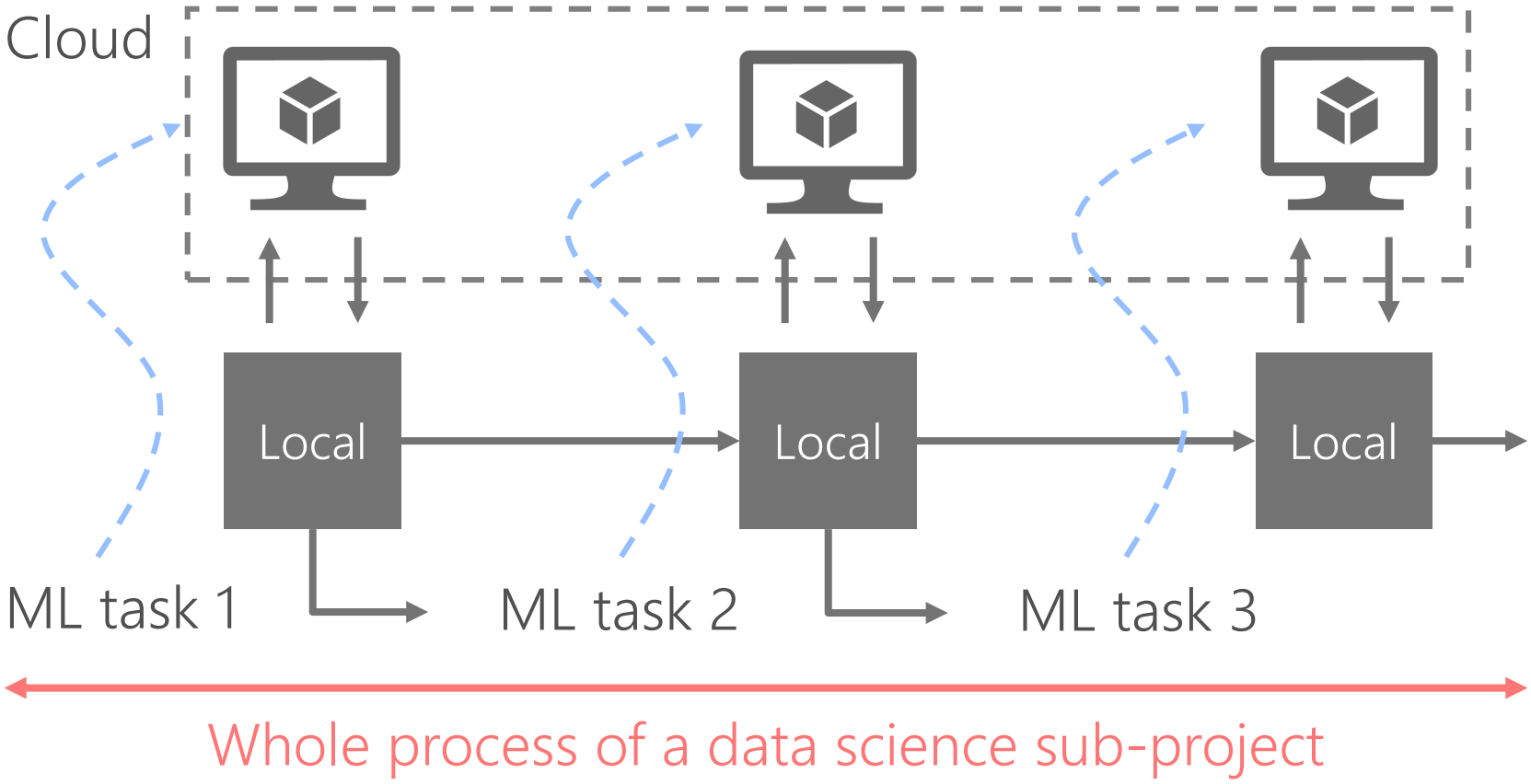
Data science project

- A data science project is often partitioned into several sub-projects.
- Elastic use of computing resources on cloud is important.



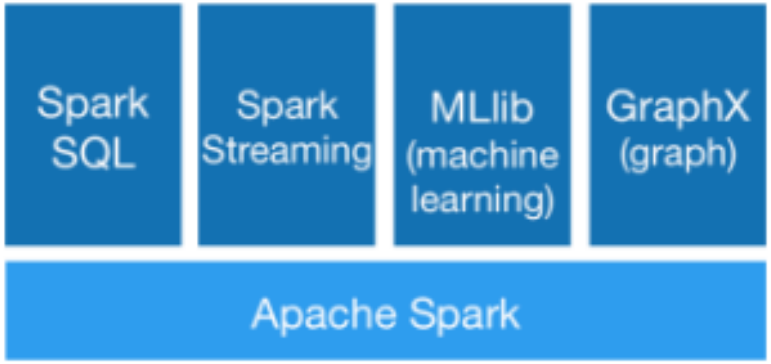
Elastic use of cloud for data science

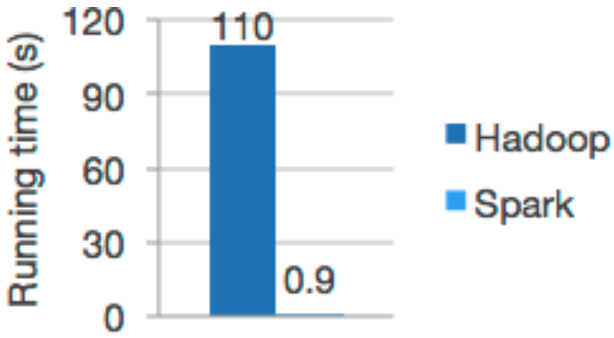
- Partition a project into sub-tasks.
- Allocate computing resource for each task.
- Invoke each computing node only when it is needed.



Apache Spark

- Fast and general engine for large scale data processing.
- Java, Scala, Python, and R.
- Combine SQL, streaming, and complex analytics.
- Standalone, Mesos, and YARN as cluster manager.

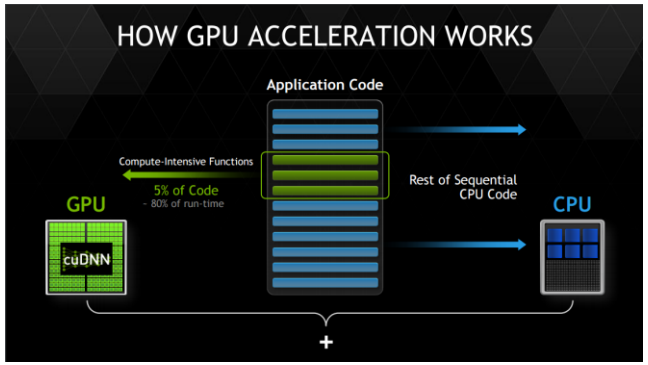




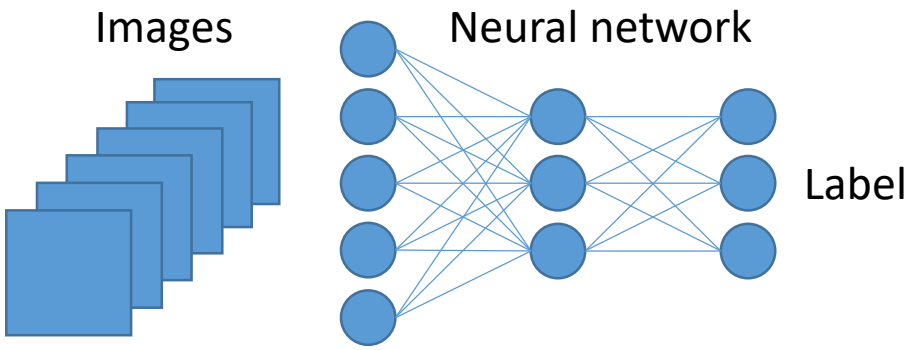
http://spark.apache.org/
http://spark.apache.org/docs/latest/cluster-overview.html

GPU-accelerated deep learning

- Graphic Processing Unit (GPU).
- Deep neural network.
- Why use GPU for deep learning?



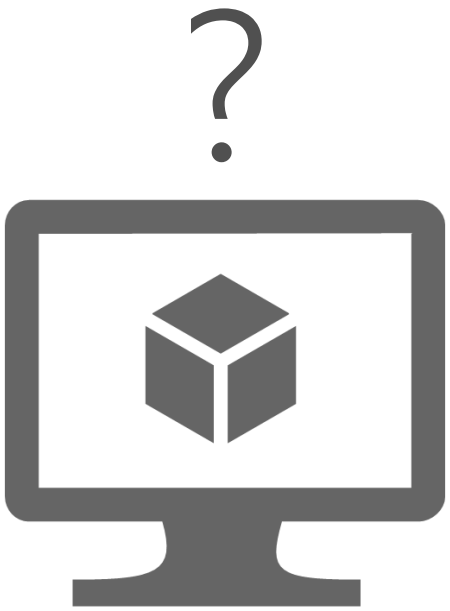
Batch Size	Training Time CPU	Training Time GPU	GPU Speed Up
64 images	64 s	7.5 s	8.5X
128 images	124 s	14.5 s	8.5X
256 images	257 s	28.5 s	9.0X



	Neural Networks	GPU
Inherently parallel	Yes	Yes
Matrix operations	Yes	Yes
FLOPS	Yes	Yes

http://www.nvidia.com/content/events/geoInt2015/LBrown_DL.pdf

A platform for all



General purpose computing, Spark, deep learning, GPU acceleration, etc.

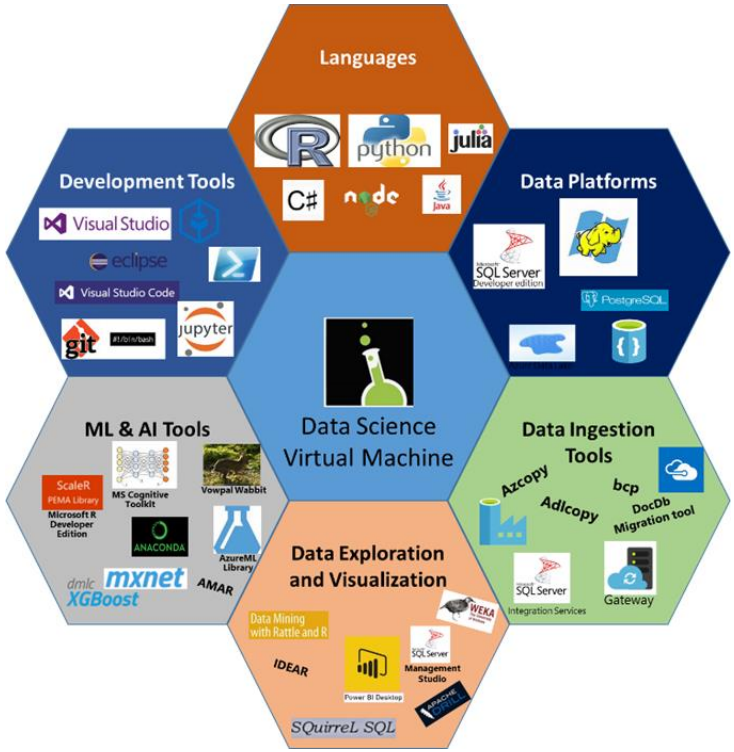
Problems in bringing all into one…

- What might bother us?
  - Installation and configuration of environment (e.g., Spark, deep learning toolkit, etc.).
  - Programming languages unfamiliar to Data Scientists.
    - Spark: Scala and Java.
    - Deep learning: CUDA, OpenCL, and C++.
    - Azure Resource Manager: Azure Command Line Interface (CLI).
    - Deployment as a web service: RESTFUL, Javascript, etc.
  - Collaborative workspace, reproducibility of results, and ease of model deployment.

Azure Data Science Virtual Machine

- Curated VM preinstalled with popular data science tools.
- Scenarios
  - Prototyping of POCs.
  - Remote working desktop for experimental analysis.
  - Data science and machine learning education.
  - Elastic computing engine for data science tasks.

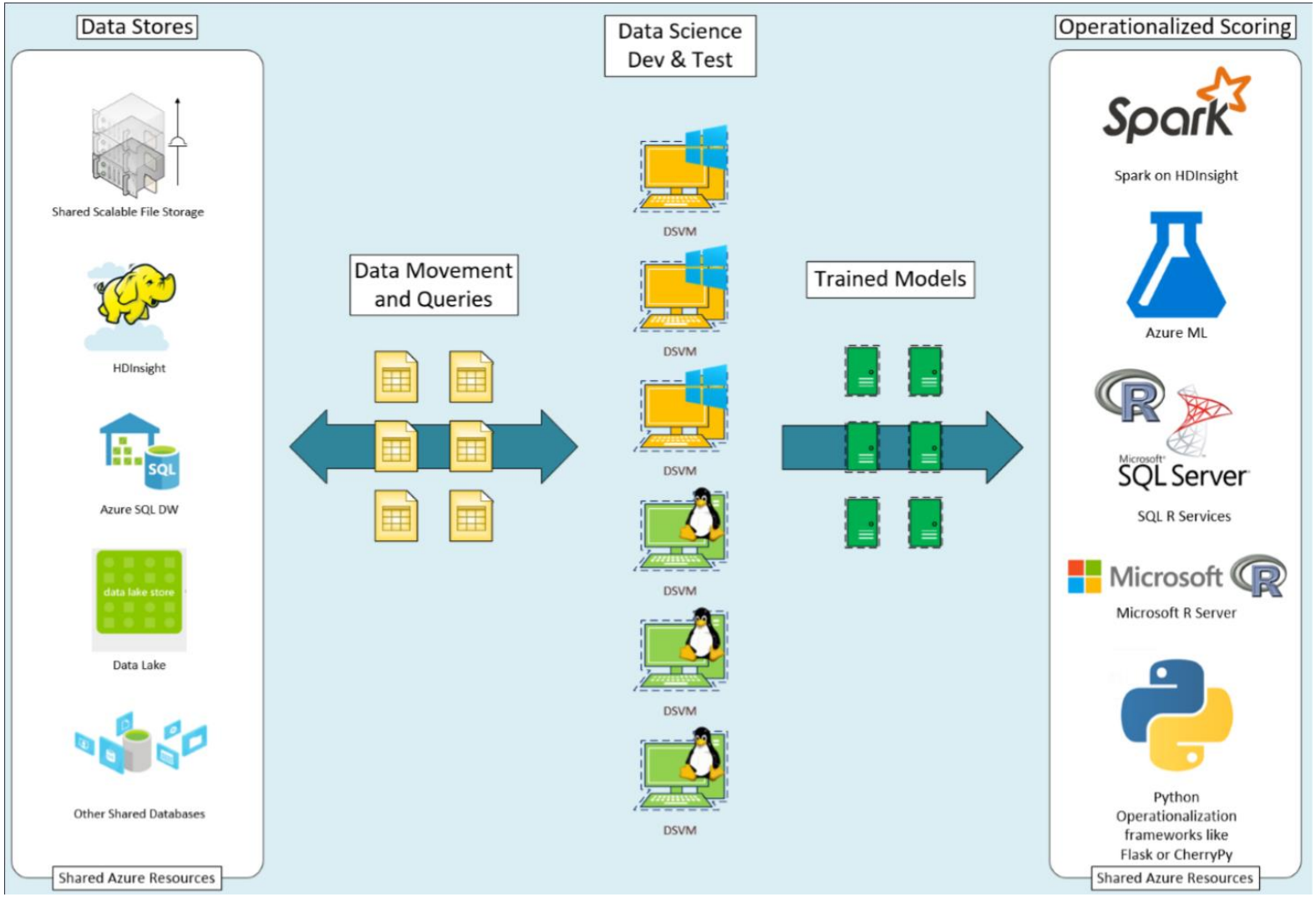
What you can do with DSVM



10 things you can do with DSVM

1. Explore data and develop models locally on the DSVM using R and Python.
2. Use a Jupyter notebook to experiment with your data on a browser.
3. Operationalize models built using R and Python on Azure Machine Learning.
4. Administer your Azure resources using Azure portal, Powershell, or R.
5. Extend storage space and share large-scale datasets / code.
6. Share code using GitHub.
7. Access various Azure services.
8. Build reports and dashboard using the Power BI Desktop.
9. Dynamically scale your DSVM.
10. Install additional tools on DSVM.

Operationalization with DSVM



DSVM version in Windows and Linux

APAC Machine Learning & Data Science Community Summit

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

Use case demo
flight delay prediction

Flight delay prediction

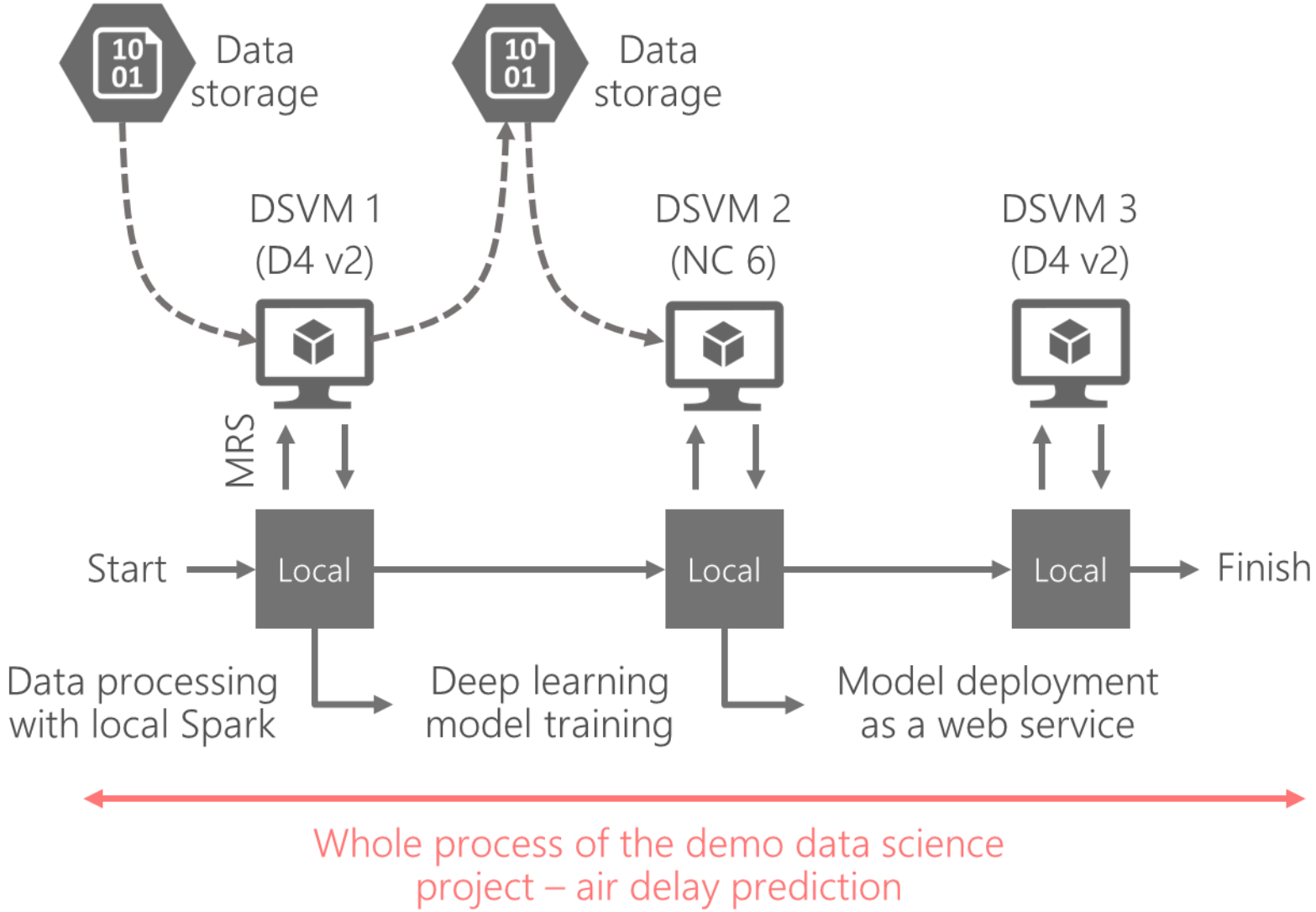
- Flight delay prediction
  - Problem statement: predict flight delay given fleet information.
  - Data.
    - Size: ~1.4 G
    - Features: day of month, day of week, origin, destination, etc.
    - Prediction target: whether or not the flight is delayed.
- Assume we are going to build a pipeline that
  - Uses a sub-sampled (1%) and aggregated version of the original data.
  - Applies Spark for data pre-processing.
  - Trains a neural network model with GPU acceleration.
  - Publishes the model as a web based service.

Operationalization with DSVM (cont'd)

- Computing resource planning.
- Data pre-processing on Spark, model training with GPU acce leration, and web-based service deployment.

DSVM name	DSVM size	OS	Description	Price
Spark	Standard D4 v2 – 8 cores with 28 GB m emory	Linux	Local standalone m ode Spark for data preprocessing and f eature engineering.	$0.585/hr
Deeplearning	Standard NC6 – 6 c ores with 56 GB me mory, and Nvidia Te sla K80 GPU	Windows	Train deep neural n etwork model with GPU acceleration.	$0.9/hr
Webserver	Standard D4 v2 – 8 cores with 28 GB m emory	Linux	Server host where web based model s ervice is published and run.	$0.585/hr

Pipeline for air delay prediction



Operationalization in R

Can we do everything in R?

Operationalization in R

Yes!

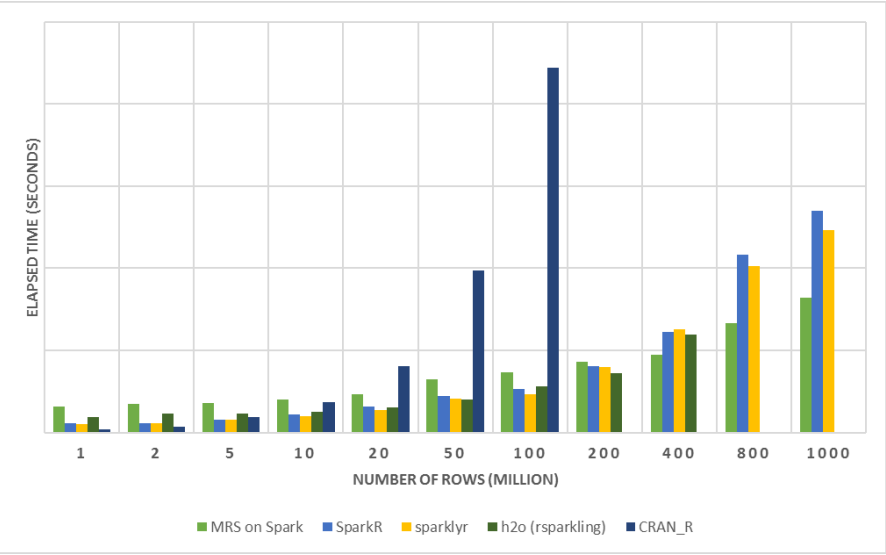
Cloud resource management

- Azure resource management in R.
  - AzureSMR
    - Managing a selection of Azure resources such as storage blobs, HDInsight, etc.
  - AzureDSVM
    - Deployment and operation on an Azure DSVM with specified size, OS, and user credentials.
    - Remote execution of script and file transfer with a Linux DSVM.
    - Retrieval of cost and expense information of using DSVM.
- Prerequisites
  - Azure subscription.
  - Initial setup for Azure Active Directory.

https://github.com/Microsoft/AzureSMR
https://github.com/Azure/AzureDSVM

Spark

- DSVM supports local standalone mode Spark.
  - For experimental and debugging purpose.
  - Up-scale code to Spark cluster in Azure HDInsight.
- R frontend for Spark
  - Microsoft R Server.
  - SparkR.
  - sparklyr.



E2E Process:

- Load Data from .csv
- Transform Features
- Split Data: Train + Test
- Fit Model: Logistic Regression (no regularization)
- Predict and Write Outputs

Configuration:

- 1 Edge Node: 16 cores, 112GB
- 4 Worker Nodes: 16 cores, 112GB
- Dataset: Duplicated Airlines data (.csv)
- Number of columns: 26

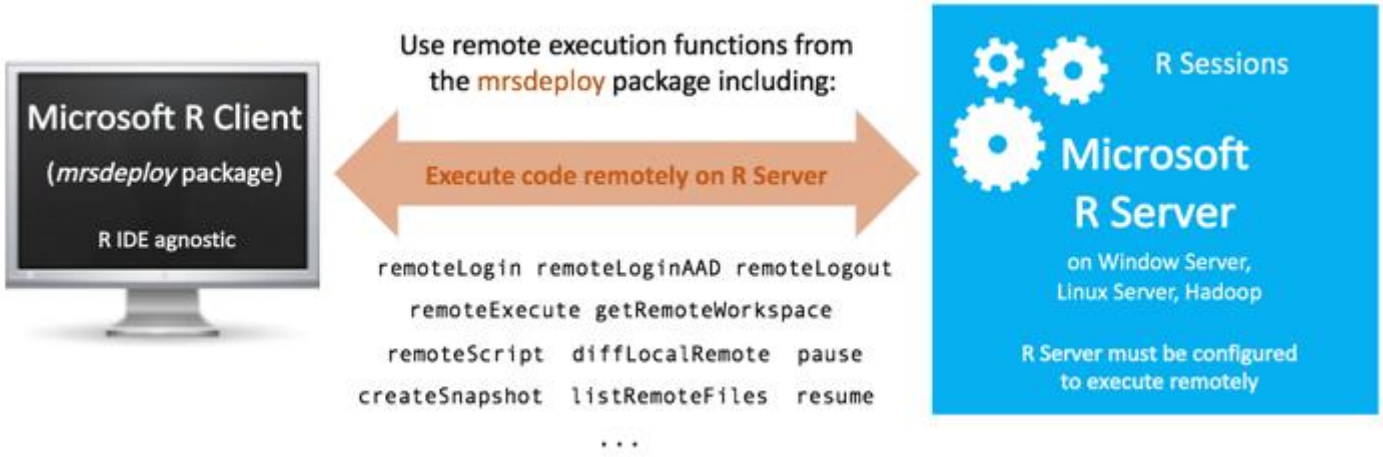
tinyurl.com/Strata2017R/Performance_Comparison

Deep learning and web service

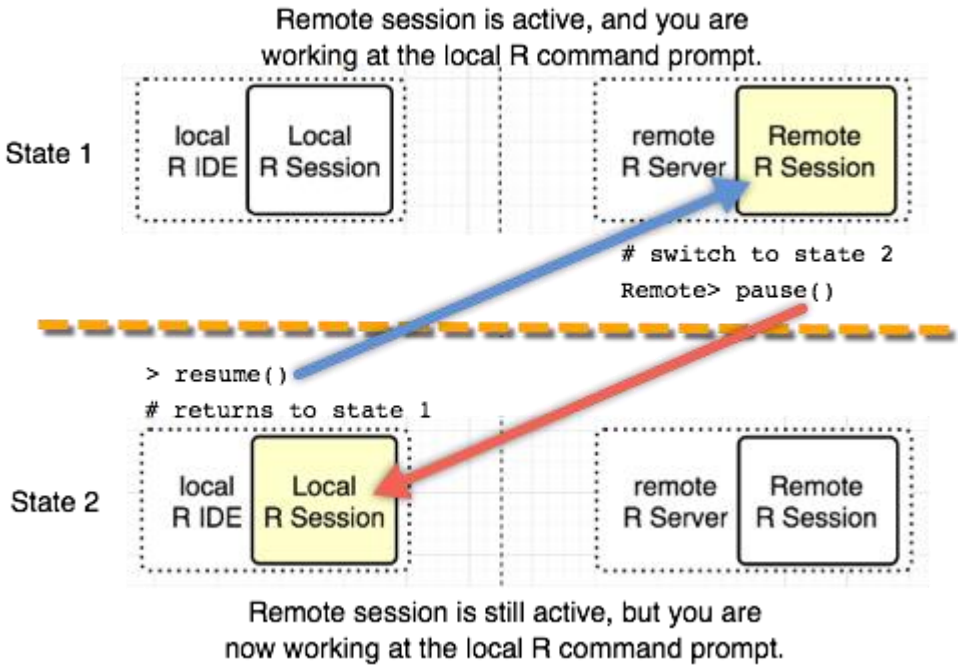
- Deep neural network in MicrosoftML package
  - rxNeuralNet() function.
  - GPU acceleration.
  - NET# language to customize network.
- Web service deployment in mrsdeploy package.
  - remoteLogin(), publishService(), getService(), etc.
  - Supports script-based and realtime based.
  - Supports Swagger.

rxNeuralNet() - https://msdn.microsoft.com/en-us/microsoft-r/microsoftml/packagehelp/neuralnet
NET# - https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-azure-ml-netsharp-reference-guide
Web service with mrsdeploy - https://msdn.microsoft.com/en-us/microsoft-r/operationalize/data-scientist-manage-services#realtime

Interaction with remote machines



- mrsdeploy() package.
- Remote execution.
- One-box configuration.
- Access control via AAD.



Try yourself

- The demo can be found at https://github.com/Microsoft/acceleratoRs/tree/master/flightDelayPredictionWithDSVM
- Prerequisites:
  - Azure subscription (**free for trials**).
  - R 3.3.x.
  - Microsoft R Server 9.x.
  - Microsoft R Client.
  - R packages
    - AzureSMR – http://github/Microsoft/AzureSMR
    - AzureDSVM – http://github/Azure/AzureDSVM

Thank you!