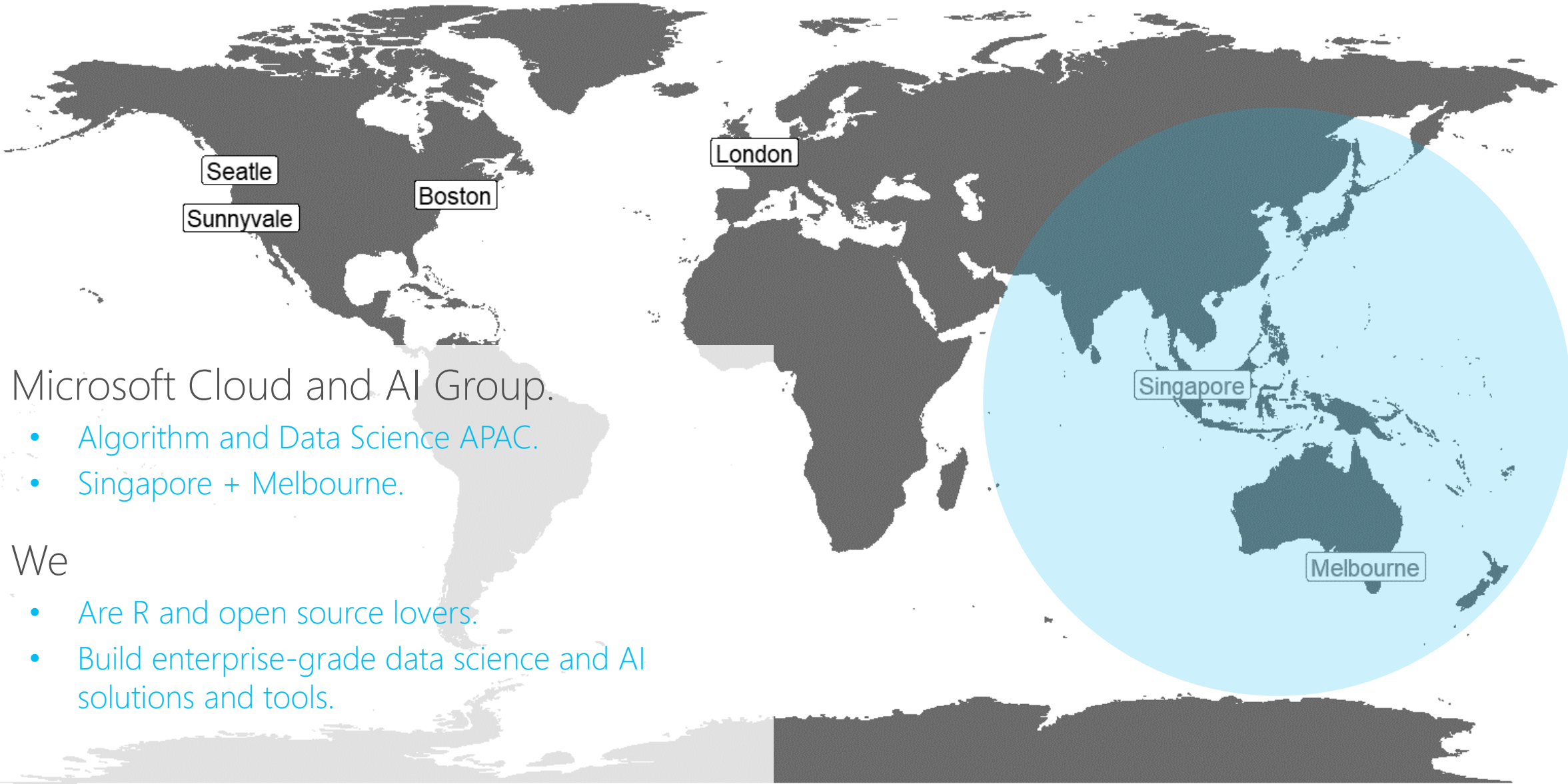
R you ready for Cloud and AI?

by Le Zhang, Ph. D., Data Scientist
Microsoft Algorithm and Data Science Asia Pacific

Microsoft

Self-introduction



- Microsoft Cloud and AI Group.
  - Algorithm and Data Science APAC.
  - Singapore + Melbourne.
- We
  - Are R and open source lovers.
  - Build enterprise-grade data science and AI solutions and tools.

A glimpse of R

Firstly appeared in the year of 1993, created by Ross Ihaka and Robert Gentleman.

Programming language and software environment for statistical computing.

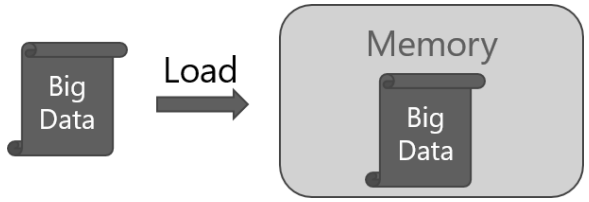
More than 11000 packages on CRAN, Github, Bioconductor, Bitbucket, etc.

Cross-platform, functional, graphical, etc.

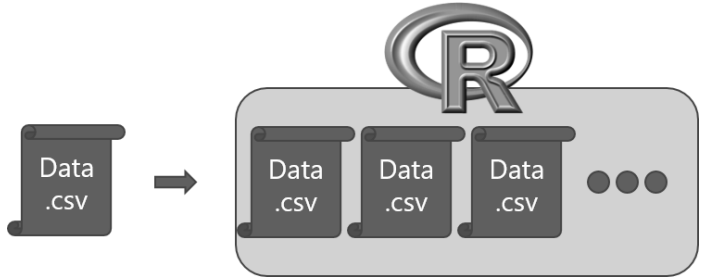
A glimpse of R

- R in the past
  - Deficiency in execution mechanisms.
  - Inadequacy of community and commercial support.

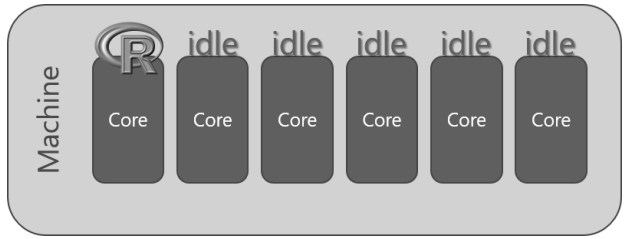
In-memory operation



Expensive data movement and duplication



Lack of parallelism



Inadequacy of community support

Lack of guaranteed support timeliness

No SLAs or support models

http://blog.revolutionanalytics.com/2016/10/tutorial-scalable-r-on-spark.html

A glimpse of R

- R in 2017
  - Scalable, parallelizable, product-ready, and commercial support from eco-system.
  - Big data, deep learning, and artificial intelligence.

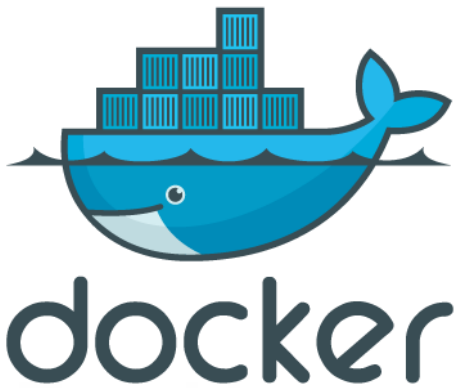
acxiom dv01 eToro PROS Powering Modern Commerce with Dynamic Pricing Science MERCK 125 YEARS HEARTLAND BANK

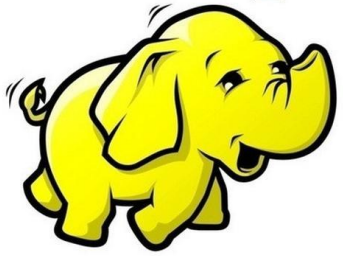
R consortium Microsoft RStudio ROpenSci software carpentry rOpenGov

https://www.slideshare.net/RevolutionAnalytics/r-then-and-now

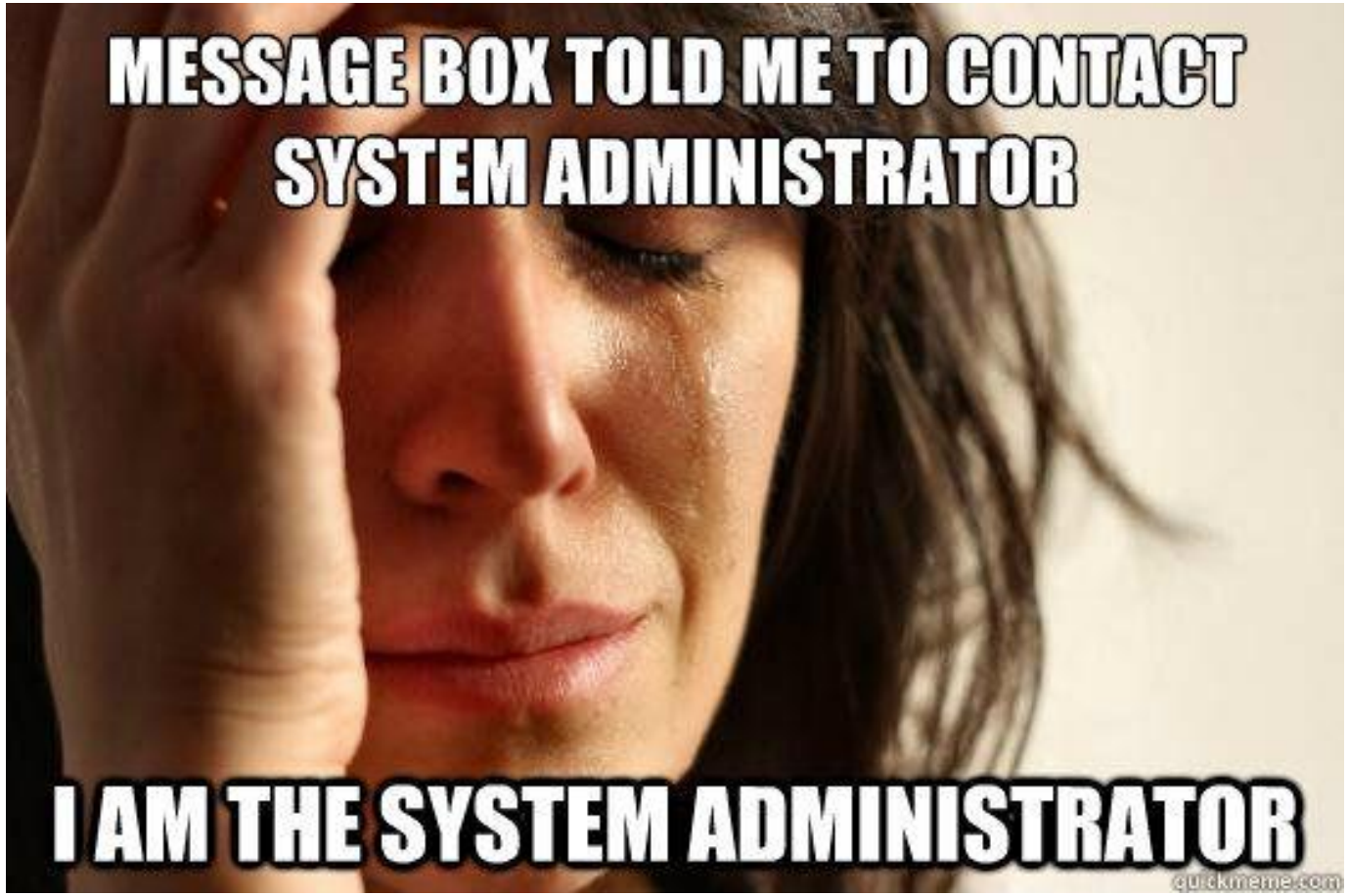
Why cloud computing

- Before practicing with the state-of-the-art technologies for your R analytical work, you probably need to
  1. Scale up/out cluster.
  2. Set up Hadoop/Spark.
  3. Install and configure software/runtime.
  4. …

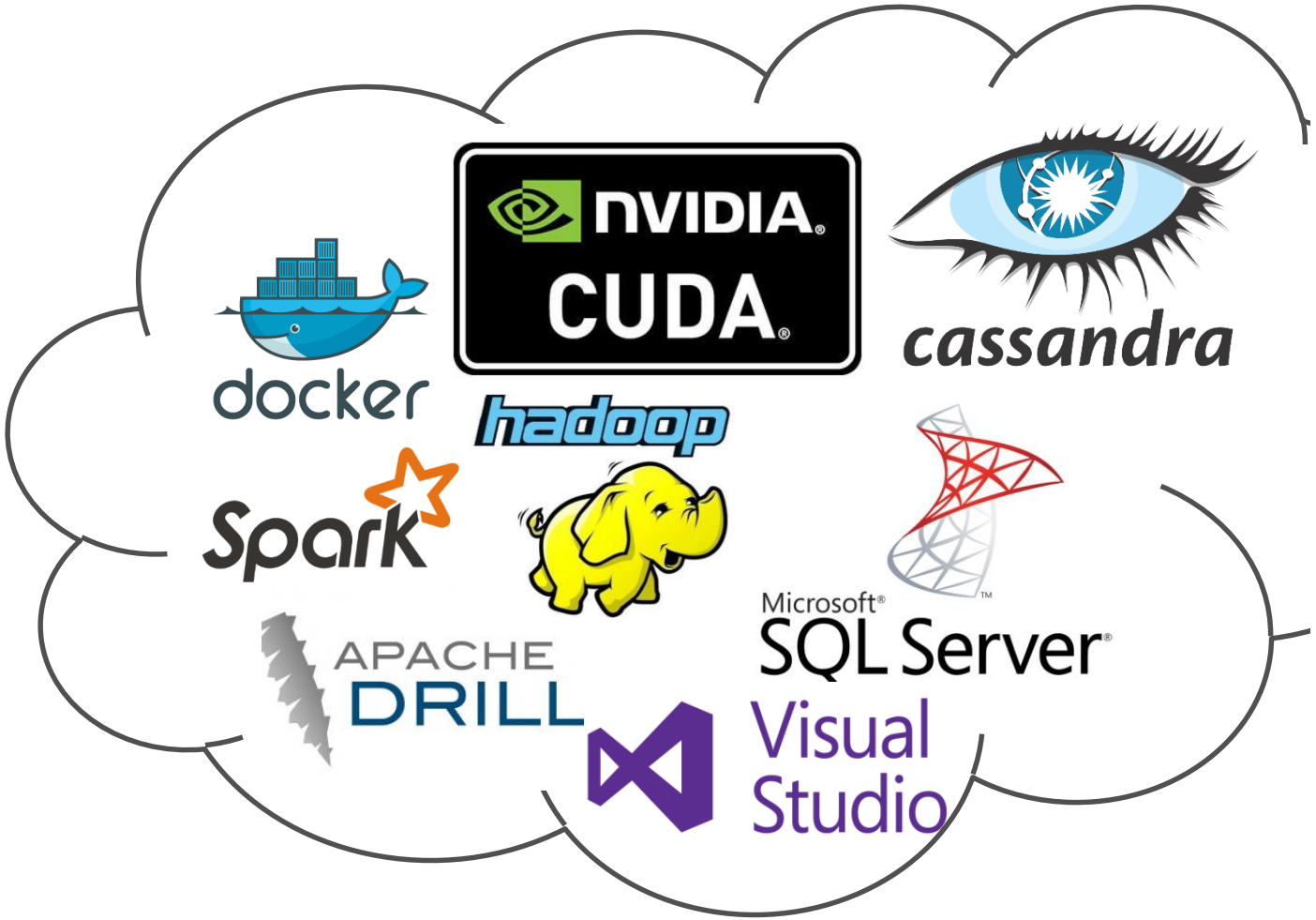




Why cloud computing



Why cloud computing



https://azure.microsoft.com/en-us/suites/cortana-intelligence/

Why cloud computing

Cloud computing with R

- R on cloud instances
  - Administration and operation of cloud resources.
  - Run large-scale analytical jobs.
  - Interactively prototype and develop data science or AI solutions.
  - Deploy applications or services.

Resource management	Scalable analytics	Interaction with remote	Application deployment
AzureSMR AzureDSVM doAzureParallel …	RevoScaleR MicrosoftML doAzureParallel dplyrXdf parallel doSNOW sparklyr SparkR …	IDE (Rstudio Server, Jupyter Notebook, etc.) Remote Desktop SSH mrsdeploy remoter …	mrsdeploy AzureML. Shiny Docker container …

Cloud computing with R – resource admin

- AzureSMR
  - Managing a selection of Azure resources such as storage blobs, HDInsight, etc.
- AzureDSVM
  - Deployment and operation of Azure Data Science Virtual Machine (DSVM).
  - Remote execution of script and file transfer with a Linux DSVM.
  - Retrieval of cost and expense information of using DSVM.

```
# Authentication.
sc <- createAzureContext(tenant="{TID}",
                         clientID="{CID}",
                         authKey="{KEY}")
azureAuthenticate()

azureListSubscriptions()
```

```
# DSVM operation – deployment.
deployDSVM(sc,
           resource.group="{resource group}",
           location="{location}",
           hostname="{DSVM name}",
           username="{user name}",
           size="{DSVM size}",
           authen="{authentication method}",
           password="{password}")
```

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

Cloud computing with R – scalable analytics

- RevoScaleR – collection of functions for practicing data science in scale
  - Stream data into RAM in blocks – "big data" in any size.
  - XDF file format – optimized file format to speed up iterative algorithm processing. Used with dplyrXdf for convenient manipulation.
  - Support for various computing contexts, local multi-core, multi-node across cluster, SQL server, and Hadoop/Spark.

```
### LOCAL COMPUTING CONTEXT ###
# Set up local environment variables.
rxSetComputeContext("localpar")

# Create Linux directory and file objects.
linuxFS <- RxNativeFileSystem( )
AirlineDataSet <- RxXdfData("airline_20MM.xdf",
                            fileSystem = linuxFS)
```

```
### LOCAL COMPUTING CONTEXT ###
# Setup of Spark environment variables.
mySparkCC <- RxSpark()

# Spark compute context.
rxSetComputeContext(mySparkCC)

# Create HDFS directory and file objects.
hdfsFS <- RxHdfsFileSystem()
AirlineDataSet <- RxXdfData("airline_20MM.xdf",
                            fileSystem = hdfsFS)
```

https://docs.microsoft.com/en-us/r-server/r/concept-what-is-revoscaler

Cloud computing with R – scalable analytics

- sparklyr – R interface to Apache Spark with a complete backend of dplyr
  - Filter and aggregate Spark datasets and then bring them into R for analysis and visualization.
  - Use Spark's distributed machine learning library from R.
  - Create extensions to call Spark API and provide interfaces to Spark packages.
- SparkR – light-weight frontend to use Apache Spark in R
  - Operations like filtering, grouping, selecting, etc., and MLib model training.

```
# sparklyr for data manipulation.
sc <- spark_connect(master="local")

iris_tbl <- copy_to(sc, iris)
flights_tbl <- copy_to(sc, flights, "flights")

iris_preview <-
   dbGetQuery(sc, "SELECT * FROM iris LIMIT 10")
flights_tbl <- filter(dep_delay == 2)
```

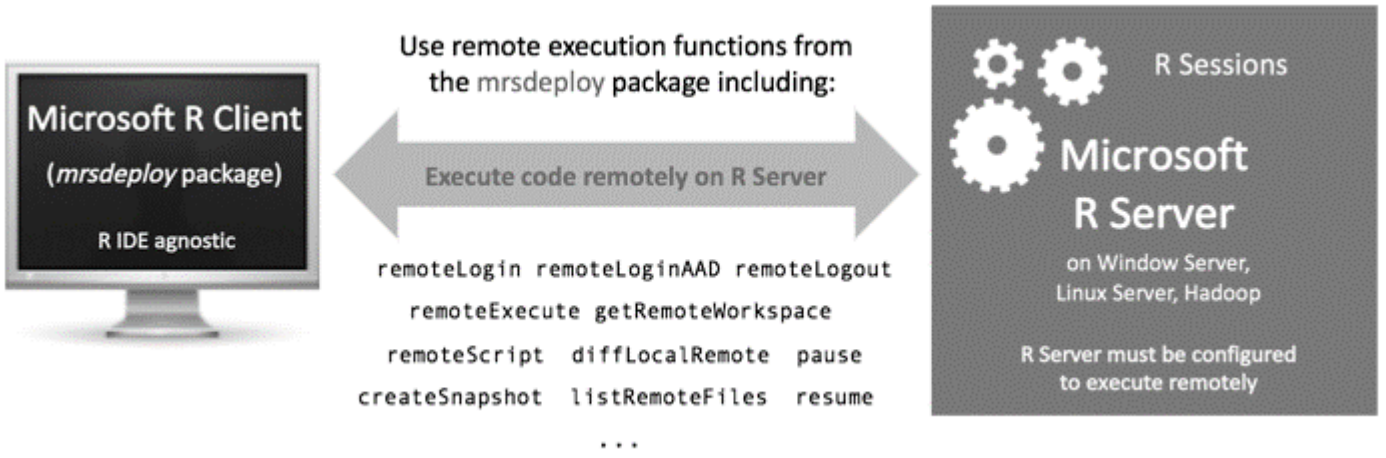
https://spark.rstudio.com/

```
# SparkR for data manipulation
sparkR.session(master="local[*]",
                  sparkConfig=list(spark.driver.memory="2g")

df <- as.DataFrame(faithful)

head(select(df, df$eruptions))
head(select(df, "eruptions"))
```

https://spark.apache.org/docs/latest/sparkr.html

Cloud computing with R – remote interaction

- IDEs
  - Rstudio Server.
  - Jupyter Notebook.
- Remote Desktop (Windows RDC or X2Go)
- R packages
  - Microsoft mrsdeploy.



https://docs.microsoft.com/en-us/r-server/r/how-to-execute-code-remotely

```
# Login to the remote R Server.
remoteLogin(deployr_endpoint, session, ...)

# Remote execution.
remoteExecute(rcode, script, ...)

# Remote interaction if session is TRUE.
# R session at remote machine will be created.
# REMOTE> x <- rnorm(100)
```

Cloud computing with R – app deployment

- Shiny
  - A web application framework for R.
  - Interactive web application with no prior knowledge of JS, HTML, or CSS required.
- mrsdeploy
  - Easy to deploy and consume web based service in R session.
  - Compatible with Swagger.

```
# Publish a standard service 'mtService' version 'v1.0.0'. Assign service to 'api' variable.
api <- publishService( "mtService",
                       code = manualTransmission,
                       model = carsModel,
                       inputs = list(hp = "numeric", wt = "numeric"),
                       outputs = list(answer = "numeric"),
                       v = "v1.0.0" )
result <- api$manualTransmission(120, 2.8)
```

https://docs.microsoft.com/en-us/r-server/r-reference/mrsdeploy/mrsdeploy-package

Cloud computing with R – app deployment

- AzureML
  - Interaction with Azure Machine Learning (AML) Studio.
  - Deployment of AML service in R session.

```
# Publish a web service on AzureML.
add <- function(x, y) {x +y}
ws <- workspace()
api <- publishWebService(ws, fun=add, name="AzureML-vignette-silly",
                                      inputSchema=list(x="numeric", y="numeric"),
                                      outputSchema=list(ans="numeric"))
```

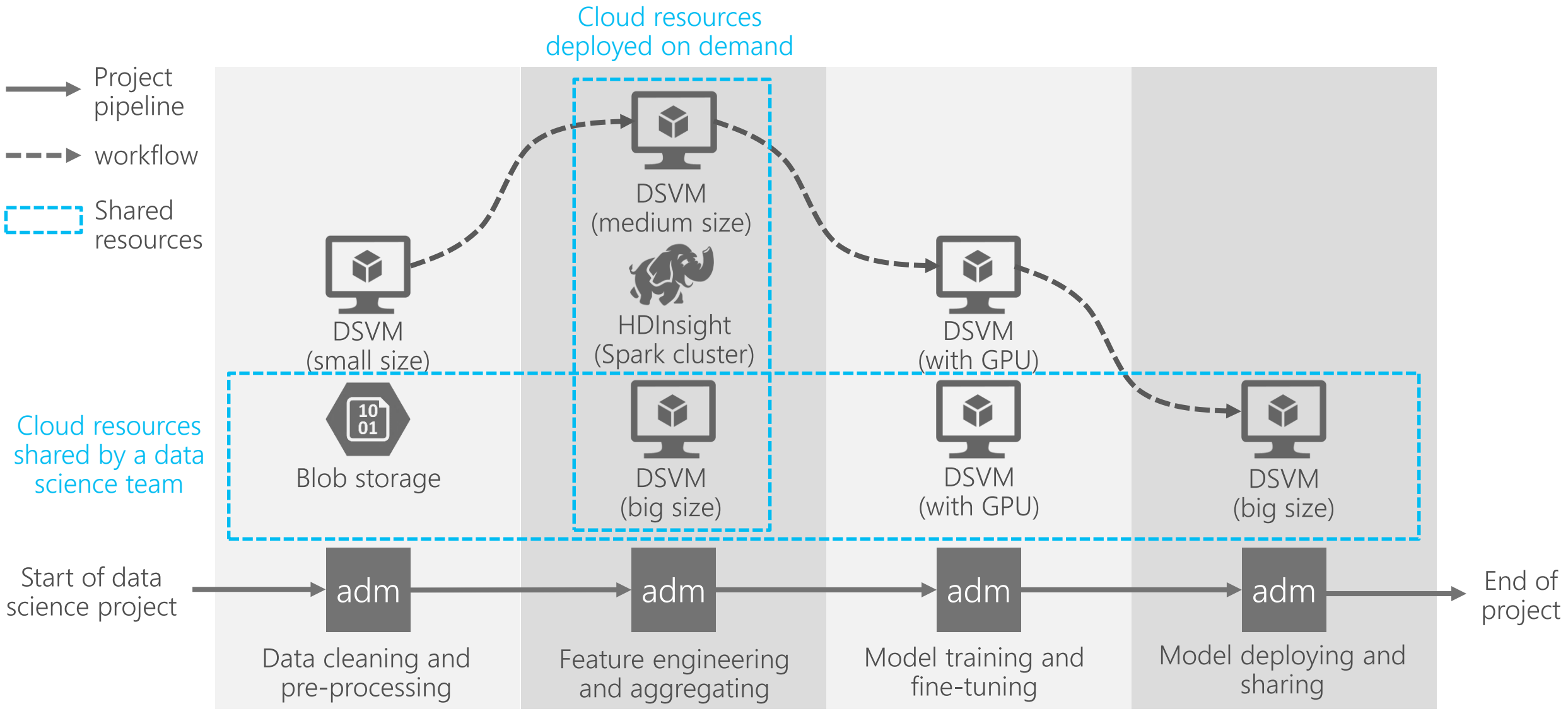
https://cran.r-project.org/web/packages/AzureML/vignettes/getting_started.html

- Docker container
  - Docker hub/Azure Container Registry.
  - Azure Container Services/Azure Container Instances.

Data science and AI pipeline on cloud

- Operationalizing, managing, and administrating an enterprise-grade data science development/production pipeline ALL BY USING R.
- Benefits inherited from Azure cloud platform
  - Elasticity of resource deployment.
  - Security and role access control.
  - Computational efficiency of on-demand high-performance and distributed data analytics.
  - Ease of modularization.
  - Cost effectiveness.
- Collaborative development SW/HW environment for R-user data scientist, data engineer, and solution architect teams – Microsoft R Server, Hadoop/Spark, GPU-incorporated virtual machine, Docker container, etc.

Operationalization of a data science pipeline



Operationalization of a data science pipeline

Elastic deployment and use of cloud resources ranging from data/file storage, single virtual machine node, to large-scale high-performance computing cluster.

Azure cloud resources

Local R packages:
- AzureSMR
- AzureDSVM

Key functions:
- deployDSVM
- deployDSVMCluster
- azureCreateHDI
- azureCreateStorageAccount
- azureDeleteResourceGroup

Azure AD Application

adm

Management and operation of Azure cloud resources.

Serverless-like data analytics execution on varieties of computing instances on cloud.

Azure cloud computing resources

Local R packages:
- MicrosoftML
- mrsdeploy
- doAzureParallel
- dplyrXDF

Key functions:
- remoteLogin
- remoteExecute
- rxExeBy
- rxFastTrees
- rxNeuralNet

adm

High-performance and distributed data analytics

AI service deployment with a micro-service framework.

Azure cloud resources

adm

Model and service deployment

Talk is cheap. Show me the code.

- *Linus Torvalds*

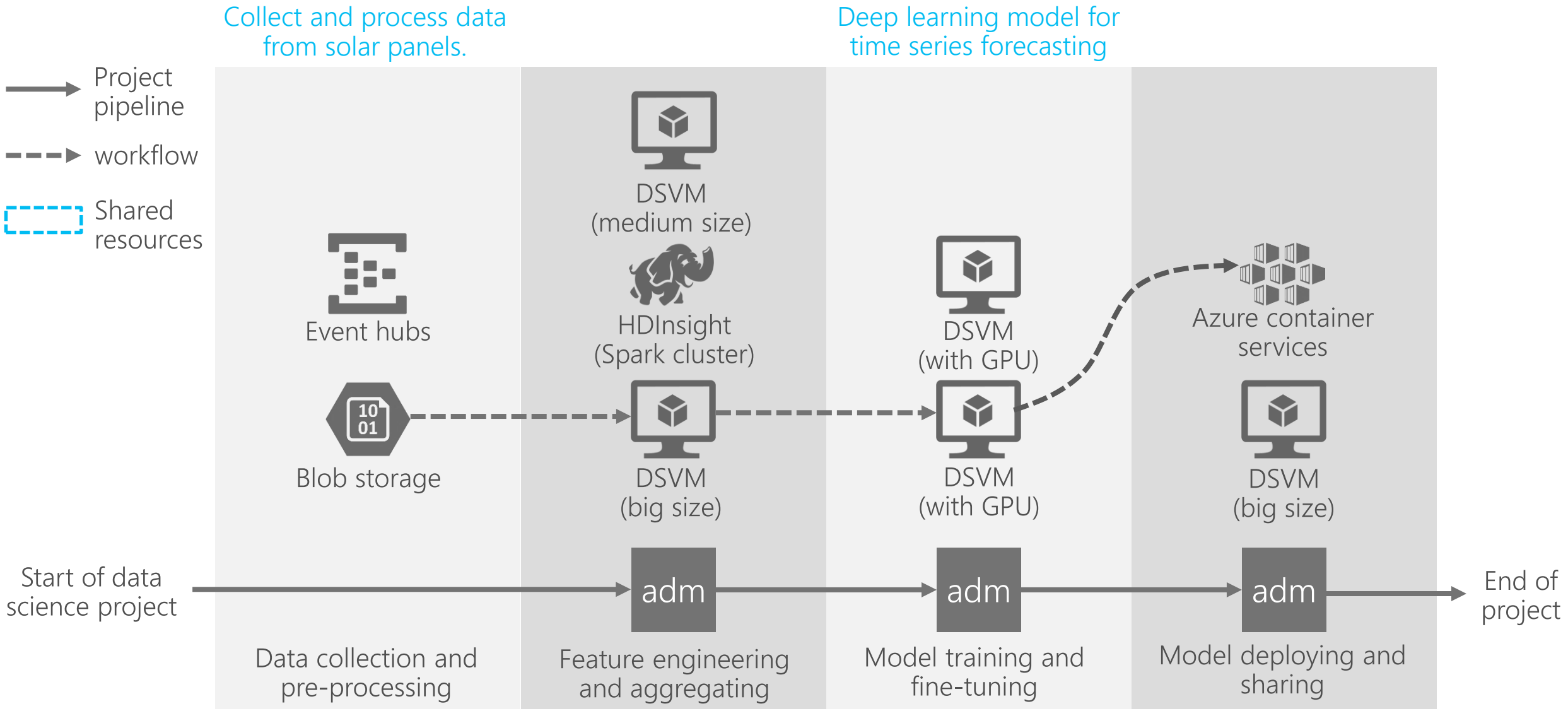
Use case 1 – Solar power forecasting

- Solar power forecasting is essential for efficient use, the management of electricity grid, and solar energy trading.
- Stochastically learning method fits in resolving the forecasting problem.
  - Modeling the time series.
  - Forecasting the long-term dependencies on the univariates.
- Deep learning based approach.
  - Algorithm: Long Short-Term Memory (LSTM).
  - Computing engine: DSVM with GPU.
  - Data set: photovoltaic systems device reading from solar panels.

https://github.com/Microsoft/CNTK/blob/master/Tutorials/CNTK_106B_LSTM_Timeseries_with_IOT_Data.ipynb

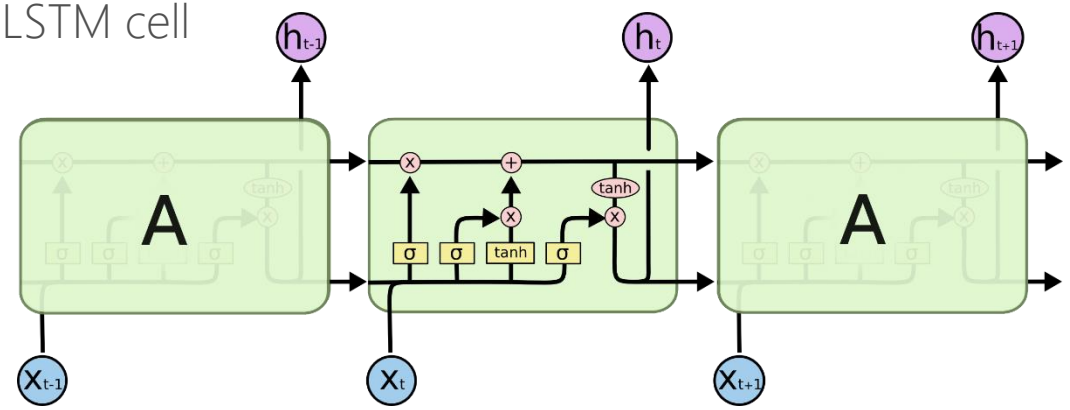
https://en.wikipedia.org/wiki/Solar_power_forecasting

Solar power forecasting – operationalization



Solar power forecasting with LSTM

LSTM cell



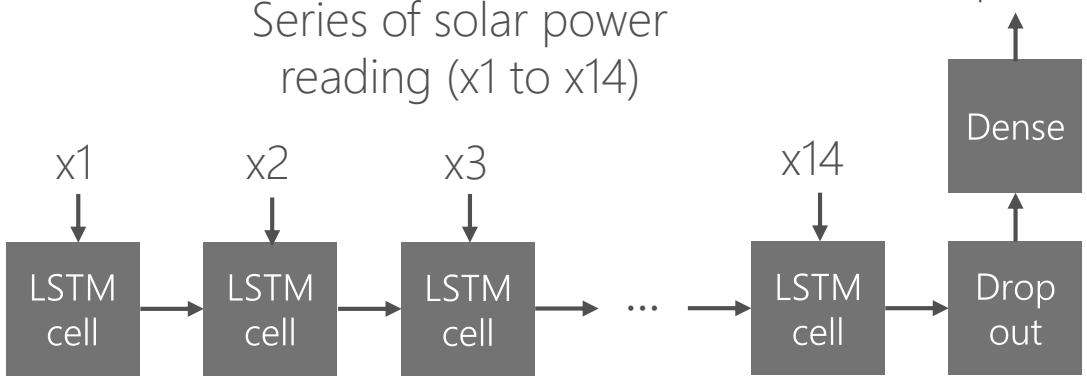
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Original data

Time	Current	Total
2013-12-01 07:00:00	6.3	1.69
2013-12-01 07:30:00	44.3	11.36
2013-12-01 08:00:00	208.0	67.50
2013-12-01 08:30:00	482.0	250.5

Predicted total solar power (y)

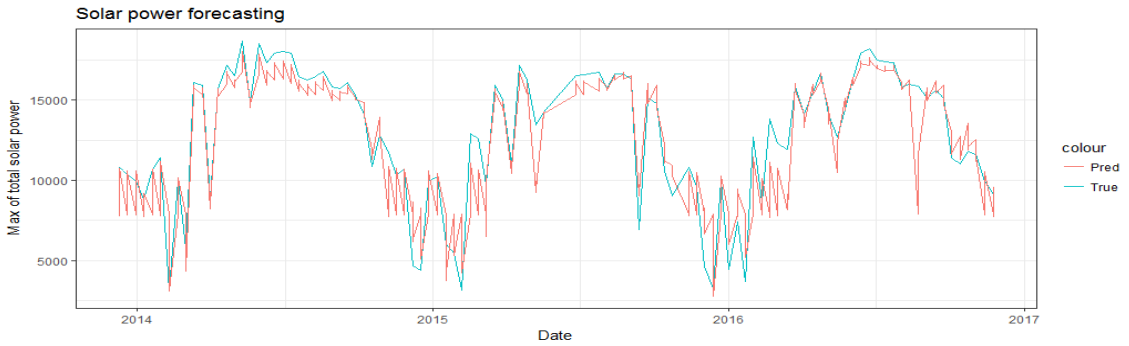
Series of solar power reading (x1 to x14)



https://github.com/Microsoft/CNTK/blob/master/Tutorials/CNTK_106B_LSTM_Timeseries_with_IOT_Data.ipynb

Prediction scenario & results

1.7,11.4 -> 10300
1.7,11.4,67.5 -> 10300
1.7,11.4,67.5,250.5 ... -> 10300



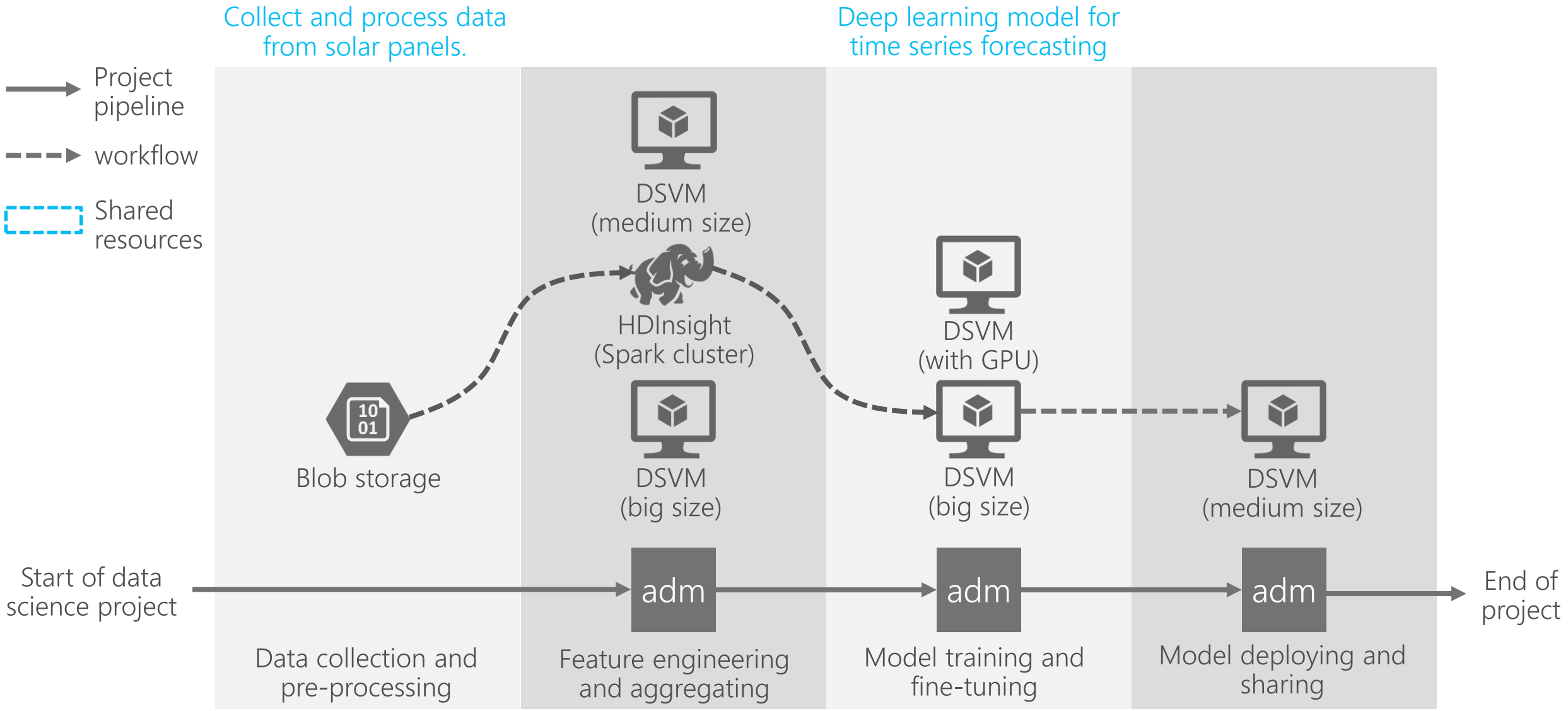
Use case 2 – Flight delay prediction

- Flight delay and cancellation induces huge loss to both airline companies and passengers. Predicting delay helps reduce cost and improve operational efficiency.
- Statistically learning the patterns in the relevant features for creating models.
  - Binary classification problem
  - Feature engineering involves exploratory studies.
- Big data analytics
  - Spark cluster: scalable analytics on large data set.
  - DSVM: model training and deploying.

http://www.moneysavingexpert.com/travel/flight-delays

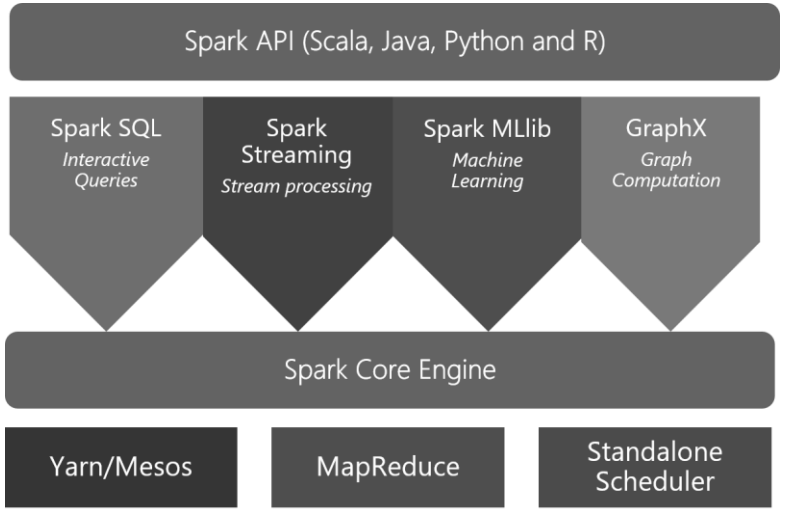
Original data source: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

Flight delay prediction – operationalization



Flight delay prediction with Spark cluster

- 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.



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

R interfaces

- RevoScaleR
- sparklyr
- SparkR

- To predict whether a flight is delayed or not given relevant features.
- Summary:
  - Size: ~1.4 GB.
  - Features: day-of-month, day-of-week, origin, etc.
- A sub-set (10%) is used.
- Spark is only illustrated for feature selection.

ArrDel15 ~
DayofMonth + DayOfWeek + Dest + Origin + DepTime

- Model is trained with a simple neural network.
  - rxNeuralNet of MicrosoftML package.
  - GPU acceleration.
  - Illustration only. Can be further optimized.

Try it yourself!

- Microsoft R packages featured for cloud based data science and AI development
  - AzureSMR – https://github.com/Microsoft/AzureSMR
  - AzureDSVM – https://github.com/Azure/AzureDSVM
  - Microsoft R (RevoScalerR, mrsdeploy, MicrosoftML, etc.) – https://docs.microsoft.com/en-us/r-server/
- acceleratoRs
  - R based data science solution accelerator suite that provides templates for prototyping, reporting, and presenting data science analytics of specific domain.
  - https://github.com/Microsoft/acceleratoRs
    - Solar panel power forecasting – https://github.com/Microsoft/acceleratoRs/tree/master/SolarPanelForecasting
    - Flight delay prediction – https://github.com/Microsoft/acceleratoRs/tree/master/flightDelayPredictionWithDSVM
    - Other accelerators for specific domains such as “Credit Risk Prediction”, “Employee Attrition Prediction”, etc.

We would love to have your feedbacks and pull requests!

- Microsoft ADS APAC team


Microsoft