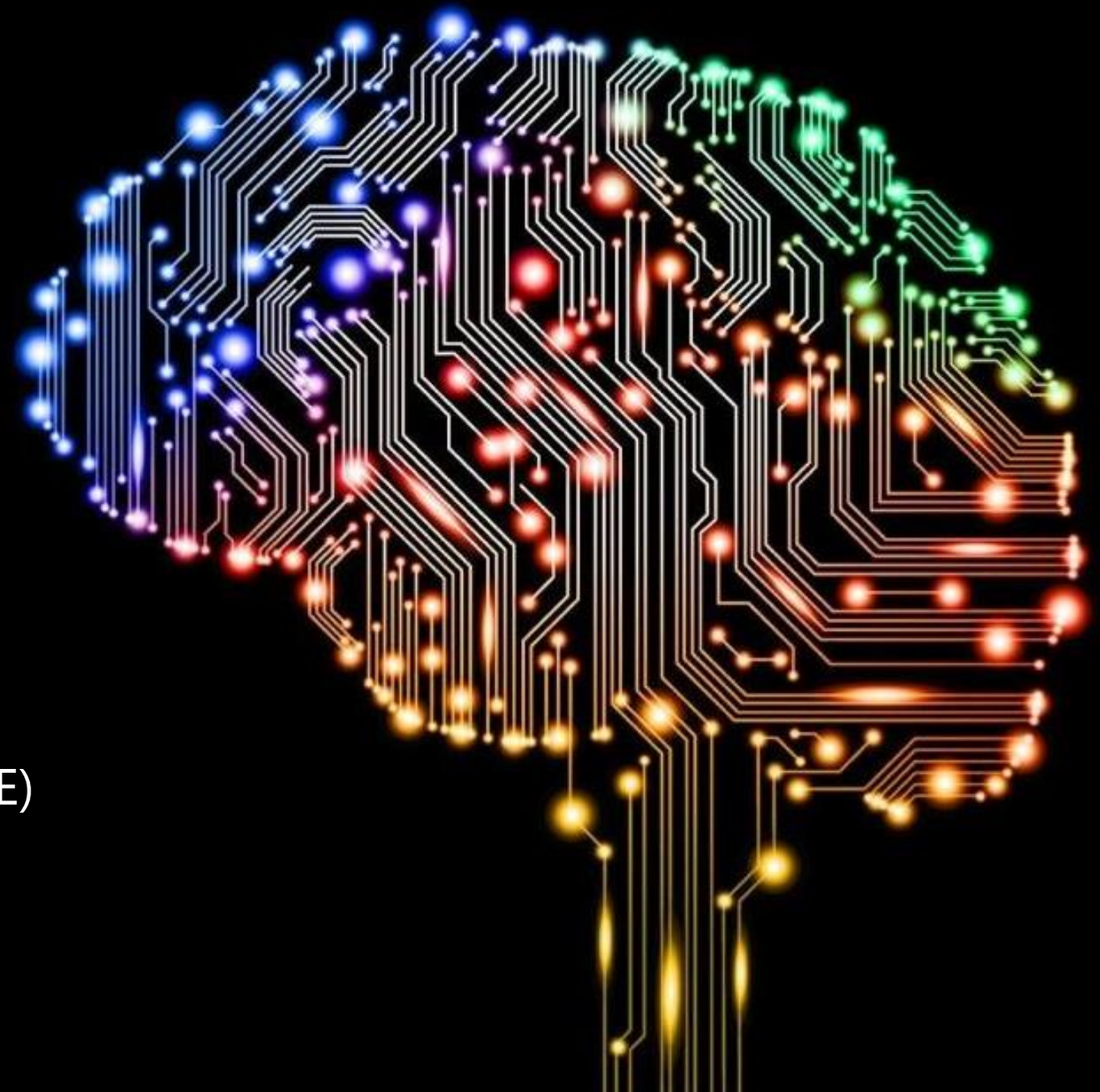
Empowering Data Analytics with R in Microsoft Azure

R User Group KL

Krit Kamtuo
Software Engineer
Commercial Software Engineering (CSE)
Microsoft Asia Pacific

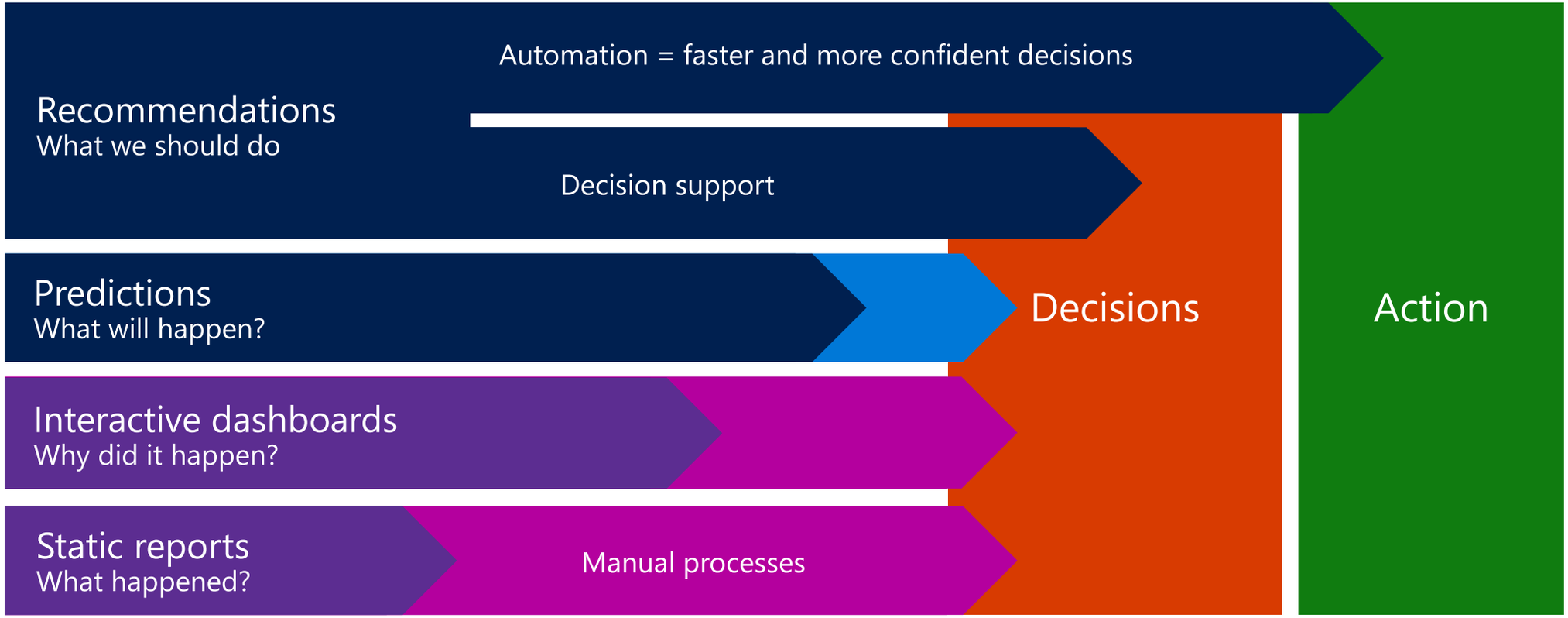


Microsoft

From data to decisions and actions

Data will drive $1.6 Trillion in additional value for businesses

*IDC 2014



Value

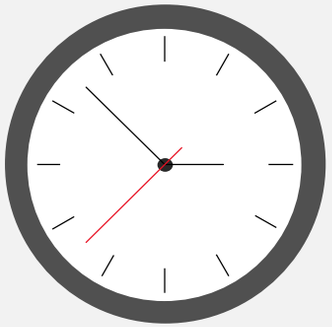
Barriers to deriving value from analytics

Cost

Legacy analytics software has a high total cost of ownership

Talent

Finding talent with the right skills is hard. Universities are not training data scientists on legacy tools



Speed

You analytics team spends a lot more time on cleaning, preparing and operationalizing than actually modelling the data (from concept to production)

Flexibility

Legacy analytics software make its hard to work with data that lives across both on premise and cloud stores because of incompatible platforms

Microsoft's approach to advanced analytics

Hybrid

Scale

Agility

Low cost

Works with what you have

AI Driven Innovation

Solutions to accelerate time to value

Microsoft advanced analytics

Solutions

Microsoft R	SQL Server 2016	Cortana Intelligence
		
R-based analytics	In database analytics	Cloud analytics

What is

#1
Language
Advanced
Analytics

- A statistics programming language
- Data analysis and visualization capabilities
- Strong ties to academia feeds ever-growing machine learning capabilities

3M+
Users

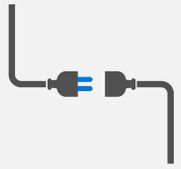
- Majority of data scientists use R
- New and recent grad's use it
- Thriving user groups worldwide

Open
Biggest
Ecosystem

- Vibrant open Source community
- 10,000 + free algorithms in CRAN
- Constantly innovating

Microsoft's investment principles for R

Enterprise grade operationalization and support



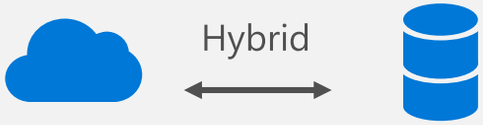
Operational in SQL or in web apps

24/7 enterprise grade support

The flexibility to work with what you have




Works across your analytics environment



Works on premise and in the cloud

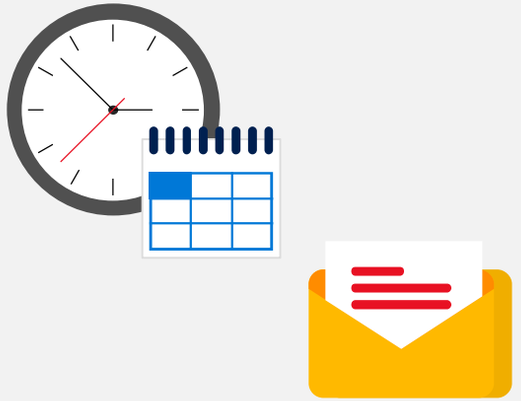
Best of Open Source and Microsoft innovation

Rapidly growing library of 10,000+ packages

+

Scalable algorithms and Microsoft Machine Learning Libraries

Accelerated time to value with solutions



Marketing Campaign Optimization and other solutions to help you derive value

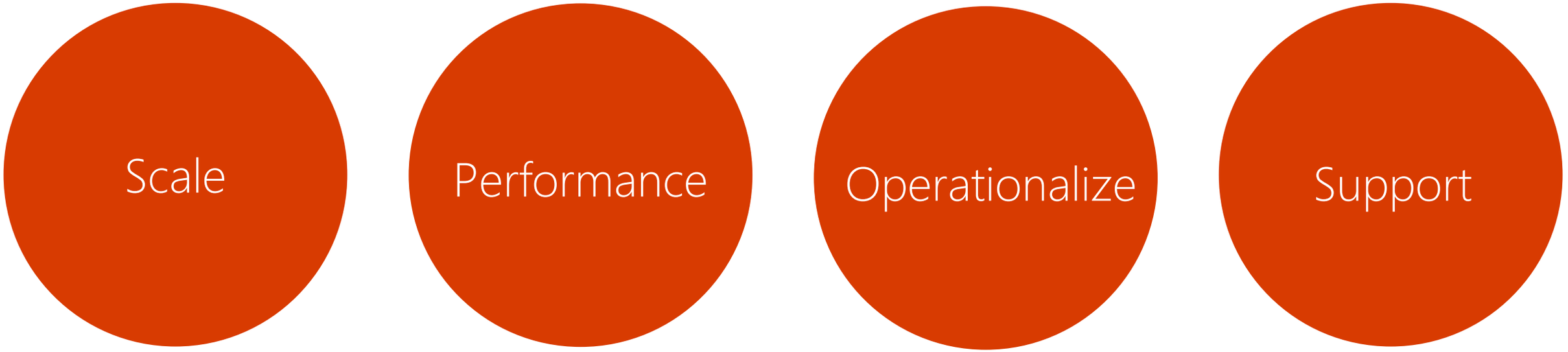
+

Built on top of Open Source R with access to all 10K+ CRAN libraries

CRAN is the "Comprehensive R Archive Network"

Open Source R challenges

Microsoft R solves open source R challenges



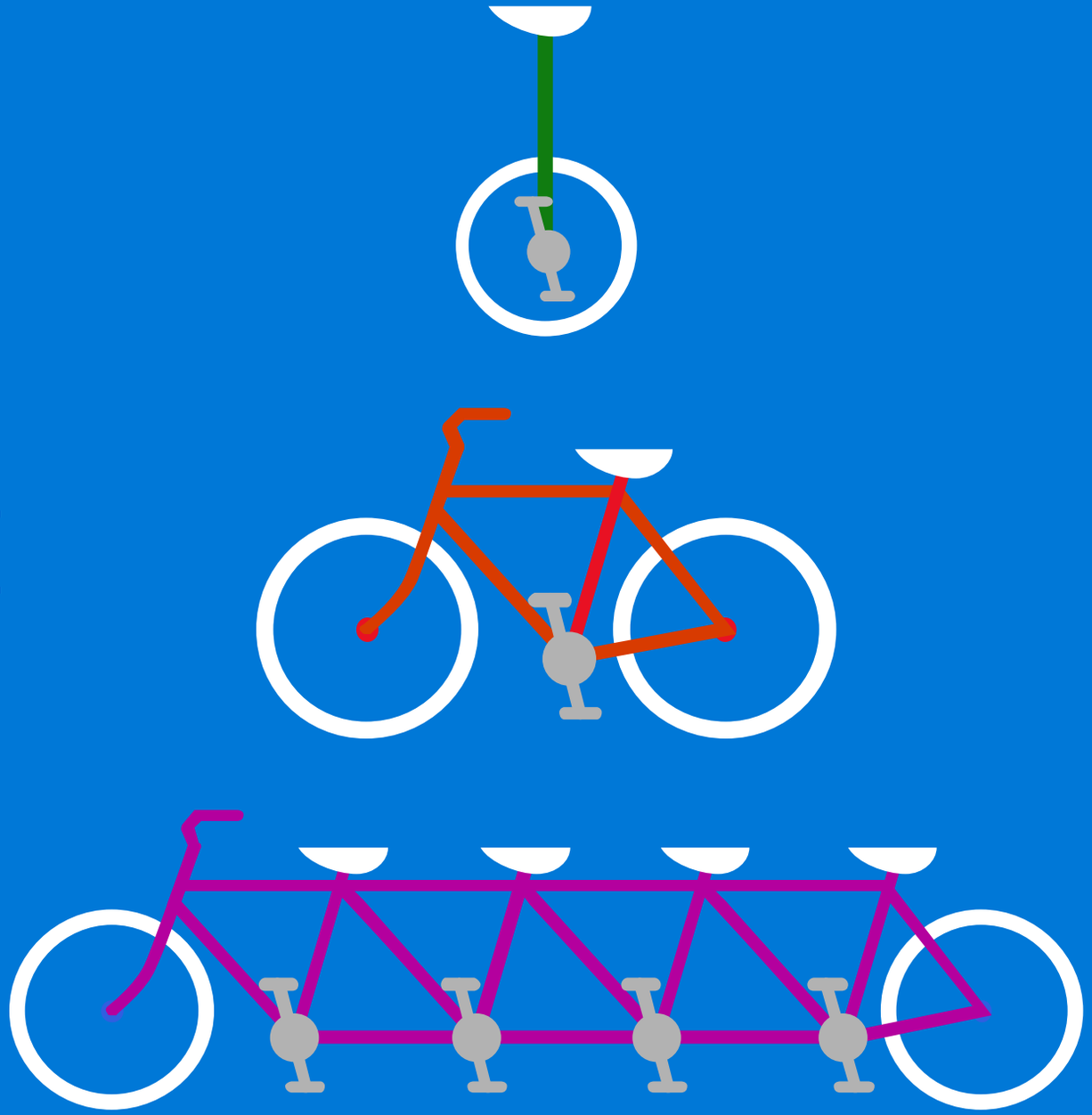
Scales from workstations to large clusters; scales to large data sizes

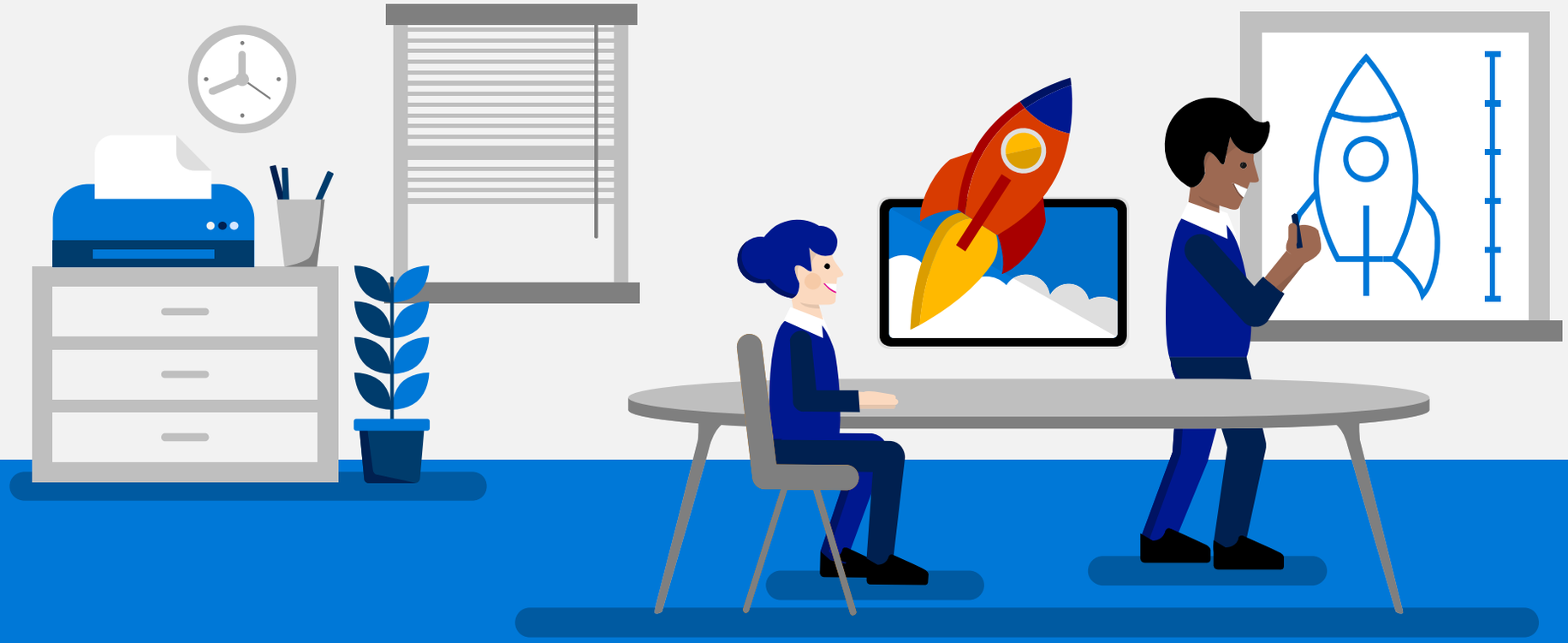
Growing portfolio of transparently parallelized and Microsoft Machine Learning algorithms

Write Once Deploy on multiple platforms—on-premises and cloud

Solution templates, enterprise level support, partner eco-system

Flexibility to work with what you have

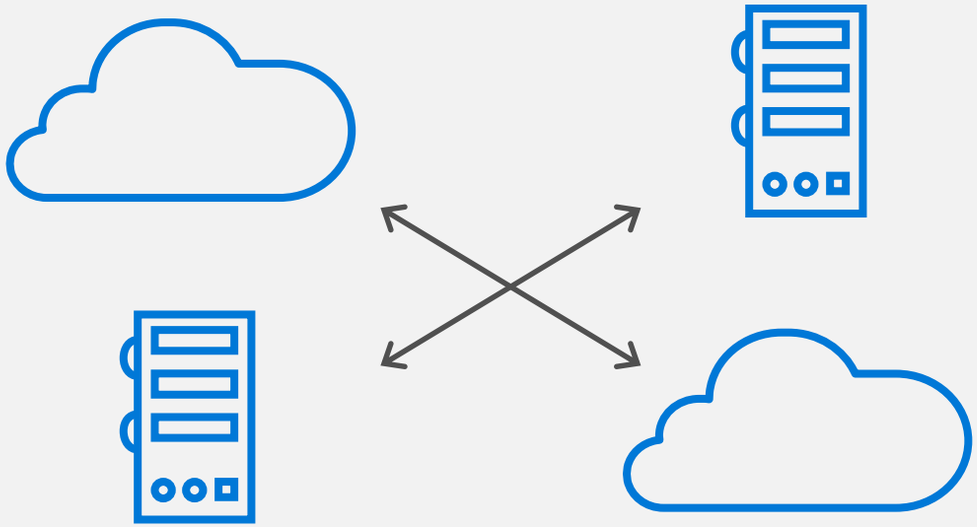




Microsoft R product overview

Flexibility of modeling and deployment

Model and deploy on premise, in the cloud or both



Deploy your R code to multiple platforms

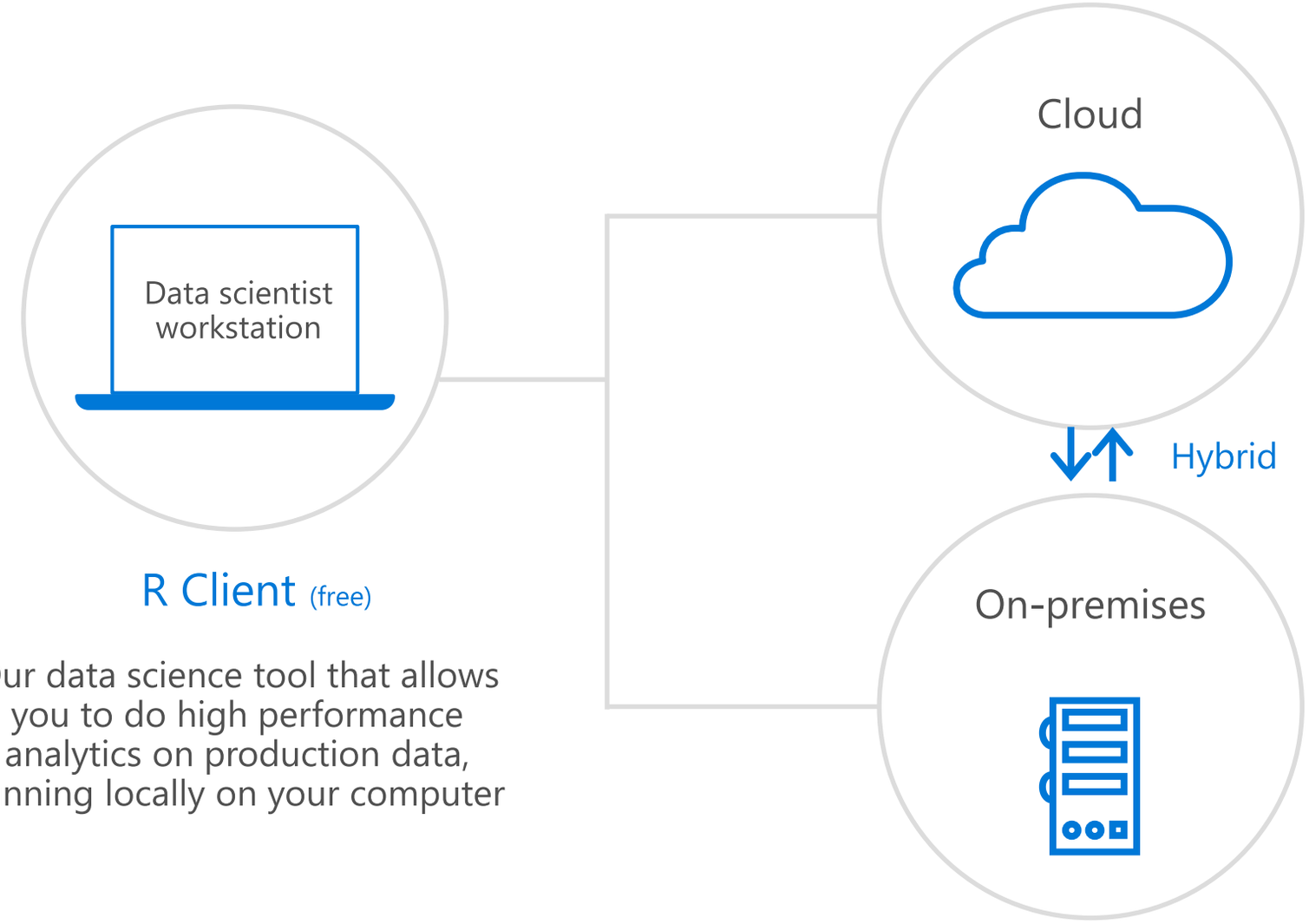
Microsoft SQL Server

TERADATA

APACHE Spark

hadoop

Microsoft R products



R Client (free)

Our data science tool that allows you to do high performance analytics on production data, running locally on your computer

Big Data
Machine Learning Server for HDInsight

Linux and Windows Servers
Virtual Machines

Cortana Intelligence Gallery
Azure Machine Learning Studio

Big Data
Machine Learning Server for Hadoop/Spark

In-Database
SQL Server 2016 (R Services)

Machine Learning Server for Teradata

Linux and Windows Servers
Machine Learning Server for Linux and Windows

Introducing Microsoft Machine Learning in R

- Create text classification models for problems such as sentiment analysis and support ticket classification
- Train deep neural nets with GPU acceleration in order to solve complex problems such as retail image classification and handwriting analysis
- Work with high-dimensional categorical data for scenarios like online advertising click-through prediction
- Solve many other common machine learning tasks such as churn prediction, loan risk analysis, and demand forecasting using state-or-the-art, fast and accurate algorithms
- Train models 2x faster than logistic regression with the Fast Linear Algorithm (SDCA)
- Train multilayer custom nets on GPUs up to 8x faster with GPU acceleration for Neural Nets
- Reduce training time up to 10x while still retaining model accuracy using feature selection

Microsoft Machine Learning package

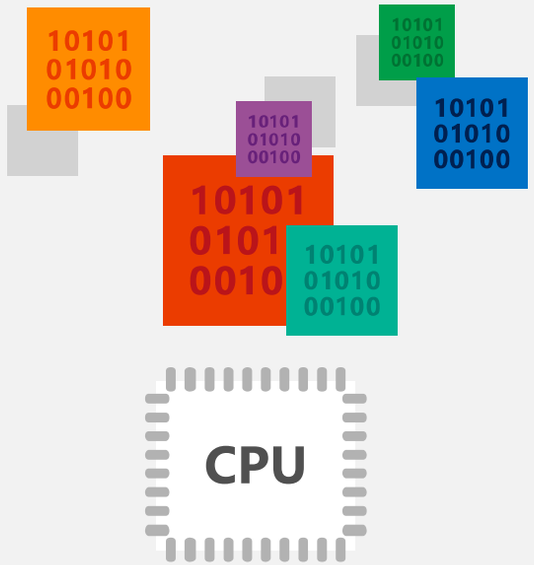
MicrosoftML adds battle tested algorithms and transforms bringing new machine learning functionality with increased speed, performance and scalability

Microsoft Machine Learning algorithms

Algorithm	ML task supported	Scalability	Application Examples
rxFastLiner() Fast Linear model (SDCA)	binary classification, linear regression	#cols: ~1B; #rows: ~1B; CPU: multi-proc	Mortgage default prediction, Email spam filtering
rxOneClassSvm() OneClass SVM	anomaly detection	cols: ~1K; #rows: RAM-bound; CPU: single-proc	Credit card fraud detection
rxFastTrees() Fast Tree	binary classification, regression	#cols: ~50K; #rows: RAM-bound; CPU: multi-proc	Bankruptcy prediction
rxFastForest() Fast Forest	binary classification, regression	#cols: ~50K; #rows: RAM-bound; CPU: multi-proc	Churn Prediction
rxNeuralNet() Neural Network	binary and multiclass classification, regression	#cols: ~10M; #rows: Inf; CPU: multi-proc CUDA GPU	Check signature recognition, OCR, Click Prediction
rxLogisticRegression() Logistic regression	binary and multiclass classification	#cols: ~100M; #rows: Inf for single-proc CPU #rows: RAM-bound for multi-proc CPU	

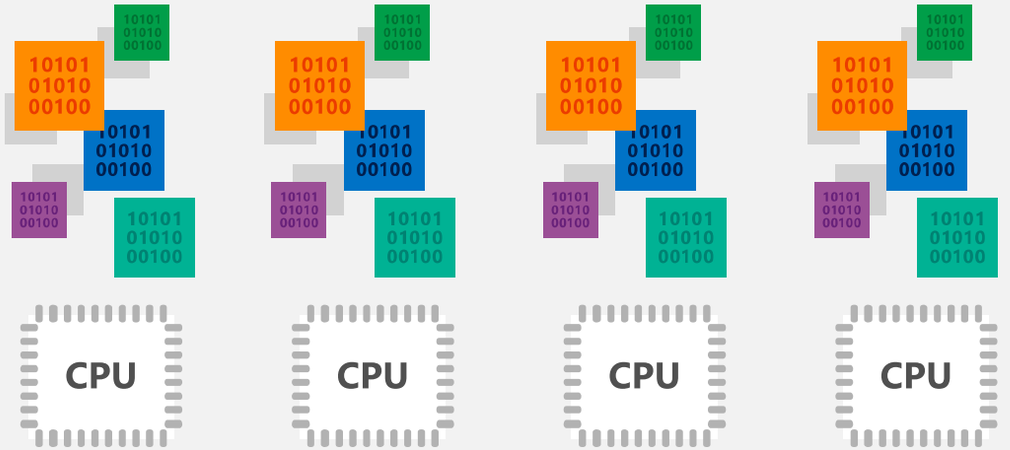
Multi-Threaded Parallelized Analytics with R

Open Source R



Open Source R is designed to use only a single thread (processor) at a time

Microsoft R



Parallel Worker tasks

Microsoft R allows you to scale your analytics with enhanced performance using multi-threaded parallelized algorithms

Hand on labs: https://github.com/kritcs18/rusergroupkl/

Write once—deploy on multiple platforms

ScaleR models can be deployed **from a server or edge node to run in Spark/Hadoop** without any functional R model re-coding

Compute context R script —sets where the model will run

Linux or Windows—compute context

```
### SETUP LOCAL ENVIRONMENT VARIABLES ###
  myLocalCC <- "localpar"

  ### LOCAL COMPUTE CONTEXT ###
  rxSetComputeContext(myLocalCC)


  ### CREATE LINUX, DIRECTORY AND FILE OBJECTS ##
  linuxFS <- RxNativeFileSystem( )
  AirlineDataSet <- RxXdfData("airline_20MM.xdf",
                              fileSystem = linuxFS)
```

In – Spark/Hadoop – Compute Context

```
### SETUP SPARK/HADOOP ENVIRONMENT VARIABLES ###
mySparkCC <- RxSpark()                    myHadoopCC <- RxHadoopMR()

### HADOOP COMPUTE CONTEXT ###
rxSetComputeContext(mySparkCC) rxSetComputeContext(myHadoopCC)

### CREATE HDFS, DIRECTORY AND FILE OBJECTS ###
hdfsFS <- RxHdfsFileSystem()
AirlineDataSet <- RxXdfData("airline_20MM.xdf",
                  fileSystem =hdfsFS)
```

Functional model R script—does not need to change to run in Spark

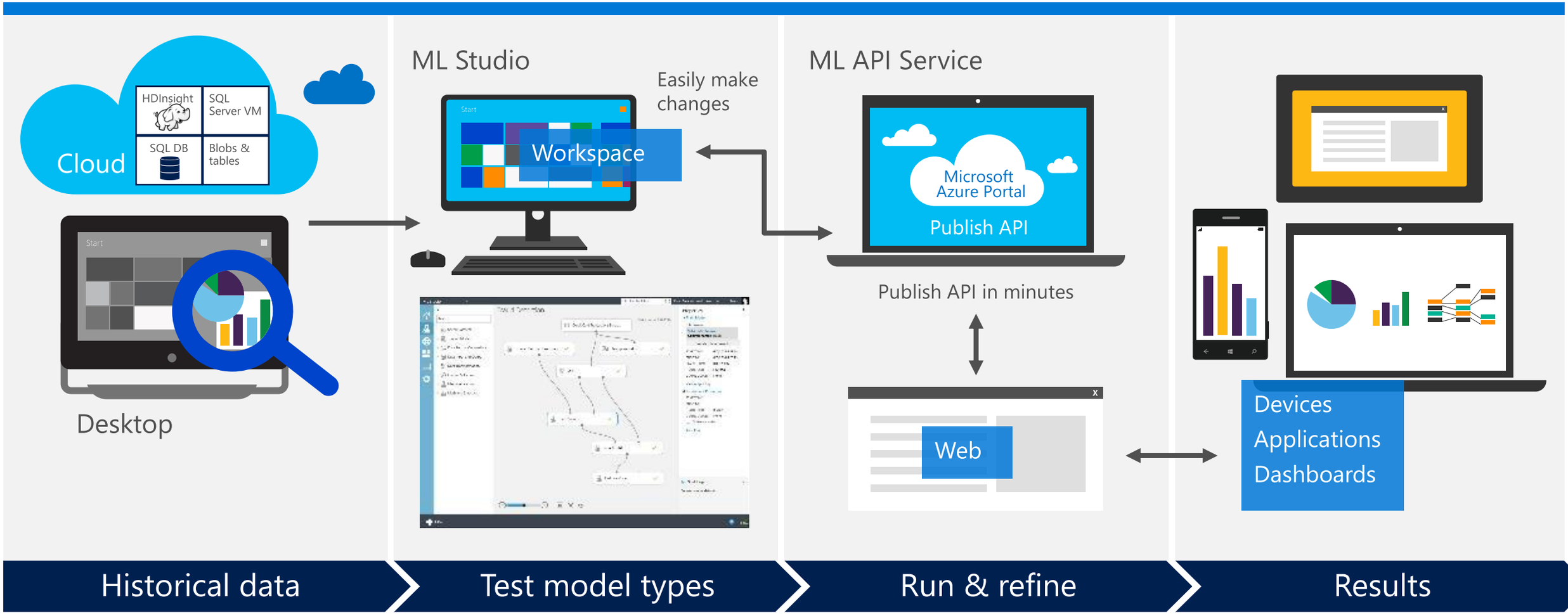
```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
    rxSummary( ~ ArrDelay + DayOfWeek, data = AirlineDataSet, reportProgress = 1)

### CrossTab the data
    rxCrossTabs(ArrDelay ~ DayOfWeek, data = AirlineDataSet, means = T)

### Linear model and plot
    hdfsXdfArrLateLinMod <- rxLinMod(ArrDelay ~ DayOfWeek + CRSDepTime, data = AirlineDataSet)
    plot(hdfsXdfArrLateLinMod$coefficients)
```

Hand on labs: https://github.com/Azure-Readiness/hol-azure-machine-learning

Microsoft Azure Machine Learning Studio



Operationalization made easy

Deploy to SQL Server

SQL Tools

Dashboards

Custom Apps

Reporting

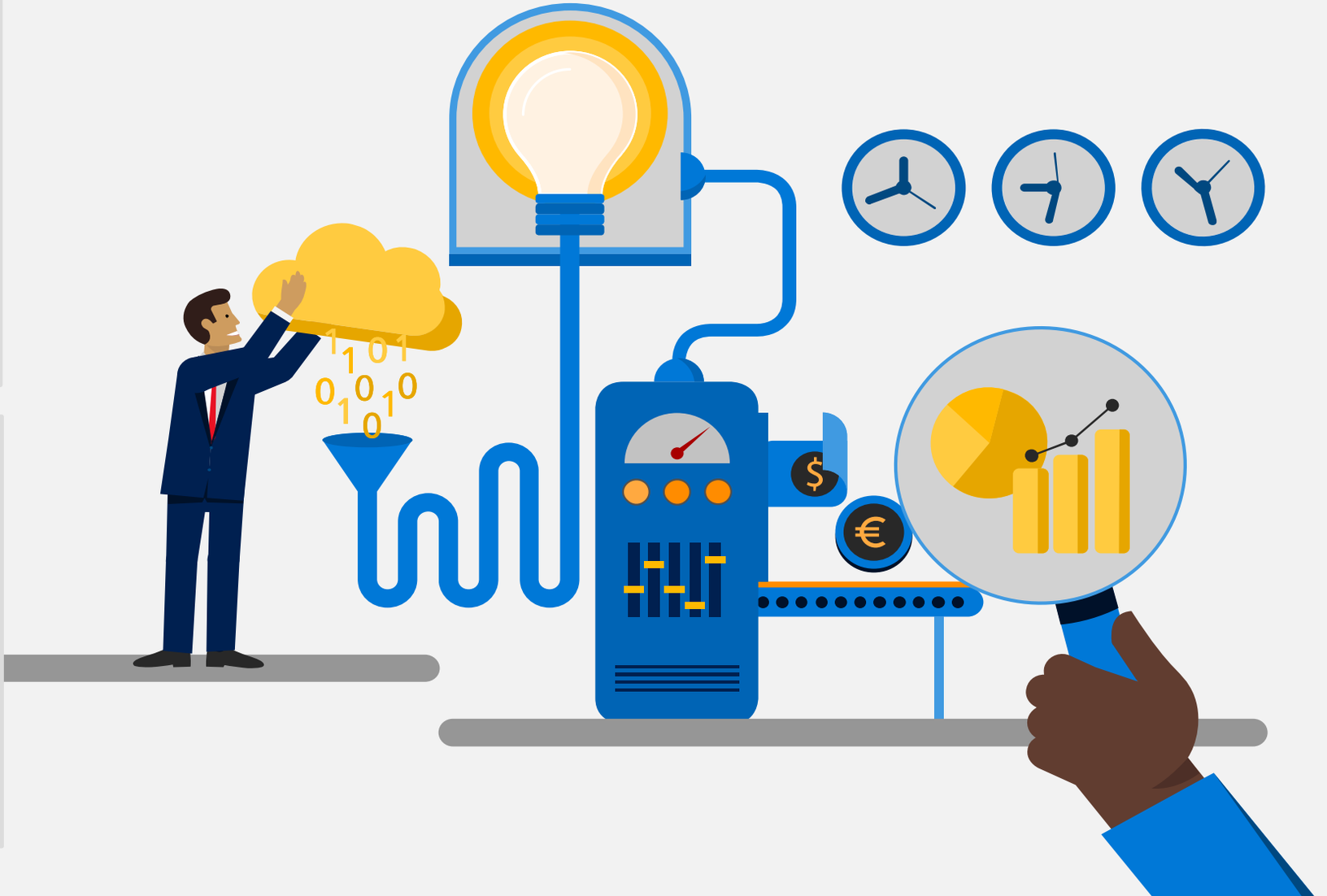
Deploy to Web Applications

Web Apps

Mobile Apps

Custom Apps

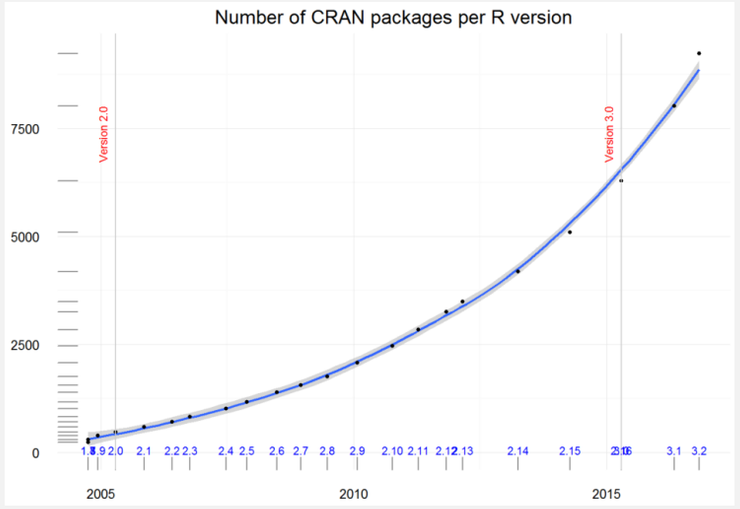
Data Viz Tools



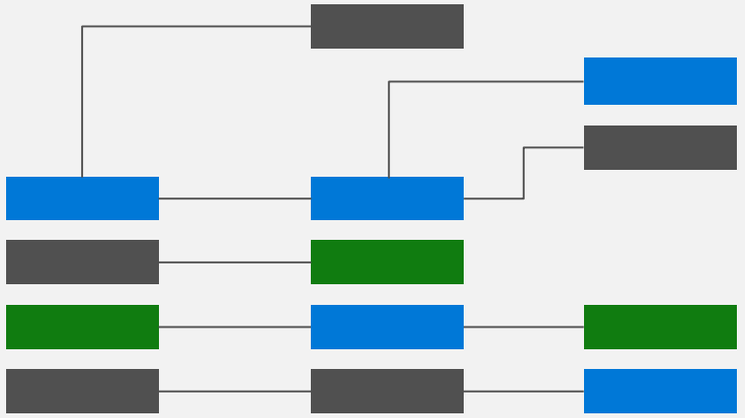
Best of Open Source and Microsoft Innovation

Best of Open Source and Microsoft innovation

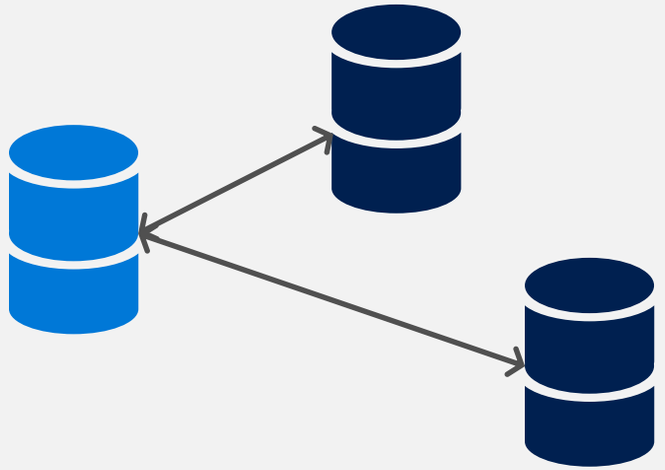
Community contribution to R packages are growing rapidly



Access Microsoft's machine learning algorithms and GPU enabled deep learning



Scale your analytics with parallelized remote executing functions



Parallelized, remote executing algorithms

Data step/ETL

Data import—delimited, fixed, SAS, SPSS, OBDC
Variable creation and transformation
Recode variables
Factor variables
Missing value handling
Sort, merge, split
Aggregate by category (means, sums)

Classification/ML

Decision forests
Gradient-boosted decision trees
Naïve Bayes
GPU-accelerated DNNs
Fast linear learner, with support for L1 and L2 regularization
Fast boosted decision tree
Fast random forest
Logistic regression, with support for L1 and L2 regularization
Binary classification using a One-Class Support Vector Machine

Simulation

Simulation (e.g., Monte Carlo)
Parallel random number generation

Statistical tests

Chi Square Test
Kendall Rank Correlation
Fisher's Exact Test
Student's t-Test

Descriptive statistics

Min/max, mean, median (approx.)
Quantiles (approx.)
Standard deviation
Variance
Correlation
Covariance
Sum of squares (cross-product matrix for set variables)
Pairwise cross tabs
Risk ratio and odds ratio
Cross-tabulation of data (standard tables and long form)
Marginal summaries of cross tabulations

Predictive Statistics

Sum of squares (cross-product matrix for set variables)
Multiple linear regression
Generalized linear models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie
Standard link functions: cauchit, identity, log, logit, probit
User defined distributions and link functions
Covariance and correlation matrices
Logistic regression
Classification and regression trees
Predictions/scoring for models
Residuals for all models

Sampling

Subsample (observations and variables)
Random sampling

Custom parallelization

PEMA-R API
rxDataStep
rxExec

Cluster analysis

K-Means

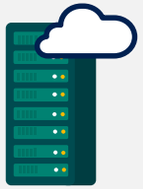
Solutions to accelerate time to value

Solution to accelerate time to value

Ready to deploy solution templates

‘Try Now’ interactive Power BI dashboards for each solution

Deploy the solution in a few clicks and get a complete walkthrough for your data scientists or analysts



Source code available to be extended and used by your team

Microsoft’s partner ecosystem to help you execute

Leverage Microsoft’s partner ecosystem with prior knowledge of building and migrating analytics solutions to implement your own solutions

Our partners

Machine Learning Server across your analytics platforms

In-database analytics with SQL Server 2016

Reduce data movement

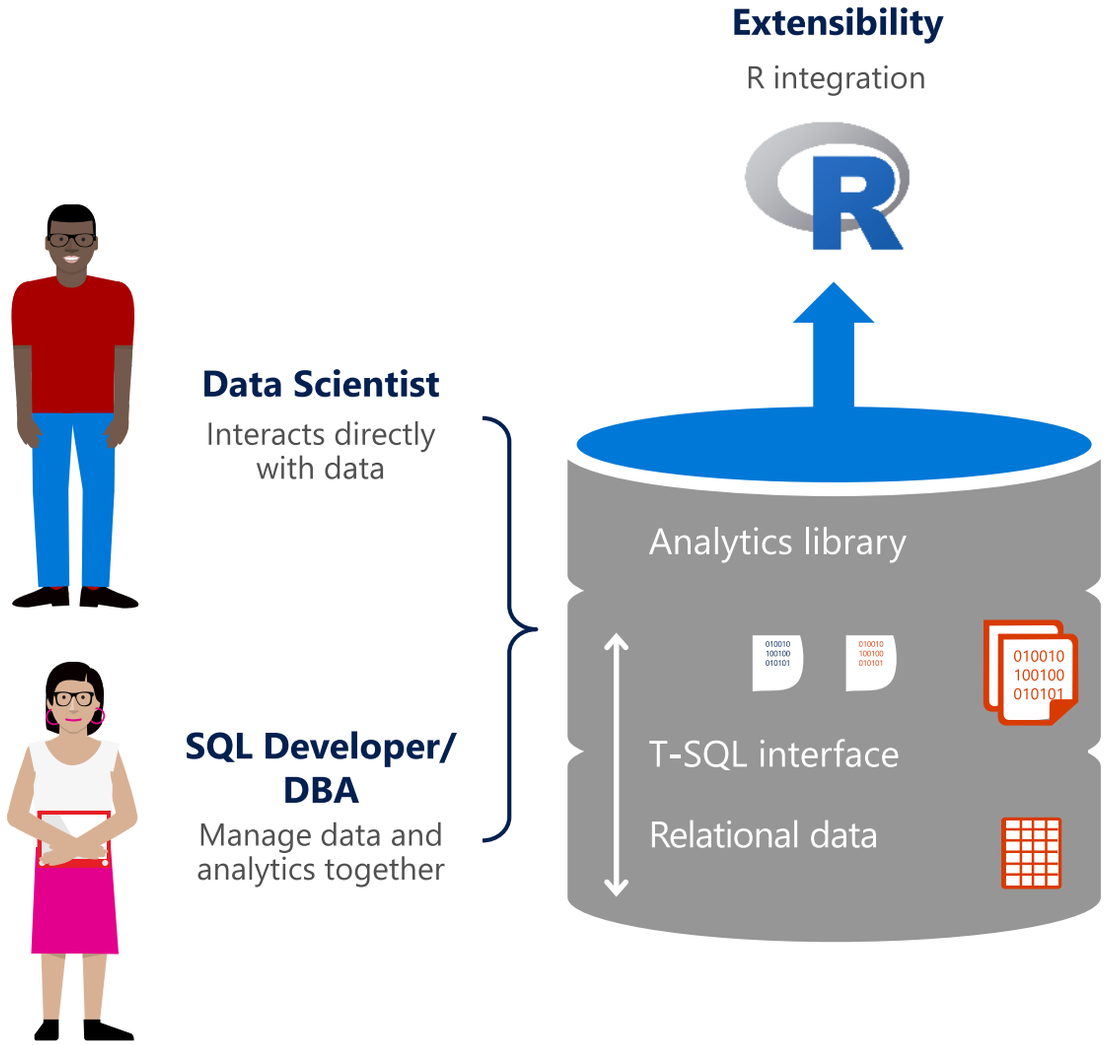
Eliminate data movement, reduce unnecessary duplication and leverage database data protections

Operationalize R scripts and models

Operationalize an R script/model over SQL Server data by calling familiar T-SQL stored procedures from your application

R with in-memory scalability

Scale your analytics with multi-threading and massive parallel processing



Big Data analytics with Microsoft R

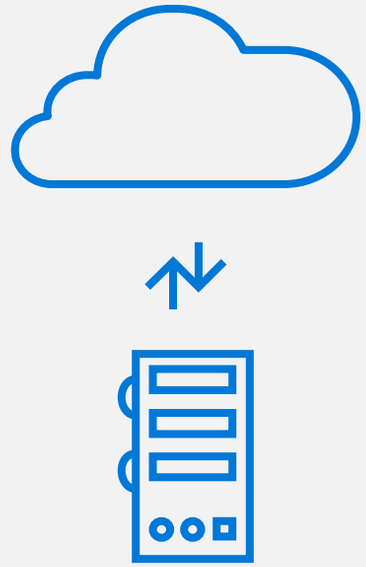
10x–50x faster analytics performance with parallelization and in-cluster execution

Work with data within the datalake

Reduce data movement and duplication

Rapid deployment in a managed cloud service. Use computation resources more efficiently

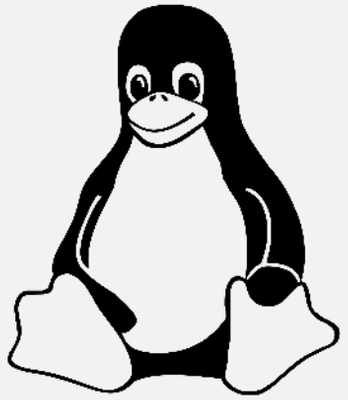
Machine Learning Server for HDInsight



Machine Learning Server for Hadoop/Spark

Microsoft R on Linux

Predictive and prescriptive analytics that makes the most of your open source investments



Microsoft R on Teradata

In-database analytics with Teradata

Bring advanced analytics to your data with Microsoft Machine Learning Server for Teradata databases

Reduce the cost of analytics

Run advanced analytics in-database on Teradata databases, rather than incurring the overhead of the traditional extract and analyze paradigm

Customer stories

Heartland Bank

New Zealand bank leads the way with innovative analytics platform based on Microsoft Machine Learning Server

New Zealand's rapidly growing Heartland Bank has more than NZD$2.34 billion (US$1.62 billion) in assets under management and provides services at 15 branches

To maintain its growth path while staying true to its customer-focused roots, the bank is replacing its existing SAS system with a Microsoft platform based on R Server and SQL Server 2016

- As a result, Heartland Bank has gained the performance, scalability, and ease of development it needs to deliver competitive solutions faster than ever
- Now, the bank can put advanced analytics at the fingertips of business users, enabling it to adapt to its rapidly changing business environment and customer needs

Products and services

Microsoft R Server
Microsoft SQL Server 2016
Microsoft Visual Studio

Organization size

Employees: 350
Assets under management: NZD$2.34 billion (US$1.62 billion)

Industry

Financial services—Banking

Country

New Zealand

Business need

Business critical

HEARTLAND BANK

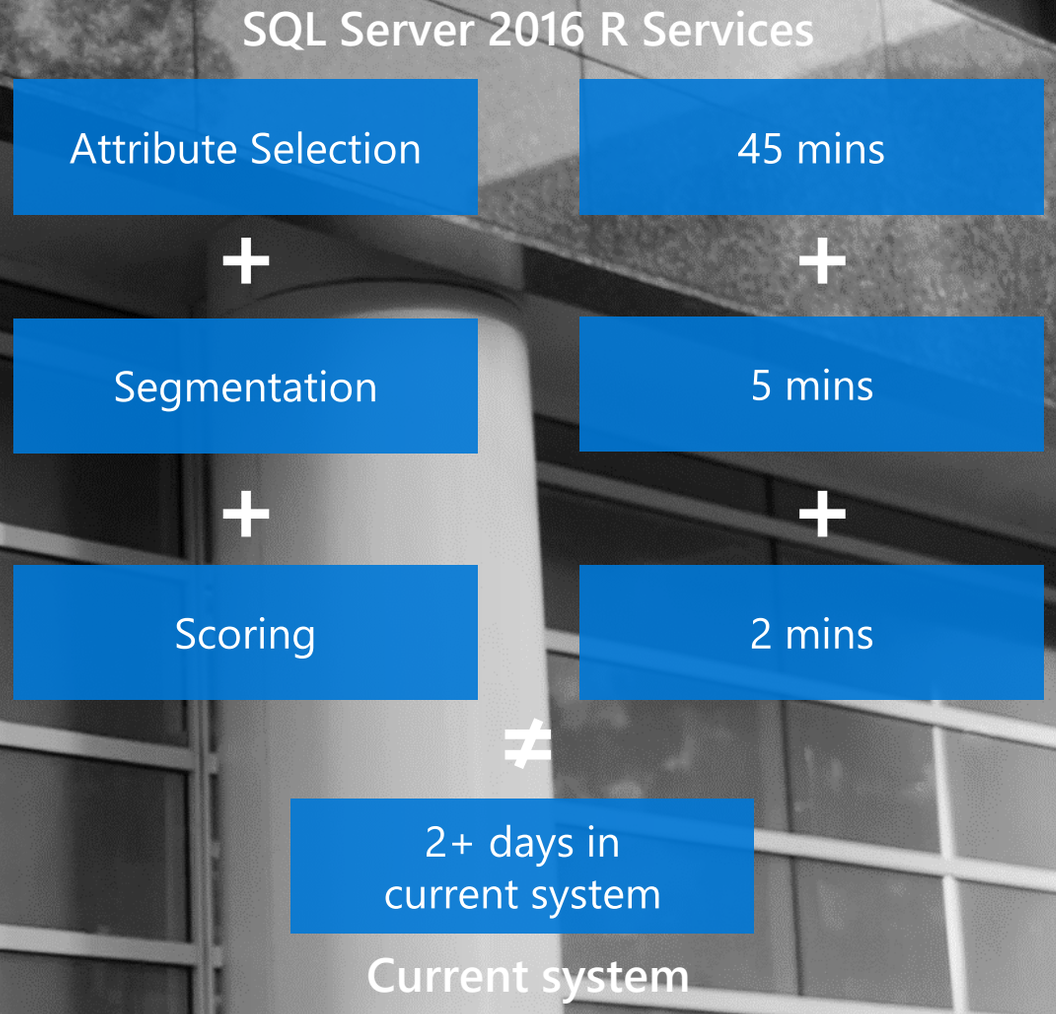
PROS

SQL Server with R Services improved performance by 100x!

PROS provides a real-time software solution platform to help companies drive pricing and sales effectiveness. Its software groups customers, products and transactions into micro-segments of similar willingness-to-pay.

We then apply optimization algorithms to target the pricing envelope 'sweet spot' in every segment. We used **SQL Server R Services** and leveraged its **'ScaleR'** libraries to drive our pricing segmentation as a test case.

What we saw was a 100x performance improvement that cut a process that took 2+ days in our current system cut down to 52 minutes.



Products and Services

Microsoft R Server
Microsoft SQL Server

Organization

Employees: 1000+
Annual Revenue: 366 Million

Industry

Sales and Pricing Software Solutions platform

Country

55+ Countries

PROS

Getting started with Microsoft R is easy

We created Microsoft R Client to help you get started and train your employees on Microsoft R

Spin Up a Machine Learning Server VM

Explore these Beginners Resources

Explore these Advanced Resources

Try our solutions with these tutorials or these ready to deploy templates


Azure Machine Learning Studio
Simple, scalable, cutting edge
Welcome to Machine Learning Studio, the Azure Machine Learning solution you've grown to love. Machine Learning Studio is a powerfully simple browser-based, visual drag-and-drop authoring environment where no coding is necessary. Go from idea to deployment in a matter of clicks.
Get started now

Microsoft