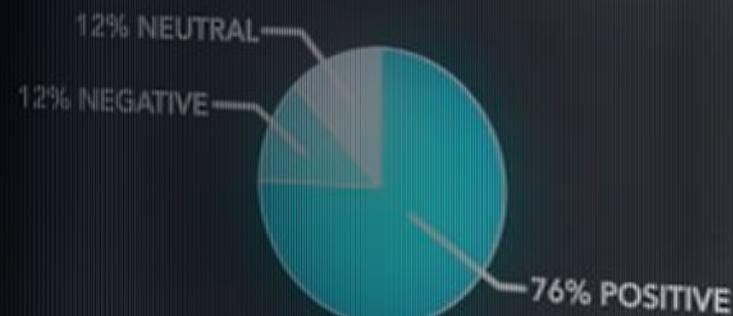


BEST SELLER: PECANS & CREAM

SOCIAL AFFINITY SEARCH



SENTIMENT ANALYSIS: BACON + PRALINES



Introduction to Microsoft Advanced Analytics

Ali Zaidi

Data Scientist, Machine Learning and Data Science Education Team



Agenda

● Trends

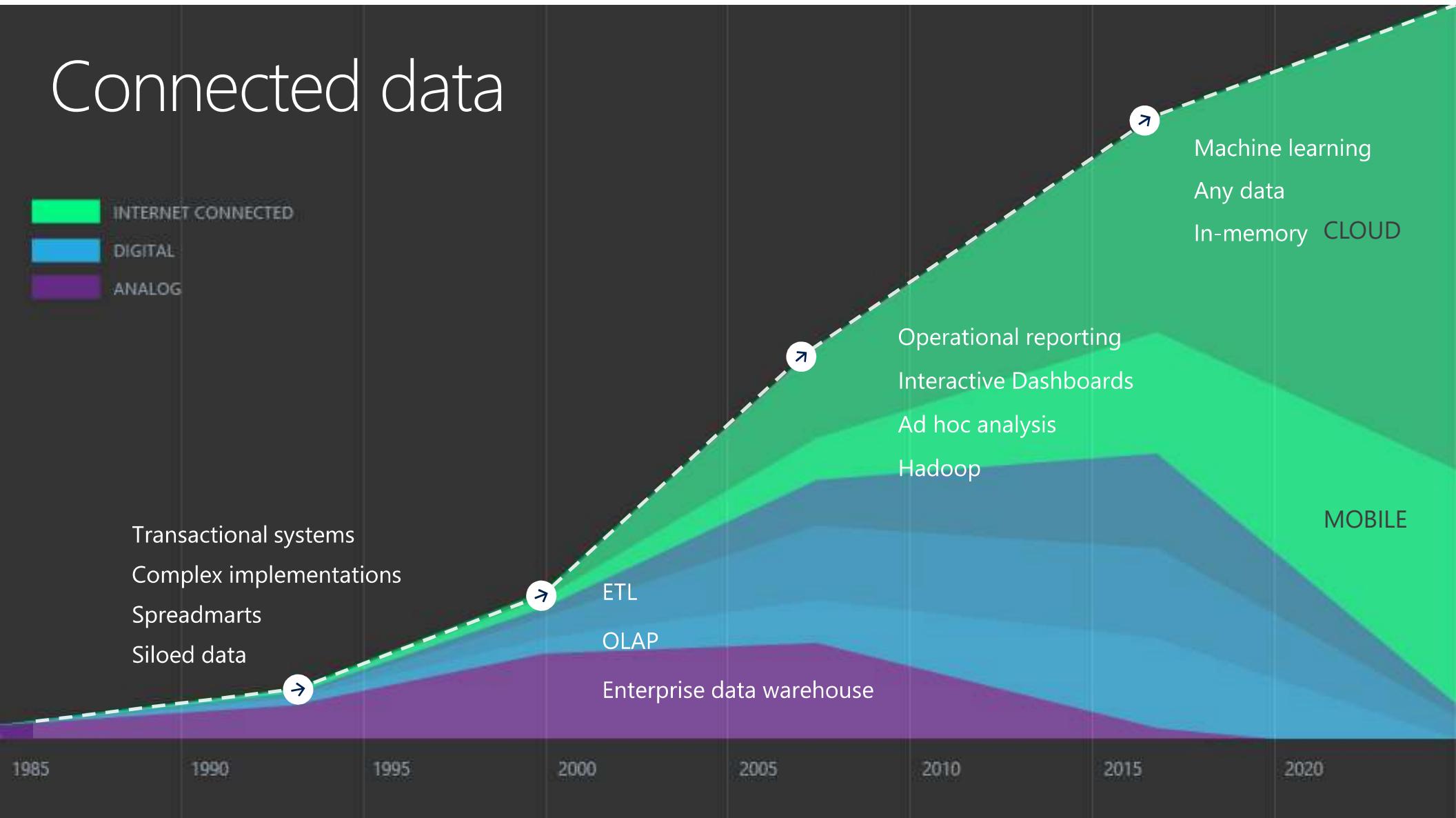
● Opportunities

● Solutions

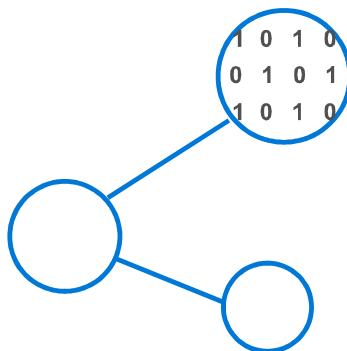
● Directions



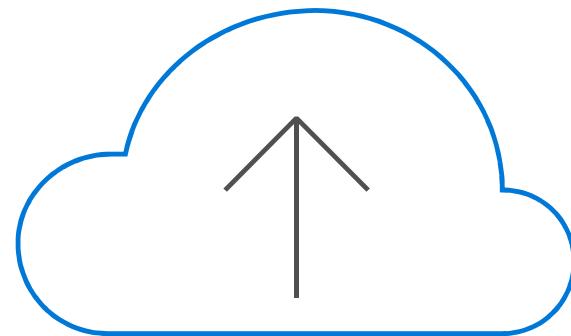
Connected data



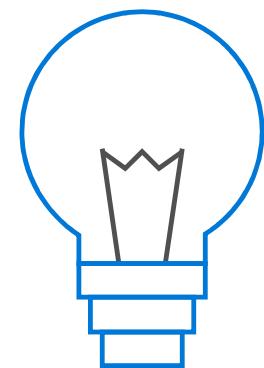
Three major trends converging



Data

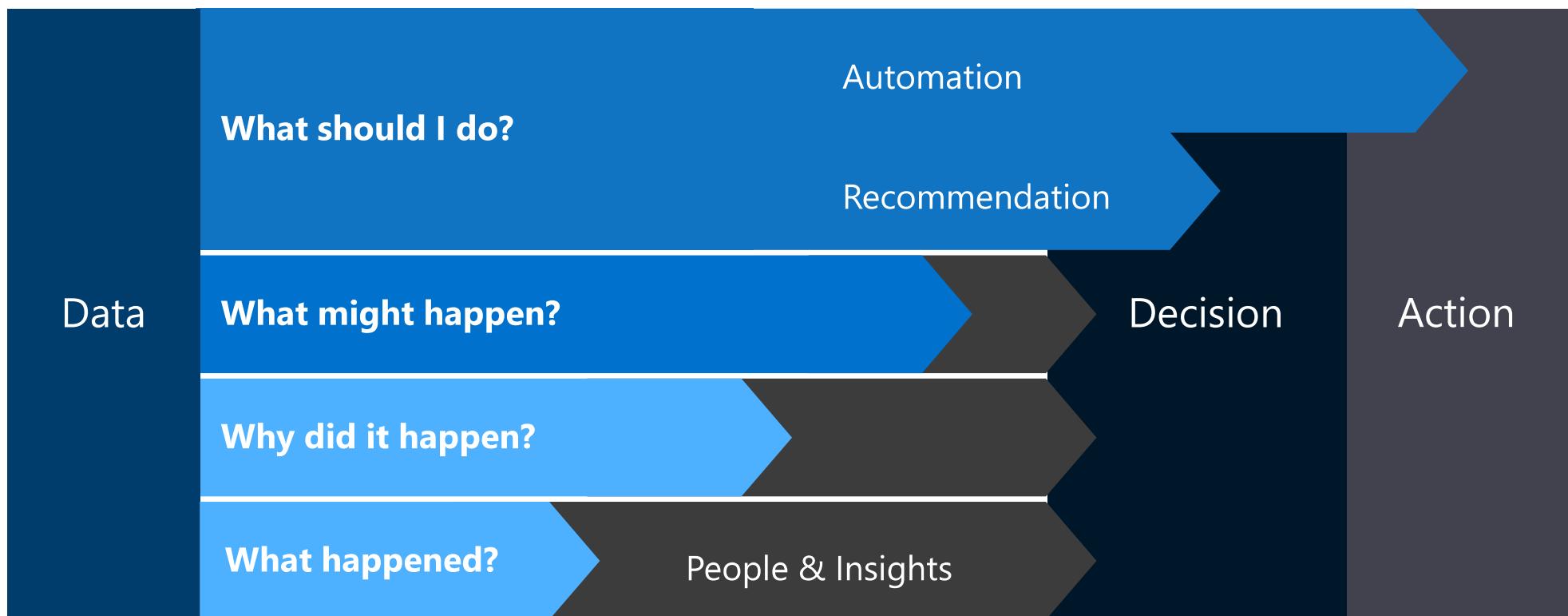


Cloud

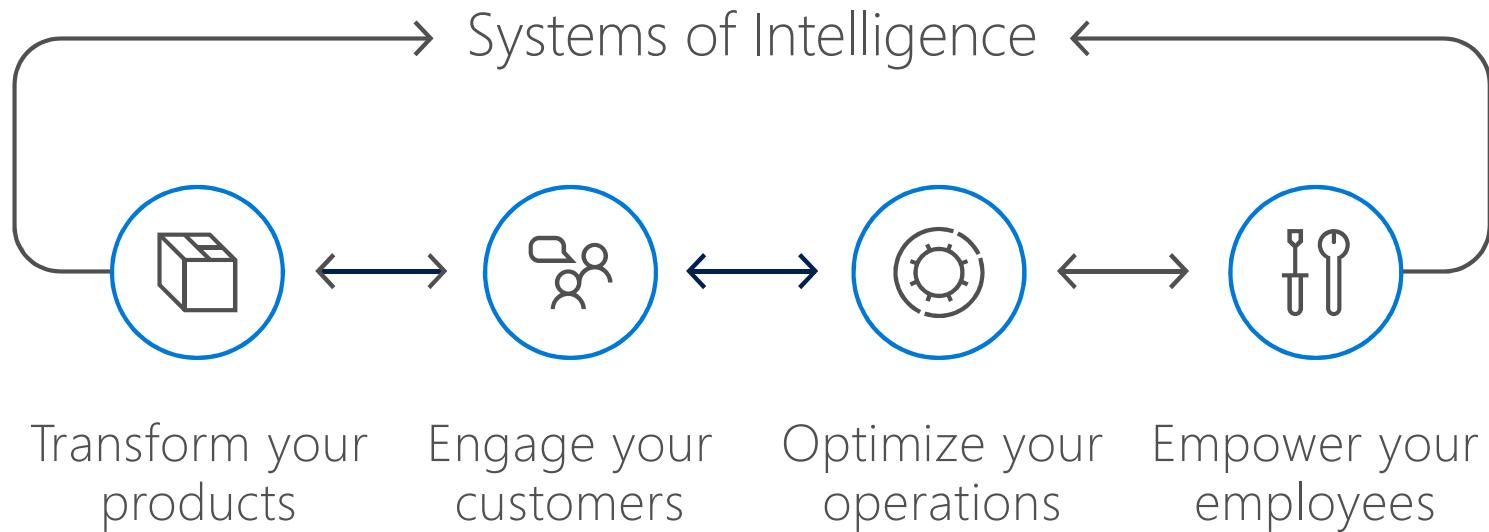


Intelligence

Big Data + Predictive Analytics = Business Value



Transforming key aspects of business



Data is a key strategic asset

\$1.6T

Additional business value captured by companies that are leaders in using data assets to their advantage

Source: IDC, 2014

10%

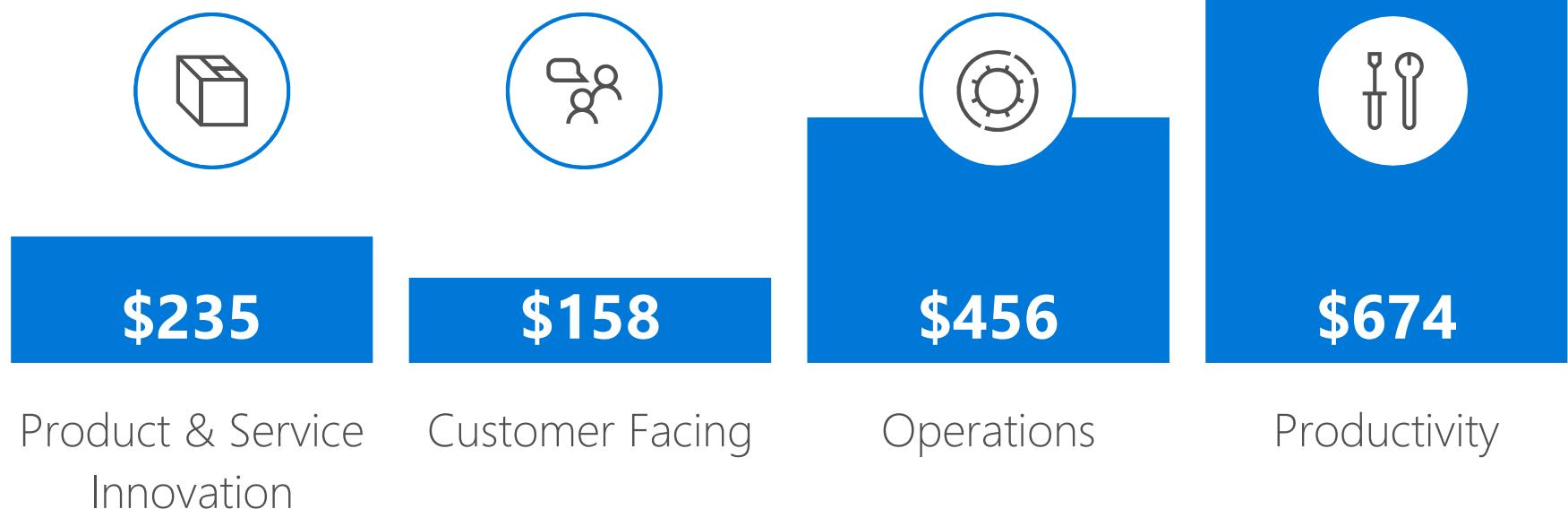
Percent of organizations expected to have a highly profitable business unit specifically for productizing and commercializing their data by 2020

Source: Gartner, 2016

Capitalizing on a 1.6 trillion \$ data dividend

Data Dividend

Incremental Gains Made by Leaders in Data and Analytics



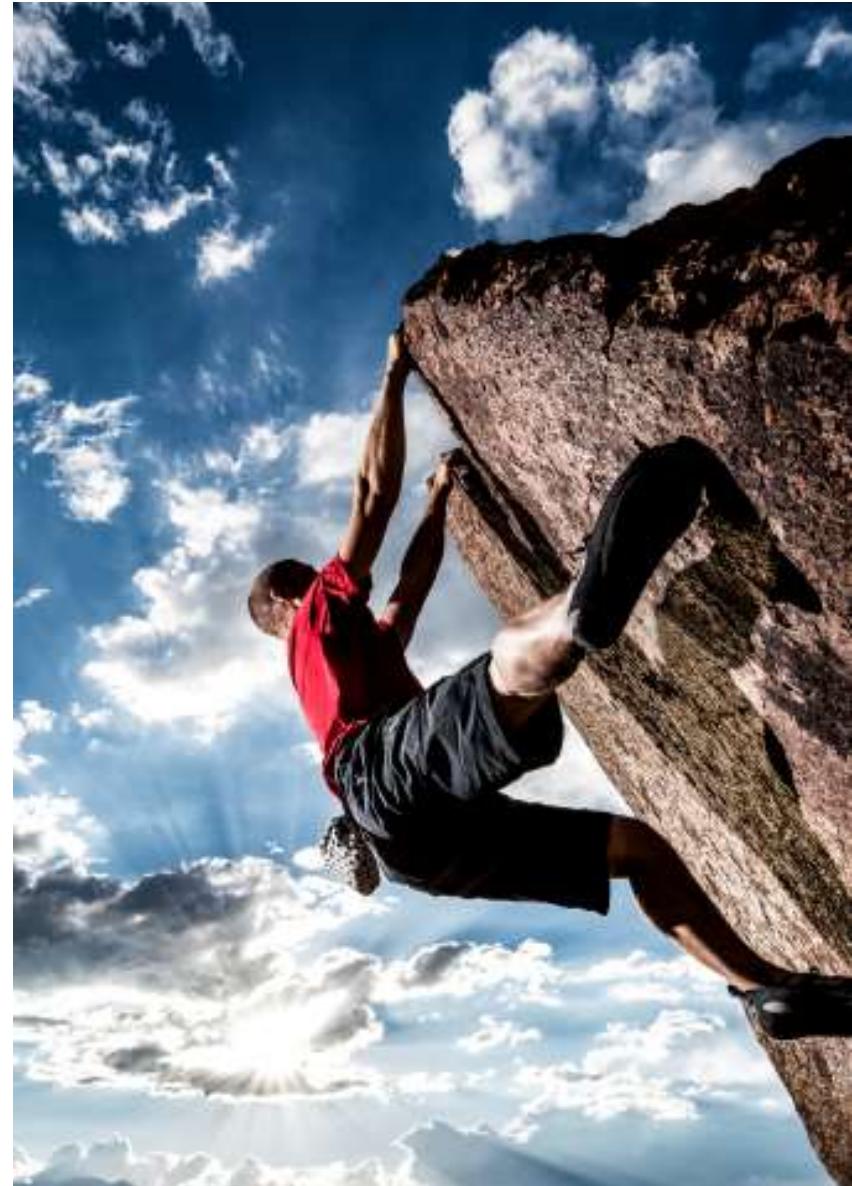
IDC Data Dividend Study and Survey, N=2,020, April 2014

Advanced Analytics scenarios

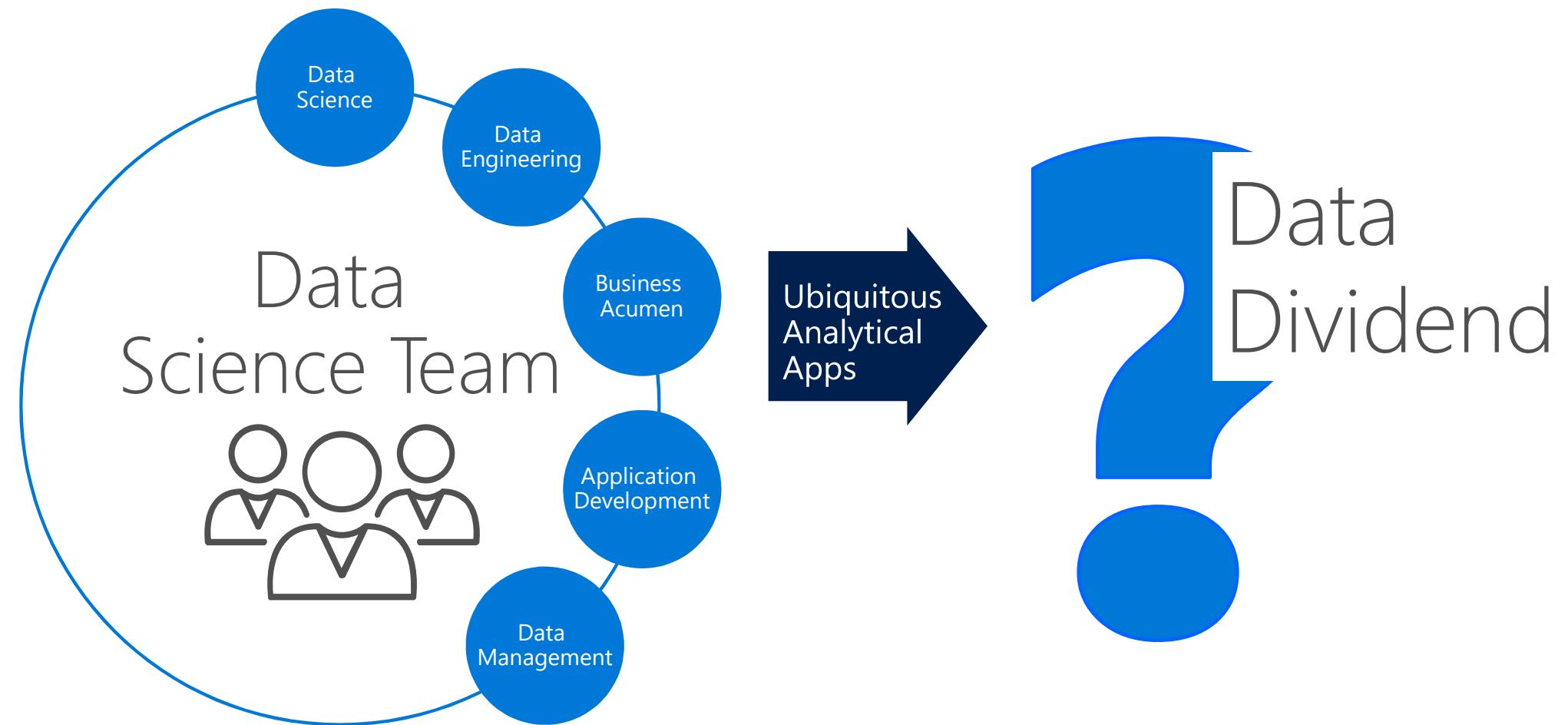
EXAMPLE SOLUTIONS				
Sales and marketing	Finance and risk	Customer and channel	Operations and workforce	
 Customer Acquisition	 Fraud detection	 Lifetime customer value	 Remote Monitoring	
 Cross-sell and upsell	 Credit risk management	 Personalized offers	 Operational efficiency	
 Loyalty programs		 Product recommendation	 Smart buildings	
 Marketing mix optimization		 Customer Service improvement	 Predictive maintenance	
 Demand forecasting			 Supply chain optimization	

Industries applying advanced analytics

-  Retail & Consumer Products
-  Financial Services & Insurance
-  Government
-  Manufacturing
-  Healthcare

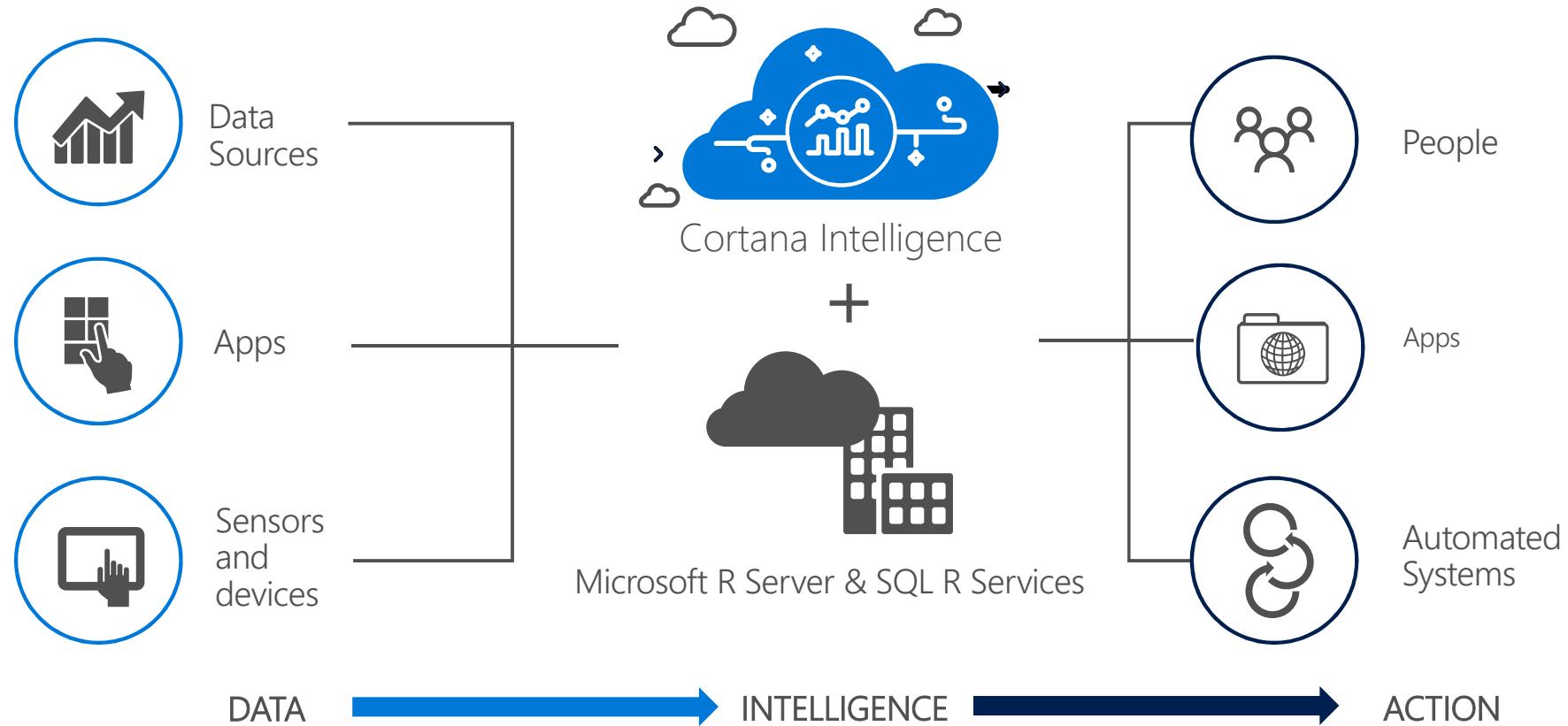


Success requires convergence of skills

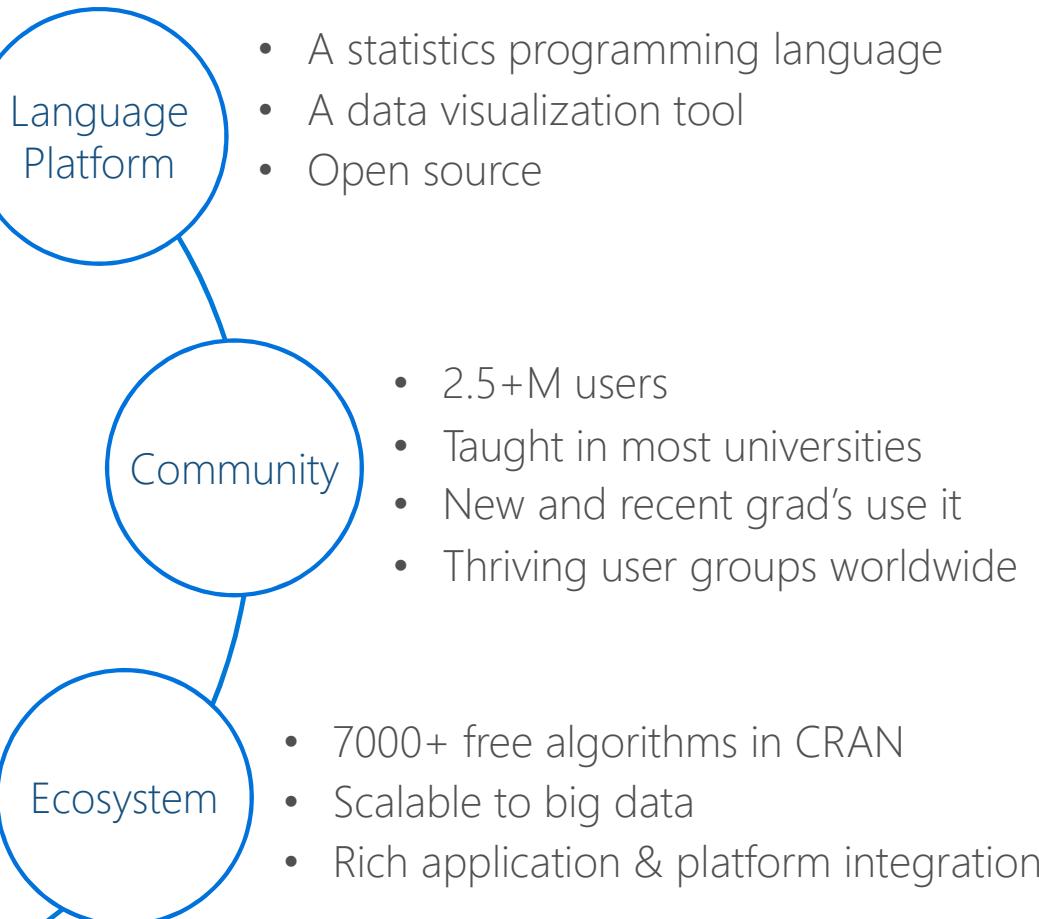


Microsoft R Server family

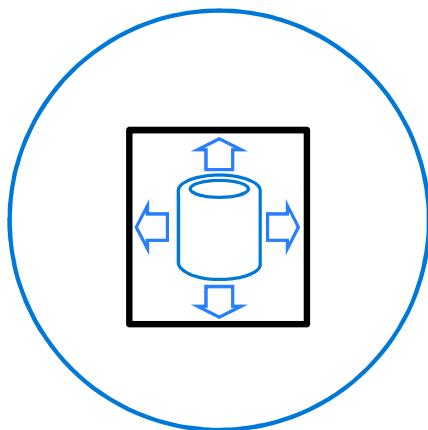
From Data To Action On Premises



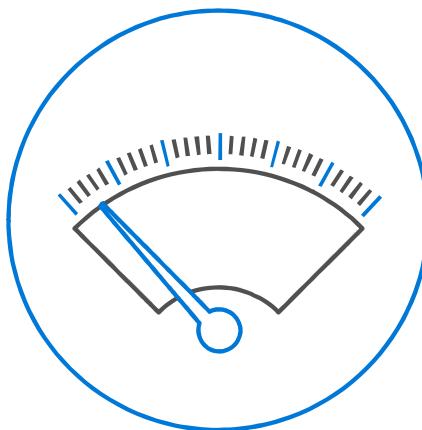
What is



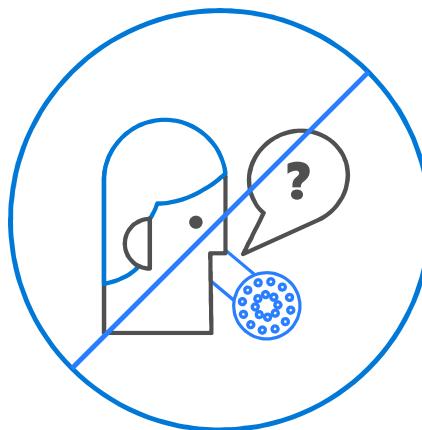
Challenges posed by open source R



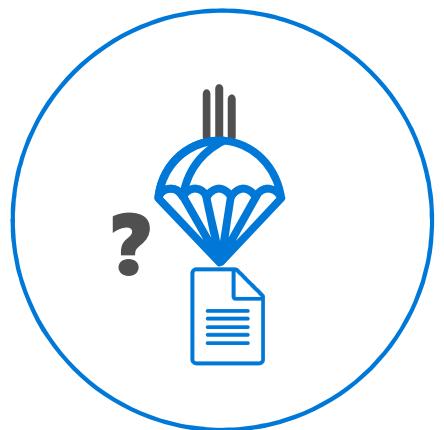
Limited
Data
Scale



Inadequate
Modeling
Performance

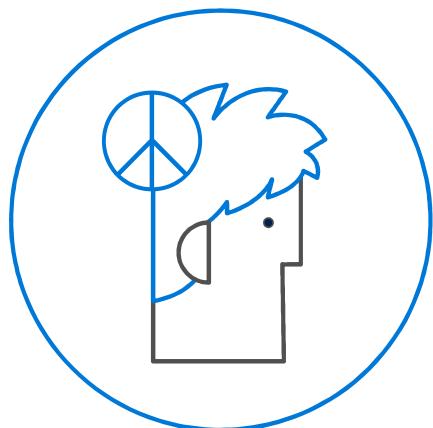


Lack of
Commercial
Support

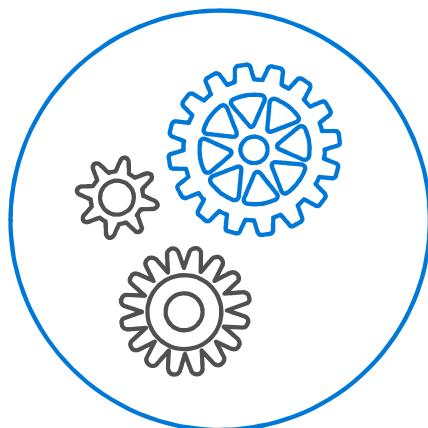


Complex
Deployment
Processes

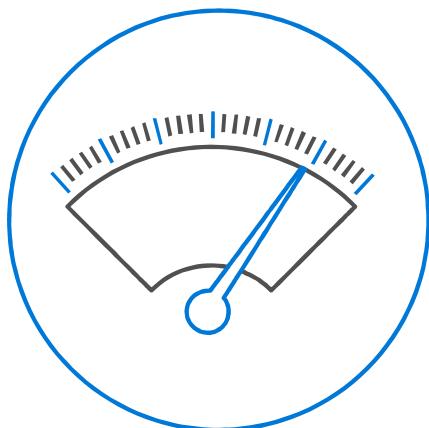
R from Microsoft brings



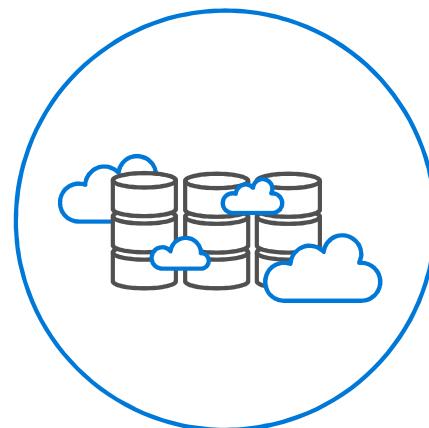
Peace of
mind



Efficiency

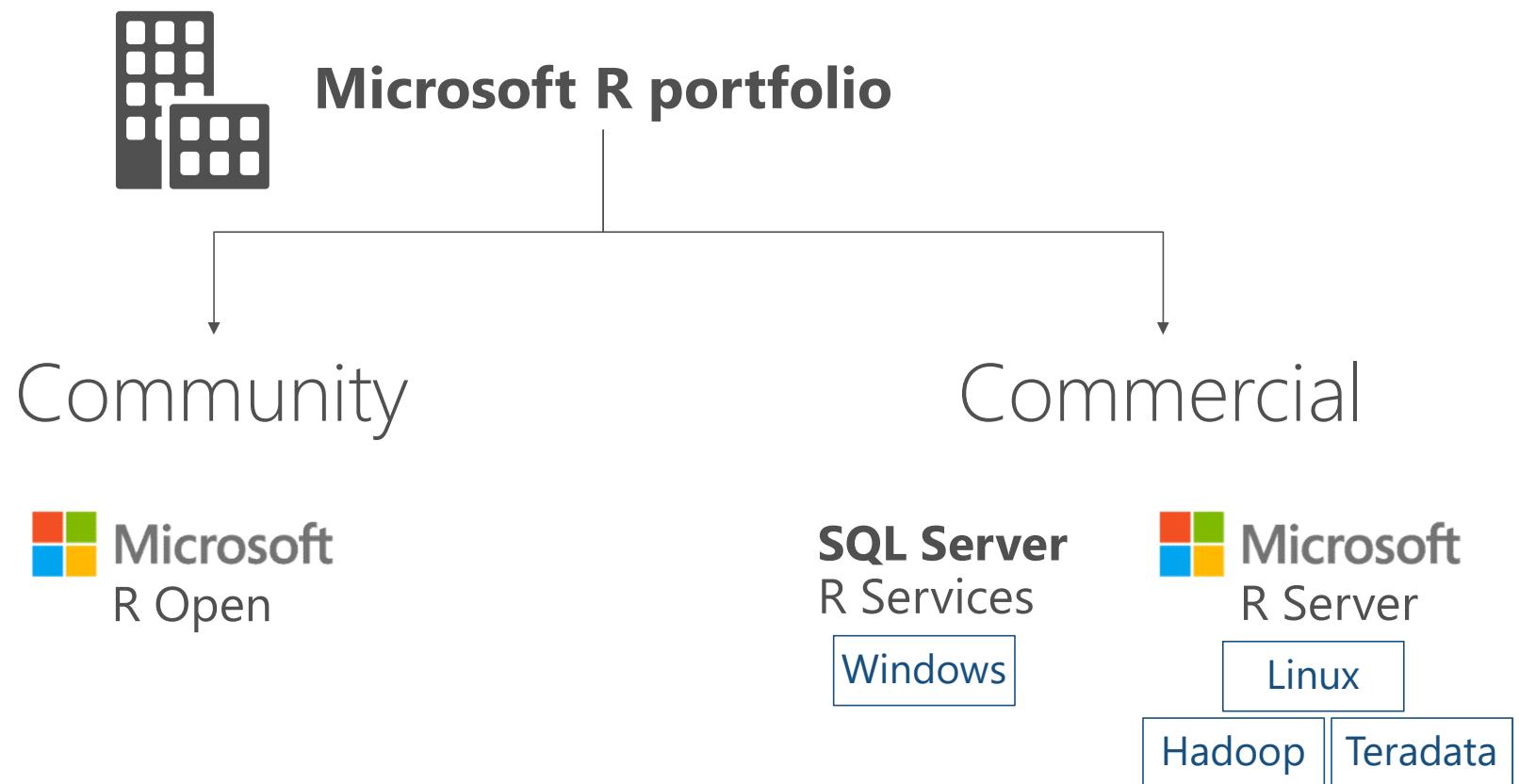


Speed and
scalability



Flexibility
and agility

Microsoft R portfolio



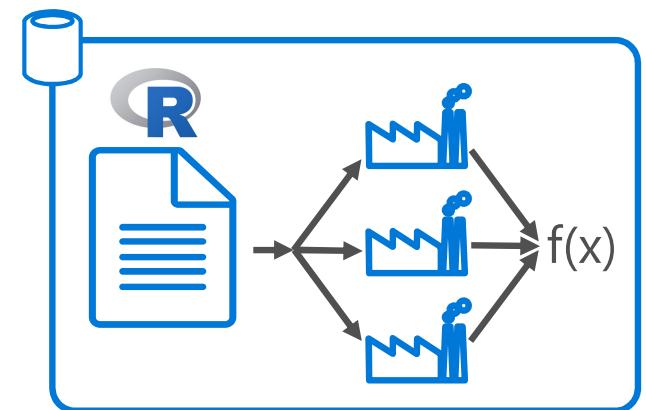
Microsoft R Scales to Big Data for Enterprises

Escapes R's traditional memory limits

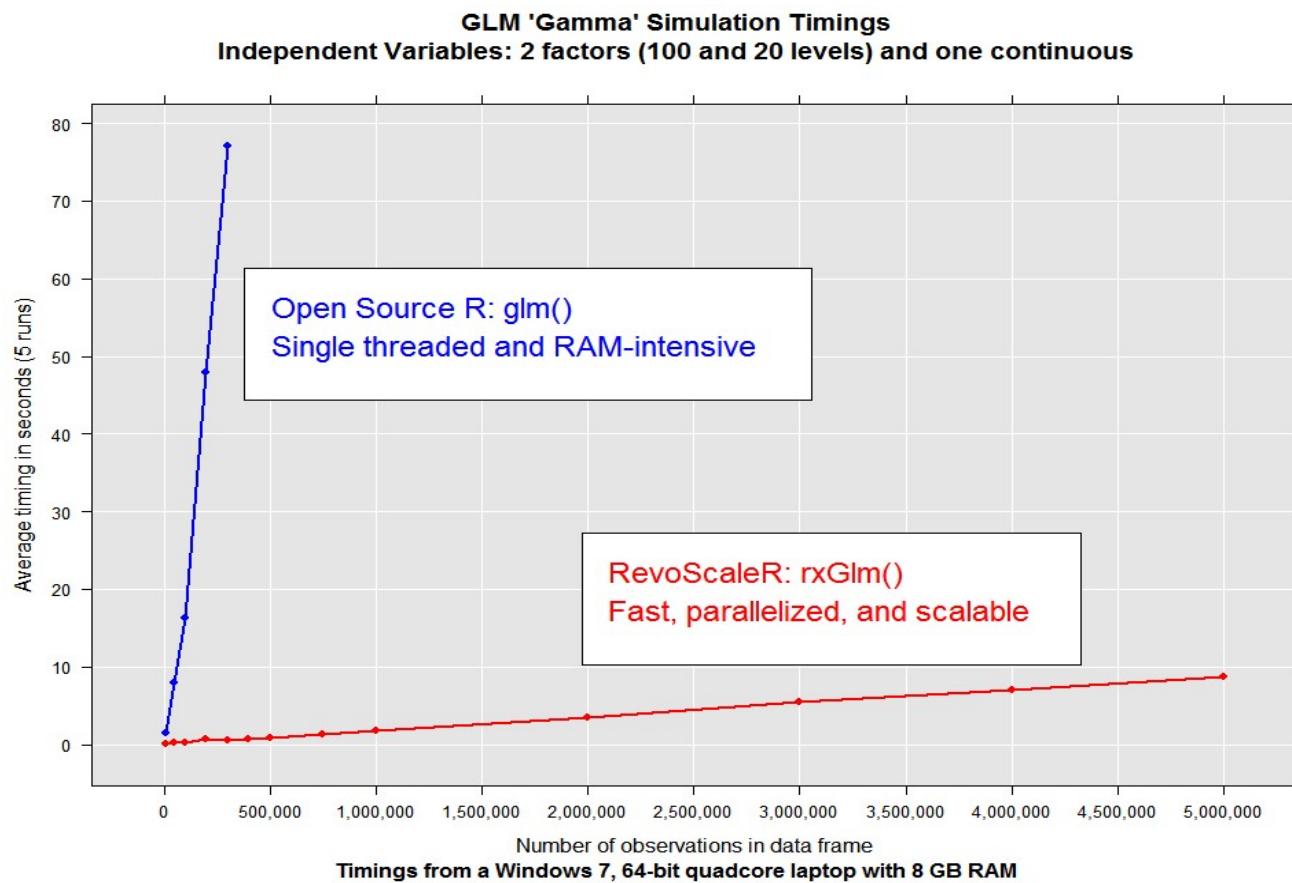
Scales predictive modeling using parallelization

Distributes computation cores & nodes

Minimizes data movement using in-database, in-MapReduce and in-Apache Spark execution



Scalable algorithms



Introducing Microsoft R Server

Linux, Windows, Hadoop & Teradata

High-performance, Scalable R

100% open source R

CRAN, Bioconductor, MRAN, GitHub compatibility

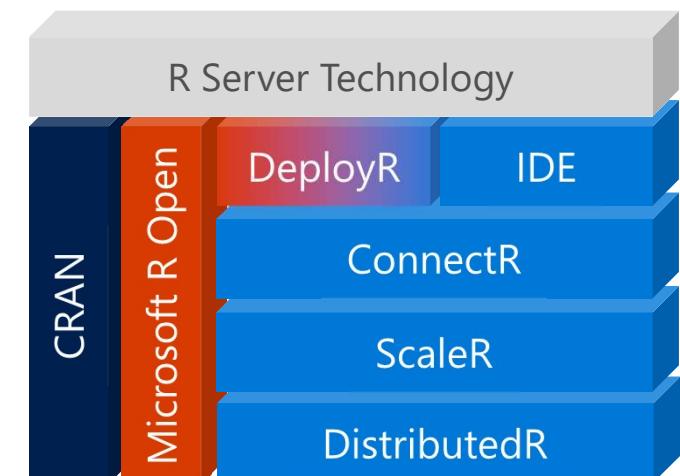
Big-data connectivity

Scalable analytics

Multi-platform

In-database, in-cluster scalability

Choice of IDE



Introducing SQL Server 2016 R services



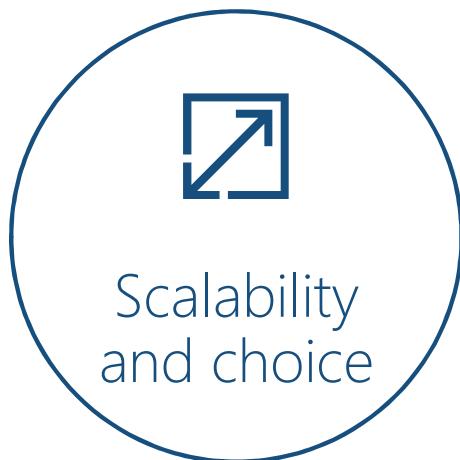
Simplicity
and agility

Enterprise speed and
scale

Near-DB analytics

Parallel threading and
processing

Reuse SQL skills for data
engineering



Scalability
and choice

In-database deployment

Memory and disk
scalability

No R memory limits
Write once, deploy
anywhere



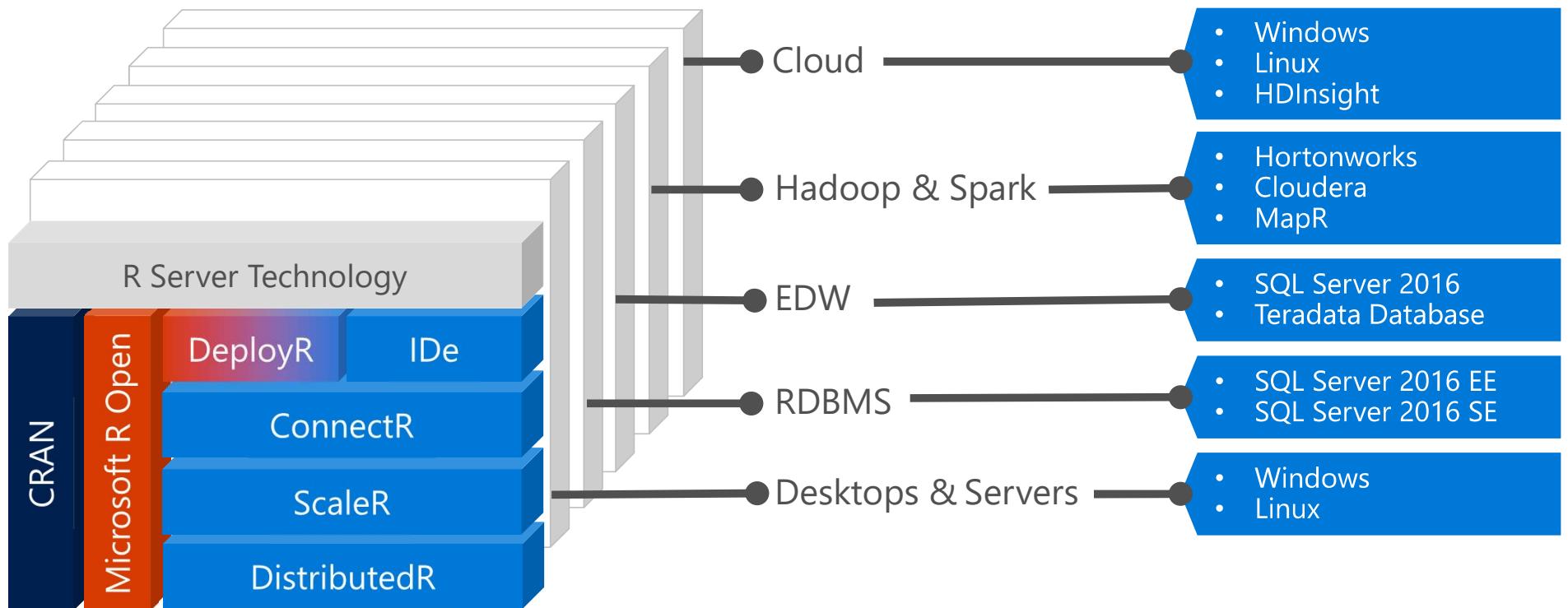
Cost
effectiveness

Included in SQL Server
2016

Reuse and optimize
existing R code

Eliminate data movement

Portability & investment assurance



Write Once – Deploy Anywhere

Microsoft R Server delivers

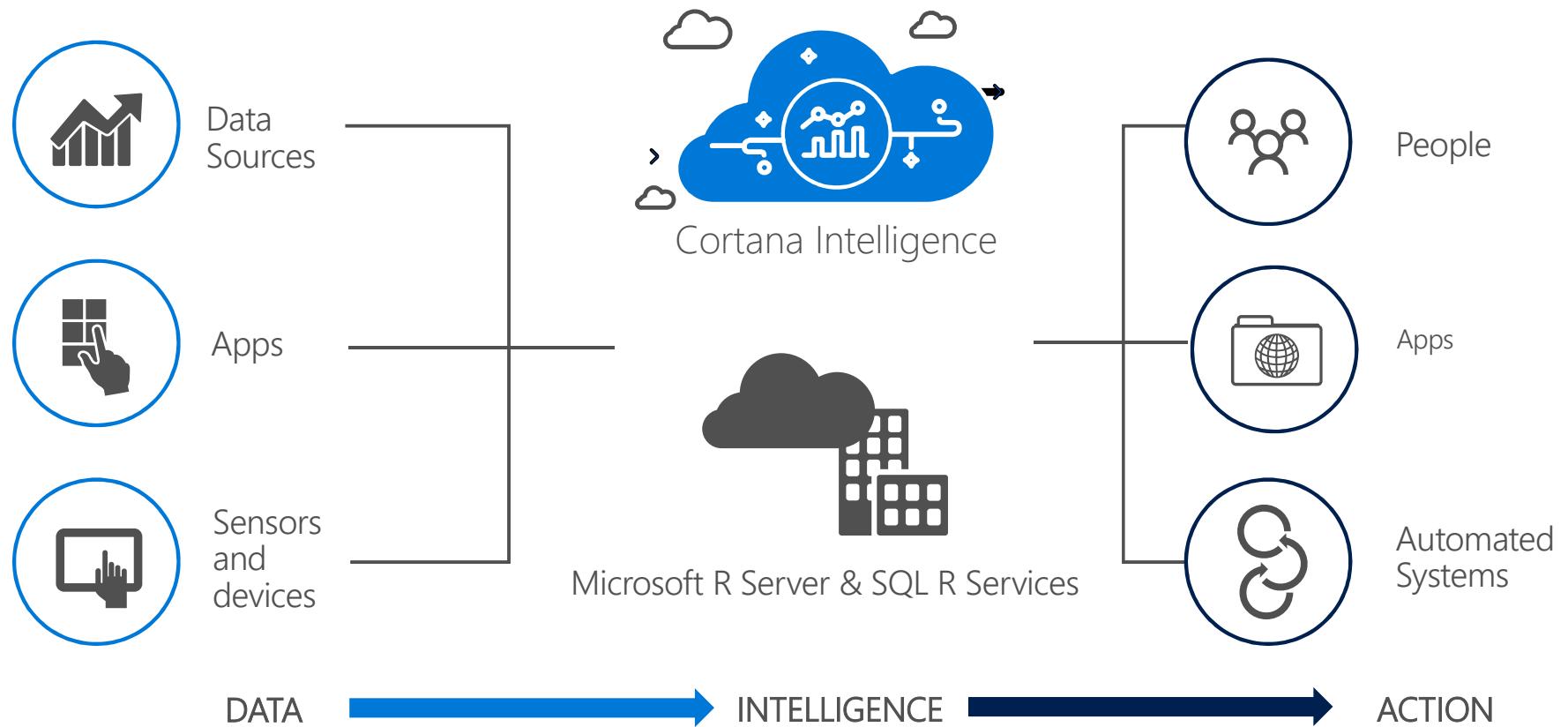
- The industry's broadest R-based platform
- Enterprise scale atop spark, Hadoop, RDBMSs & EDWs
- Freedom from memory limits
- Choice of Windows and Linux IDEs
- Stable deployment
- Write-once-deploy-anywhere portability
- Investment protection
- Hybrid cloud evolution

Microsoft R Roadmap

MICROSOFT CONFIDENTIAL
Subject to Change Without Notice

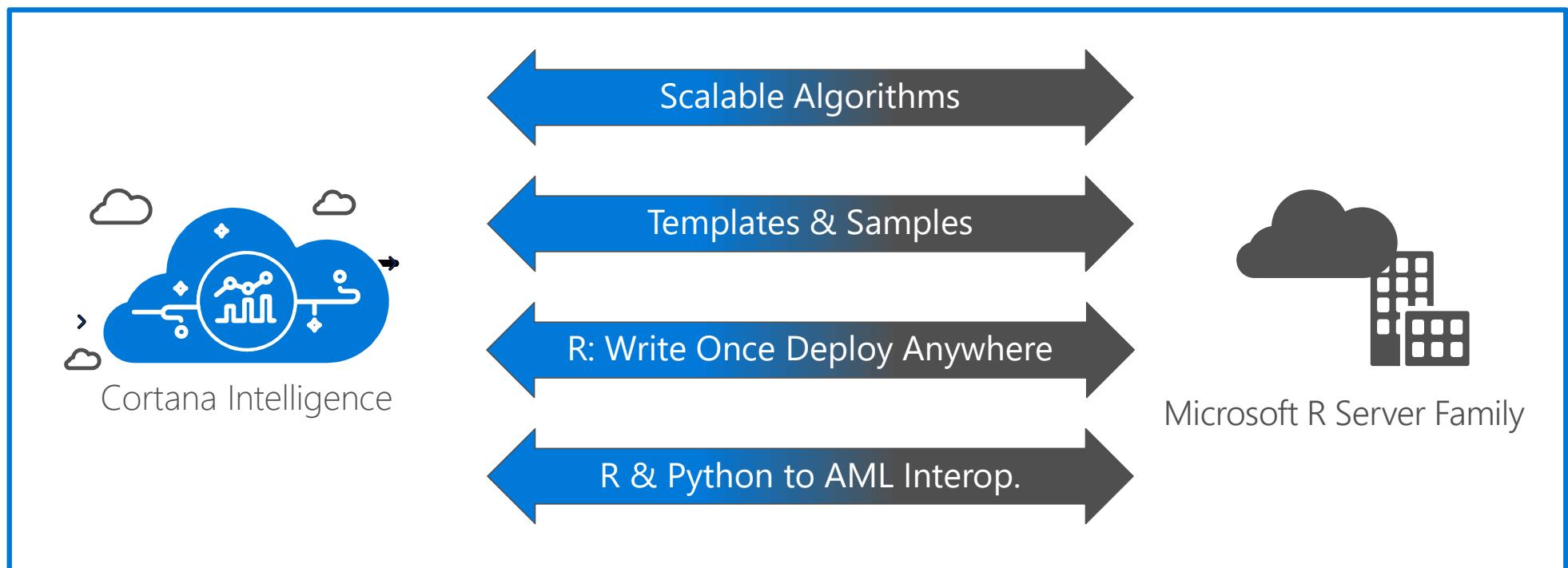
FY16	Q1	Q2	Q3	Q4
On Premises	<ul style="list-style-type: none"> Remediated RRE (v 7.4.1) for: <ul style="list-style-type: none"> Windows RedHat and SuSE Linux Hadoop (HortonWorks, Cloudera, MapR) on RedHat Teradata EDW HPC Pack 2008/2012+ (last feature release) 	<ul style="list-style-type: none"> SQL Server R Services (CTP3) New RRE features (All Platforms) <ul style="list-style-type: none"> ODBC Write Fuzzy Matching Microsoft R Open 3.2.3 	<ul style="list-style-type: none"> Microsoft R Server SKUs on VL Microsoft R Server Developer Edition in VS Dev Essentials R Tools for Visual Studio (Preview) 	<ul style="list-style-type: none"> SQL Server 2016 GA <ul style="list-style-type: none"> R Services (In database) R Server (Standalone) Microsoft R Open 3.2.4, 3.2.5, 3.3.0 Microsoft R Server for Hadoop: Install Experience++, Spark (GA), SUSE, Hadoop platform upgrades Microsoft R Server for Teradata (GA) Microsoft R Server for Red Hat Linux (GA); 7.1 Microsoft R Server for SUSE Linux (GA) DeployR: DB choices, security fixes, new install exp. New ML Algorithms for Windows (TAP) Microsoft R Client (Windows) R Tools for Visual Studio (GA)
Cloud	<ul style="list-style-type: none"> RRE VMs on Azure – Preview (Windows, CentOS) 	<ul style="list-style-type: none"> RRE VMs on Azure w/Datalake Access (Preview) (Windows, CentOS) RRE on Azure HDI w/Datalake Access (Preview) (Windows, Ubuntu) 	<ul style="list-style-type: none"> Azure VM for Microsoft R Server on Windows, Linux (GA) Azure Data Science VM for Windows (GA) Microsoft R Server for HDInsight, with Spark Support (Preview) Azure ML Jupyter R notebooks (Preview) 	<ul style="list-style-type: none"> Azure HDInsight Premium with Microsoft R Server for Hadoop, Spark Support (GA) Azure ML: <ul style="list-style-type: none"> Jupyter R notebooks (GA)

From data to decisions to action with Microsoft



Hybrid analytics platform

Convergence with Flexibility



Microsoft Advanced Analytics address barriers

Broaden The Talent Pool	<ul style="list-style-type: none">• Democratize Data Science• Skill Re-Use
Increase Productivity	<ul style="list-style-type: none">• Transparent Scaling• Facilitate Collaboration
Modernize Infrastructure	<ul style="list-style-type: none">• Decouple Data Science from Platforms• Leverage Hybrid Cloud Architecture
Maximize Innovation	<ul style="list-style-type: none">• Accelerate Experimentation• Streamline Deployment
Drive Down TCO	<ul style="list-style-type: none">• Embrace Open Source• Evolutionary Path to Cloud

Microsoft's roadmap for analytics

In 3 Years, we will help you achieve:

- Analytics-driven decision making
- Accelerated analytics lifecycle
- Dramatically lower analytics TCO
- Innovative uses of machine learning
- Continuity across cloud / on-prem environments





© 2016 Microsoft Corporation. All rights reserved.

Retail & CPG



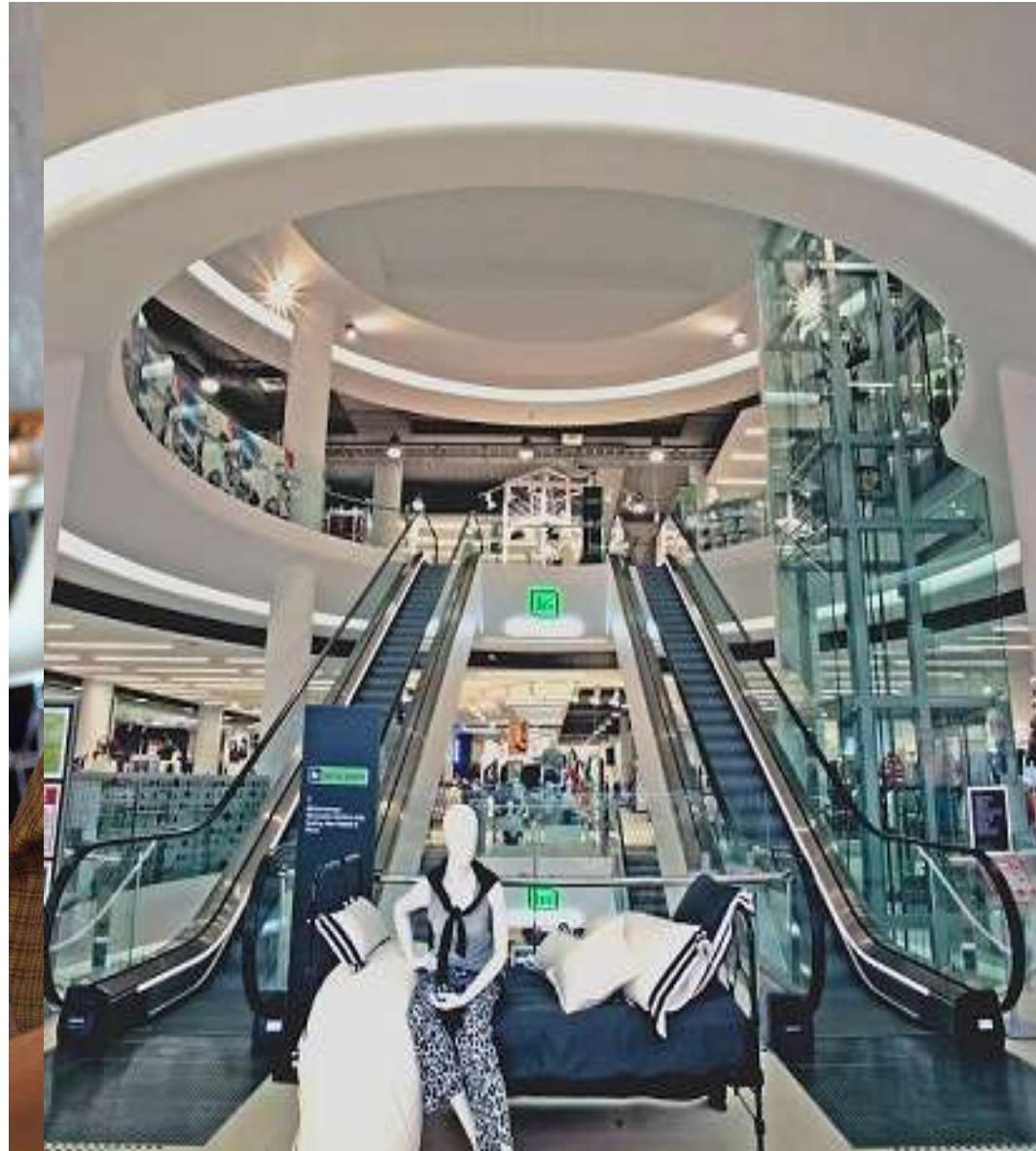
Execution Optimization



Shrinkage & Loss
Prevention



Customer 360 Experience



Manufacturing



Predictive Maintenance



Supply Chain Optimization



Quality



Financial Services



Risk Analytics



Fraud Prevention



Customer Experience



Government



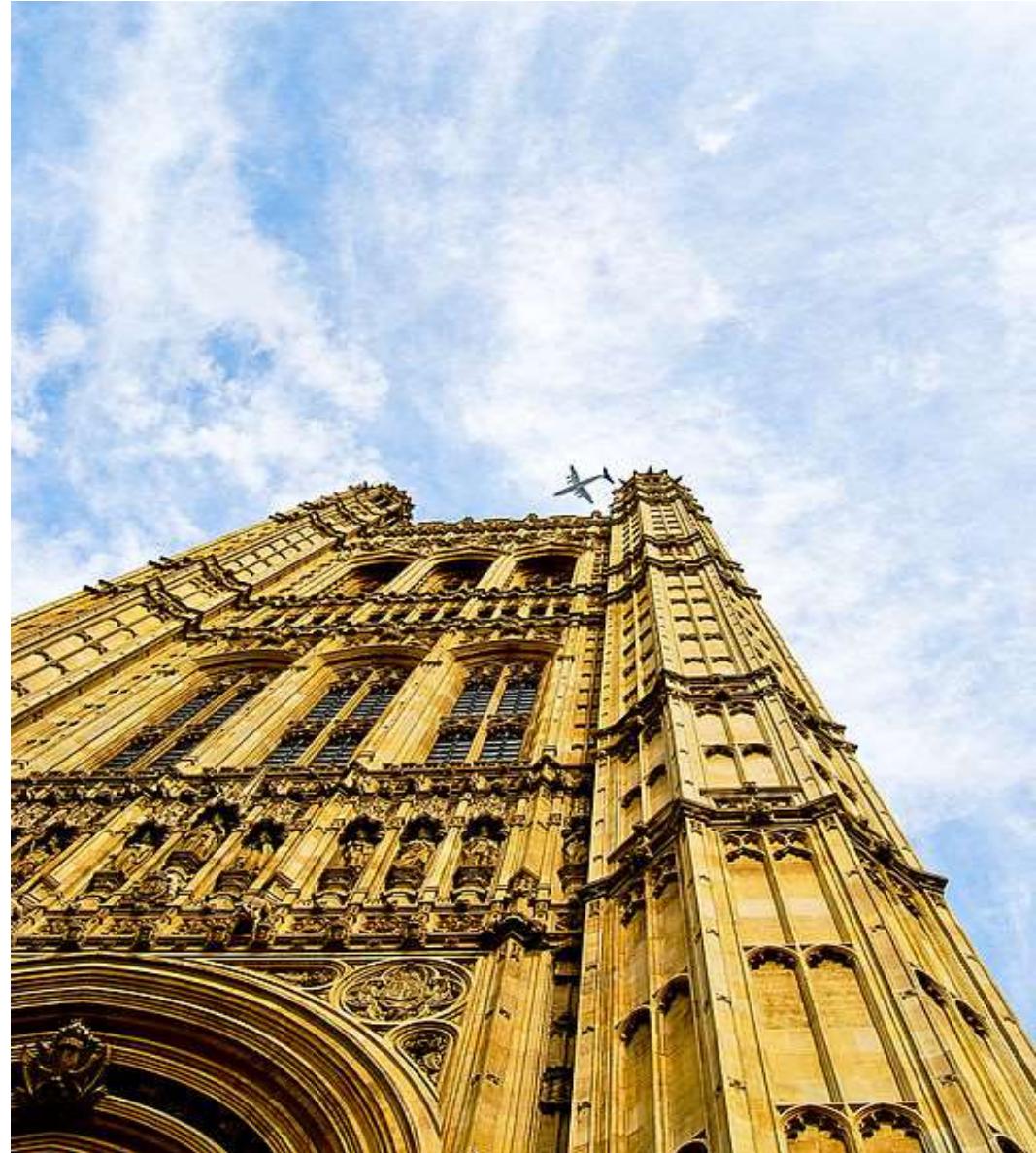
Public Health



Transportation Optimization



Revenue Assurance



HealthCare



Fraud Prevention



Care Optimization



Bioinformatics



Seizing the Opportunity

Supply Chain



Background

- Integrichain is a cloud-based Supply Chain Analytics Company disrupting the pharmaceuticals industry with faster integrated forecasting and fulfillment



Problem

- Massive computational burdens associated with an expanding product and customer base threaten to impede IntegriChain's ability to serve their customers



Solution

- Ingest 15 separate data sources into Hadoop data storage.
- Run 6.5 Billion predictions to predict order plans for 25,000 locations and 50 SKUs using 6 models
- Complete in < 4 hours permitting nightly delivery
- Drive innovation at a cost not attainable before R Server Hadoop

" The amount of analytic horsepower required for this application cannot be supported in traditional means; it would require millions of dollars of hardware. R + Hadoop is allowing us to have the compute capacity to run 6.5 billion computations on nightly basis to generate order plans for our clients."

VP of Application Development

Better Forecasting. Accurate Delivery. Customer Satisfaction

Transformational trends

data
explosion

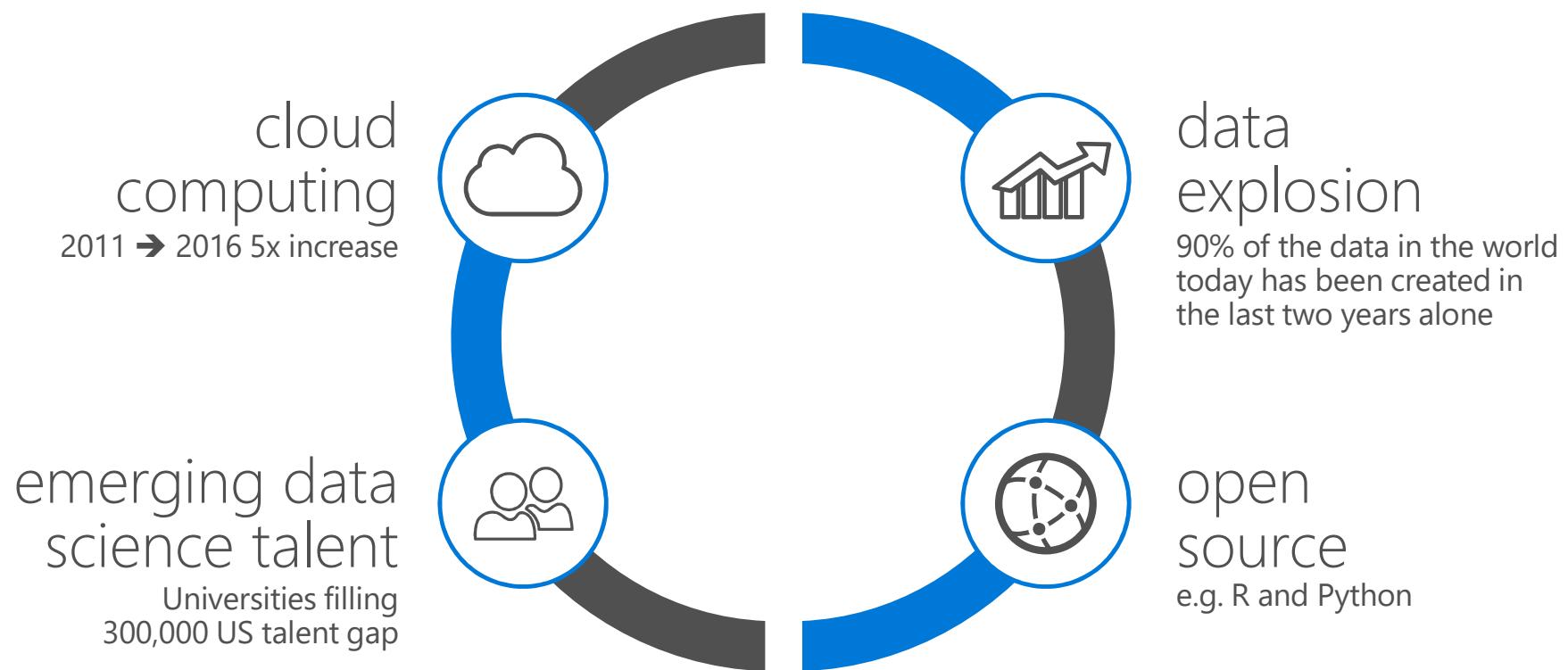
90% of the data in the world today has been created in the last two years alone

cloud
computing
2011 → 2016 5x increase

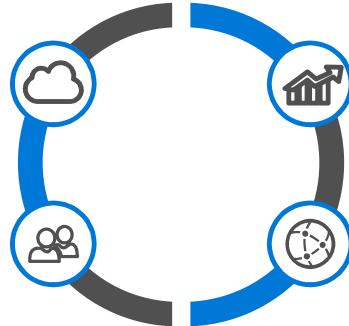
Machine
Learning



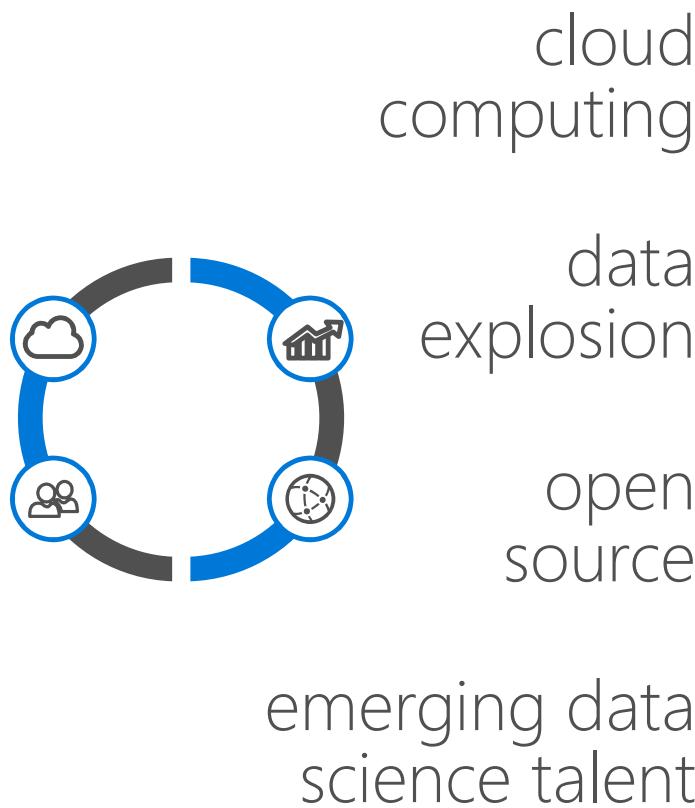
Transformational Trends



Transformational Trends

- 
- cloud computing
 - Elastic compute, storage and streaming platforms
 - Access to data born in the cloud
 - Incremental capability beyond on-premises systems
 - data explosion
 - Expanded data about people, events, machines etc.
 - Increased level of detail
 - Rapid increase in prediction accuracy
 - open source
 - Freely available platform and tool innovations
 - Combine tool sets to increase capability
 - Bridge on-prem and cloud assets
 - emerging data science talent
 - Immediate effectiveness
 - Reduce reliance on expensive staff
 - Embrace latest innovations

Microsoft's Advanced Analytics Platforms



- Flexible Path to the Cloud
- Mature, Geographically Dispersed Infrastructure
- Hybrid clouds with cloud-on-prem interoperability
- Elastic Scale On-Prem & In the Cloud
- Perceptual Algorithms
- Leading-edge Machine Learning
- Comprehensive R Integration
- Full Commercial Support
- Integration with Linux and Hadoop
- Choice of IDEs & Graphical Tools
- Collaboration Across Stakeholders
- Template & Sample Galleries

April 6, 2015



Get Technical Support



Community Applications Products AdviseR Resources Company

Microsoft Completes Acquisition of Revolution Analytics

MICROSOFT BLOG

REVOLUTIONS BLOG



"This acquisition will help customers use advanced analytics within Microsoft data platforms."

A brief history of R

1993

Research project
in Auckland, NZ

1997

R-core

2003

R Foundation

2009

New York Times

1995

Open source



Photo credit: Robert Gentleman

2000

R-1.0.0

2004

First UseR!

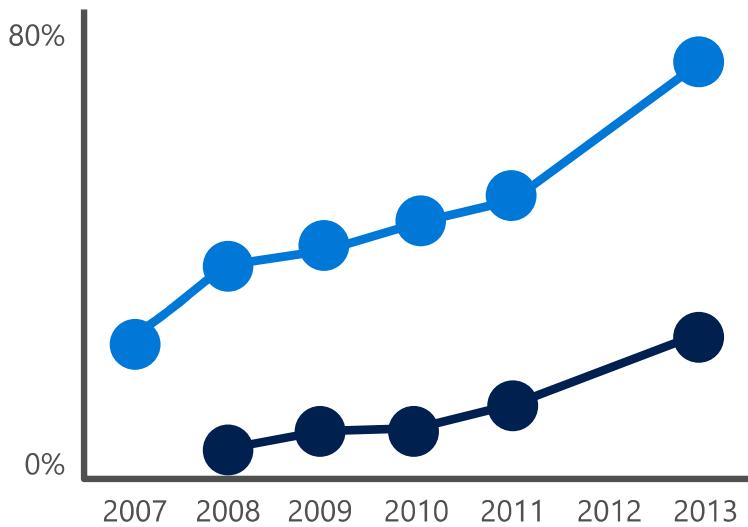
2015

R-3.2.0
R Consortium

R's popularity is growing rapidly

R Usage Growth

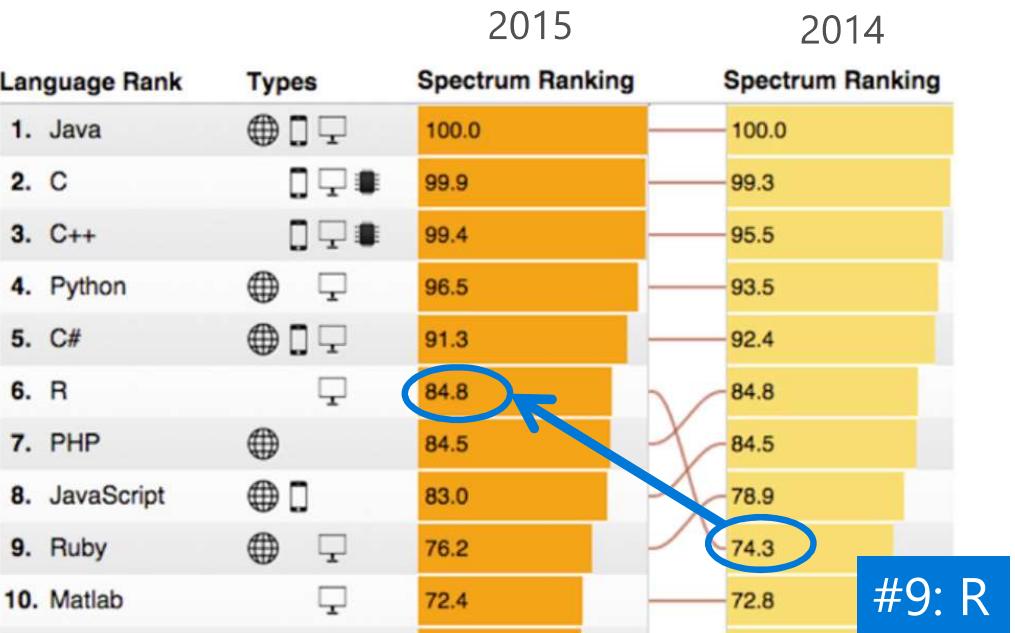
Rexer Data Miner Survey, 2007-2013



[Rexer Data Miner Survey](#)

Language Popularity

IEEE Spectrum Top Programming Languages



[IEEE Spectrum July 2015](#)

CRAN

The Comprehensive R Archive Network

CRAN Task Views

CRAN Task Views are guides to the packages and functions useful for certain disciplines and methodologies. Many long-term R users I know have no idea they exist. As an effort to make them more widely known I thought I'd jazz up the index page. Images are free to use, and got from [SXC stock photo site](#). Visual puns are mine. Task View links go to the [cran.r-project.org](#) site and not a mirror.



Bayesian Inference

Applied researchers interested in Bayesian statistics are increasingly attracted to R, because of the ease of which one can code algorithms to sample... [\[more\]](#)



Chemometrics and Computational Physics

Chemometrics and computational physics are concerned with the analysis of data arising in chemistry and physics experiments, as well as the simulation of... [\[more\]](#)



Clinical Trial Design, Monitoring, and Analysis

This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including... [\[more\]](#)



Cluster Analysis & Finite Mixture Models

This CRAN Task View contains a list of packages that can be used for finding groups in data and modelling unobserved cross-sectional heterogeneity. Many... [\[more\]](#)



Probability Distributions

Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many... [\[more\]](#)



Computational Econometrics

This Task View contains information about using R to analyse ecological and environmental data. Please feel free to suggest enhancements... [\[more\]](#)



Analysis of Ecological and Environmental Data

This task view collects information on R packages for experimental design and analysis of data from experiments. Please feel free to suggest enhancements... [\[more\]](#)



Empirical Finance

This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic... [\[more\]](#)



Statistical Genetics

Great advances have been made in the field of genetic analysis over the last years. The availability of millions of single nucleotide polymorphisms (SNPs)... [\[more\]](#)



Natural Language Processing

This CRAN task view contains a list of packages useful for natural language processing... [\[more\]](#)



Analysis of Pharmacokinetic Data

The primary goal of pharmacokinetic (PK) data analysis is to determine the relationship between the dosing regimen and the body's exposure to the drug as... [\[more\]](#)



Official Statistics & Survey Methodology

This CRAN task view contains a list of packages that includes methods typically used in official statistics and survey methodology. Many packages provide... [\[more\]](#)



Phylogenetics, Especially Comparative Methods

The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical... [\[more\]](#)



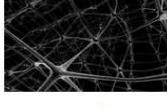
Multivariate Statistics

Base R contains most of the functionality for classical multivariate analysis... somewhere. There are a large number of packages on CRAN which extend this... [\[more\]](#)



Optimization and Mathematical Programming

This CRAN task view contains a list of packages which offer facilities for solving optimization problems. Although every regression model in statistics... [\[more\]](#)



Machine Learning & Statistical Learning

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually... [\[more\]](#)



Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic,... [\[more\]](#)



High-Performance and Parallel Computing with R

This CRAN task view contains a list of packages, grouped by topic, that are useful for high-performance computing (HPC) with R. In this context, we are... [\[more\]](#)



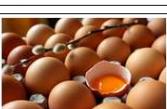
Medical Image Analysis

This task view is for input, output, and analysis of medical imaging files... [\[more\]](#)



Analysis of Spatial Data

Base R includes many functions that can be used for reading, visualising, and analysing spatial data. The focus in this view is on "geographical" spatial... [\[more\]](#)



Survival Analysis

Survival analysis, also called event history analysis in social science, or reliability analysis in engineering, deals with time until occurrence of an... [\[more\]](#)



Time Series Analysis

Base R ships with a lot of functionality useful for time series, in particular in the stats package. This is complemented by many packages on CRAN, which are... [\[more\]](#)



Robust Statistical Methods

Robust ("resistant") methods for statistics modeling have been available in S from the start, in R in package stats (e.g., median(), mean(*, trim = ...). [\[more\]](#)



Statistics for the Social Sciences

Social scientists use a wide range of statistical methods. To make the burden carried by this task view lighter, I have suppressed detail in some areas that... [\[more\]](#)



Graphical Models in R

Wikipedia defines a graphical model as a graph that represents independencies among random variables by a graph in which each node is a random variable, and... [\[more\]](#)

In addition to CRAN, Bioconductor, GitHub, others distribute R packages

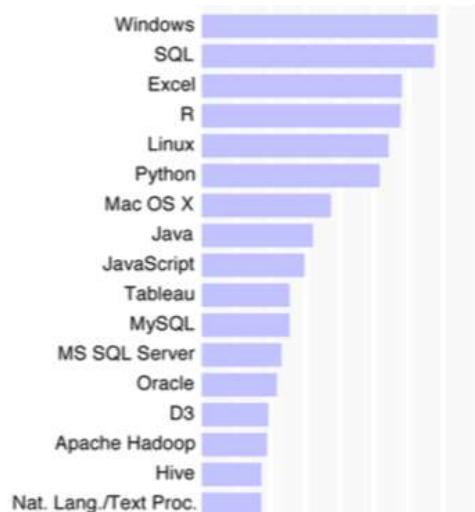
Standing on the Shoulders of Giants



A Vast Community of R Users Share Rich Repositories of Pre-Built Solutions

Tool Use for Data Science

O'Reilly Data Science Survey 2014
(max=80%)



CRAN

The Comprehensive R Archive Network
Resources For All Fields of Analysis

CRAN Task Views

CRAN Task Views are guides to the packages and functions useful for certain disciplines and methodologies. Many long-term R users I know here have no idea they exist. As an effort to make them more widely known I thought I'd just lay up the index page. Images are free to use, and give from [Attribution](#) stock photo site. Visuals are mine. Task View links go to the cran.r-project.org site and not a mirror.

Bayesian Inference
Applied researchers interested in Bayesian statistics are particularly interested in R because of the ease of which one can implement it for analysis. [\[index\]](#)

Chemometrics and Computational Physics
Chemometrics and computational physics are concerned with the development of algorithms for solving problems involving in chemistry and physics experiments, as well as the calculation of physical properties. [\[index\]](#)

Clinical Trial Design, Monitoring, and Analysis
This task view gathers software for the design and analysis of clinical trials. It includes software for dose finding, monitoring, and analysis of clinical trials. It focuses on software for [\[index\]](#)

Cluster Analysis & Finite Mixture Models
For many of the classical distributions, there is a procedure for estimating the parameters of the distribution by maximum likelihood or related techniques. In addition, there are procedures for estimating the number of clusters in a data set and for calculating cluster membership. Many [\[index\]](#)

Computational Economics
R uses S with a set of functionality useful for computational economics. It provides facilities for the solution of differential equations (ode), [\[index\]](#) and [\[index\]](#)

Design of Experiments (DoE)
This task view collects information on R packages for the design and analysis of experiments. Please feel free to suggest enhancements. [\[index\]](#)

Analysis of Ecological and Environmental Data
This Task View contains software designed to analyse ecological and environmental data. [\[index\]](#)

Empirical Finance
This task view collects information on R packages for empirical work in finance. Please feel free to suggest enhancements. [\[index\]](#)

Statistical Genetics
Genetic analysis have been made in R for analysis of association between the genotype and phenotype. The availability of millions of single nucleotide polymorphisms (SNPs) [\[index\]](#)

Official R Task Views
Official R Task Views are guides to the packages and functions useful for certain disciplines and methodologies. Many long-term R users I know here have no idea they exist. As an effort to make them more widely known I thought I'd just lay up the index page. Images are free to use, and give from [Attribution](#) stock photo site. Visuals are mine. Task View links go to the cran.r-project.org site and not a mirror.

Natural Language Processing
This CRAN task view contains a list of packages for natural language processing. [\[index\]](#)

Analytical Pharmacokinetic Data
The primary goal of pharmacokinetics (PK) data analysis is to determine the interaction of the body's response to the drug in. [\[index\]](#)

Official R Task & Survey Methods
Official R Task & Survey Methods packages that includes methods typically used in survey sampling, including methods like: [\[index\]](#)

Phylogenetic Especially Com. & Methods
The history of life is reflected within a phylogenetic tree. Comparative methods are used to analyze the relationships among living organisms. [\[index\]](#)

Iivariate Statistics
This task view contains some of the best packages for classical multivariate analysis and data mining. It also contains a collection of packages for CRAN which extend this. [\[index\]](#)

Optimization Math., Stats., and Programming
This CRAN task view contains a list of packages which often facilitate the solving of optimization problems. Optimization problems are common in regression model estimation. [\[index\]](#)

Machine Learning & Statistical Learning
All CRAN packages for machine learning and methods developed at the heliotrope research group are included in this task view. This field of research is rapidly growing. [\[index\]](#)

High-Level Data and Graphics Utilities
A collection of packages for developing high-level data and graphics utilities. [\[index\]](#)

Medical Image Analysis
This task view is for input, output, and analysis of medical imaging files. [\[index\]](#)

Psychometric Models and Methods
Psychometrics is concerned with the design and analysis of research and the application of statistical methods. Psychometrics packages have also worked. [\[index\]](#)

Survival Analysis
Statistical analysis, also called event history analysis or reliability analysis, is used in the social sciences, engineering, and medicine to estimate the time until events occur given a set of covariates. [\[index\]](#)

Time Series Analysis
R now has a lot of functionality useful for statistical modeling and analysis for time series data. It is currently the most used package in this field, which are: [\[index\]](#)

Robust Statistical Methods
Robust (or "resistant") methods for statistical modeling have been available in R for some time. In particular, in the stats package. This is a collection of robust methods which are: [\[index\]](#)

Statistics for the Social Sciences
Social scientists use a wide range of statistical methods. We have tried to collect some of the most commonly used ones in this task view. If you have any suggestions, please add them. [\[index\]](#)

Graphical Models in R
Graphical models are graphical models in which nodes represent variables and edges represent dependencies among those variables. A graph is a directed acyclic graph (DAG). [\[index\]](#)

Reproducible Research
The goal of reproducible research is to be able to recreate the results of a study using the same inputs. Reproducible research actions & graphics model a graph that represents independence among model variables by a graph is a directed acyclic graph (DAG). [\[index\]](#)

6,500+ Packages

+ Bioconductor
+ Github

Enterprises Press the Limits of Open Source R

Enterprise Data Scale Can Exceed R's Capacity

Memory-Based Data Access Model

Lack of Parallel Computation

Requires Data Movement Prior to Analysis

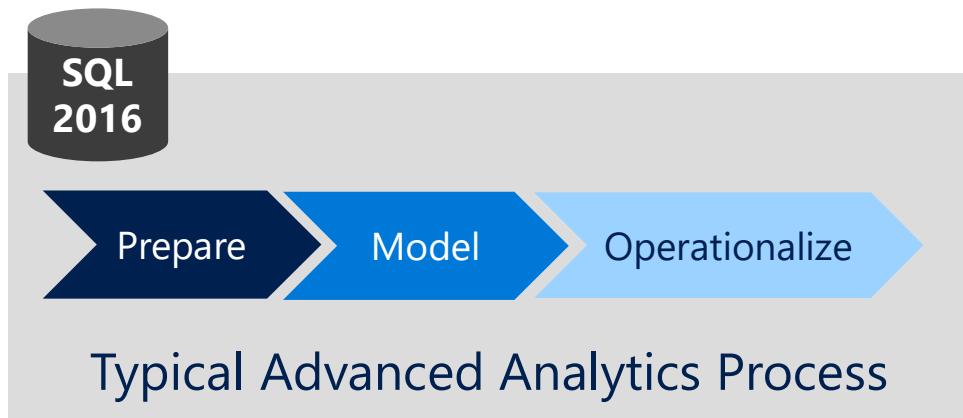
Community Support Limits Enterprise Utilization

Community Support Insufficient for Mission Critical Solutions

Governance Rules Often Prohibit Unsupported Open Source

Using SQL Server R Services

Enterprise R Analytics in SQL Server 2016



Model & Deploy In SQL16

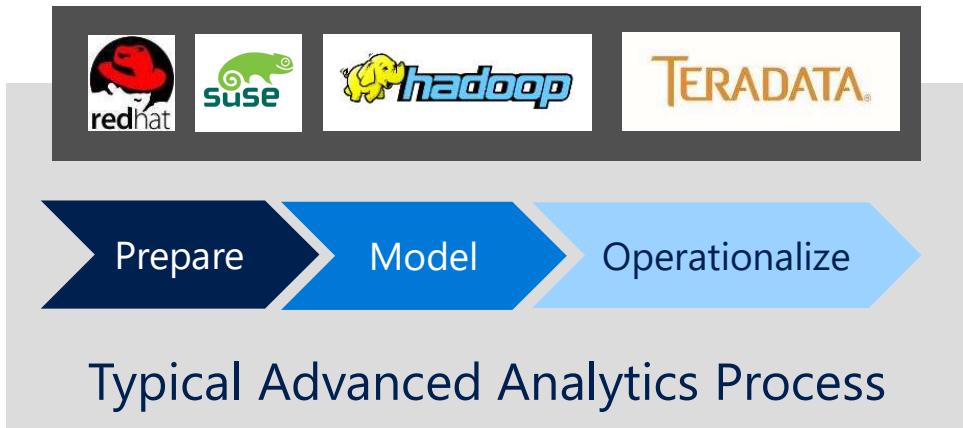
Support Entire Analytics Lifecycle
Run R Inside SQL 2016 using IDE
Embed R in T-SQL or run stored proc's

Advantages

- Scale By Eliminating Movement
- Scale Using Parallelized Computation
- Reduce Security Exposure
- Reuse SQL Skills for Data Engineering
- Reuse SQL Skills in App Development teams
- Maximize Operational Stability for Applications
- Supported 100% R

Microsoft R Server

Enterprise R Analytics for Linux, Hadoop & Teradata



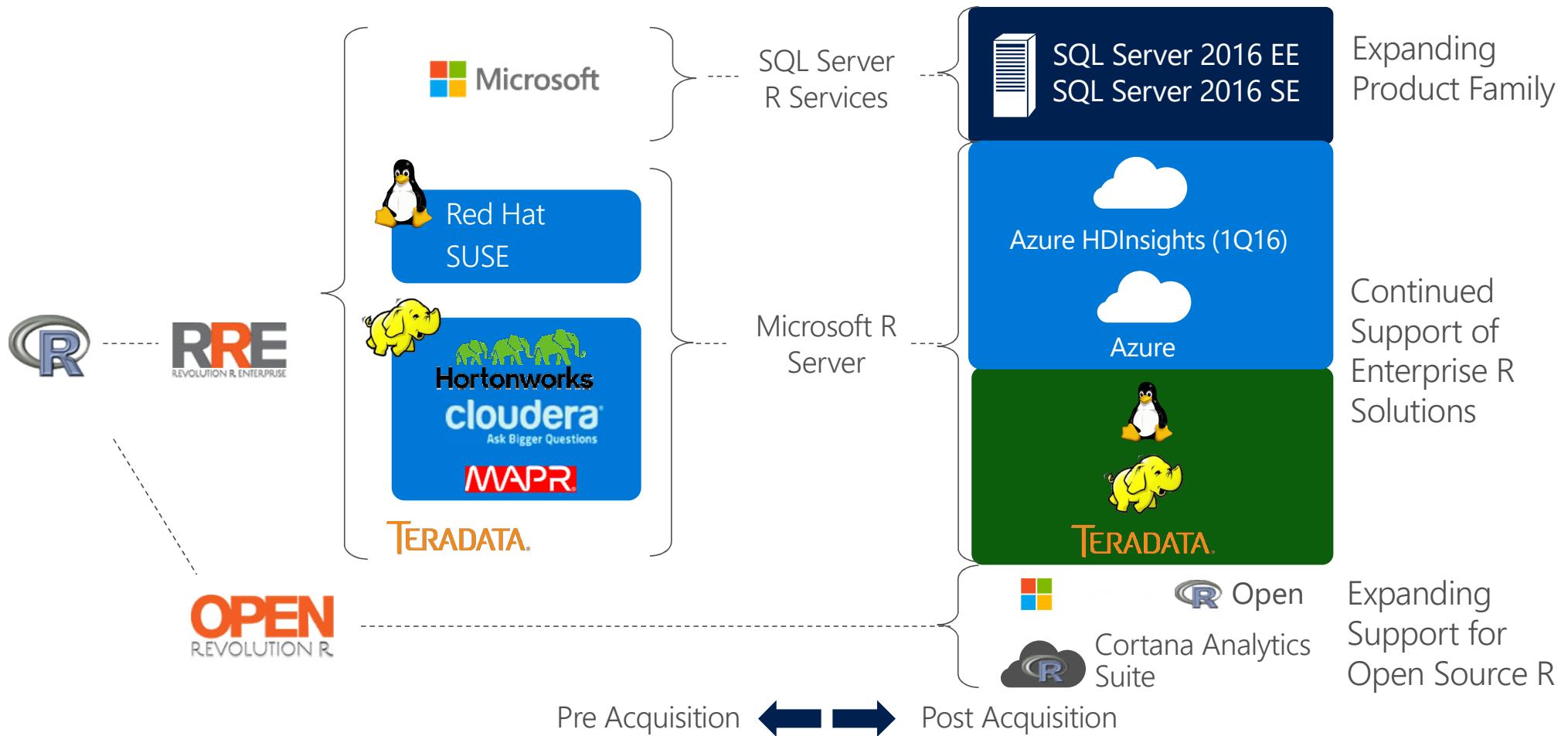
Multi-Platform Scalable R

Support Entire Analytics Lifecycle
Big Data Scale & Compute
Remote Execution in Hadoop, Teradata

Advantages

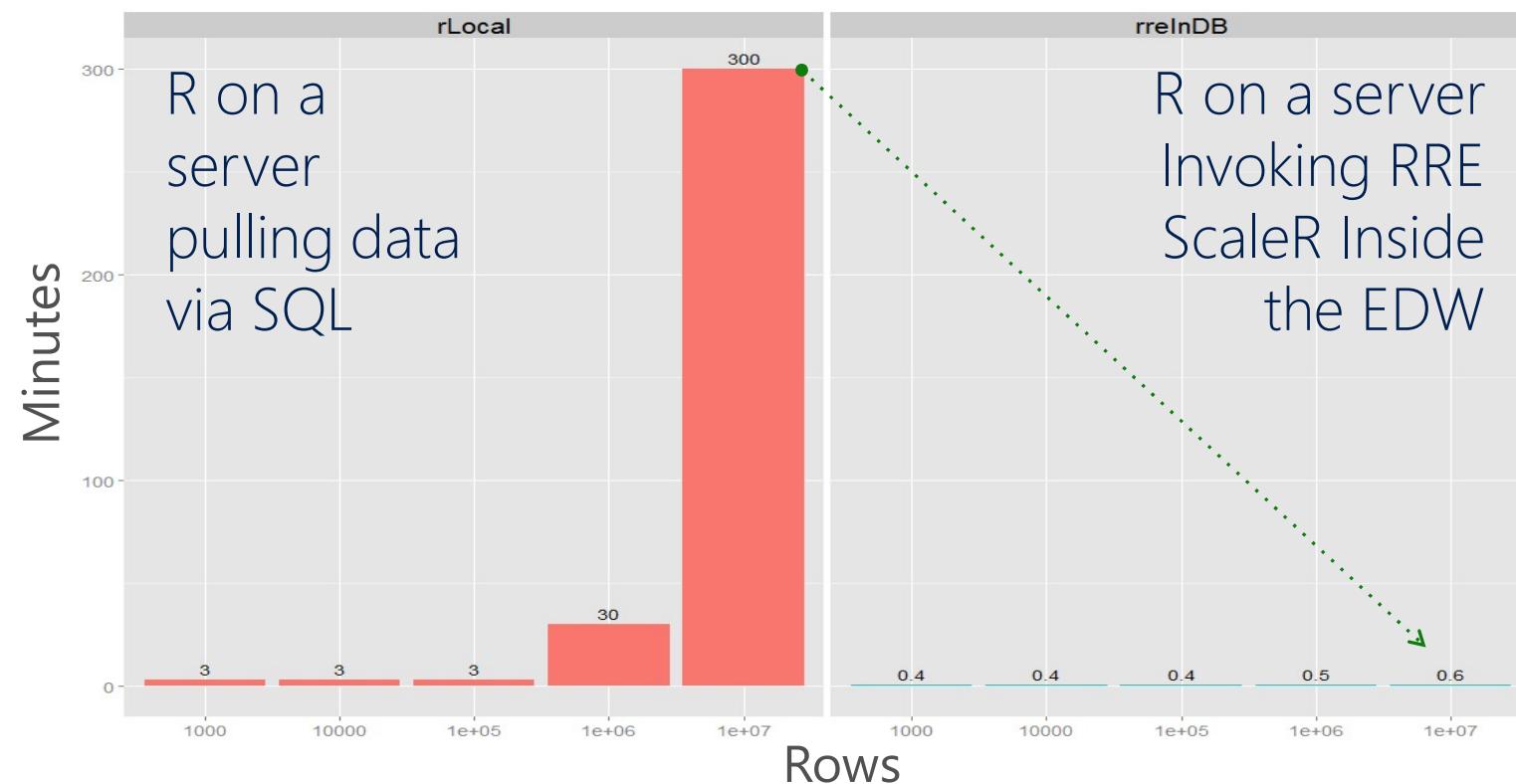
Supported 100% R
Deploy via Web Services
Reduced Security Exposure

Growing Beyond Revolution Analytics



Eliminate Data Movement

In-Database Example: From 5+ hours to 40 seconds





Parallelized, Remote Execution Algorithms

Data Step	Statistical Tests	Variable Selection
Data import – Delimited, Fixed, SAS, SPSS, OBDC	Chi Square Test	Stepwise Regression
Variable creation & transformation	Kendall Rank Correlation	
Recode variables	Fisher's Exact Test	
Factor variables	Student's t-Test	
Missing value handling		
Sort, Merge, Split		
Aggregate by category (means, sums)		
Descriptive Statistics	Sampling	Simulation
Min / Max, Mean, Median (approx.)	Subsample (observations & variables)	Simulation (e.g. Monte Carlo)
Quantiles (approx.)	Random Sampling	Parallel Random Number Generation
Predictive Models	Cluster Analysis	Classification
Standard Deviation		K-Means
Variance		
Correlation		
Covariance		
Sum of Squares (cross product matrix for set variables)	Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.	
Pairwise Cross tabs	Covariance & Correlation Matrices	
Risk Ratio & Odds Ratio	Logistic Regression	
Cross-Tabulation of Data (standard tables & long form)	Classification & Regression Trees	
Marginal Summaries of Cross Tabulations	Predictions/scoring for models	
	Residuals for all models	
Combination	 rxDataStep	 New rxExec
		PEMA-R API Custom Algorithms

Microsoft R Server Components

Marcello Benati, MCM
Advanced Analytics Product Manager



Barriers to analytics adoption

Talent Scarcity	<ul style="list-style-type: none">• Academic Rigor• Talent Competition
Low Productivity	<ul style="list-style-type: none">• Integration Complexity• Tool, Skill & Culture Gaps
Complex Infrastructure	<ul style="list-style-type: none">• Data Volume, Diversity• Security & Governance Constraints• Rapid Platform Evolution
Slow Innovation	<ul style="list-style-type: none">• Low Experimentation Rate• Complex Operationalization
High Cost	<ul style="list-style-type: none">• Legacy Products• Irregular Workload

Earning our credibility

We needed to leverage data and analytics to grow our products.

Key Innovation...

More experiments by more people!

So we...

Built an Exabyte-scale data lake for everyone to put their data.

Built tools approachable by any developer.

Built machine learning tools for collaborating across large experiment models.



Using vastly accelerated experimentation cycles:



MICROSOFT DOUBLES SEARCH SHARE



30% BETTER IN SPEECH AND GESTURE RECOGNITION



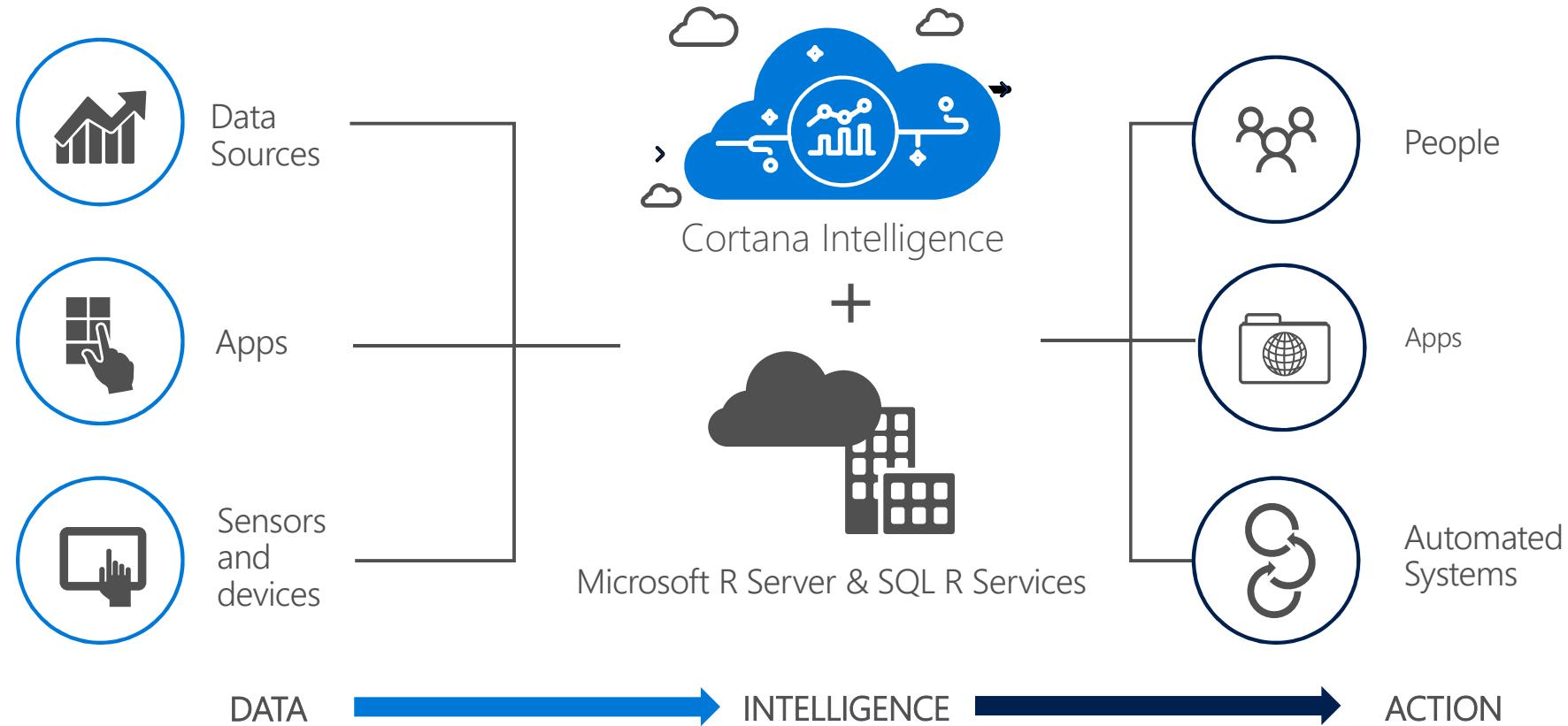
NEW CAPABILITIES LIKE OFFICE GRAPH & CLUTTER



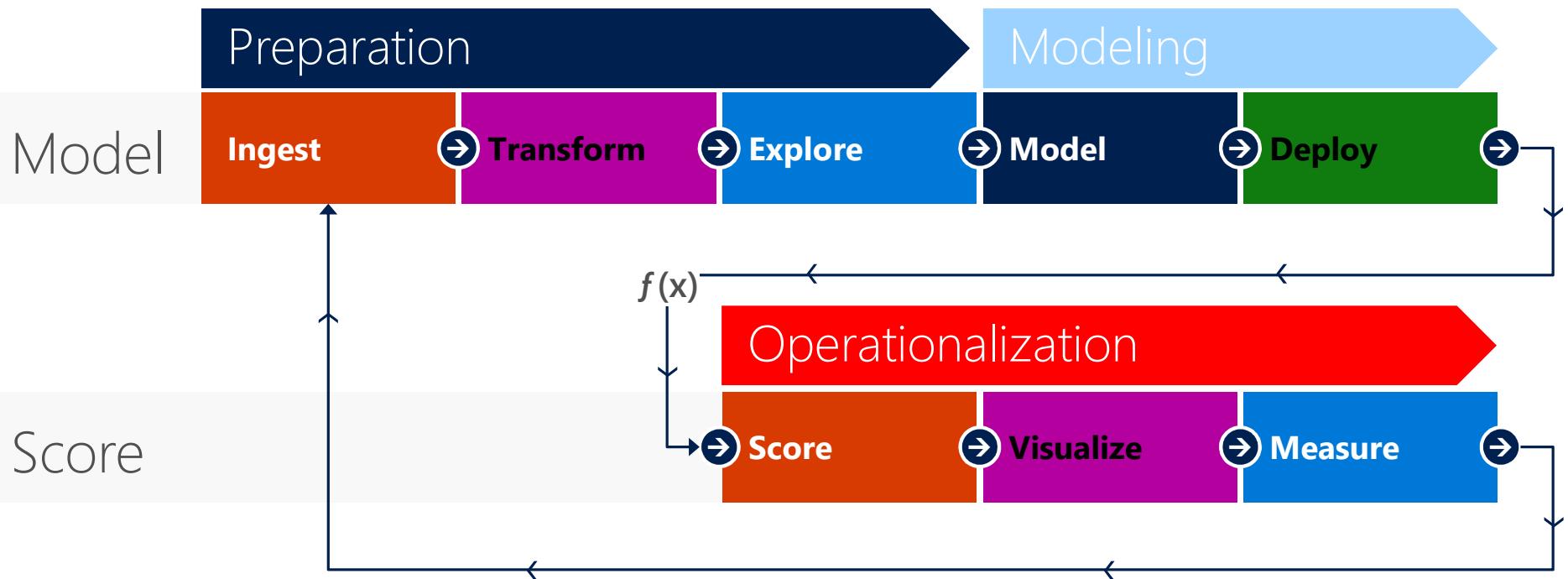
SKYPE TRANSLATE INTRODUCTION

From data to intelligence to action

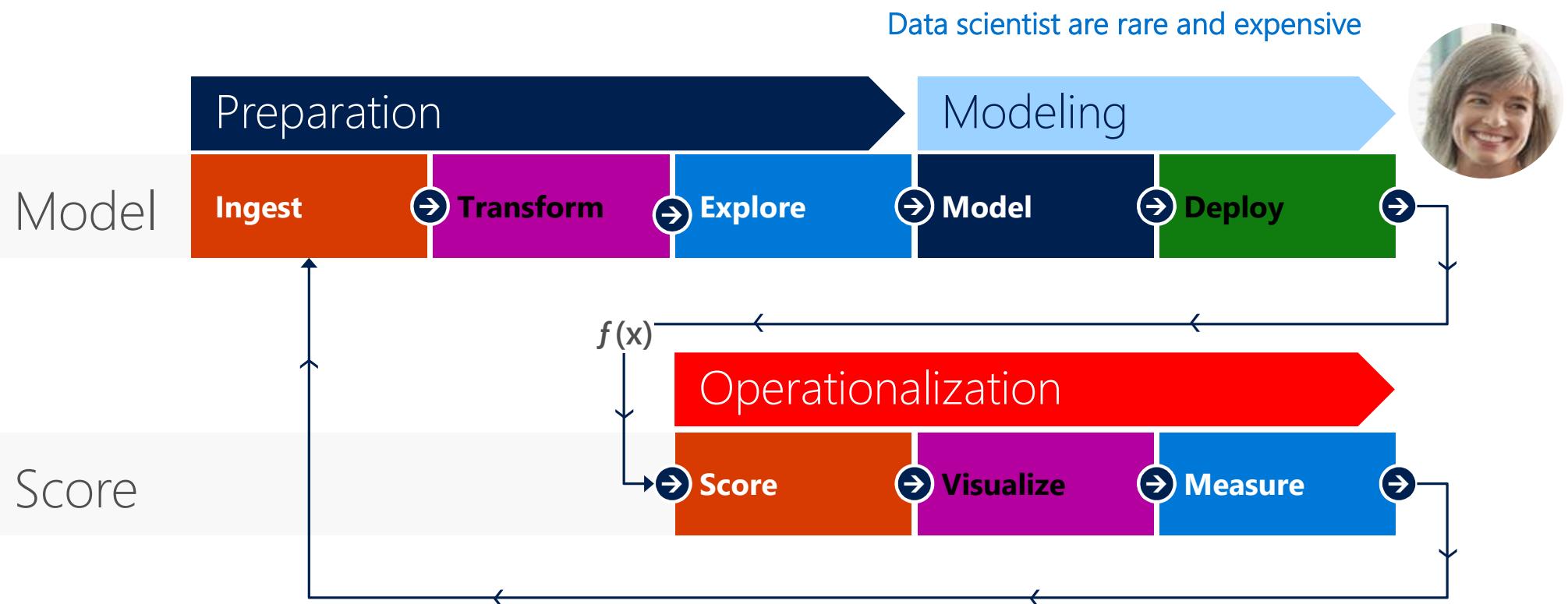
On Prem or in the Cloud



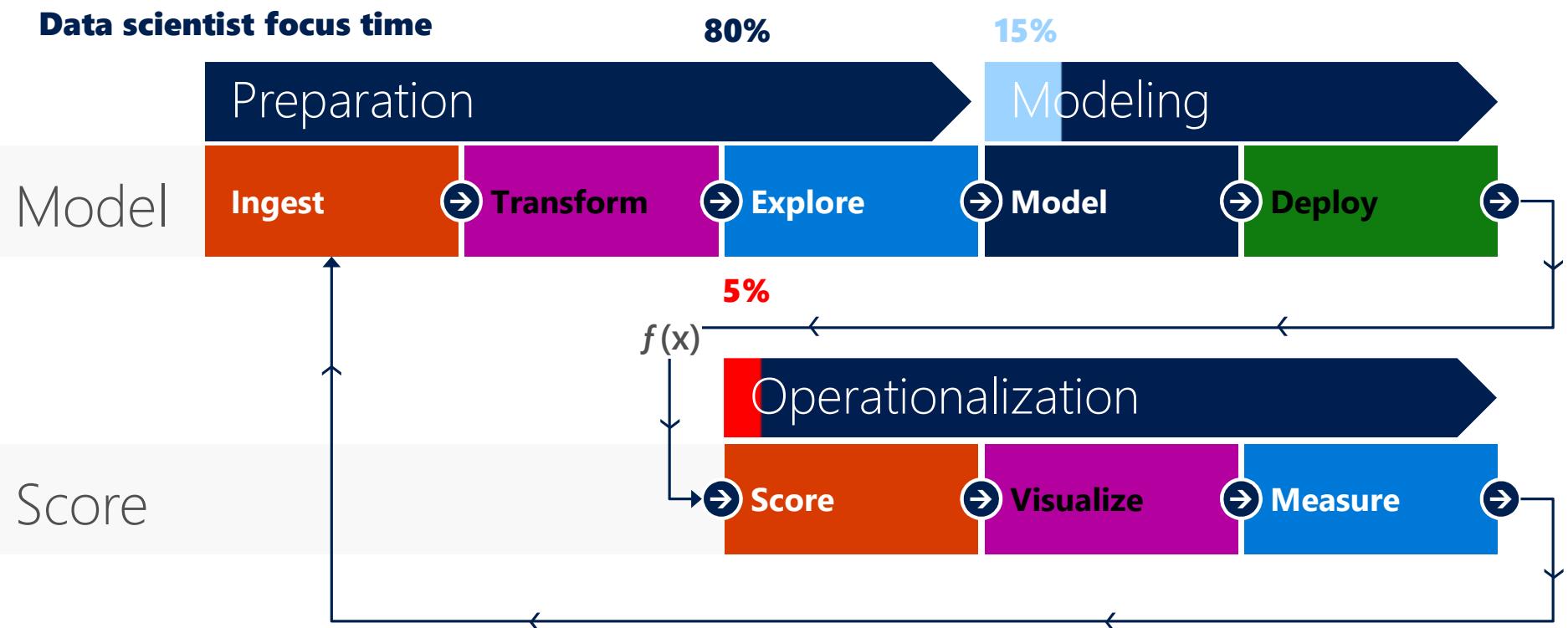
Typical advanced analytics lifecycle



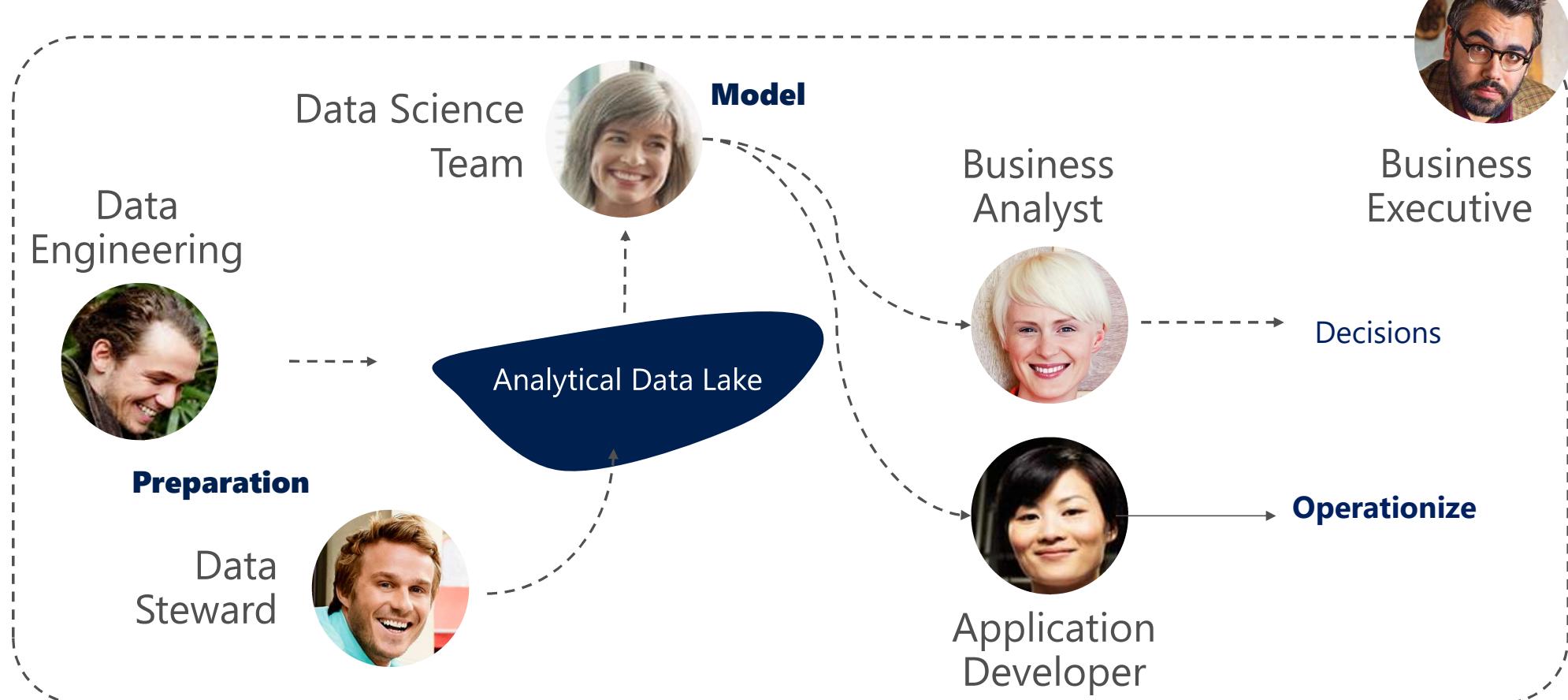
Data Scientist should be creating / testing models



But the reality is different ...



Making Advanced Analytics a team sport



Easy Scaling of R Analytics on Hadoop

- **Remote Execution**

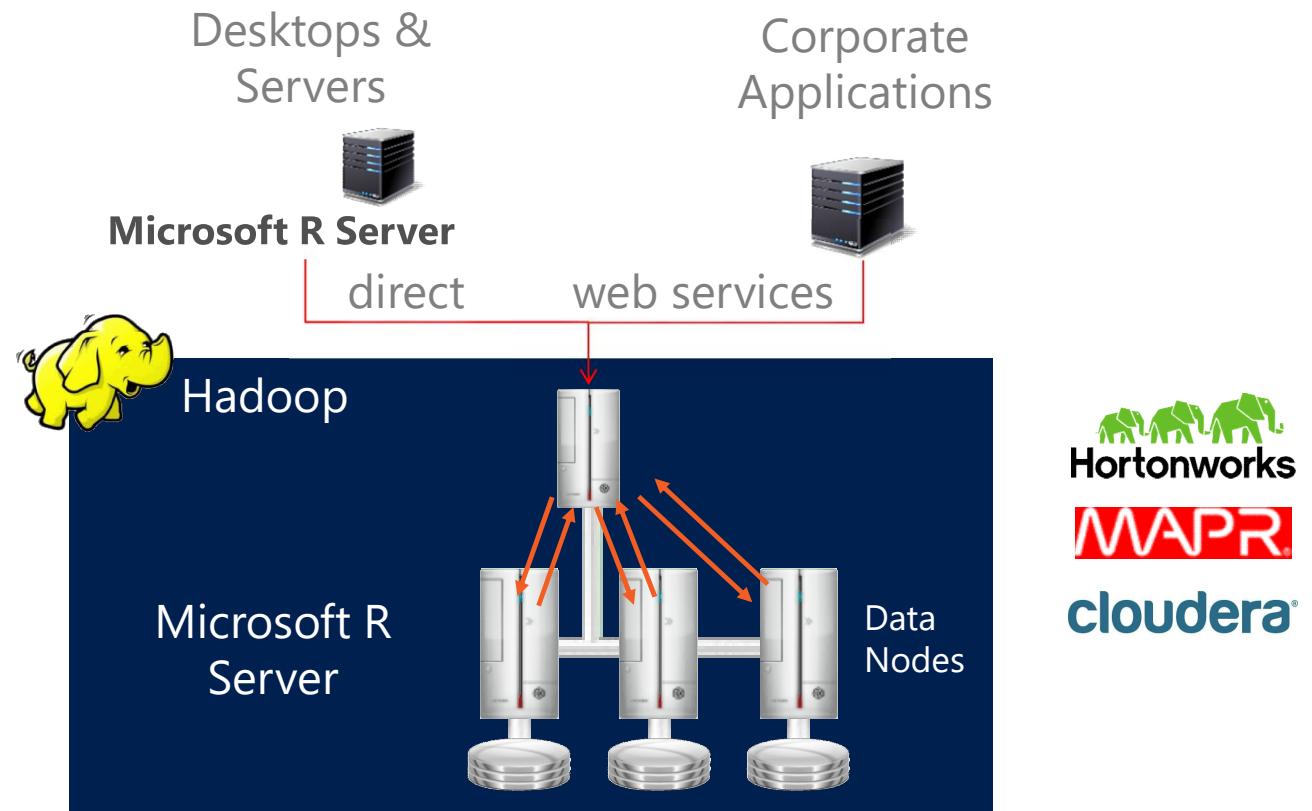
- Run Analytics from Workstations, Servers or Web Services APIs

- **Transparent Parallelization:**

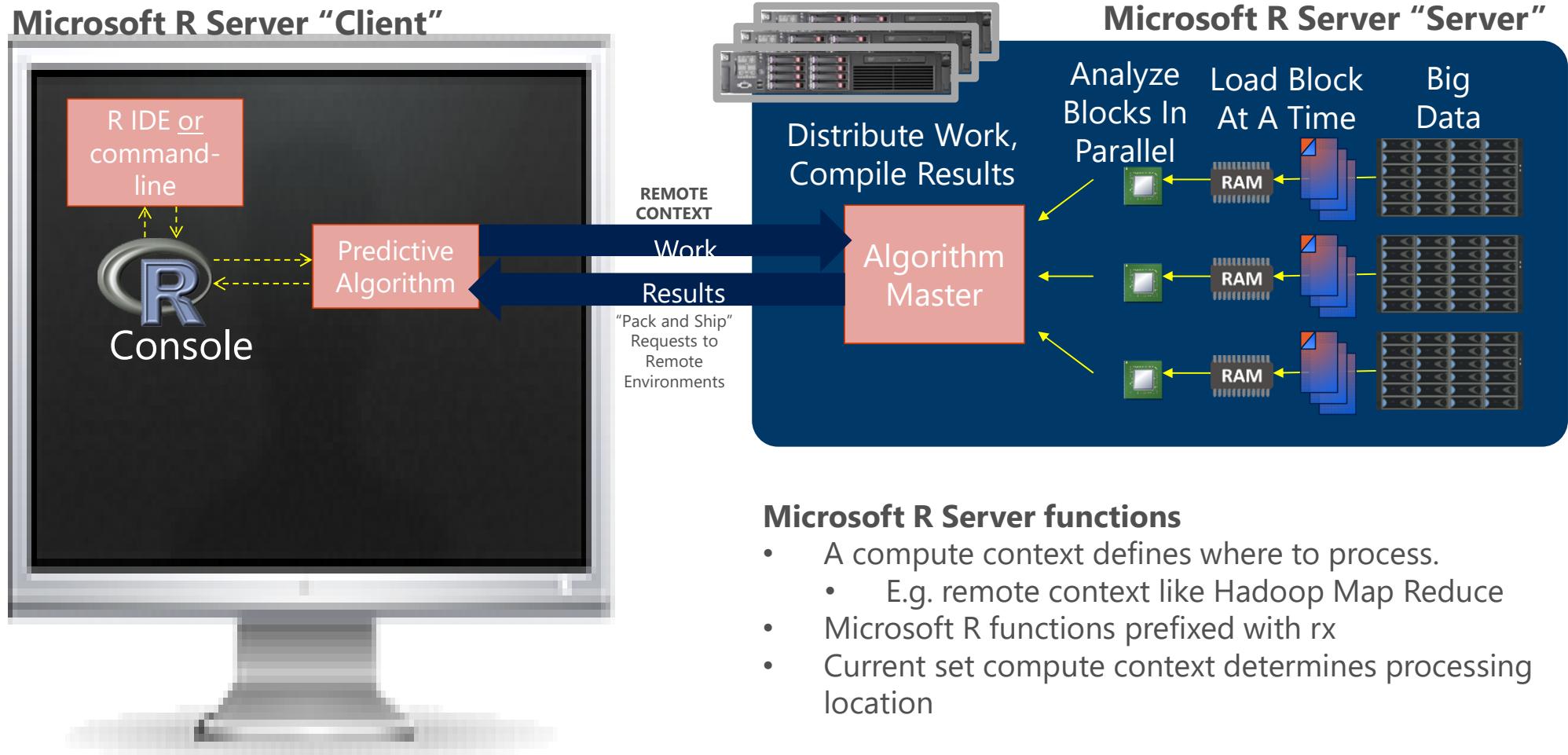
- YARN / MapReduce
- Spark

- **Shared Resource Management**

- Disk-based Data Chunking



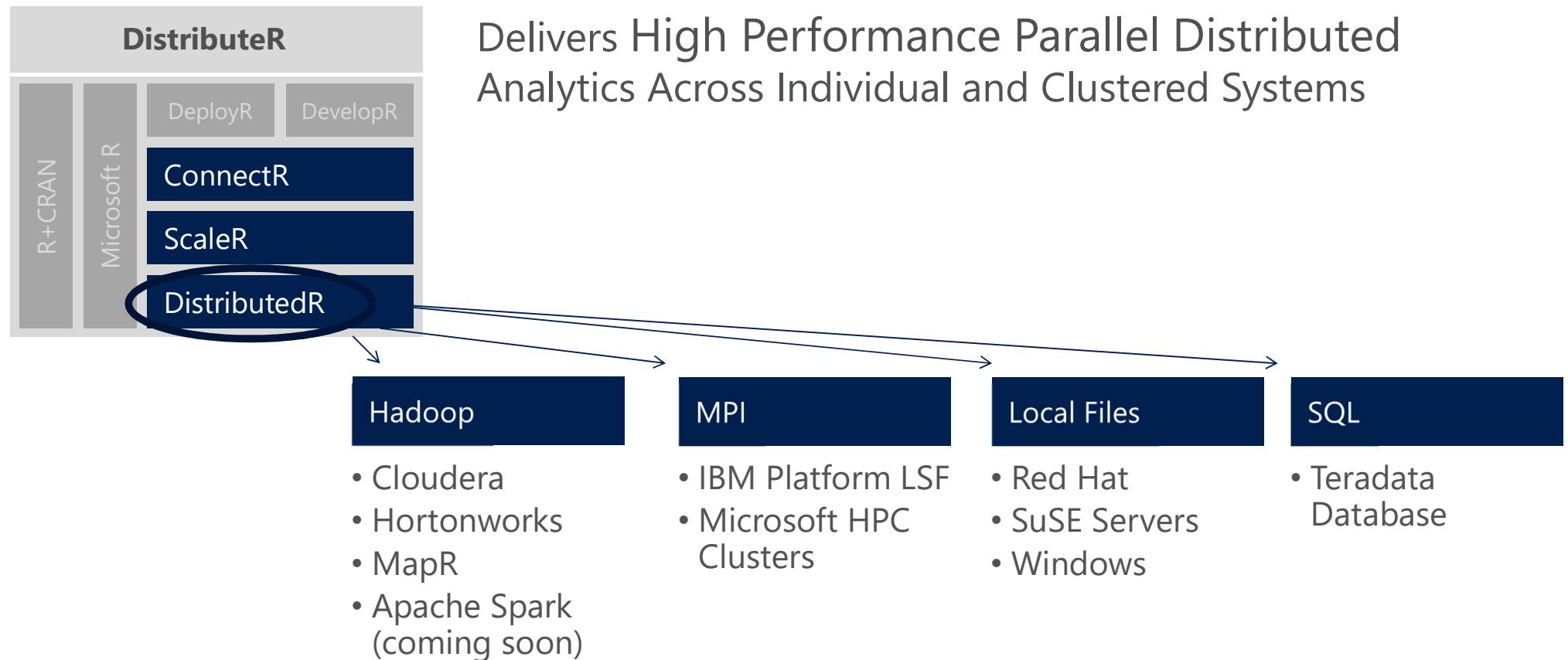
Distributed R - How Does Remote Compute Context?



Microsoft R Server functions

- A compute context defines where to process.
 - E.g. remote context like Hadoop Map Reduce
- Microsoft R functions prefixed with rx
- Current set compute context determines processing location

DistributedR



DistributedR - Revolution Code Portability

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

Compute context R script – sets where the model will run

Local Parallel processing – **Linux or Windows**

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

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

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

In – **Hadoop**

```
### SETUP HADOOP ENVIRONMENT VARIABLES ###
myHadoopCC <- RxHadoopMR()

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

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

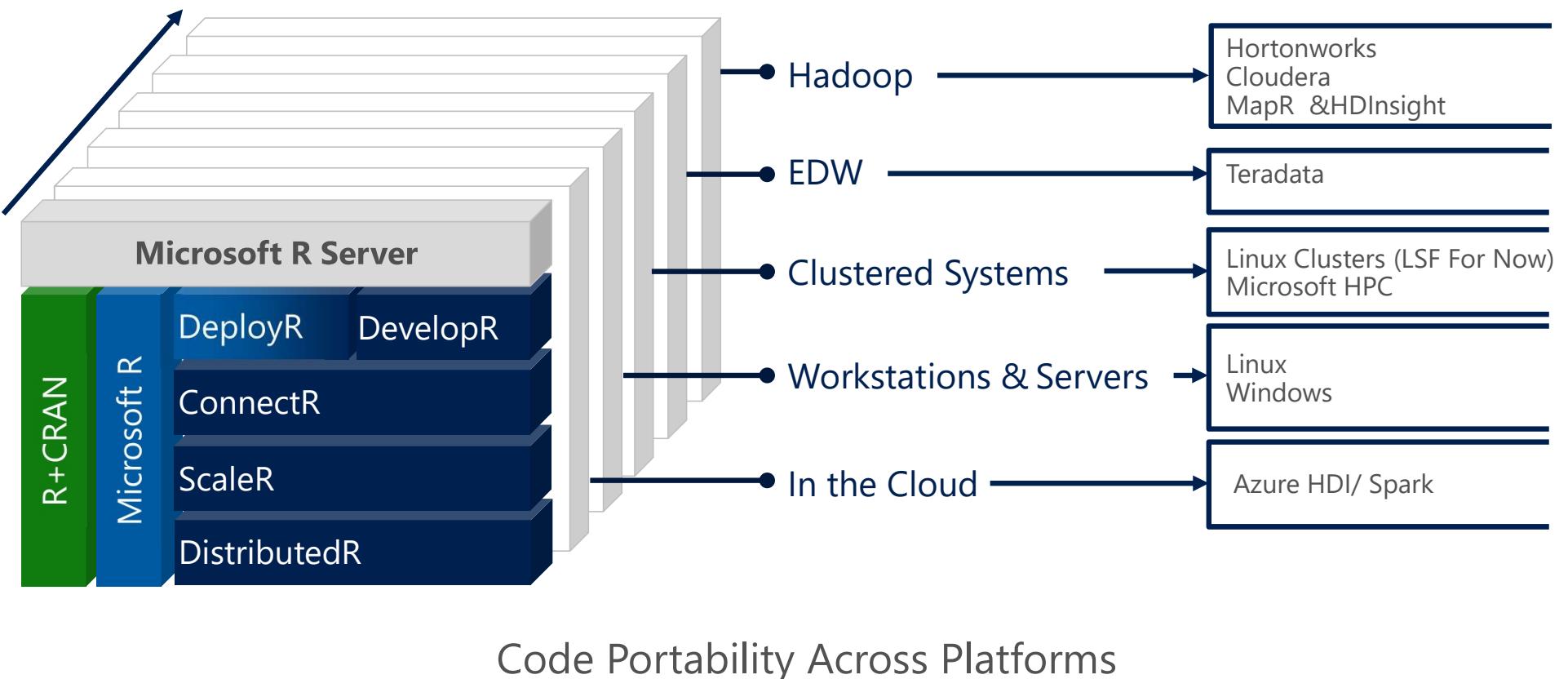
Functional model R script – does not need to change to run in Hadoop

```
### 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 + 0 , data = AirlineDataSet)
plot(hdfsXdfArrLateLinMod$coefficients)
```

Write Once. Deploy Anywhere.



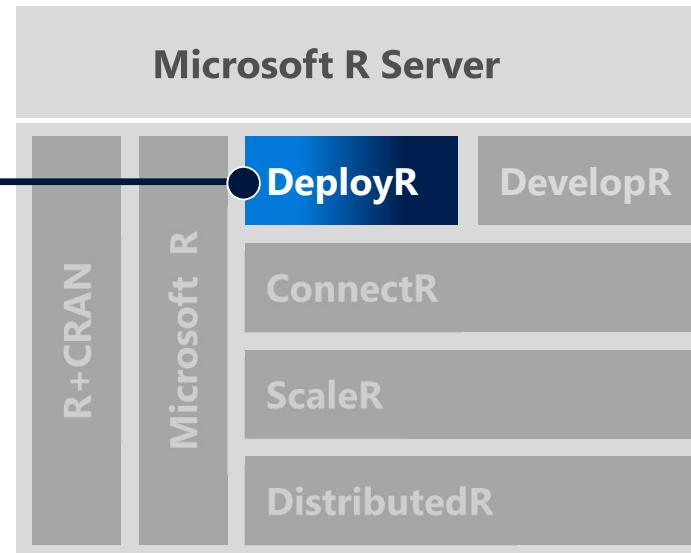
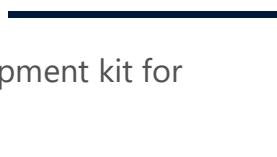
DeployR

DeployR

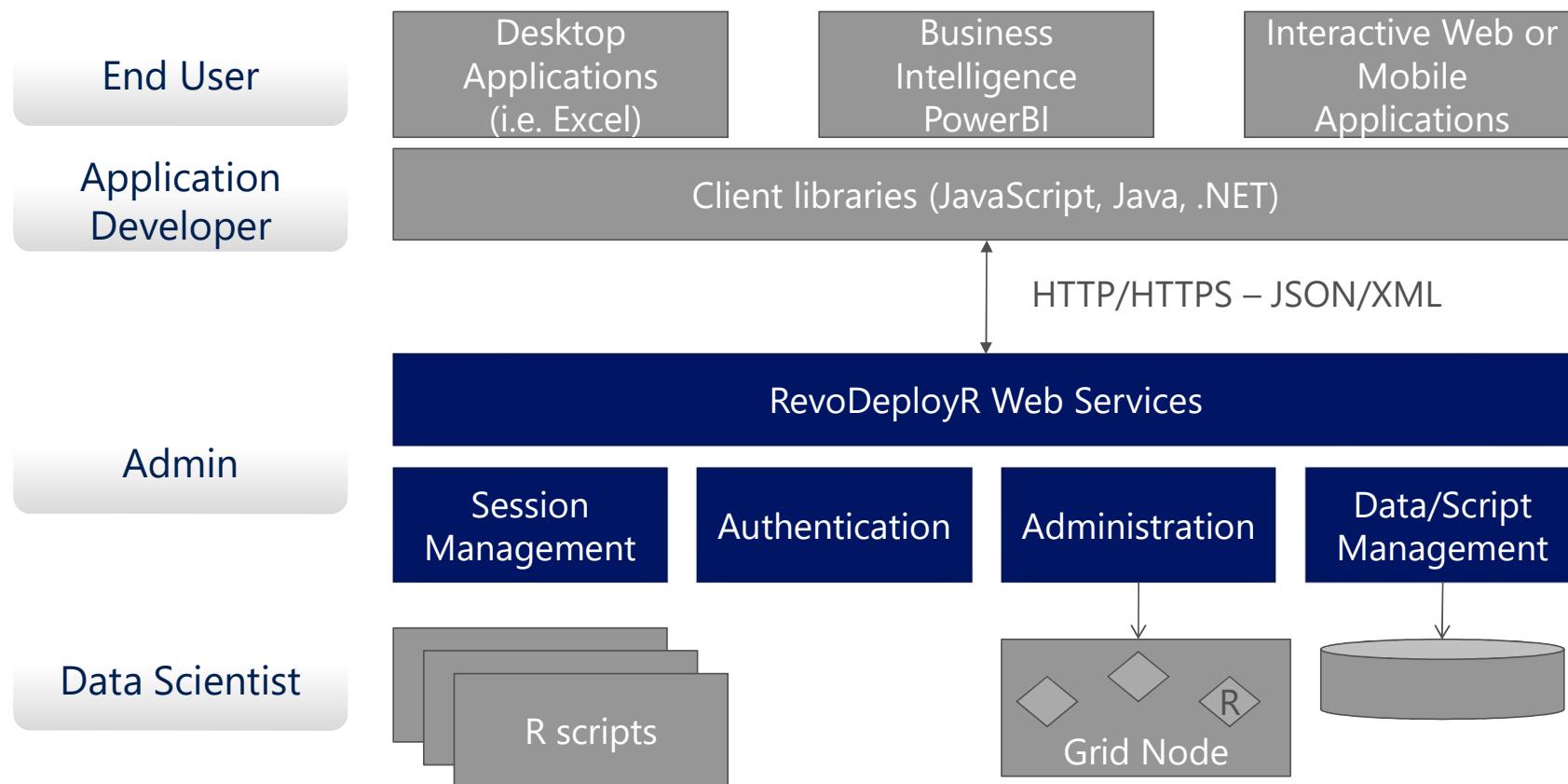
- Web services software development kit for integration analytics via APIs :
 - Java
 - JavaScript
 - .NET Integrates R Into application infrastructures

Capabilities:

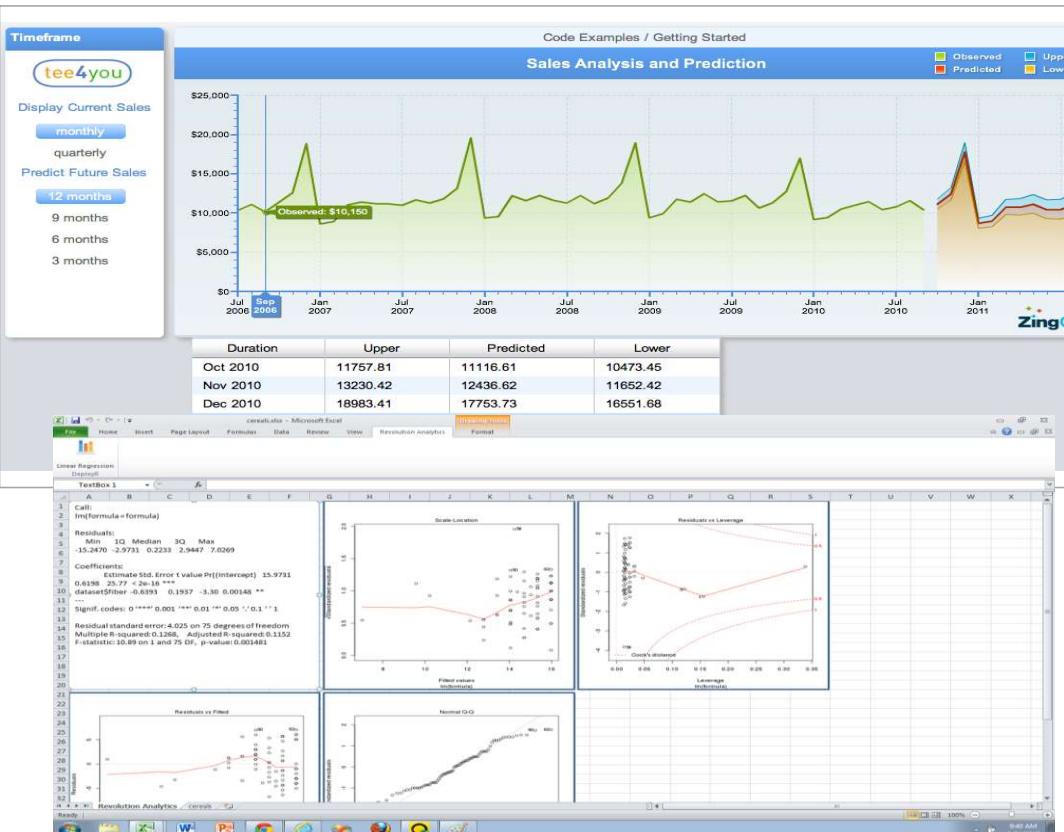
- Enterprise authentication & security
- Horizontal scaling
- Invokes R Scripts from web services calls
- RESTful interface for easy integration
- Works with:
 - Web & mobile apps
 - Leading BI & Visualization tools
 - Business rules and streaming engines



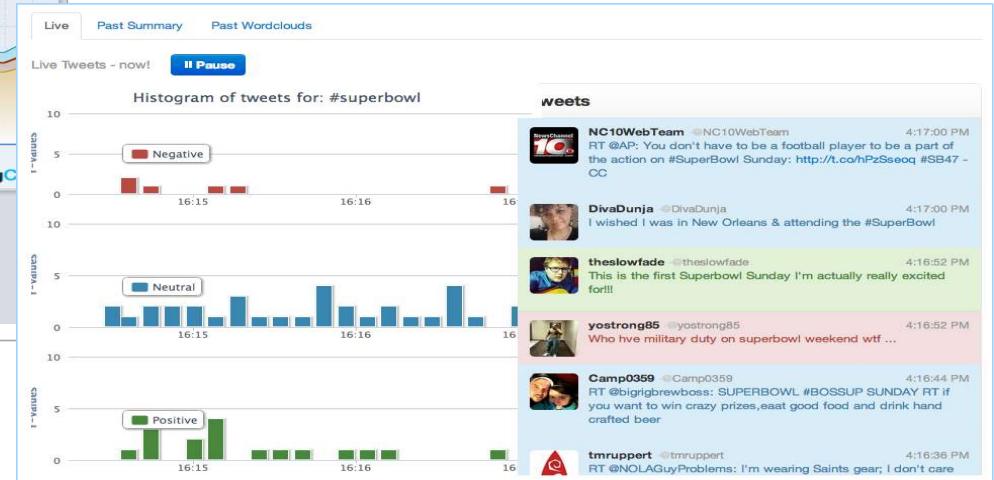
DeployR: Architecture Overview



DeployR: Examples of On-Demand Analytics



On-demand sales forecasting



Leveraging the power of Office365

Real-time social media analysis

Features of Microsoft R Server on Hadoop

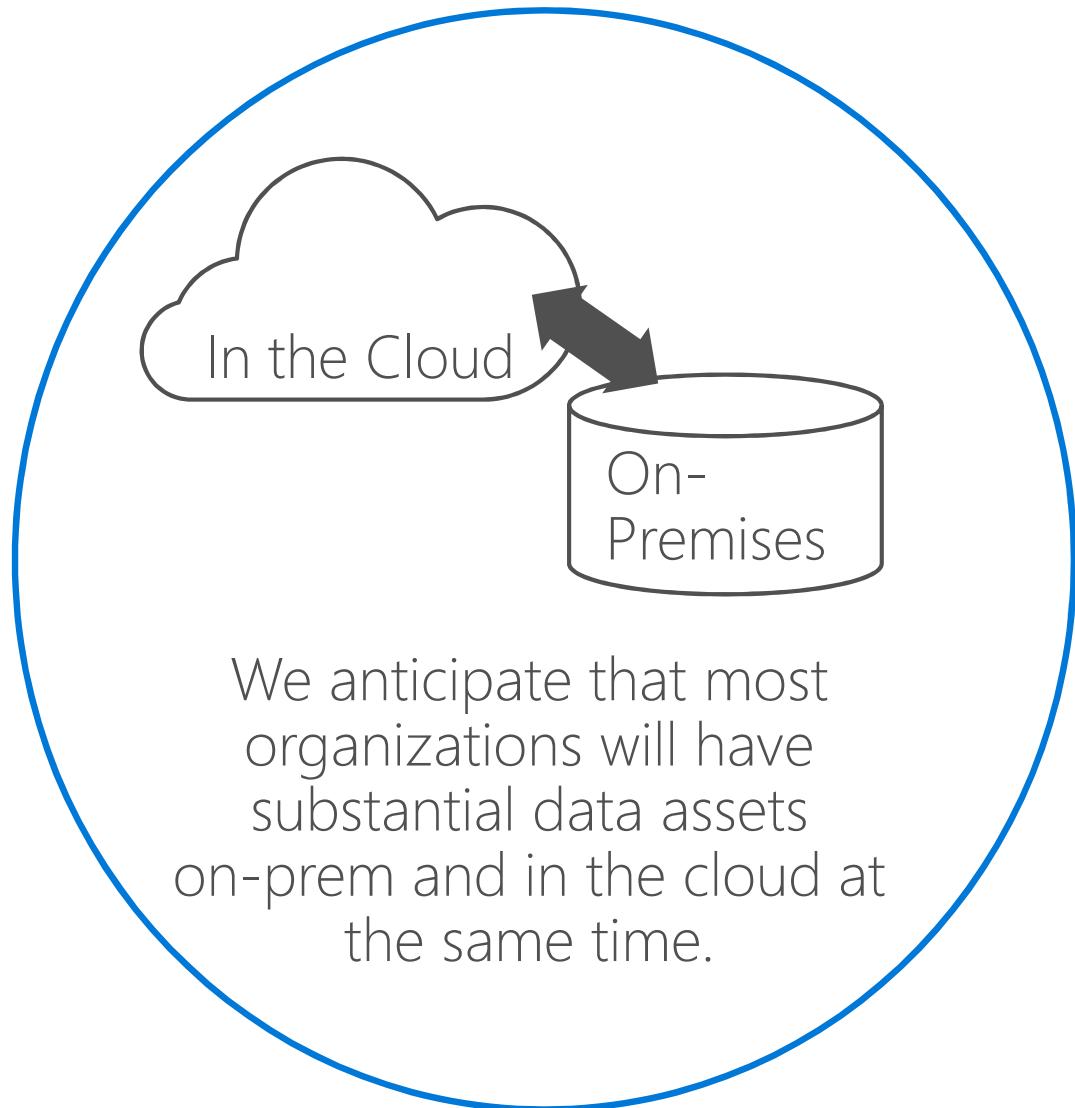
- Supports text delimited data files & Composite (sharded) XDF files
 - Reference directory path containing many files of same format!
- Utilises full parallelism of Hadoop (Mappers & Reducers) for fine-grained parallelism
 - Solve many model execution problems with full cluster parallelism (via rxExec)
- Operational
 - Configure and use YARN resource limits in Compute Context Definition
 - Supports Spark for Job-scheduling
 - Supports HDFS Caching for improved data read/write performance
- Processing platform flexibility – change compute-context and/or data source
 - Local compute context with ODBC to Hive/Impala/SparkSQL etc
 - Local compute context with local file copy
 - Local compute context with direct HDFS read (streaming)
 - Map-Reduce compute context with HDFS files/directories
 - High-Speed data extract support via Teradata Parallel Transporter
 - ODBC for small datasets
- Full Spark Compute Context (Roadmap 1H2016)
 - RDD's as a Data-Source
 - Populate RDDs from Hadoop data sources (e.g. Parquet, ORC, Avro, Composite XDF)
 - Integration and wrappers for existing open-source parallel modelling functions

Leveraging the cloud

Several Alternatives

Evolutionary, blended,
hybrid approaches will
help to avoid big
disruptions

Embracing the cloud
from On-Premises is
complicated.



Why Hybrid Cloud Analytics?

Intelligently Embrace The Cloud

Reduce Entry Barriers & Smooth Transitions

Balance Innovation with Stability & Security

Maximize Cost & Agility

End-To-End, Pre-Integrated Platforms

Independent Storage and Compute Pricing

Elastic Computational & Storage Capability

Broadest Embrace of Open Source Technologies and Pricing

Connect to Cloud-Born Data Streams

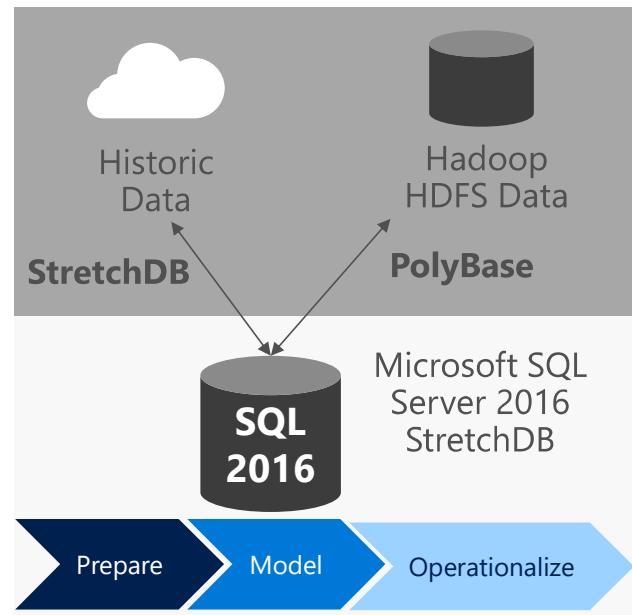
Mobile-First World

Customer-Facing Applications

Rapidly Emerging IoT

Leveraging Cloud and On-Prem Data As One

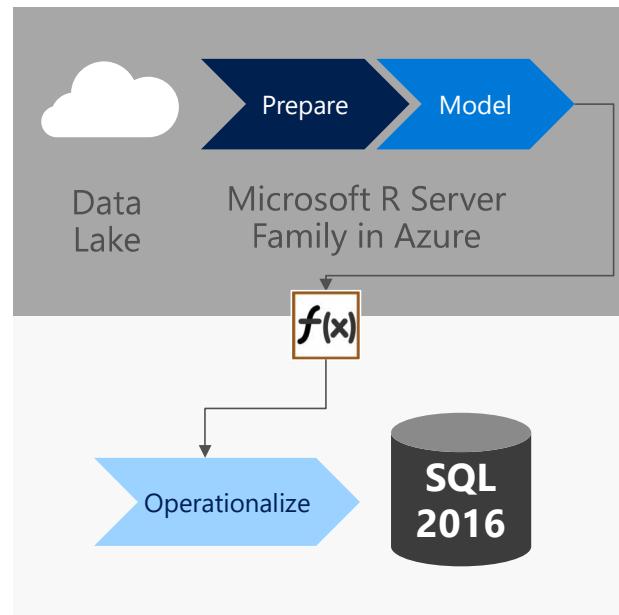
Extend SQL 2016 Into the Cloud



StretchDB Example: Cost Reduce Growth of EDW

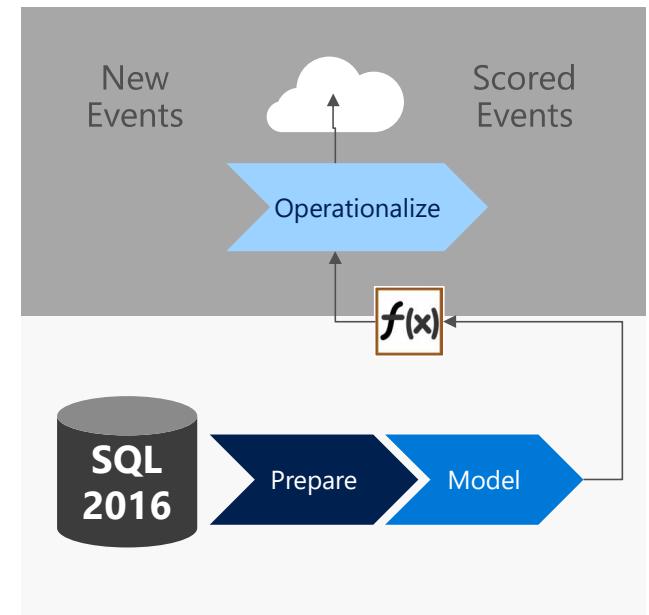
Polybase Example: Federate CRM Data with Cloud-Borne Demographics

Model using a Cloud-Based Data Lake



On-Prem Scoring Example: Purchase Propensity Prediction

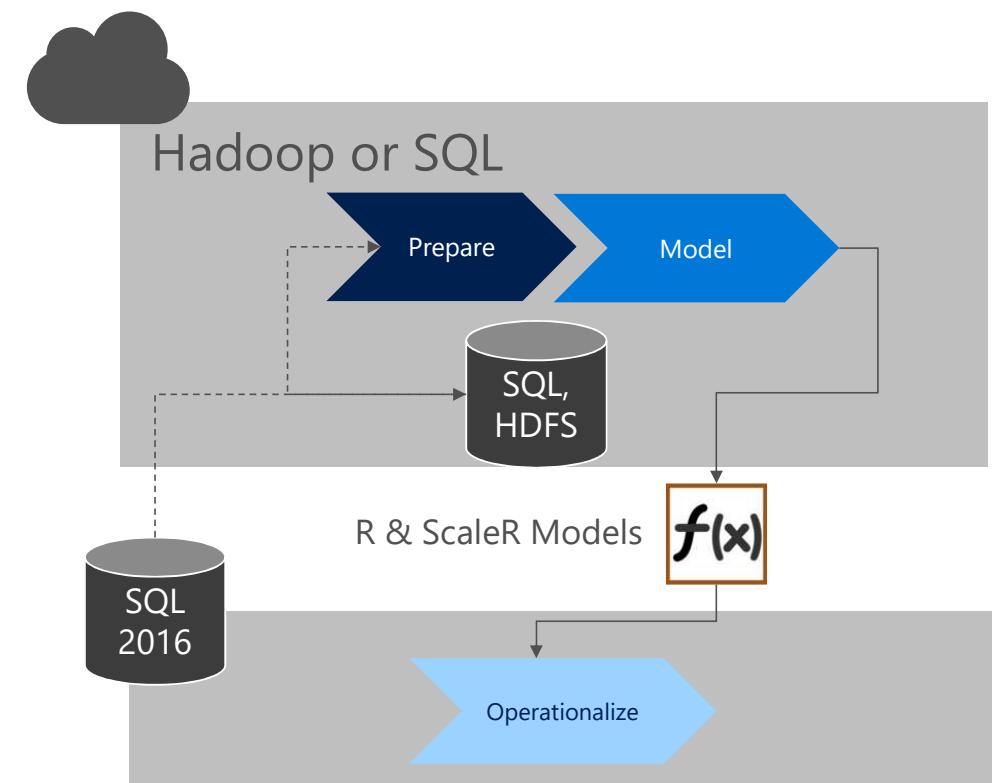
Score Streams In The Cloud



On-Prem Scoring Example: Near-Real-Time Fraud or Anomaly Detection

Deploying Hybrids

Cloud Modeling; On-Prem Deployment



Model in Azure

Capture in Data Lake
Explore & Transform in R
Deploy to SQL Server On-Prem.
SQL Scoring and BI Visualizations
Expose Web Services

Advantages

Cloud Economics & Scale for Big Data
SQL Server Stability, Privacy for Deploy

Examples

Manufacturing Optimization
Point-of-Sale Anomaly (fraud) Detection

Seizing the Opportunity: <vertical>

<Logo>



Background

*" Example graphic or
customer quote...."*



Problem

Customer



Solution