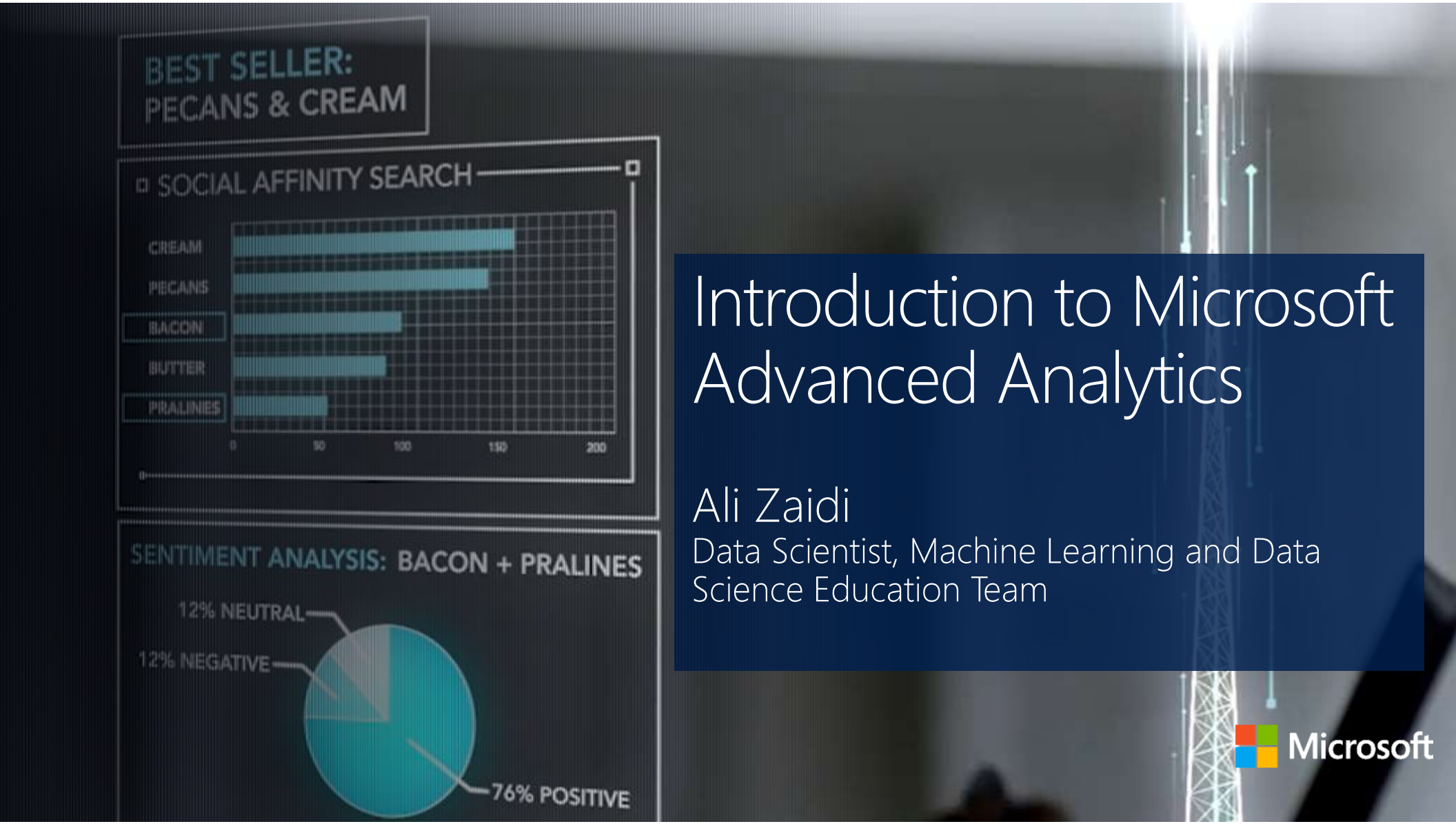


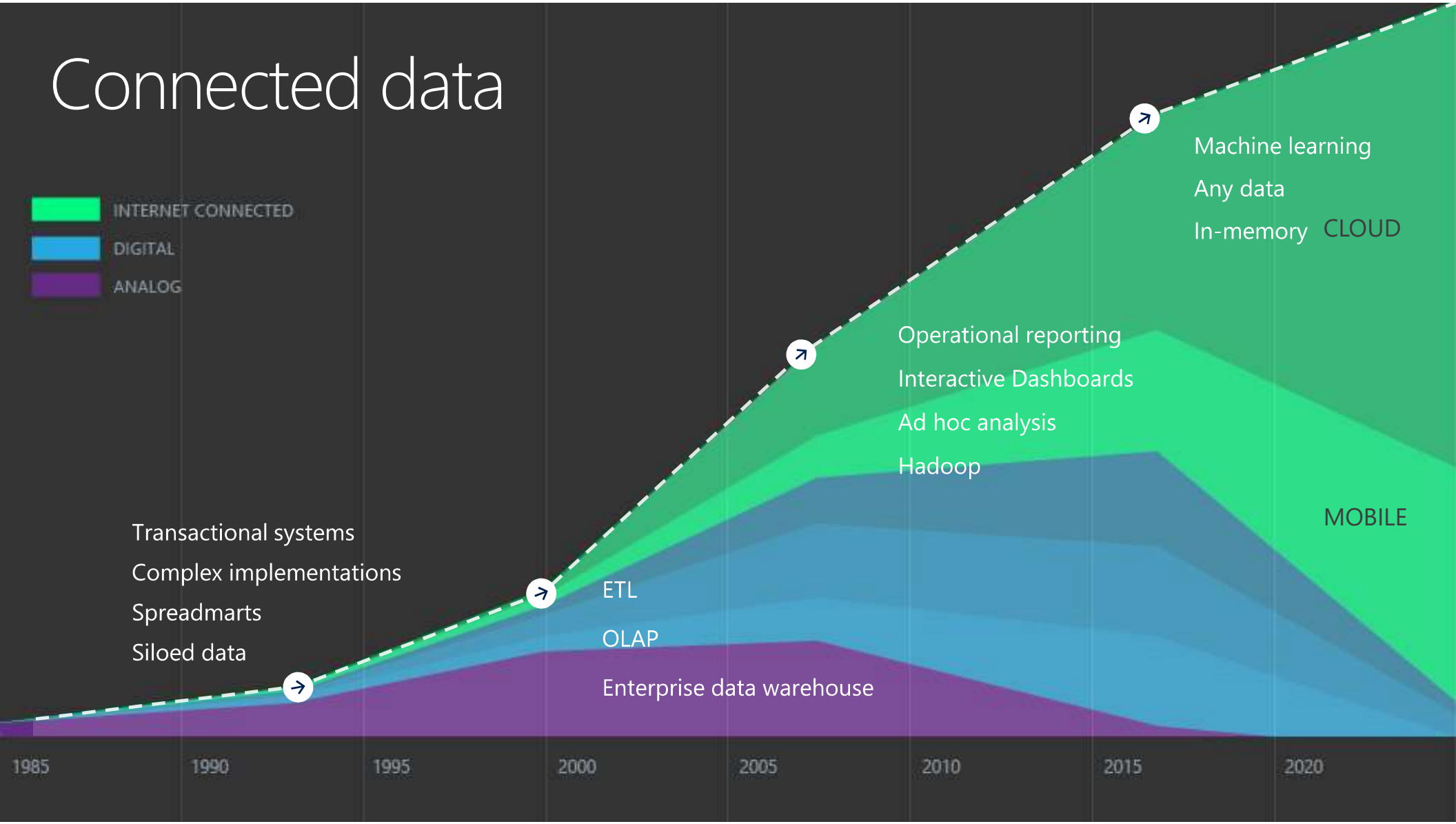

# Introduction to Microsoft Advanced Analytics

Ali Zaidi
Data Scientist, Machine Learning and Data Science Education Team

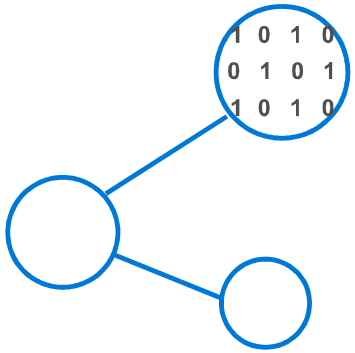
Microsoft

# Agenda

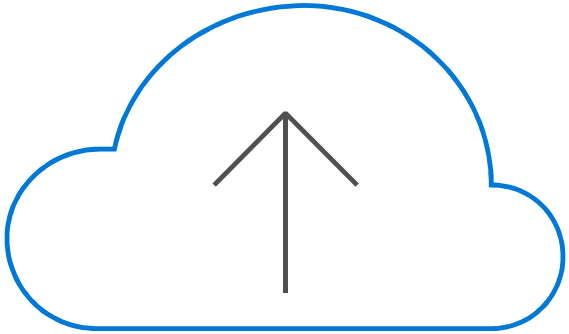
- Trends
- Opportunities
- Solutions
- Directions


Connected data
INTERNET CONNECTED
DIGITAL
ANALOG
Transactional systems
Complex implementations
Spreadmarts
Siloed data
ETL
OLAP
Enterprise data warehouse
Operational reporting
Interactive Dashboards
Ad hoc analysis
Hadoop
Machine learning
Any data
In-memory
CLOUD
MOBILE
1985
1990
1995
2000
2005
2010
2015
2020

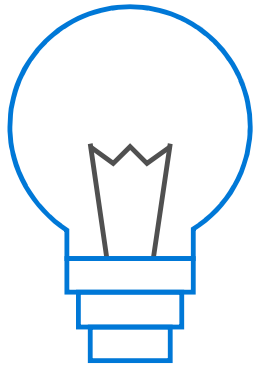
# Three major trends converging




Data

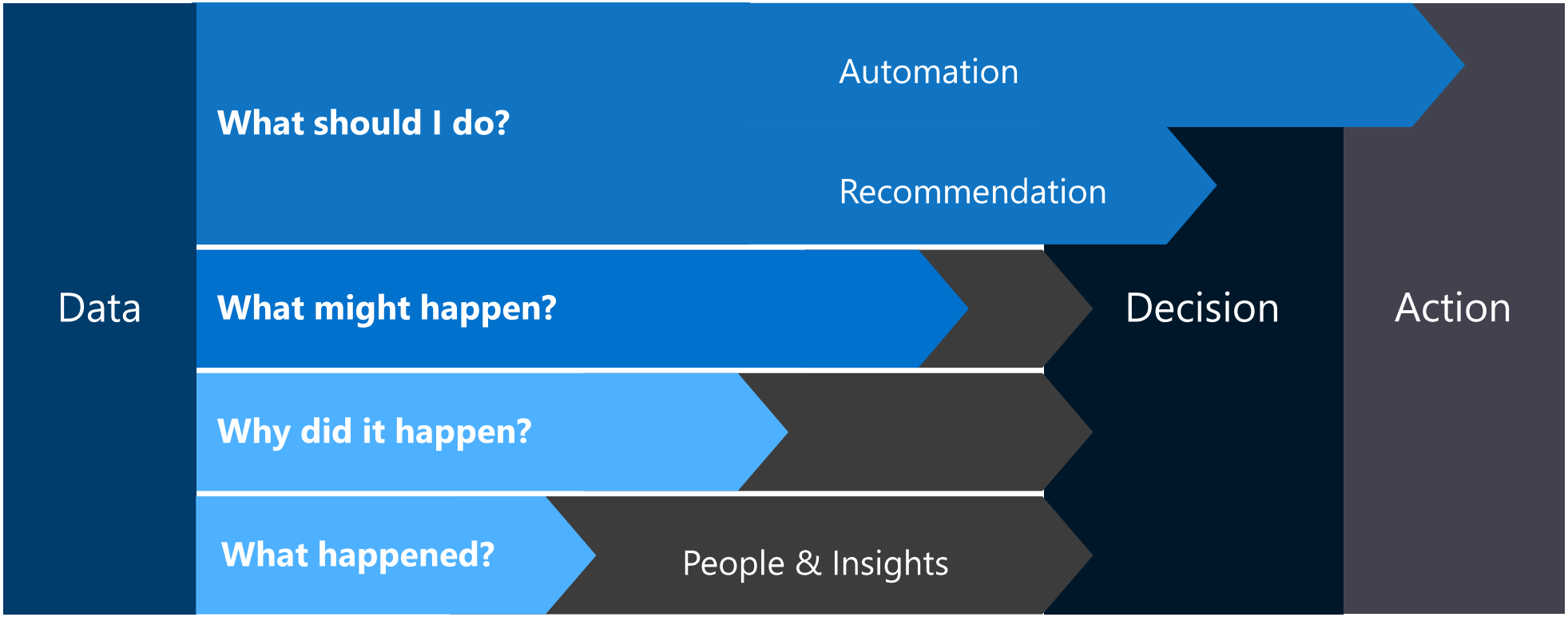


Cloud

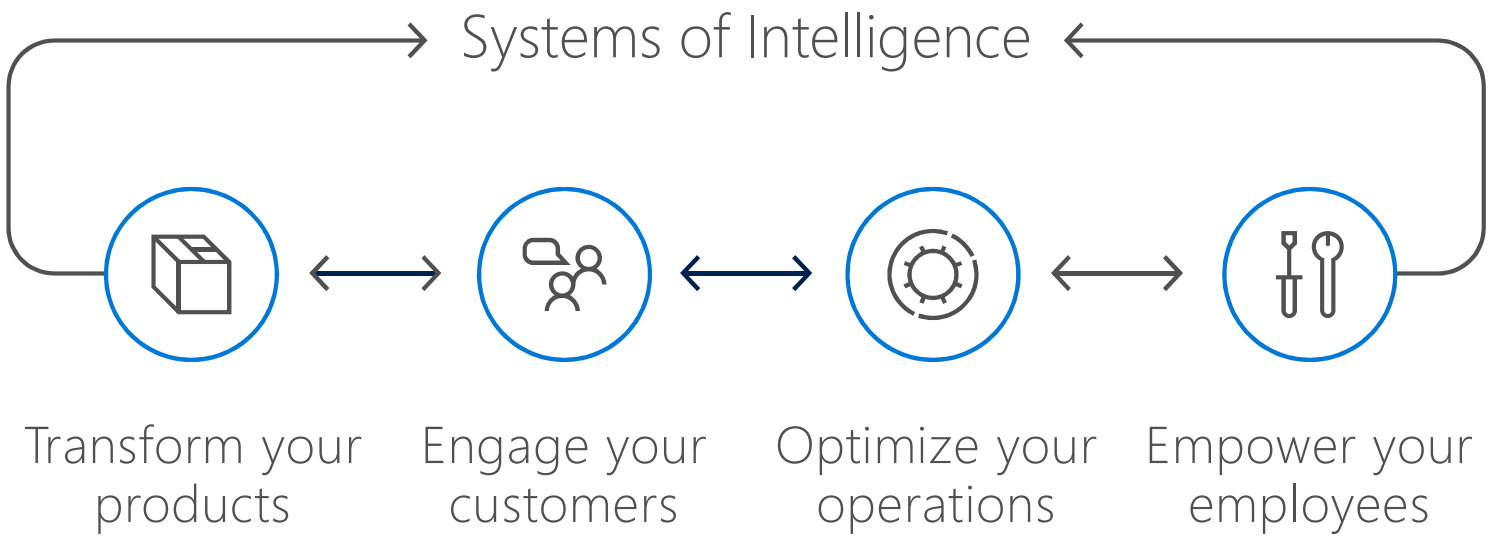


Intelligence

# 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%

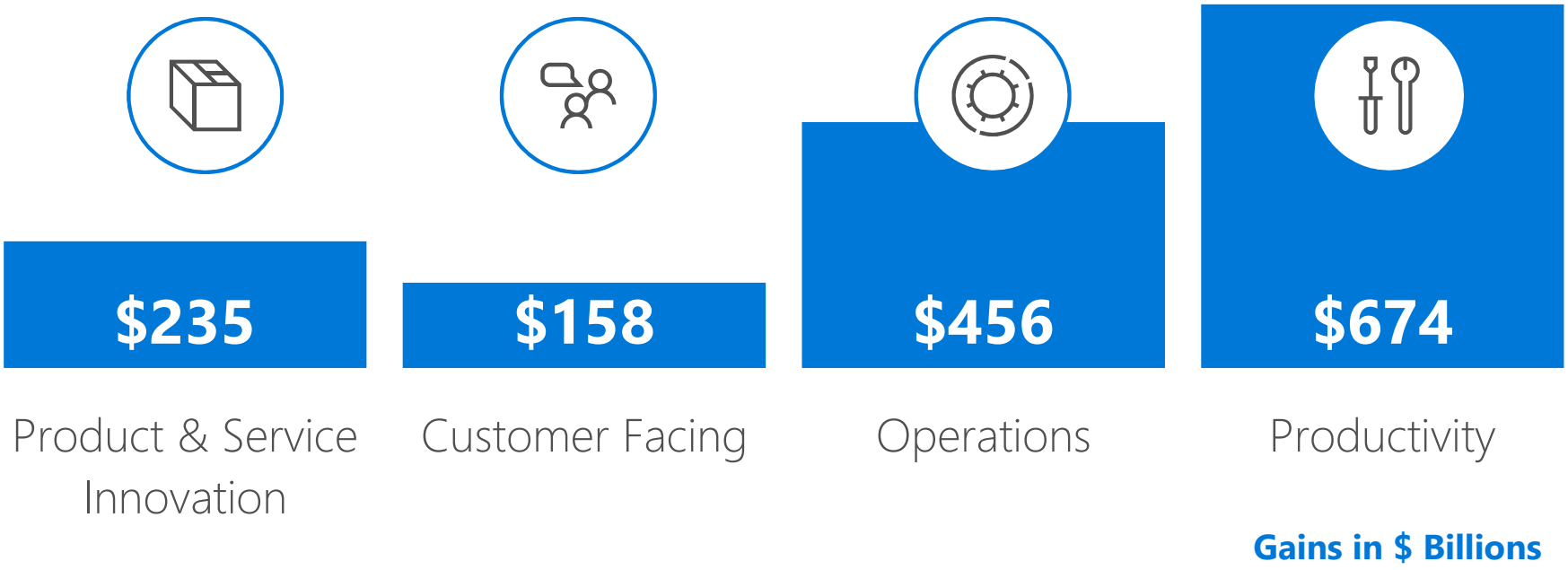
**Percent of organizations expected to have a highly profitable business unit specifically for productizing and commercializing their data by 2020**

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

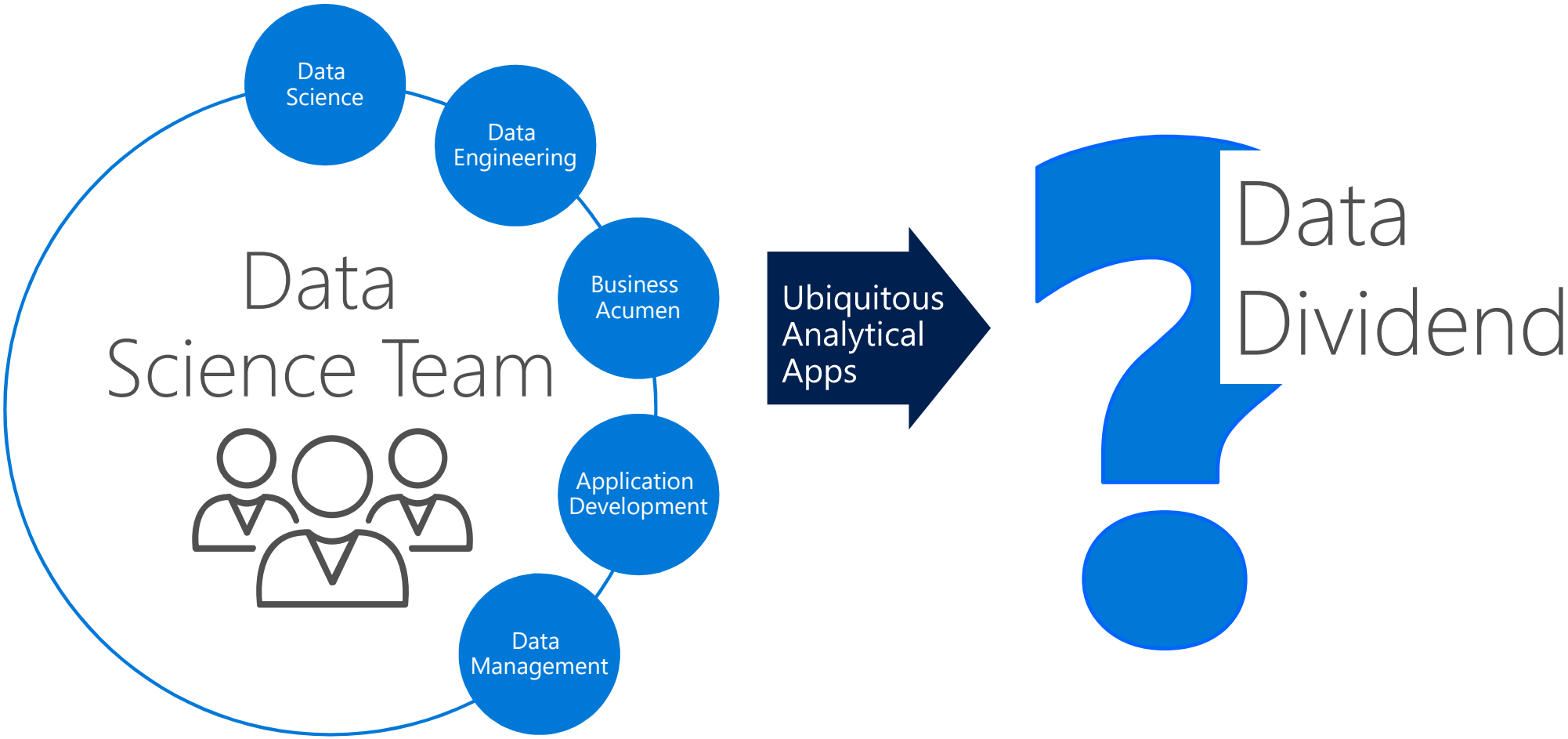
## EXAMPLE SOLUTIONS

| 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

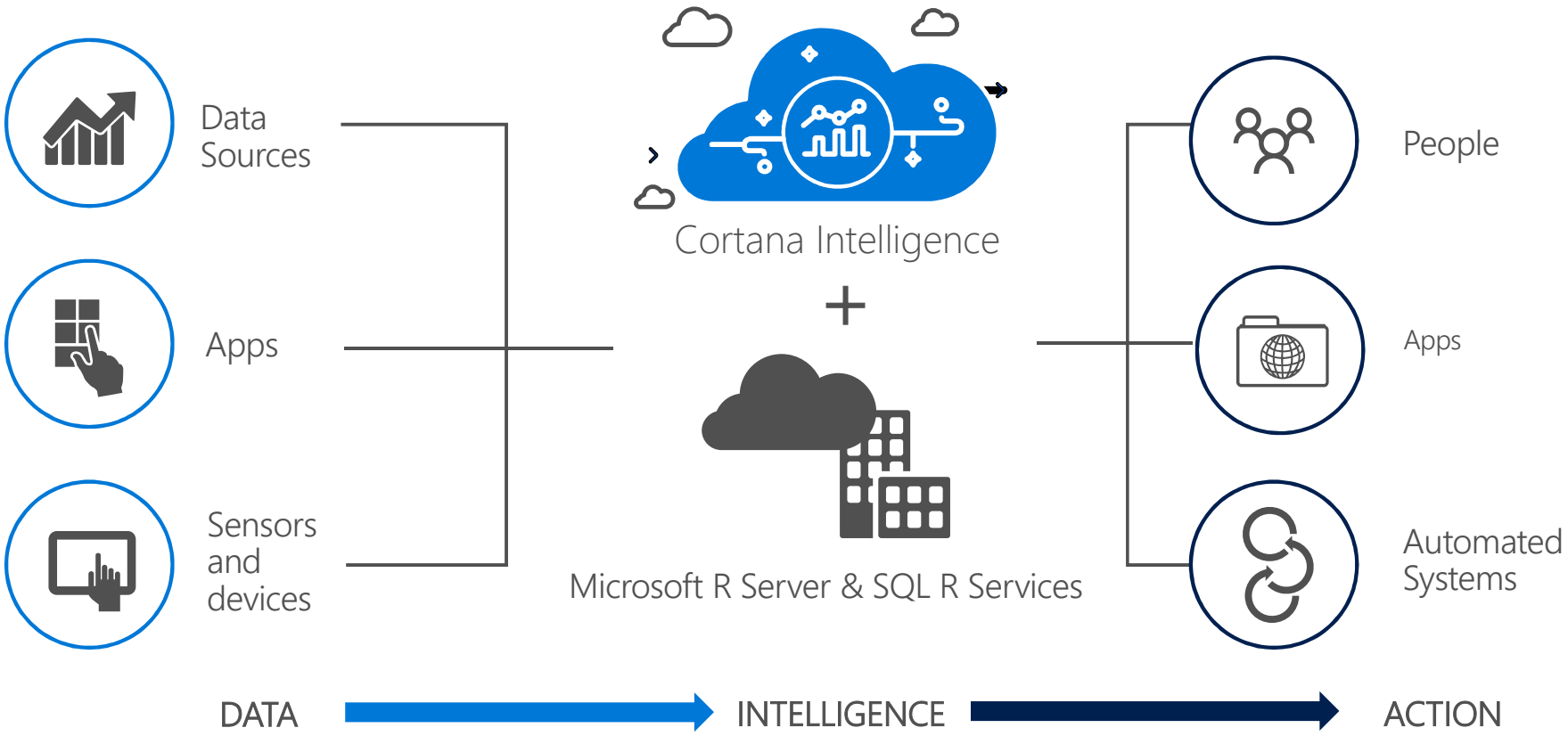
- Retail & Consumer Products
- Financial Services & Insurance
- Government
- Manufacturing
- Healthcare

# Success requires convergence of skills



# Microsoft R Server family

## From Data To Action On Premises



# What is

## Language Platform

- A statistics programming language
- A data visualization tool
- Open source

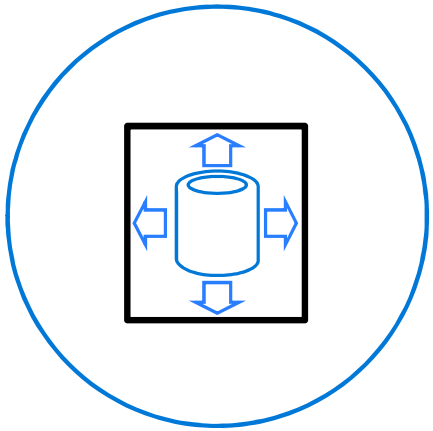
## Community

- 2.5+M users
- Taught in most universities
- New and recent grad's use it
- Thriving user groups worldwide

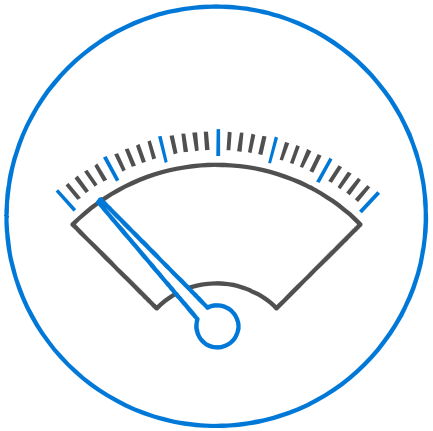
## Ecosystem

- 7000+ free algorithms in CRAN
- Scalable to big data
- Rich application & platform integration

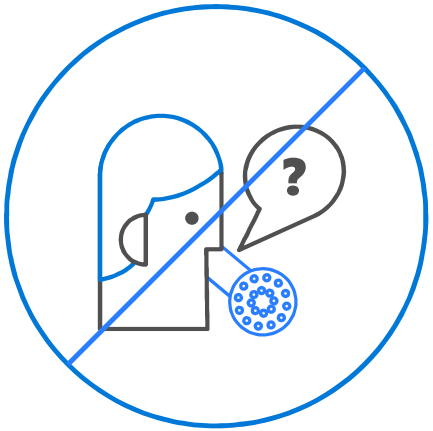
# Challenges posed by open source R



Limited Data Scale



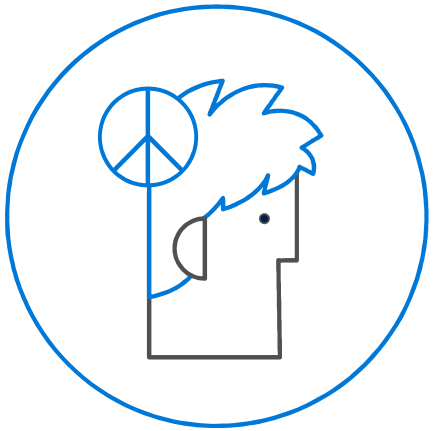
Inadequate Modeling Performance



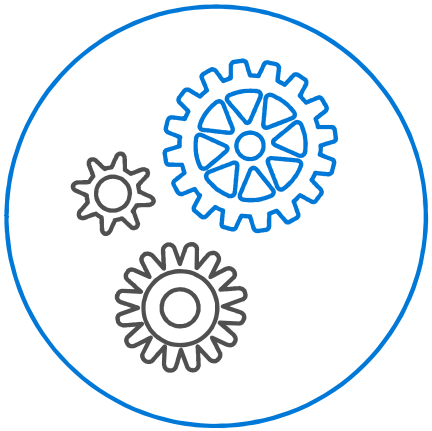

Lack of Commercial Support


Complex Deployment Processes

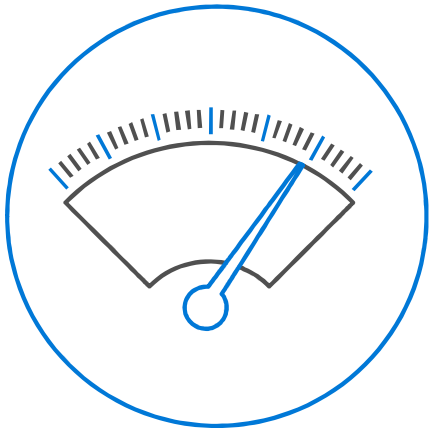
# R from Microsoft brings



Peace of mind



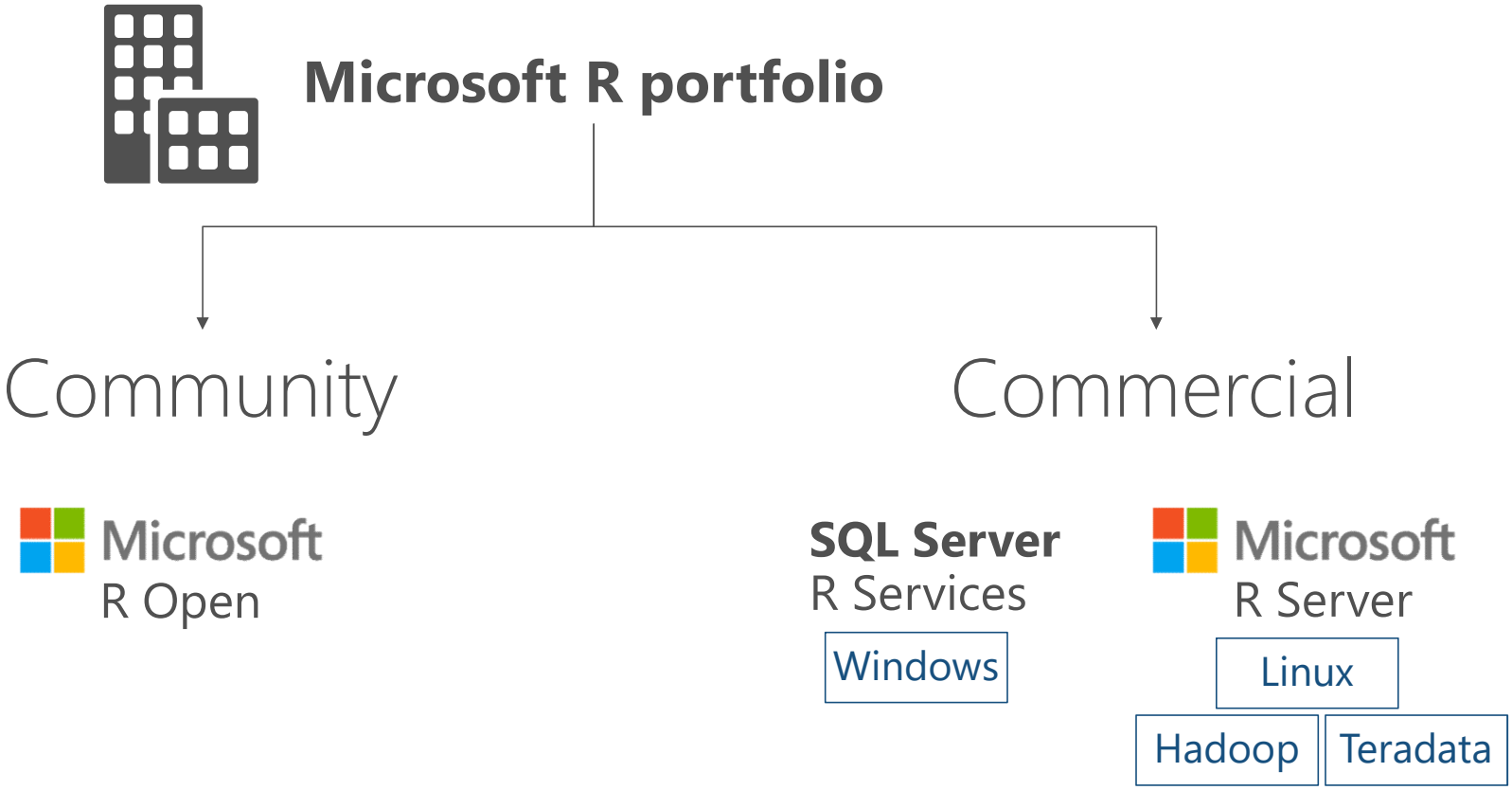
Efficiency



Speed and scalability

Flexibility and agility

# Microsoft R portfolio



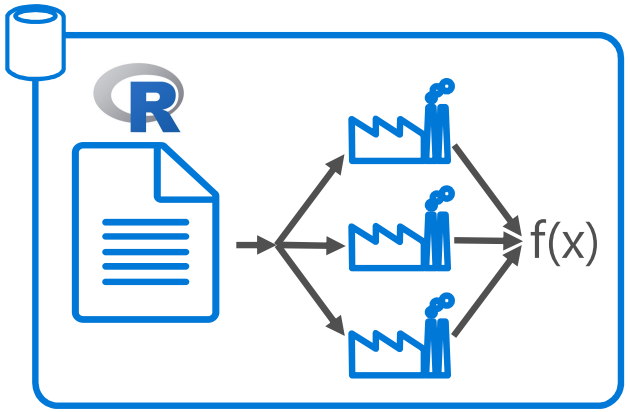
# Microsoft R Scales to Big Data for Enterprises

Escapes R's traditional memory limits

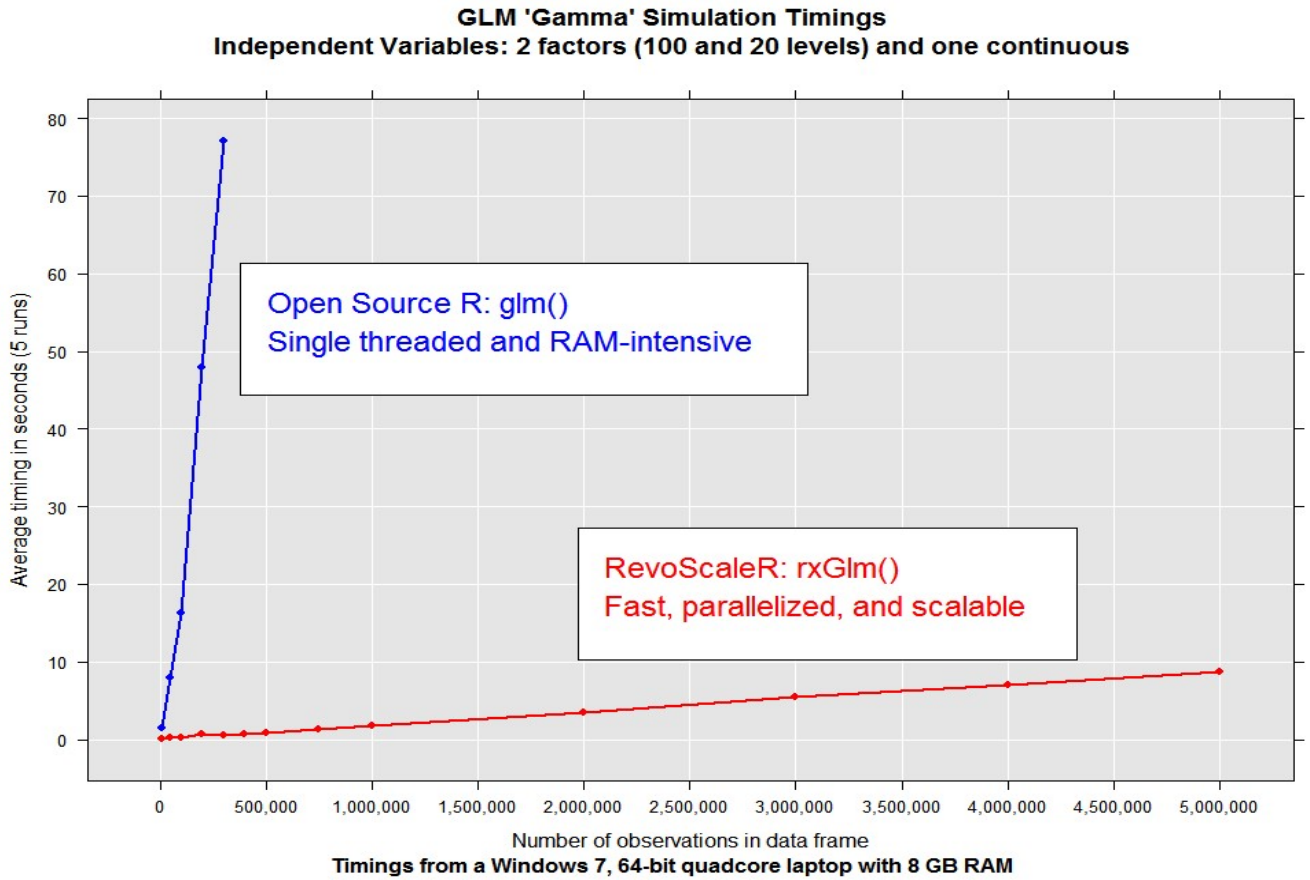
Scales predictive modeling using parallelization

Distributes computation cores & nodes

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



# Scalable algorithms

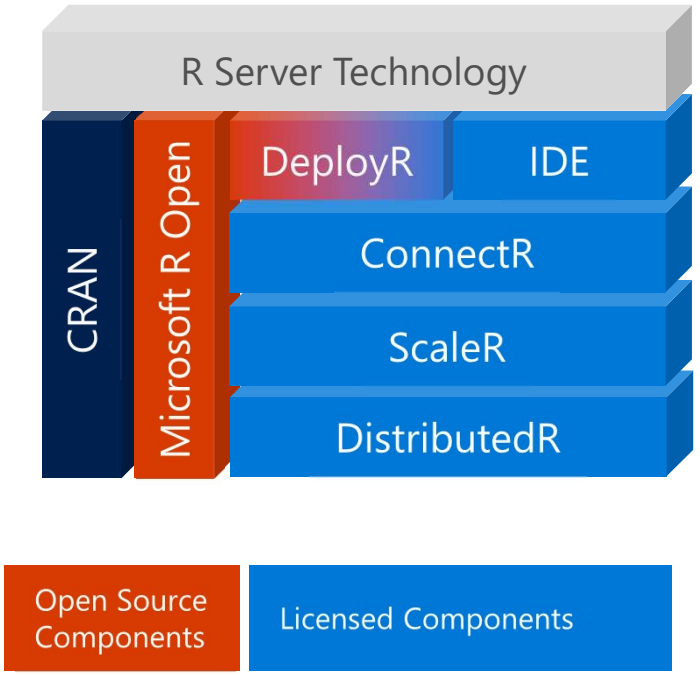


# Introducing Microsoft R Server

## 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

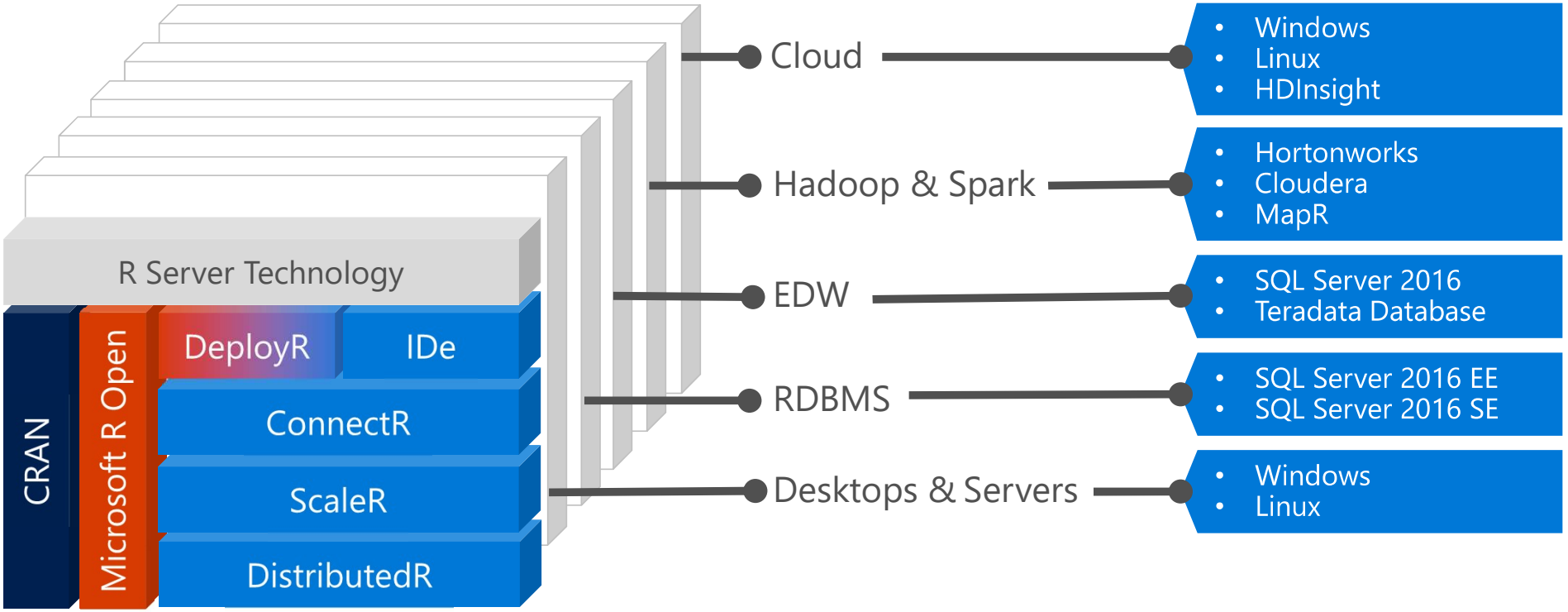
## Cost effectiveness

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

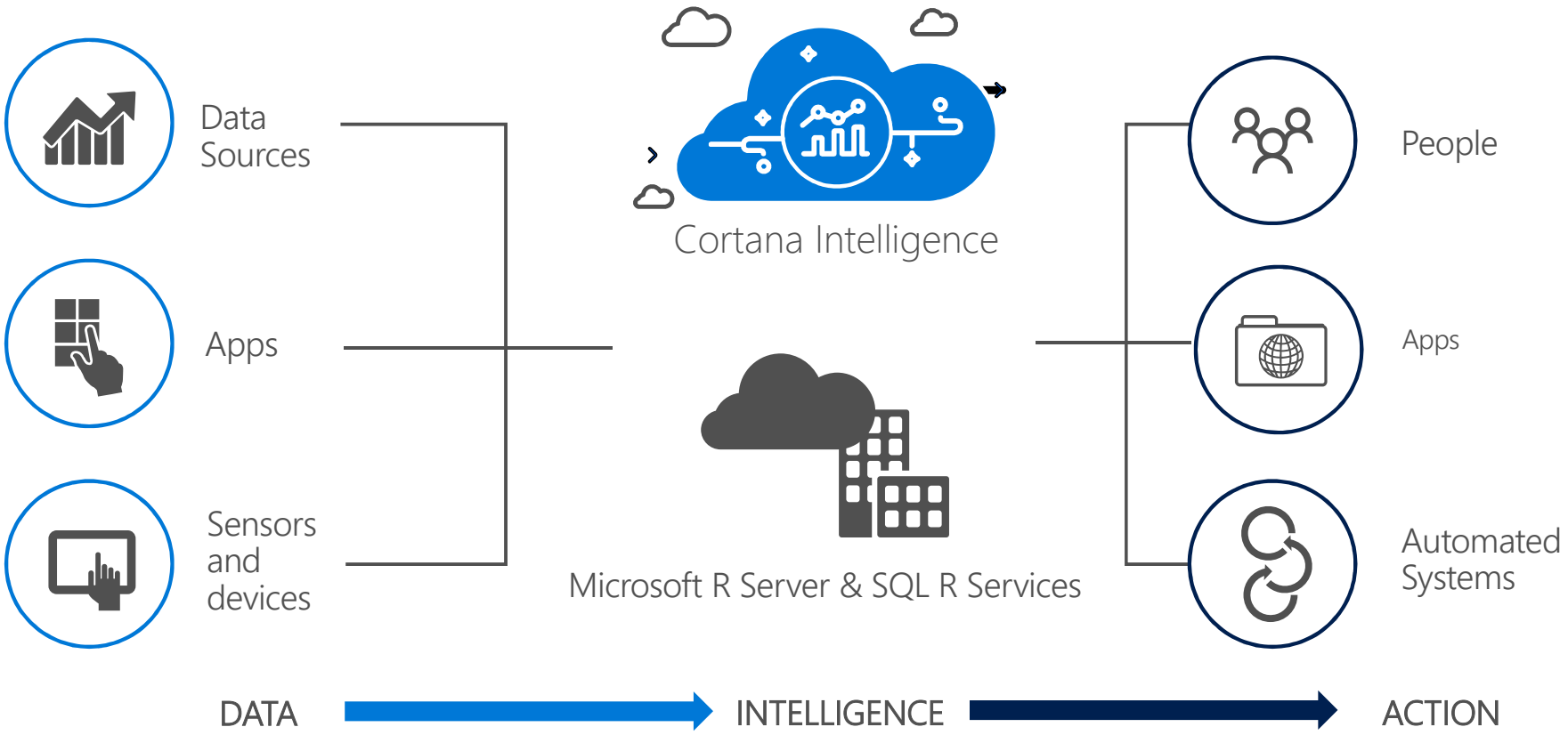
- The industry's broadest R-based platform
- Enterprise scale atop spark, Hadoop, RDBMSs & EDWs
- Freedom from memory limits
- Choice of Windows and Linux IDEs
- Stable deployment
- Write-once-deploy-anywhere portability
- Investment protection
- Hybrid cloud evolution

# Microsoft R Roadmap

**MICROSOFT CONFIDENTIAL**
**Subject to Change Without Notice**

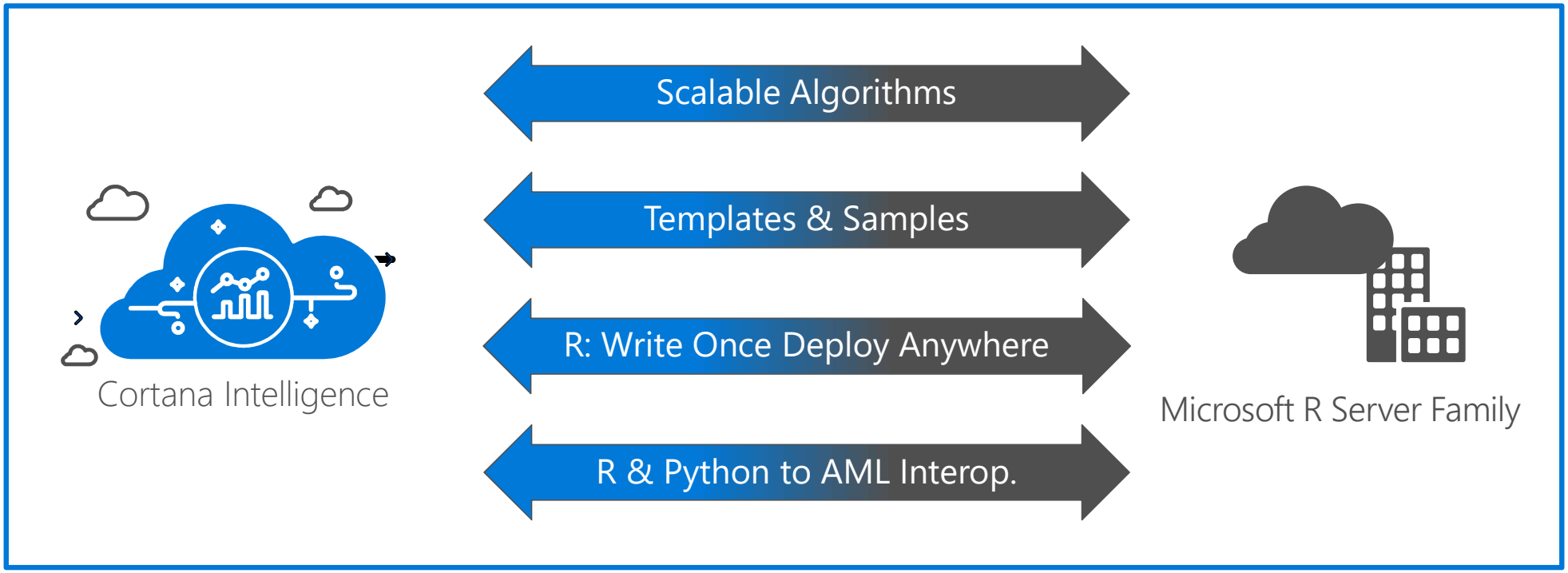
| FY16 | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| On Premises | • Remediated RRE (v 7.4.1) for:<br>  • Windows<br>  • RedHat and SuSE Linux<br>  • Hadoop (HortonWorks, Cloudera, MapR) on RedHat<br>  • Teradata EDW<br>  • HPC Pack 2008/2012+ (last feature release) | • SQL Server R Services (CTP3)<br>• New RRE features (All Platforms)<br>  • ODBC Write<br>  • Fuzzy Matching<br>• Microsoft R Open 3.2.3 | • Microsoft R Server SKUs on VL<br>• Microsoft R Server Developer Edition in VS Dev Essentials<br>• R Tools for Visual Studio (Preview) | • SQL Server 2016 GA<br>  • R Services (In database)<br>  • R Server (Standalone)<br>• Microsoft R Open 3.2.4, 3.2.5, 3.3.0<br>• Microsoft R Server for Hadoop: Install Experience++, Spark (GA), SUSE, Hadoop platform upgrades<br>• Microsoft R Server for Teradata (GA)<br>• Microsoft R Server for Red Hat Linux (GA); 7.1<br>• Microsoft R Server for SUSE Linux (GA)<br>• DeployR: DB choices, security fixes, new install exp.<br>• New ML Algorithms for Windows (TAP)<br>• Microsoft R Client (Windows)<br>• R Tools for Visual Studio (GA) |
| Cloud | • RRE VMs on Azure – Preview (Windows, CentOS) | • RRE VMs on Azure w/Datalake Access (Preview) (Window, CentOS)<br>• RRE on Azure HDI w/Datalake Access (Preview) (Windows, Ubuntu) | • Azure VM for Microsoft R Server on Windows, Linux (GA)<br>• Azure Data Science VM for Windows (GA)<br>• Microsoft R Server for HDInsight, with Spark Support (Preview)<br>• Azure ML Jupyter R notebooks (Preview) | • Azure HDInsight Premium with Microsoft R Server for Hadoop, Spark Support (GA)<br>• Azure ML:<br>  • Jupyter R notebooks (GA) |

# From data to decisions to action with Microsoft



# Hybrid analytics platform

## Convergence with Flexibility



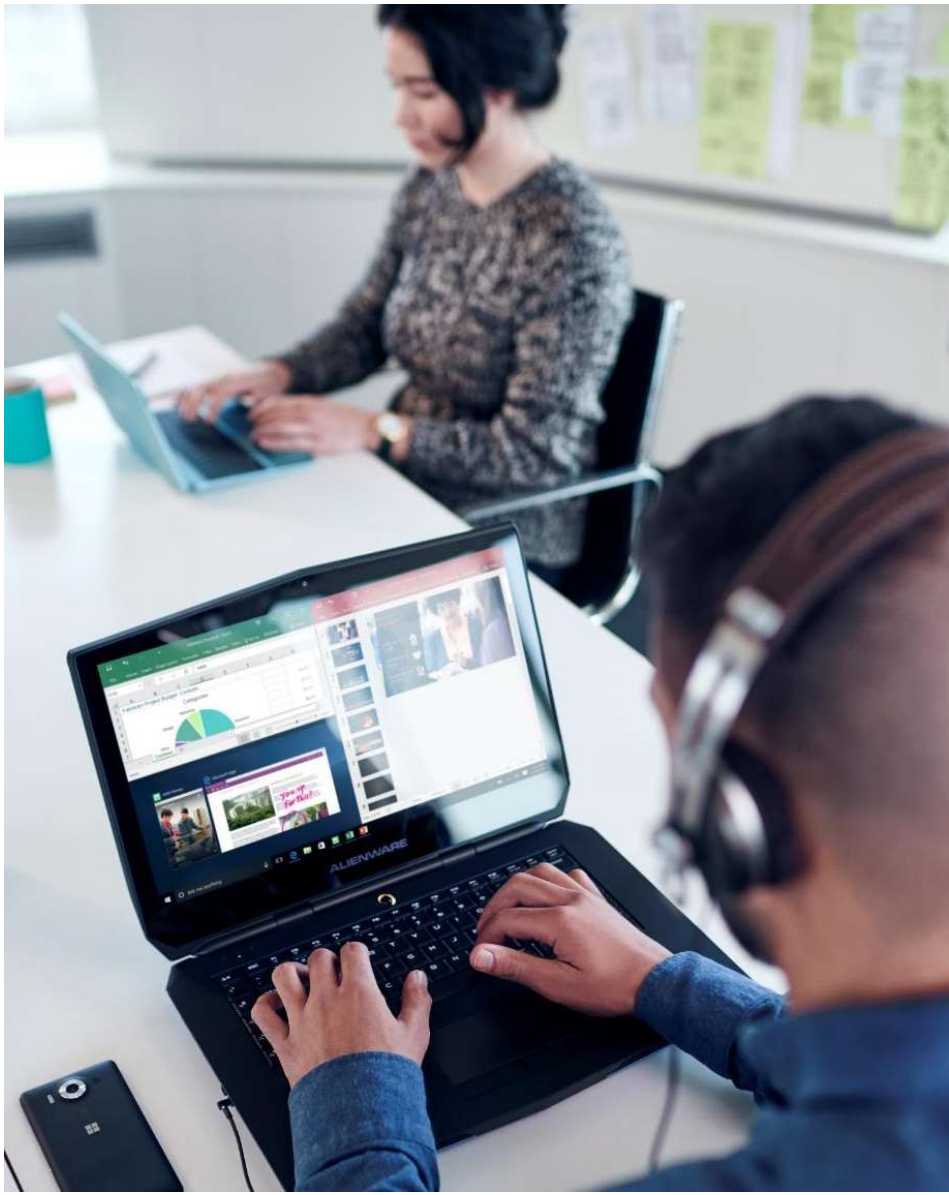
# Microsoft Advanced Analytics address barriers

| | |
|---|---|
| Broaden The Talent Pool | • Democratize Data Science<br>• Skill Re-Use |
| Increase Productivity | • Transparent Scaling<br>• Facilitate Collaboration |
| Modernize Infrastructure | • Decouple Data Science from Platforms<br>• Leverage Hybrid Cloud Architecture |
| Maximize Innovation | • Accelerate Experimentation<br>• Streamline Deployment |
| Drive Down TCO | • Embrace Open Source<br>• Evolutionary Path to Cloud |

# Microsoft's roadmap for analytics

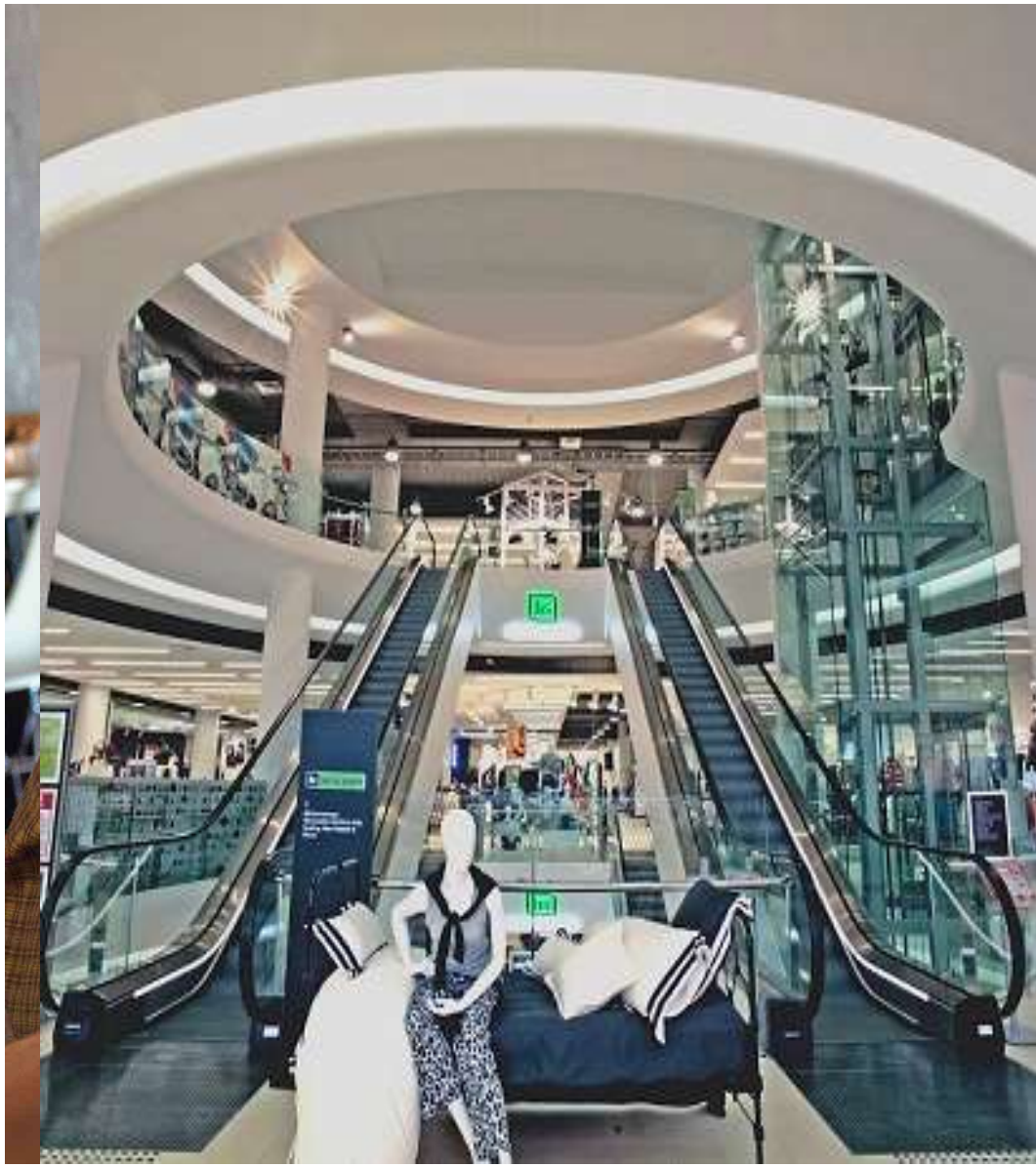
## In 3 Years, we will help you achieve:

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



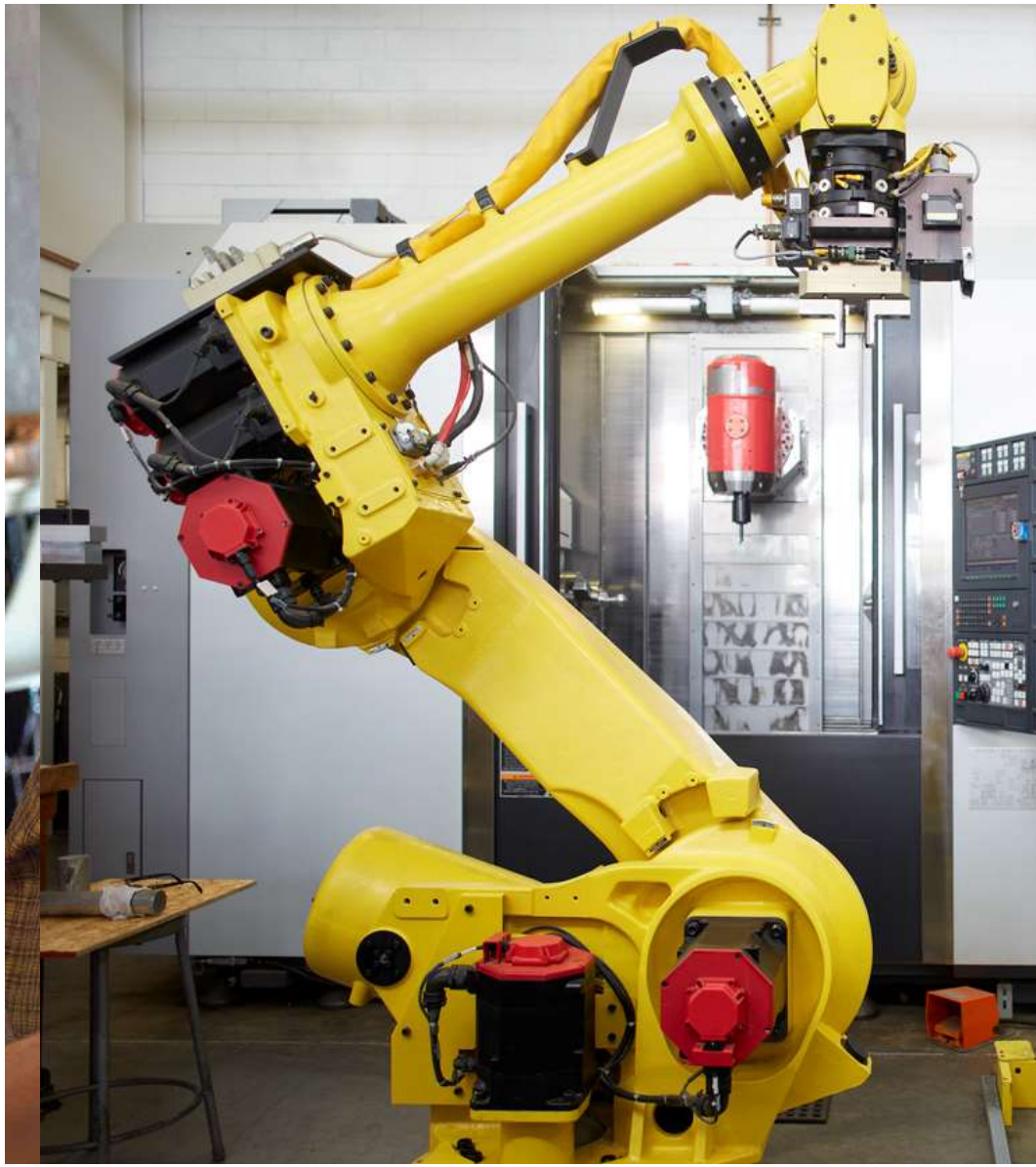
# Retail & CPG

- Execution Optimization
- Shrinkage & Loss Prevention
- 360 Customer 360 Experience



# Manufacturing

- Predictive Maintenance
- Supply Chain Optimization
- Quality



# Financial Services

- Risk Analytics
- Fraud Prevention
- 360 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 server 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

IntegriChain®

*" 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."*

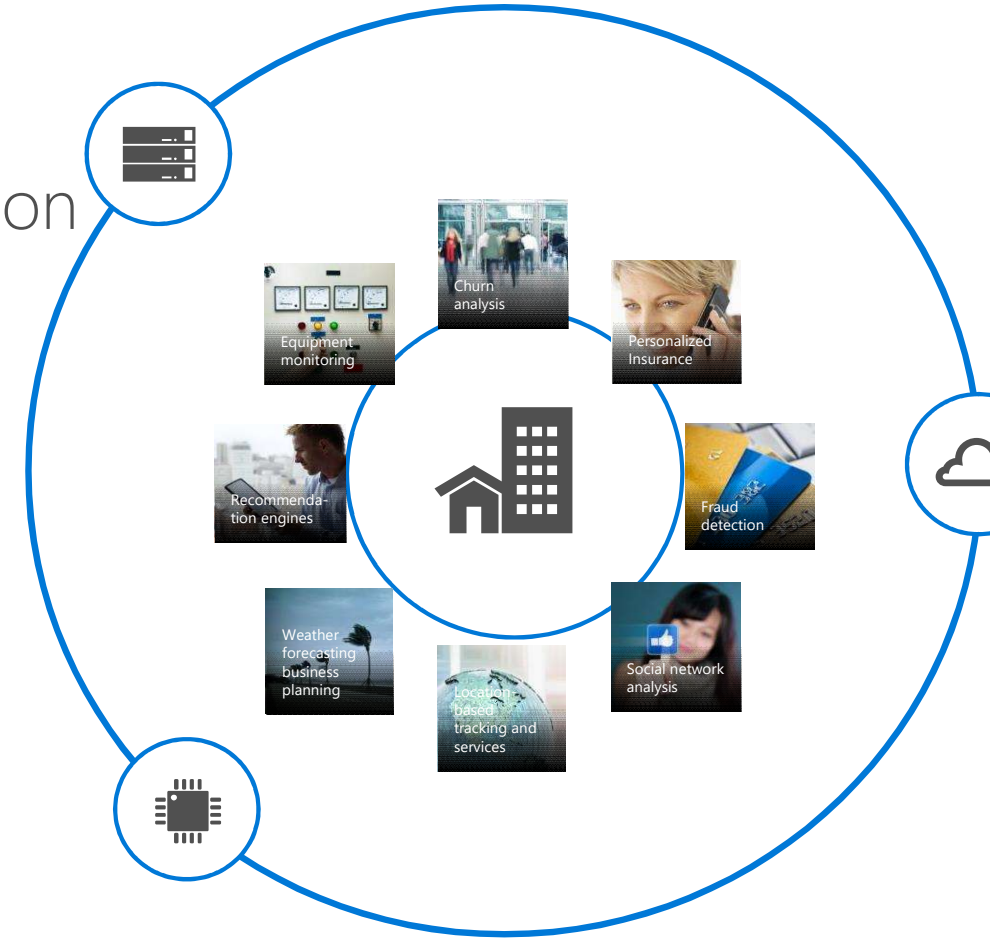
*VP of Application Development*

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

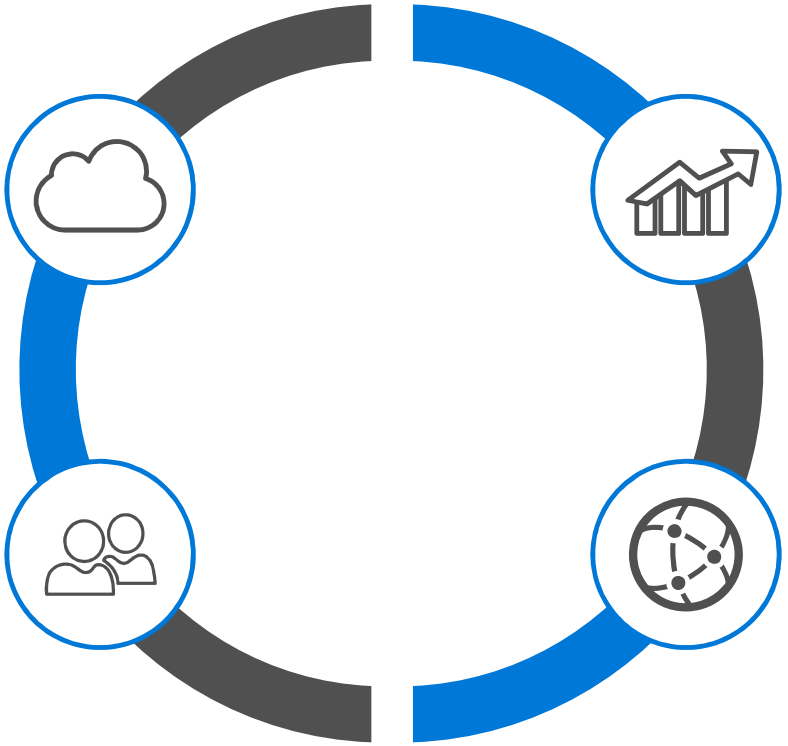
# Transformational Trends

## cloud computing

2011 ➔ 2016 5x increase

## data explosion

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



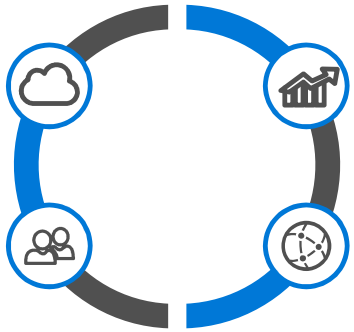
## emerging data science talent

Universities filling 300,000 US talent gap

## open source

e.g. R and Python

# 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

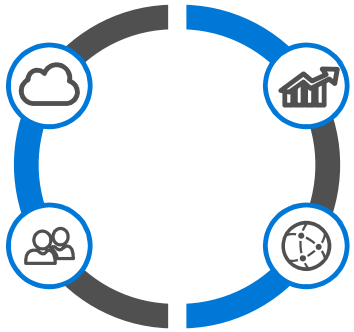
**open source**

- 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



**cloud computing**

- Flexible Path to the Cloud
- Mature, Geographically Dispersed Infrastructure
- Hybrid clouds with cloud-on-prem interoperability

**data explosion**

- Elastic Scale On-Prem & In the Cloud
- Perceptual Algorithms
- Leading-edge Machine Learning

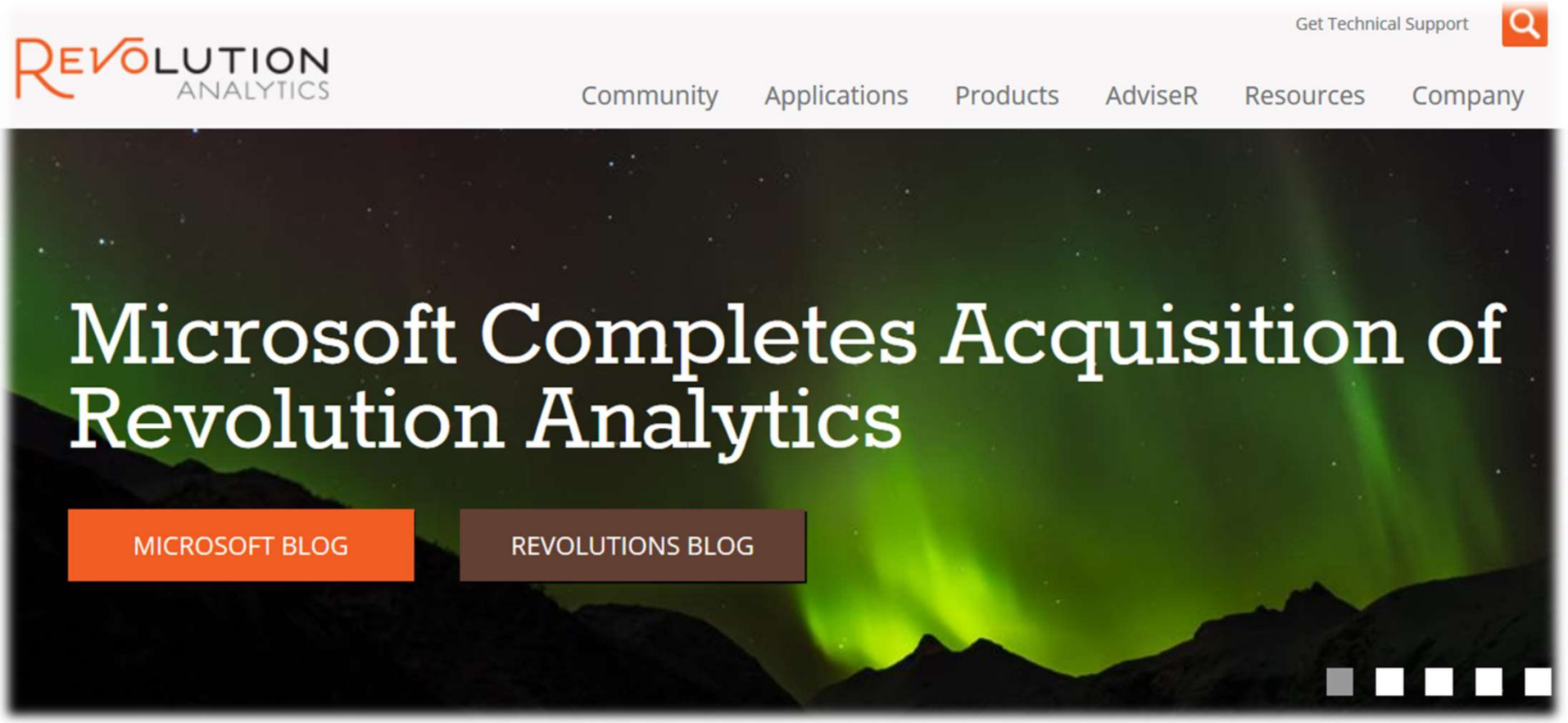
**open source**

- Comprehensive R Integration
- Full Commercial Support
- Integration with Linux and Hadoop

**emerging data science talent**

- Choice of IDEs & Graphical Tools
- Collaboration Across Stakeholders
- Template & Sample Galleries

**April 6, 2015**




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

# A brief history of R

**1993**
Research project in Auckland, NZ

**1995**
Open source

**1997**
R-core

**2000**
R-1.0.0

**2003**
R Foundation

**2004**
First UseR!

**2009**
New York Times

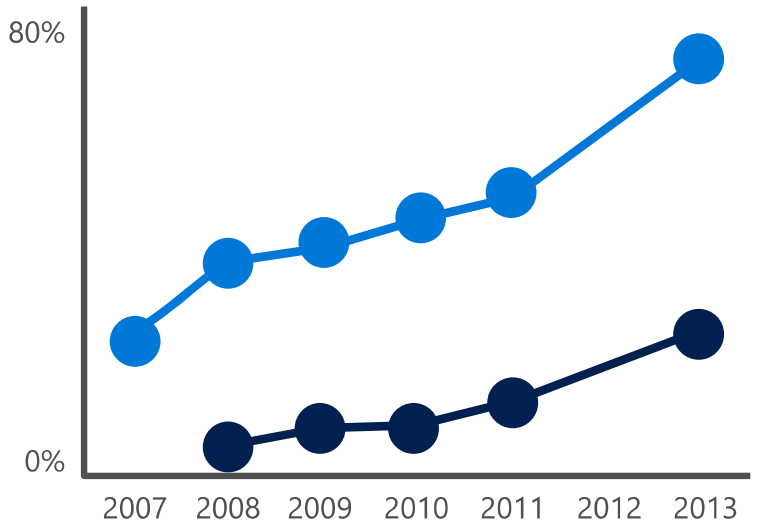
**2015**
R-3.2.0
R Consortium

Photo credit: Robert Gentleman

# R's popularity is growing rapidly

**R Usage Growth**

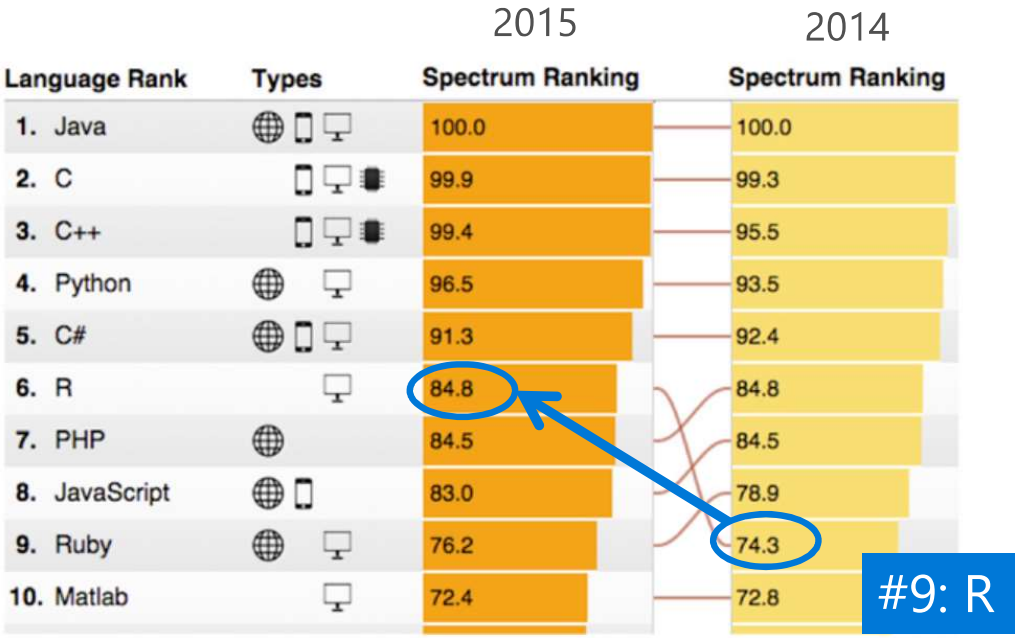
Rexer Data Miner Survey, 2007-2013




[Rexer Data Miner Survey]

**Language Popularity**

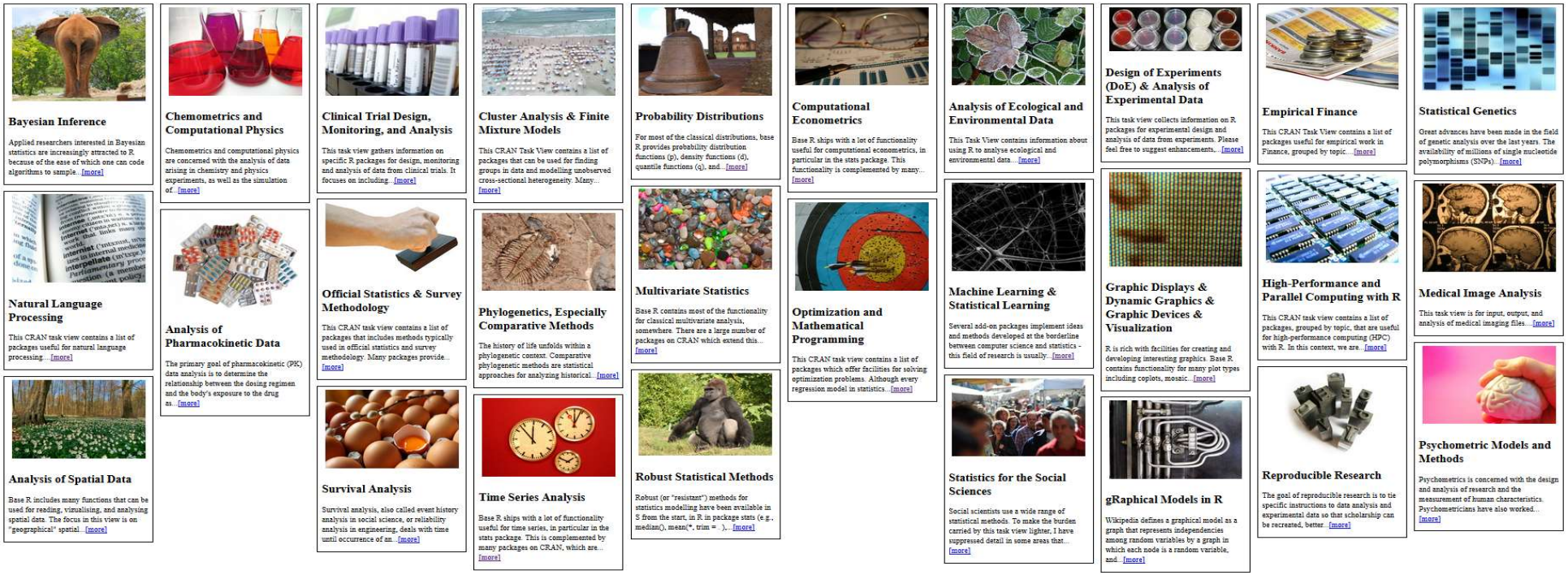
IEEE Spectrum Top Programming Languages




[IEEE Spectrum July 2015]

# CRAN

## The Comprehensive R Archive Network



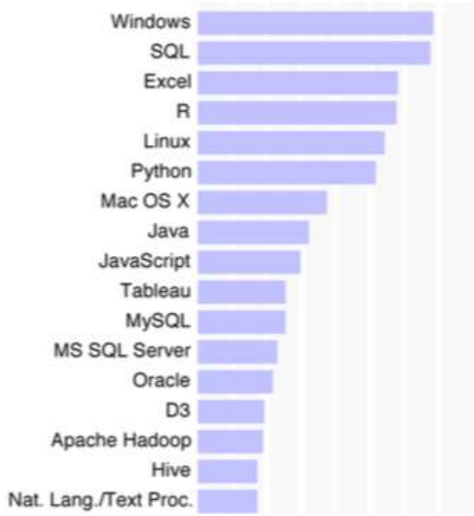

In addition to CRAN, Bioconductor, GitHub, others distribute R packages

# Standing on the Shoulders of Giants

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

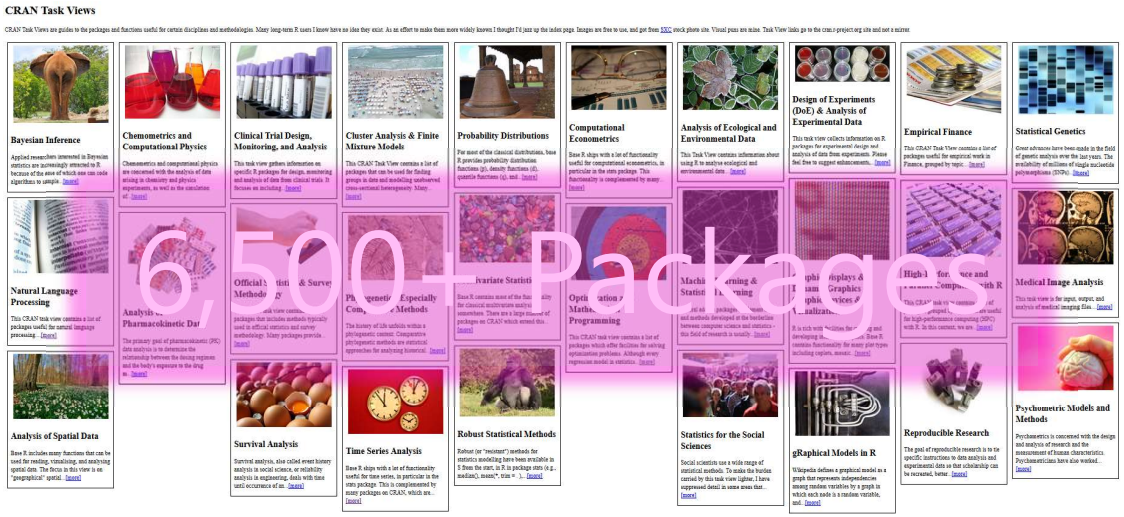
### Tool Use for Data Science

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




### CRAN
### The Comprehensive R Archive Network

Resources For All Fields of Analysis




**+ Bioconductor**
**+ Github**

# Enterprises Press the Limits of Open Source R

## Enterprise Data Scale Can Exceed Can R's Capacity

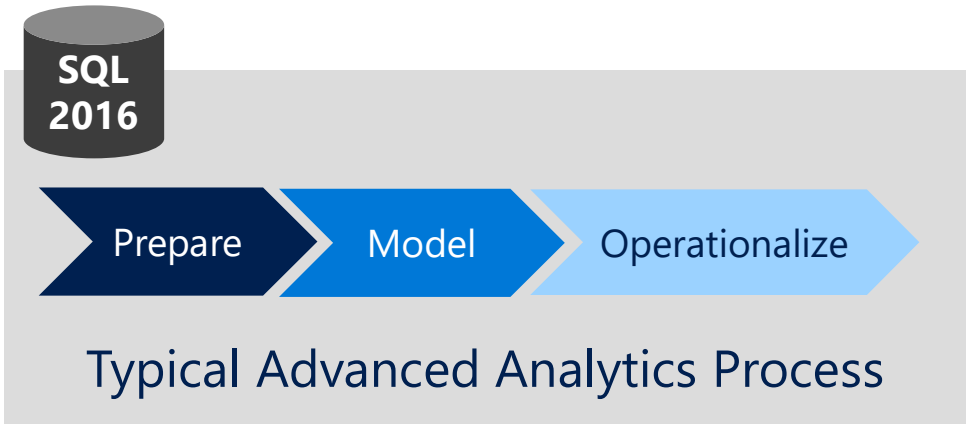
Memory-Based Data Access Model
Lack of Parallel Computation
Requires Data Movement Prior to Analysis

## 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




Typical Advanced Analytics Process

## Model & Deploy In SQL16

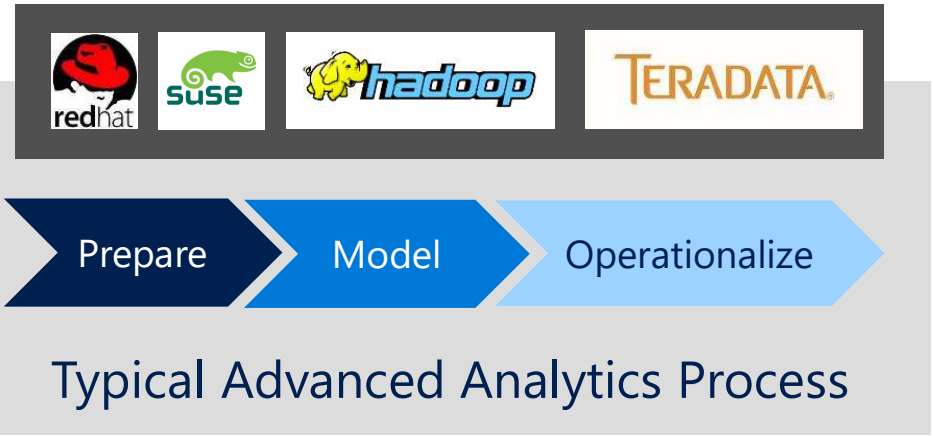
Support Entire Analytics Lifecycle
Run R Inside SQL 2016 using IDE
Embed R in T-SQL or run stored proc's

## 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




Typical Advanced Analytics Process

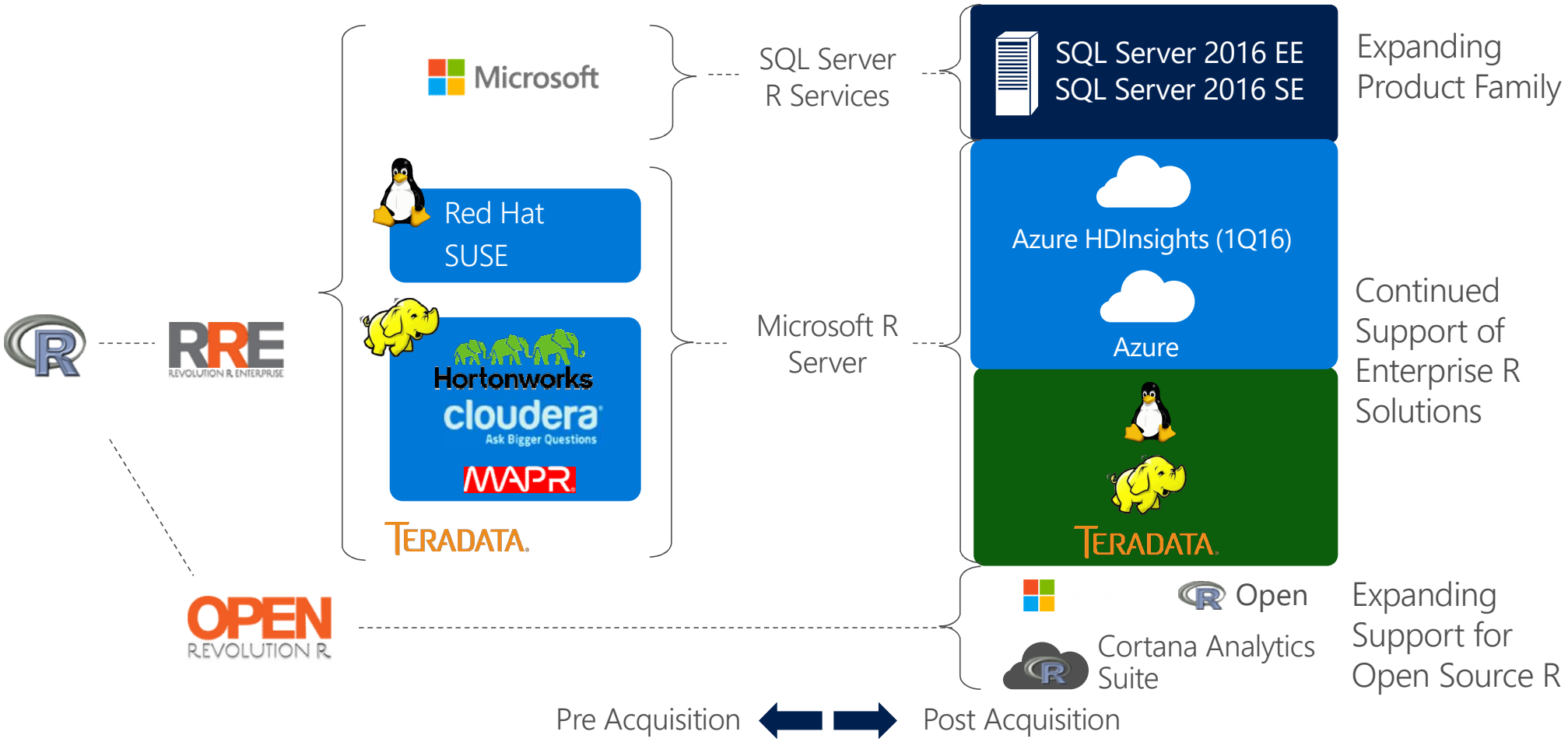
## Multi-Platform Scalable R

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

## Advantages

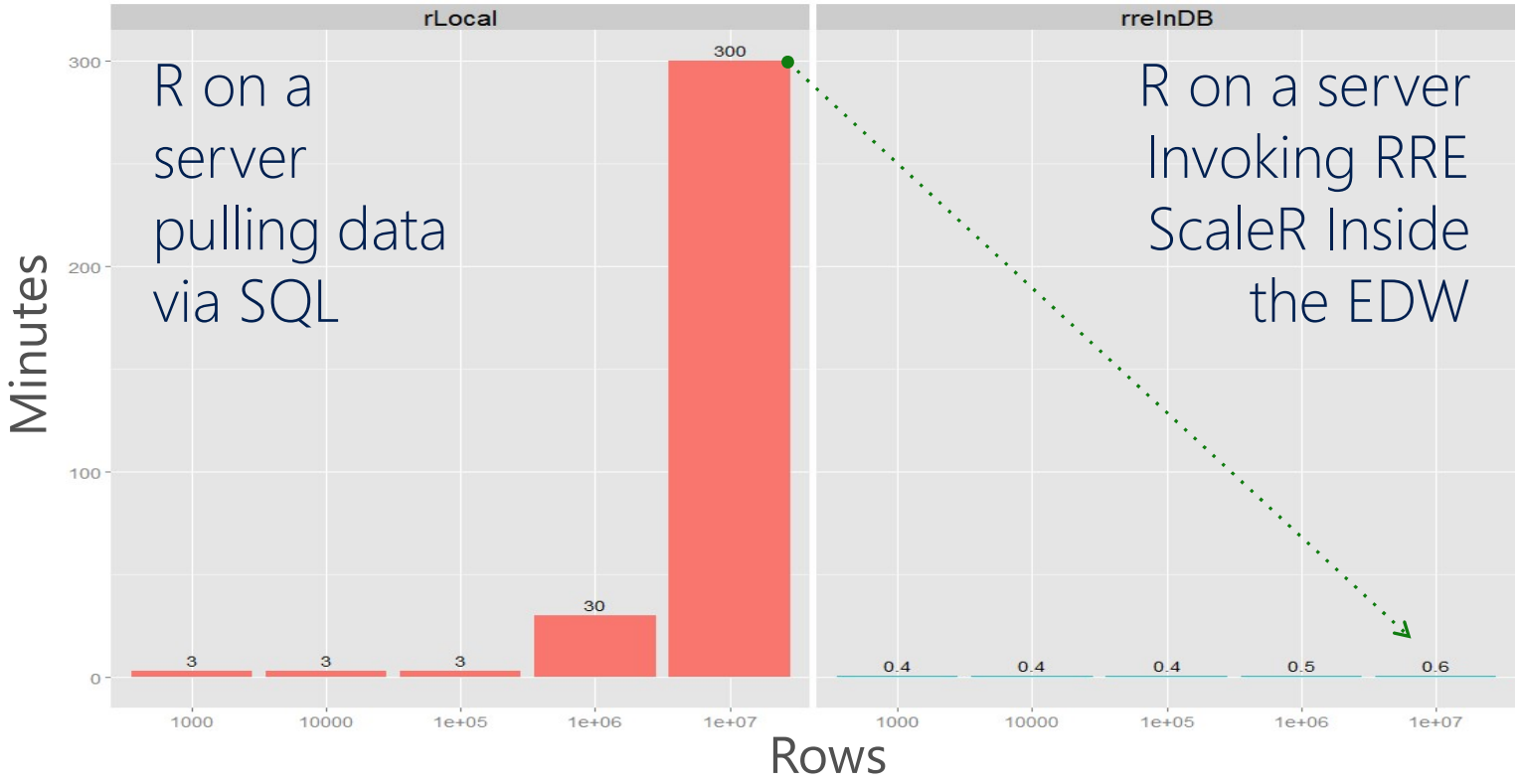
Supported 100% R
Deploy via Web Services
Reduced Security Exposure

# Growing Beyond Revolution Analytics



# Eliminate Data Movement

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



# Parallelized, Remote Execution Algorithms

## Data Step

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

## 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 & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

## Statistical Tests

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

## Sampling

- Subsample (observations & variables)
- Random Sampling

## Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Quantiles (approx.)
- 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.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

## Variable Selection

- Stepwise Regression

## Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

## Cluster Analysis

- K-Means

## Classification

- Decision Trees
- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes

## Combination

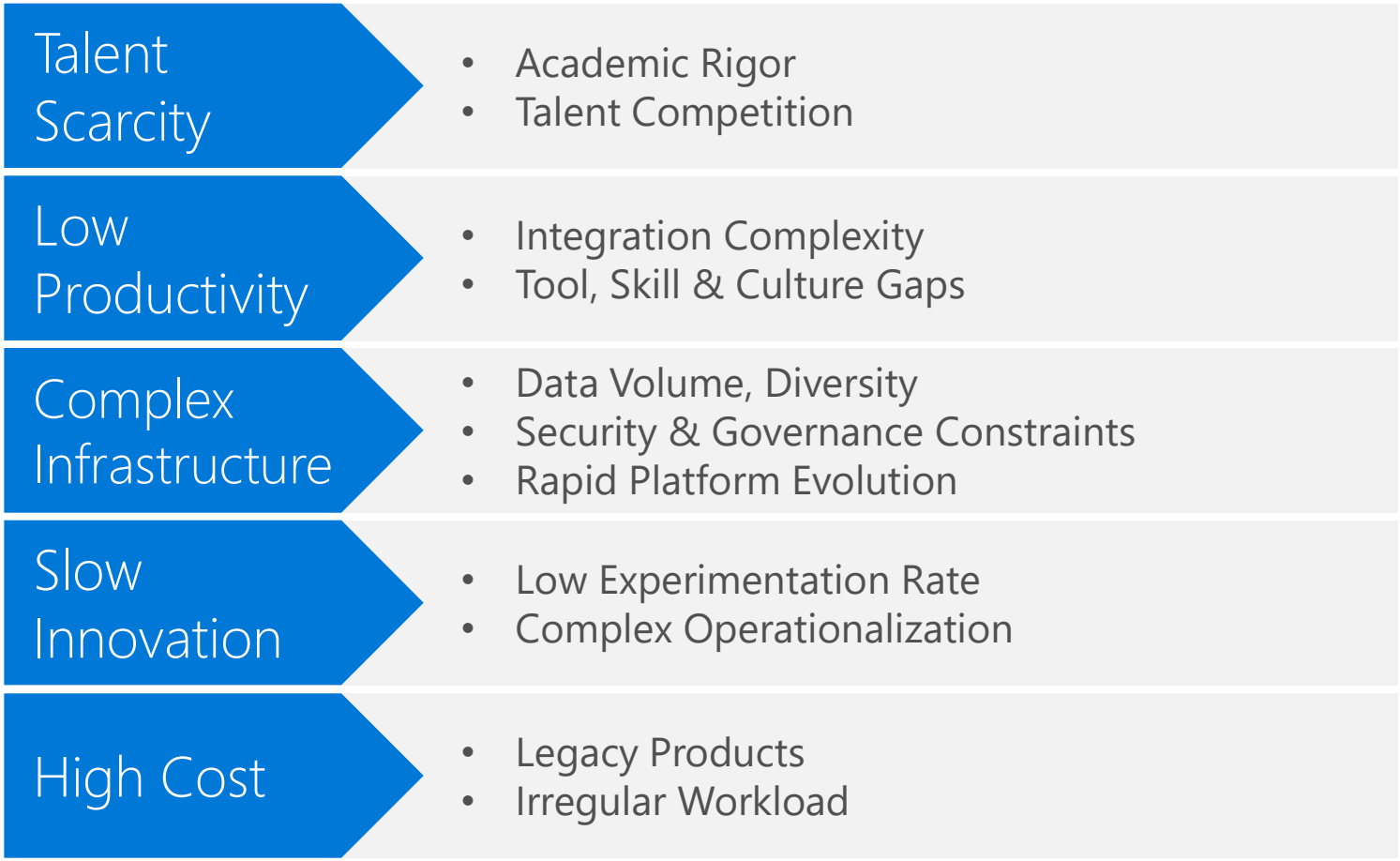

- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms

# Microsoft R Server Components

Marcello Benati, MCM
Advanced Analytics Product Manager

Microsoft

# Barriers to analytics adoption



# 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:

bing™

**MICROSOFT DOUBLES SEARCH SHARE**

XBOX

**30% BETTER IN SPEECH AND GESTURE RECOGNITION**

Office 365

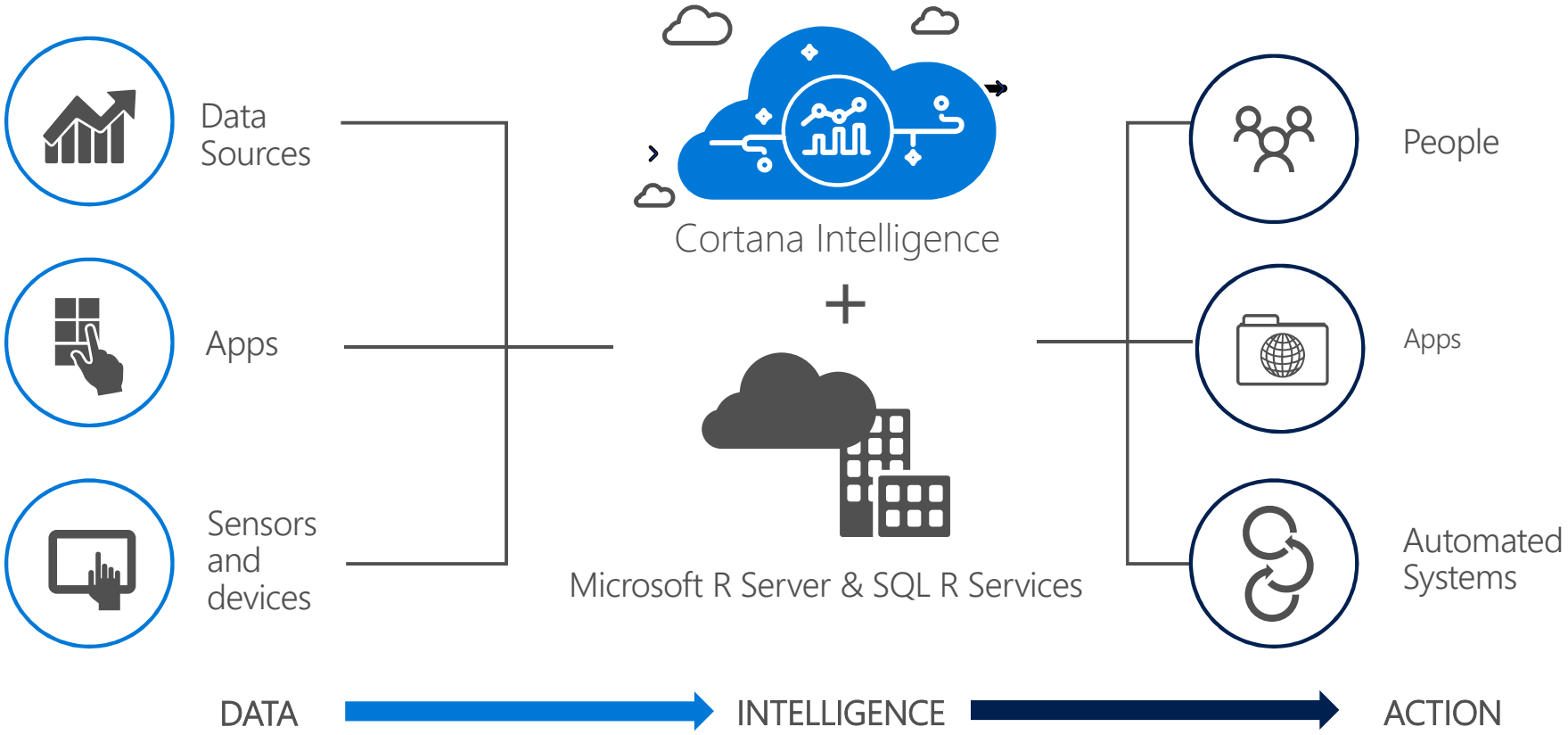
**NEW CAPABILITIES LIKE OFFICE GRAPH & CLUTTER**

skype™

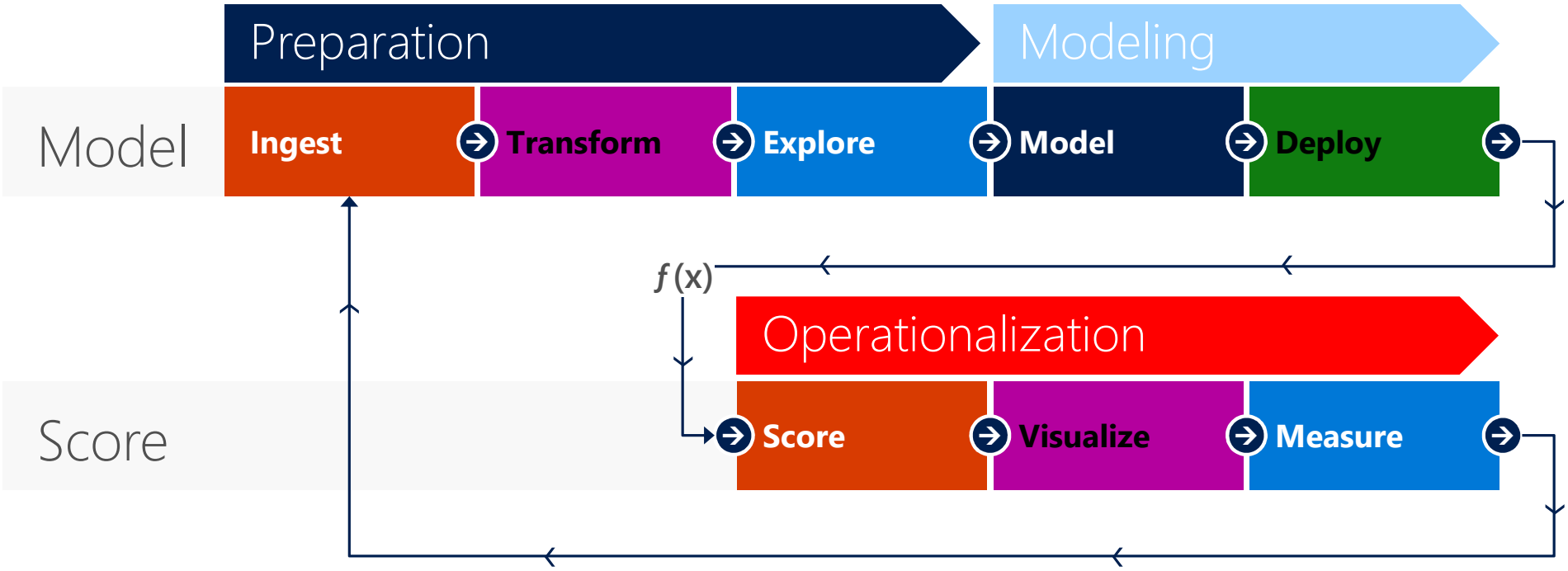
**SKYPE TRANSLATE INTRODUCTION**

# From data to intelligence to action

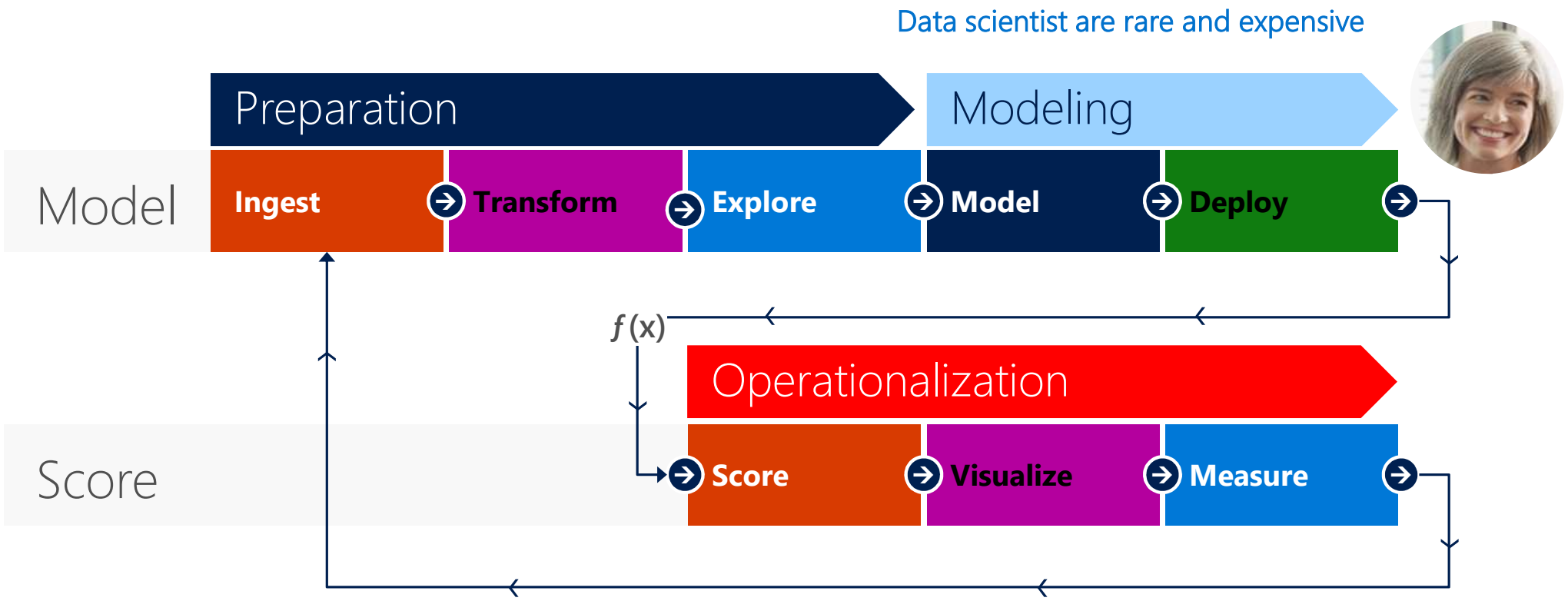
## On Prem or in the Cloud



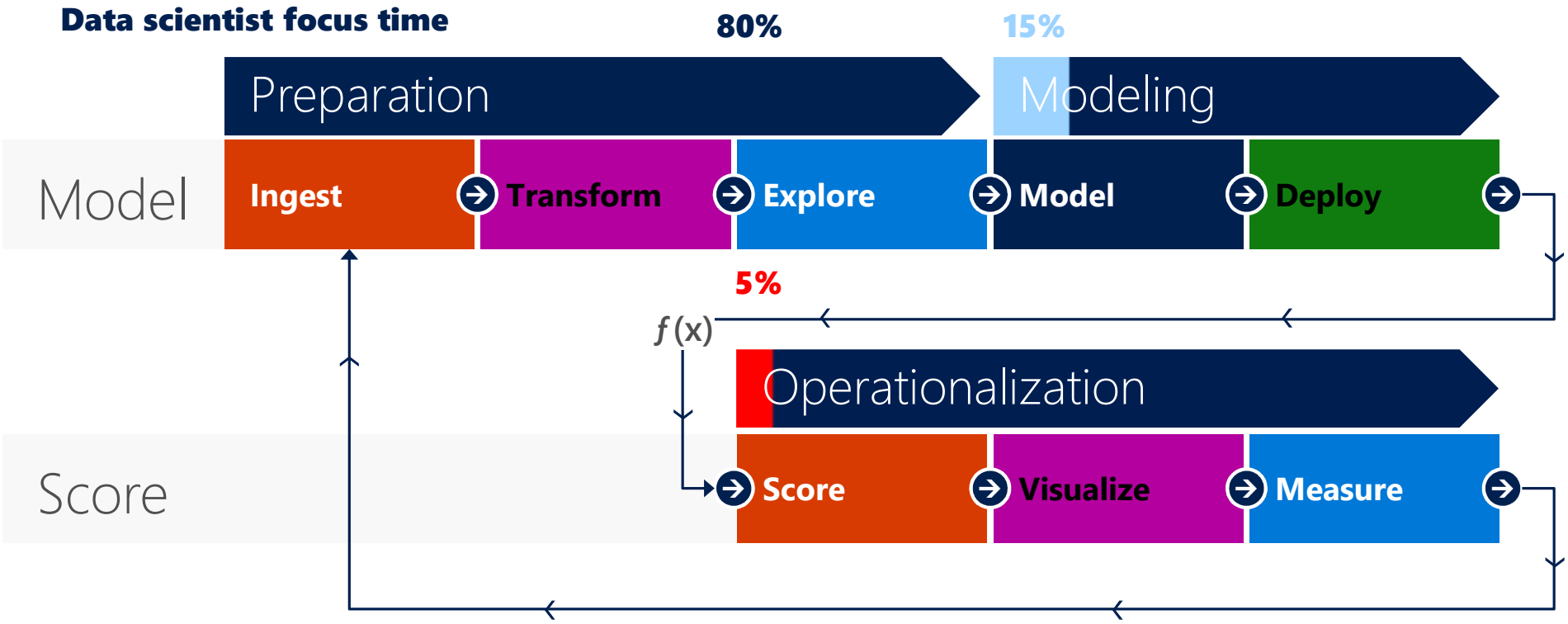
# Typical advanced analytics lifecycle



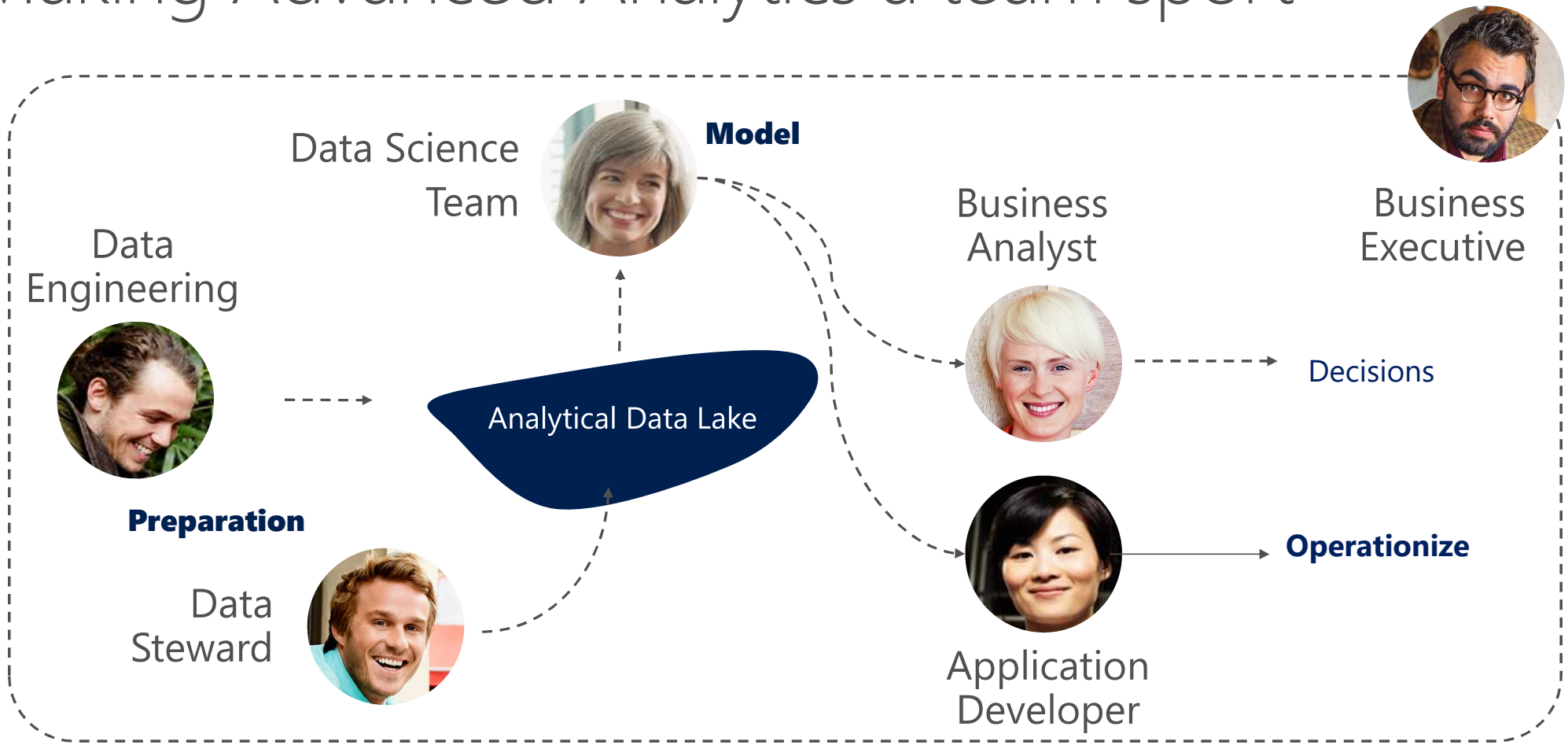
# Data Scientist should be creating / testing models



# But the reality is different …

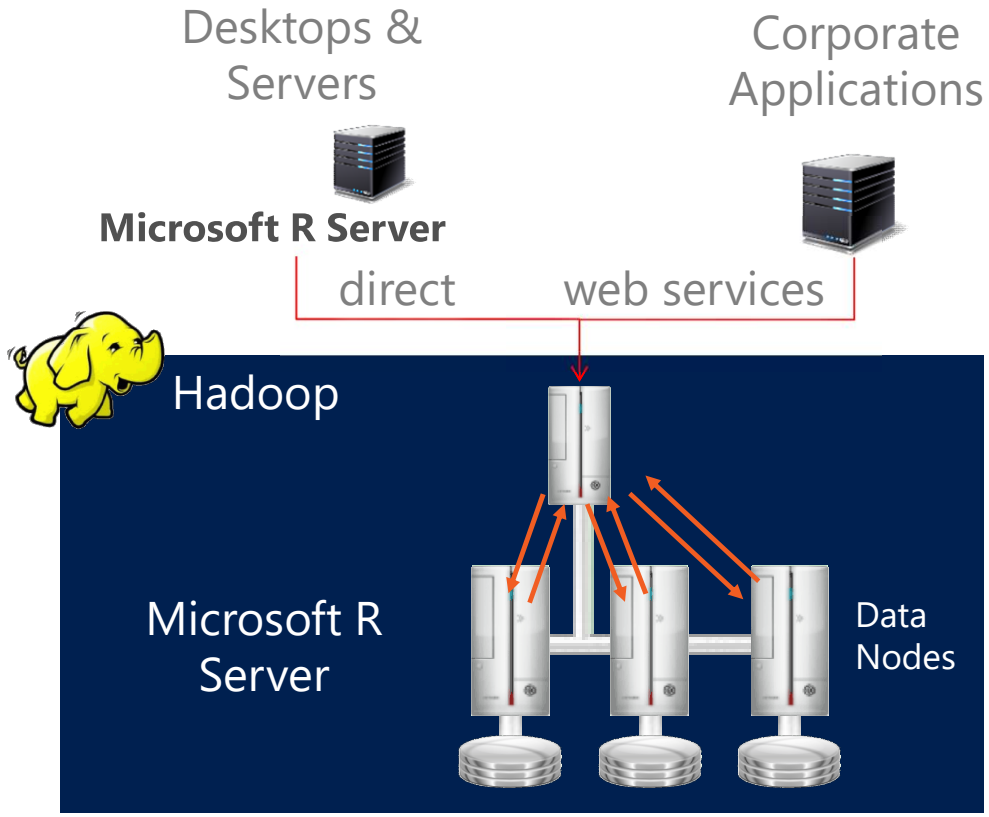


# Making Advanced Analytics a team sport



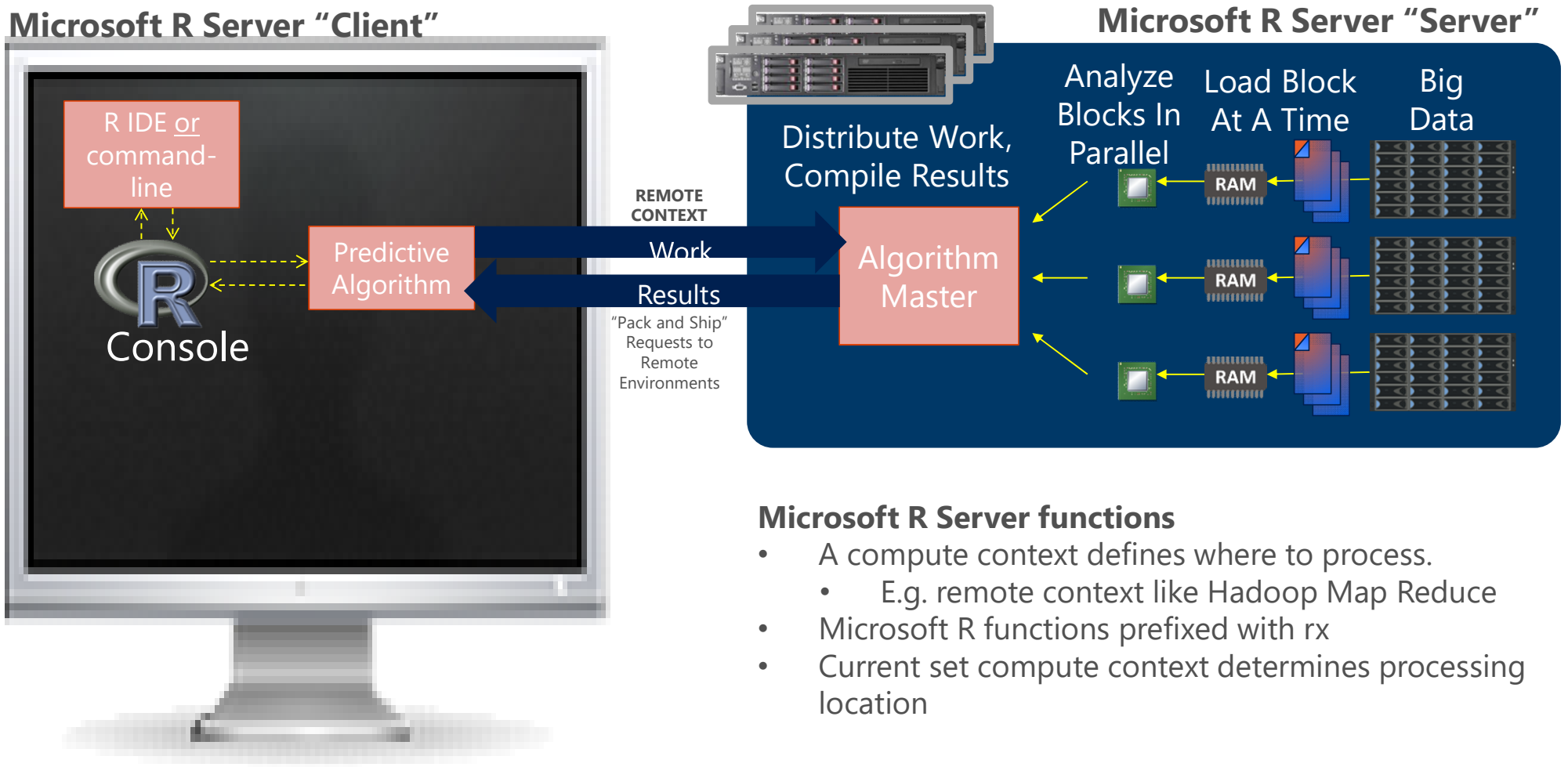
# Easy Scaling of R Analytics on Hadoop

- **Remote Execution**
  - Run Analytics from Workstations, Servers or Web Services APIs
- **Transparent Parallelization:**
  - YARN / MapReduce
  - Spark
- **Shared Resource Management**
  - Disk-based Data Chunking




Hortonworks
MAPR
cloudera

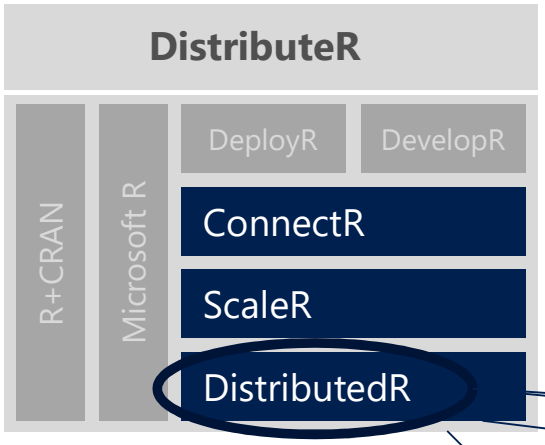
# Distributed R - How Does Remote Compute Context ?




**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




Delivers High Performance Parallel Distributed Analytics Across Individual and Clustered Systems

**Hadoop**

- Cloudera
- Hortonworks
- MapR
- Apache Spark (coming soon)

**MPI**

- IBM Platform LSF
- Microsoft HPC Clusters

**Local Files**

- Red Hat
- SuSE Servers
- Windows

**SQL**

- Teradata Database

# 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)
```

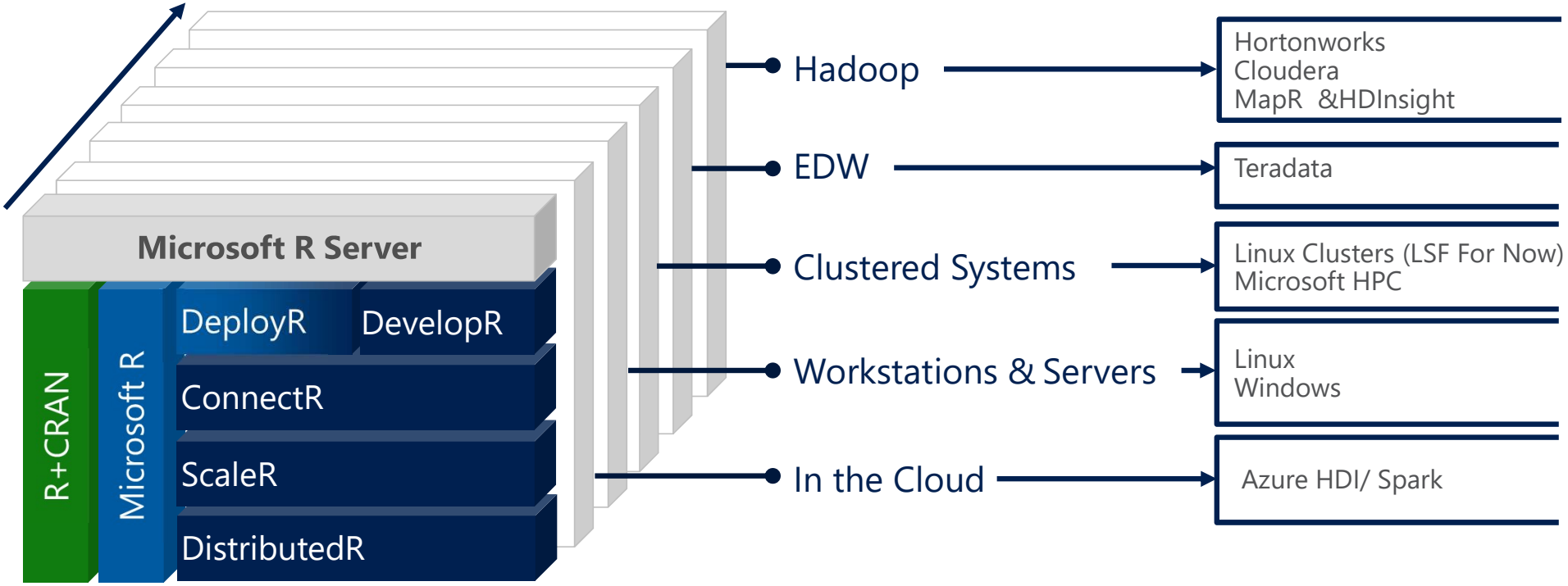
Functional model R script – does not need to change to run in Hadoop

```
### 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.




Code Portability Across Platforms

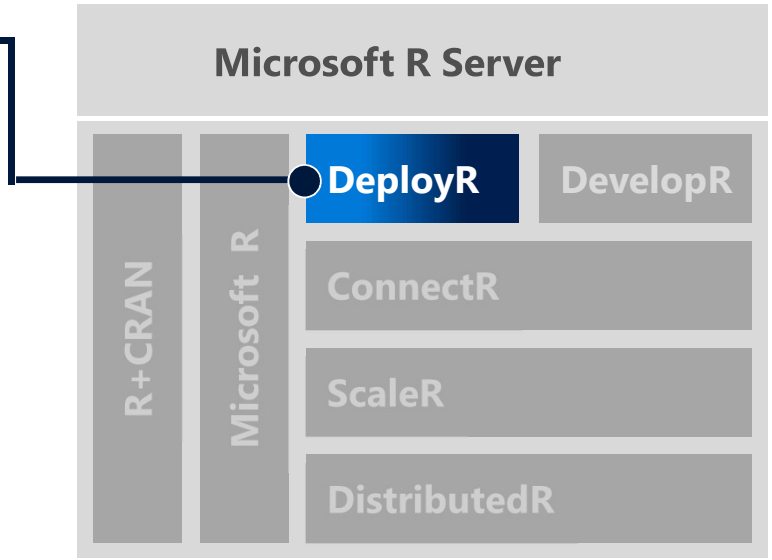
# DeployR

## DeployR

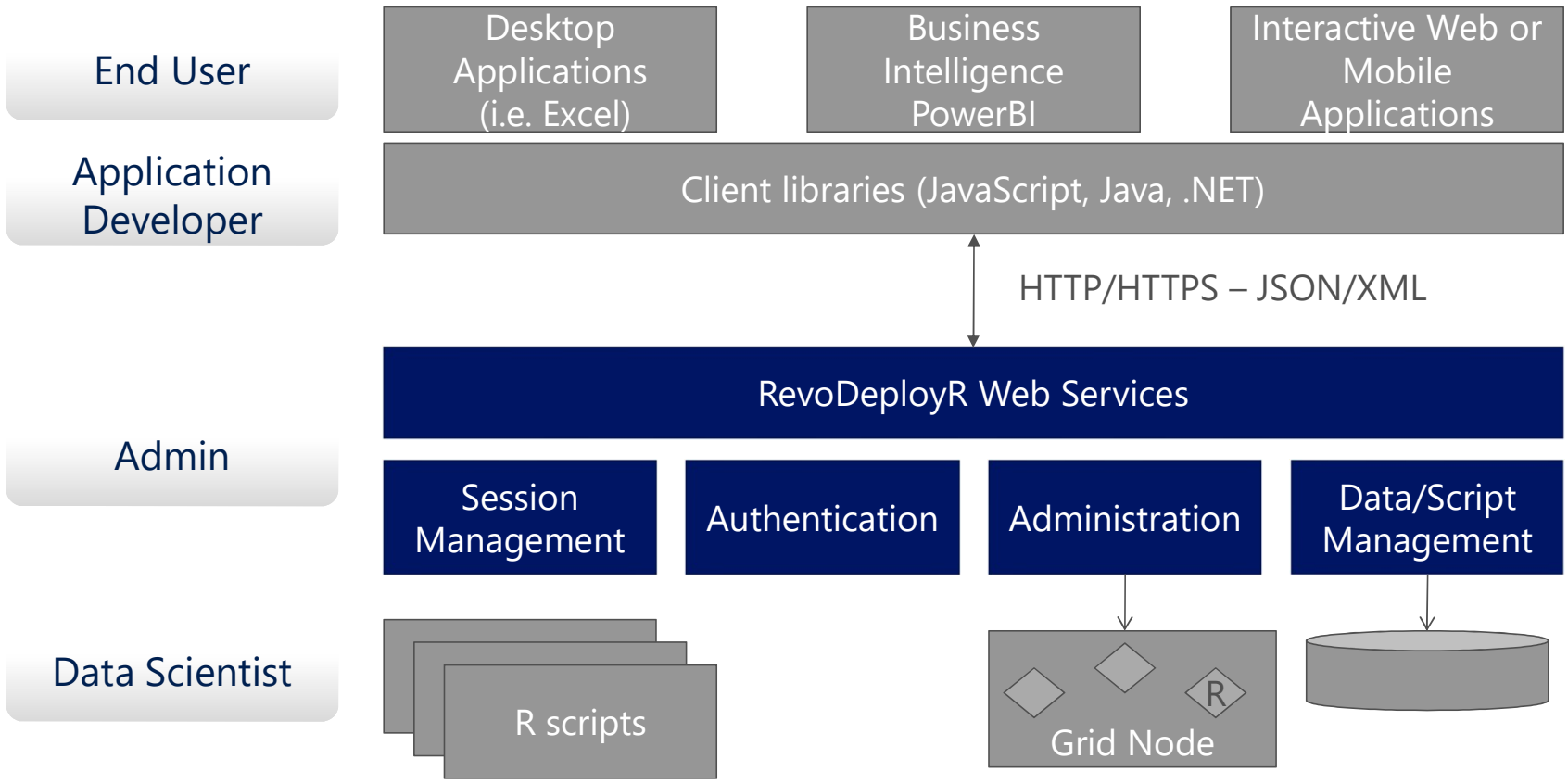
- Web services software development kit for integration analytics via APIs :
  - Java
  - JavaScript
  - .NET Integrates R Into application infrastructures

### 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

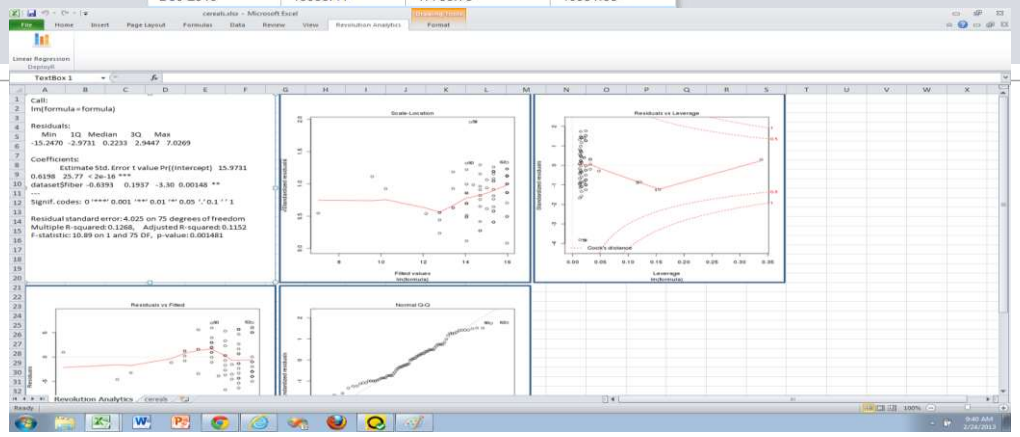


# DeployR: Architecture Overview

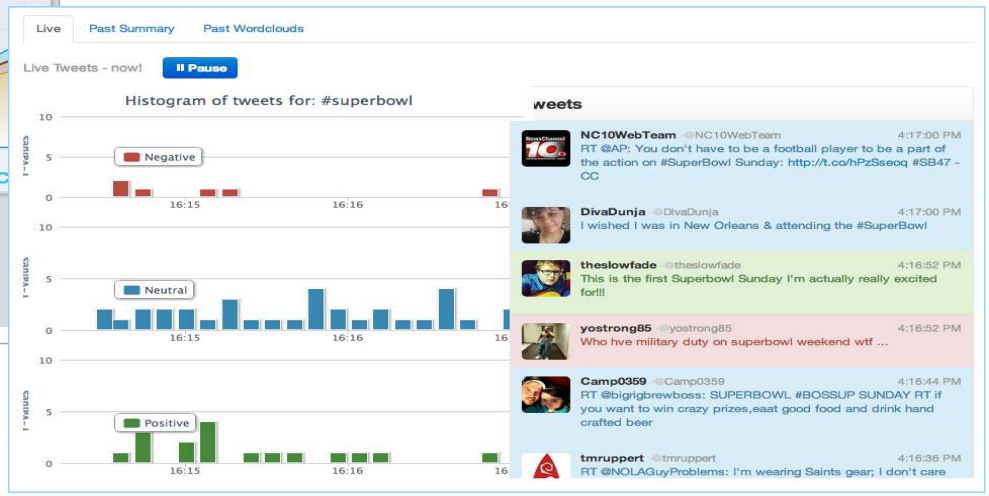


# DeployR: Examples of On-Demand Analytics


| Duration | Upper | Predicted | Lower |
|---|---|---|---|
| Oct 2010 | 11757.81 | 11116.61 | 10473.45 |
| Nov 2010 | 13230.42 | 12436.62 | 11652.42 |
| Dec 2010 | 18983.41 | 17753.73 | 16551.68 |



*On-demand sales forecasting*




*Real-time social media analysis*

*Leveraging the power of Office365*

# Features of Microsoft R Server on Hadoop

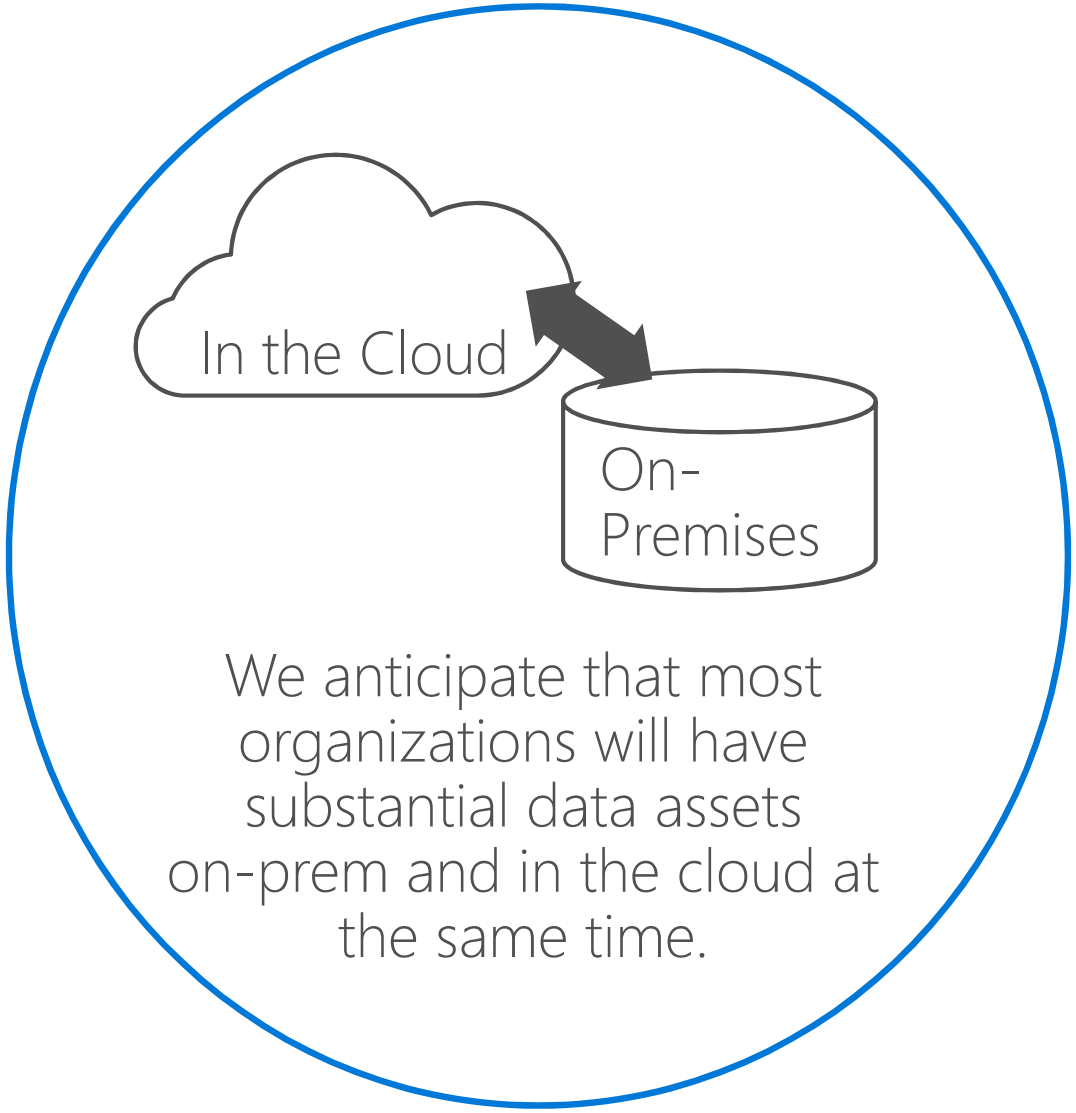
- Supports text delimited data files & Composite (sharded) XDF files
  - Reference directory path containing many files of same format!
- Utilises full parallelism of Hadoop (Mappers & Reducers) for fine-grained parallelism
  - Solve many model execution problems with full cluster parallelism ( via rxExec )
- 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

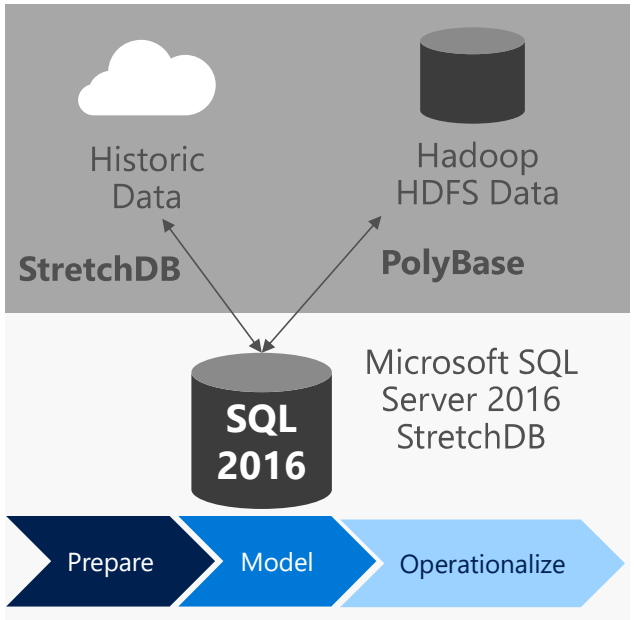
End-To-End, Pre-Integrated Platforms
Independent Storage and Compute Pricing
Elastic Computational & Storage Capability
Broadest Embrace of Open Source Technologies and Pricing

## Connect to Cloud-Born Data Streams

Mobile-First World
Customer-Facing Applications
Rapidly Emerging IoT

# Leveraging Cloud and On-Prem Data As One

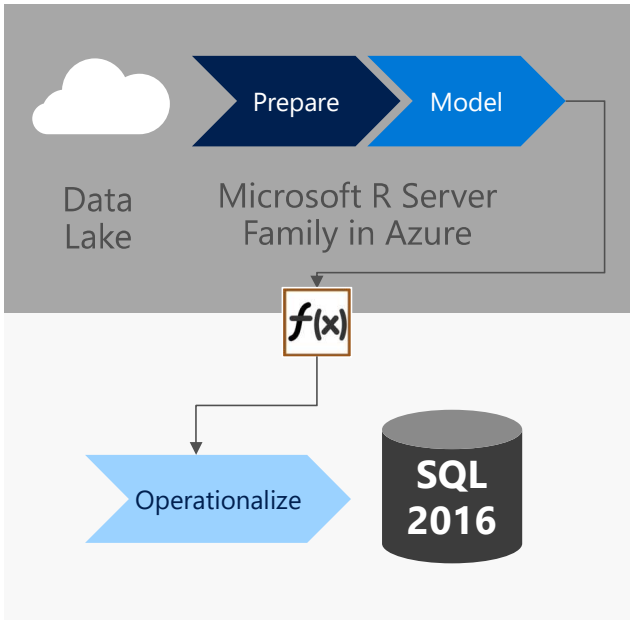
## Extend SQL 2016 Into the Cloud




StretchDB Example: Cost Reduce Growth of EDW

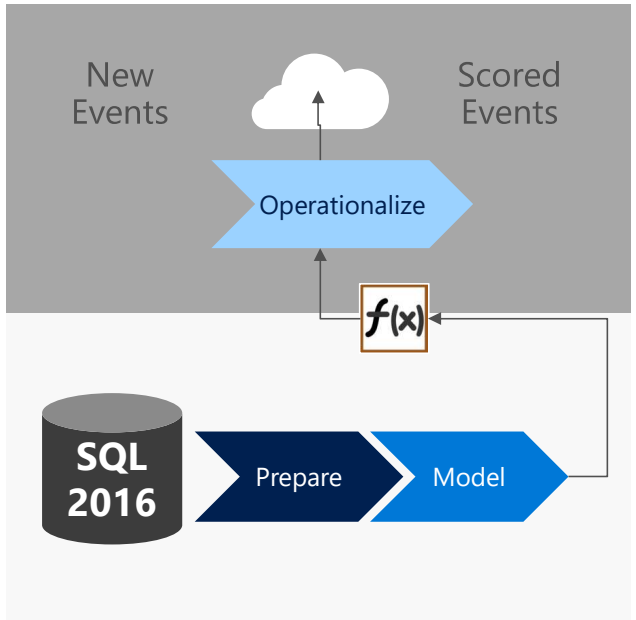
Polybase Example: Federate CRM Data with Cloud-Borne Demographics

## Model using a Cloud-Based Data Lake




On-Prem Scoring Example: Purchase Propensity Prediction

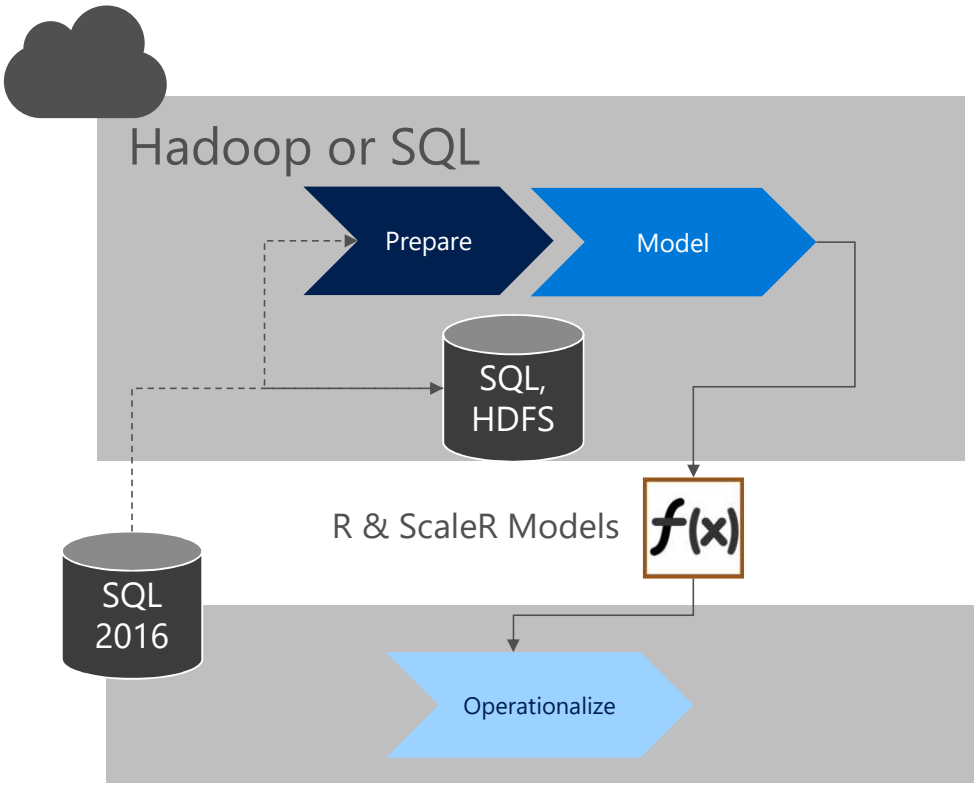
## Score Streams In The Cloud




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 Visualzations
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

## Problem

## Solution

*" Example graphic or customer quote…."*

Customer