Microsoft

The Future of Analytics

With Microsoft

Alain Dormehl

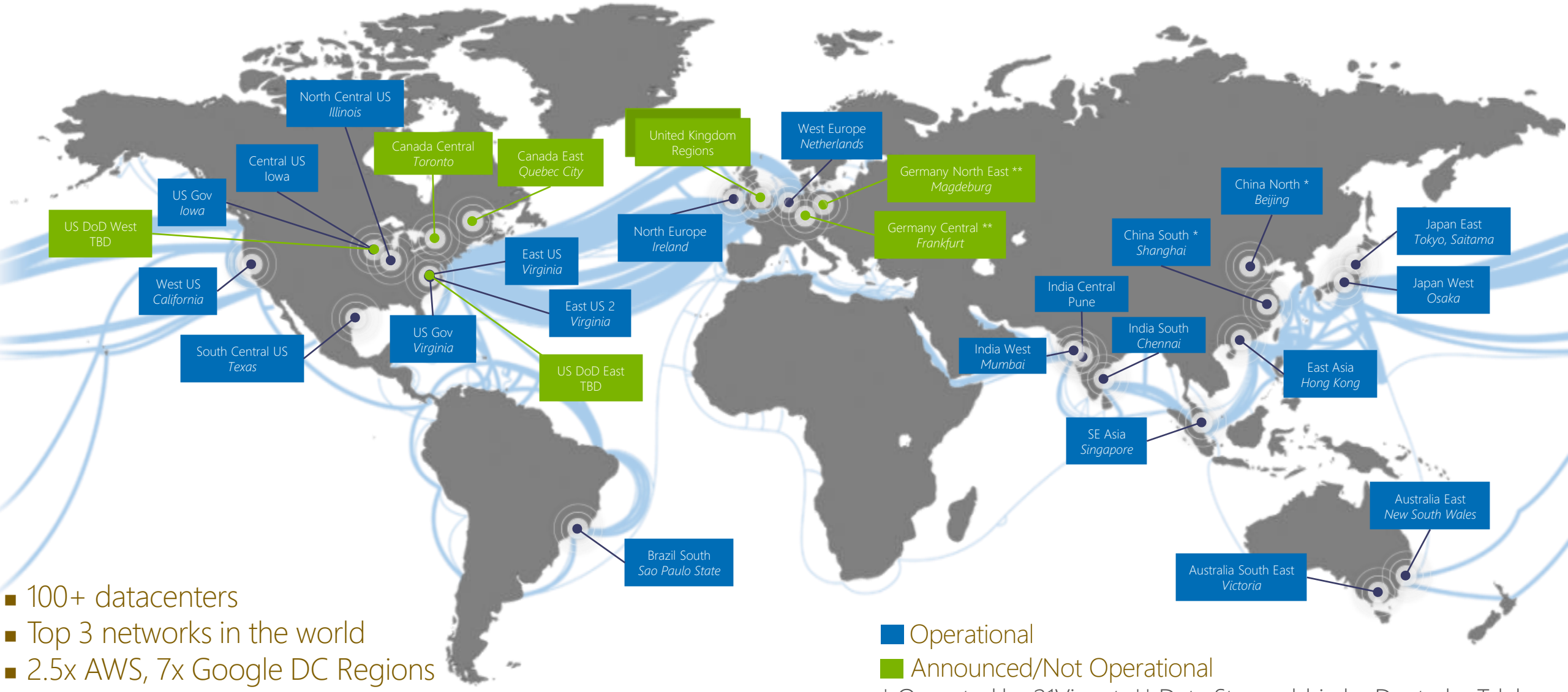
Data Solution Architect, Microsoft

Industry Leader in Operational Database Management Systems

Industry Leader in Business Intelligence and Analytics Platforms

Hyper scale Infrastructure is the enabler

30 Regions Worldwide, 22 Generally Available…



- 100+ datacenters
- Top 3 networks in the world
- 2.5x AWS, 7x Google DC Regions
- G Series – Largest VM in World, 32 cores, 448GB Ram, SSD…

Operational

Announced/Not Operational

* Operated by 21Vianet ** Data Stewardship by Deutsche Telekom

Platform Services

Security & Management

- Portal
- Azure Active Directory
- Azure AD B2C
- Multi-Factor Authentication
- Automation
- Scheduler
- Key Vault
- Store/ Marketplace
- VM Image Gallery & VM Depot

Services Compute

- Cloud Services
- Service Fabric
- Batch
- RemoteApp

Integration

- Storage Queues
- BizTalk Services
- Hybrid Connections
- Service Bus

Media & CDN

- Media Services
- Content Delivery Network (CDN)

Web and Mobile

- Web Apps
- API Apps
- Mobile Apps
- Logic Apps
- API Management
- Notification Hubs

Developer Services

- Visual Studio
- Azure SDK
- VS Online
- App Insights

Data

- SQL Database
- Data Warehouse
- DocumentDB
- Redis Cache
- Azure Search
- Storage Tables

Analytics & IoT

- HDInsight
- Machine Learning
- Stream Analytics
- Data Lake
- Data Factory
- Event Hubs
- Data Catalog
- IoT Hub
- Mobile Engagement

Hybrid Operations

- Azure AD Health Monitoring
- AD Privileged Identity Management
- Domain Services
- Backup
- Operational Analytics
- Import/Export
- Azure Site Recovery
- StorSimple

Infrastructure Services

OS/Server Compute

- Virtual Machines
- Container Service

Storage

- BLOB Storage
- Azure Files
- Premium Storage

Networking

- Virtual Network
- Load Balancer
- DNS
- Express Route
- Traffic Manager
- VPN Gateway
- App Gateway

Datacenter Infrastructure (30 Regions, 22 Online)

Azure Compliance

The largest compliance portfolio in the industry

ISO 27001

SOC 1 Type 2

SOC 2 Type 2

PCI DSS Level 1

Cloud Controls Matrix

ISO 27018

Content Delivery and Security Association

Shared Assessments

FedRAMP JAB P-ATO

HIPAA / HITECH

FIPS 140-2

21 CFR Part 11

FERPA

DISA Level 2

CJIS

IRS 1075

ITAR-ready

Section 508 VPAT

European Union Model Clauses

EU Safe Harbor

United Kingdom G-Cloud

China Multi Layer Protection Scheme

China GB 18030

China CCCPPF

Singapore MTCS Level 3

Australian Signals Directorate

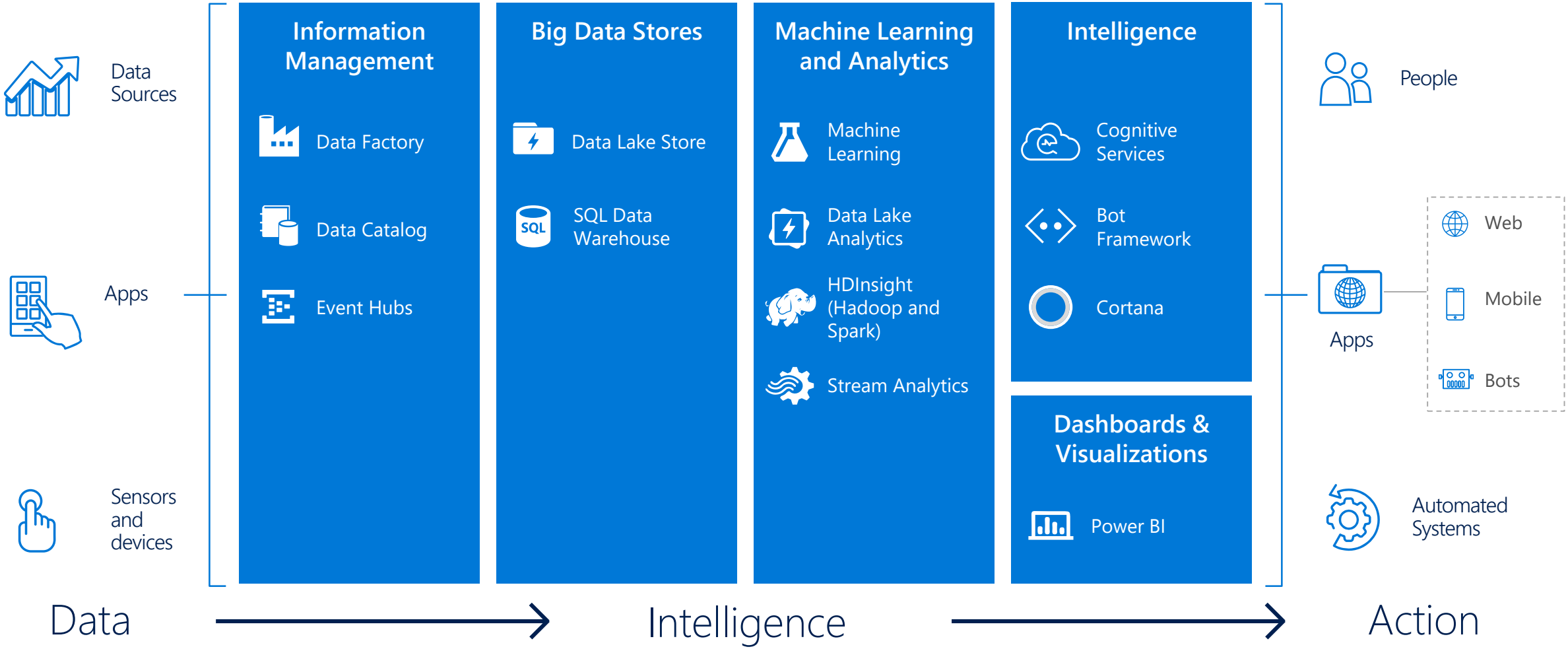
New Zealand GCIO

Japan Financial Services

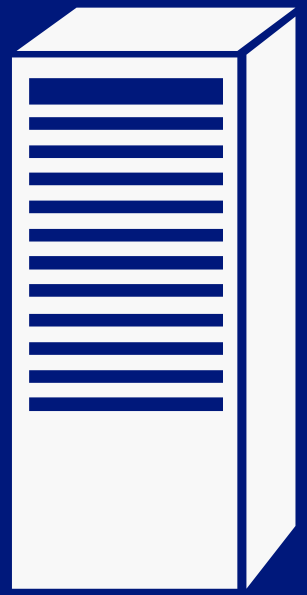
ENISA IAF

Azure is an open cloud

Transform data into intelligent action



From data to intelligent action



SQL Server 2016

- Advanced Analytics as a new workload
- Enables a new class of scenarios in SQL Server
- In-DB and in-memory analytics for performance
- Parallel execution for scale (Revolution Analytics)
- Train and test on-prem, operationalize in the cloud

SQL Server 2016

Industry leadership in Mission Critical OLTP

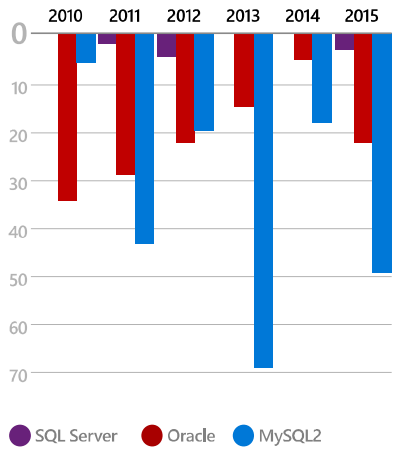
Operational DBMS

Business Intelligence

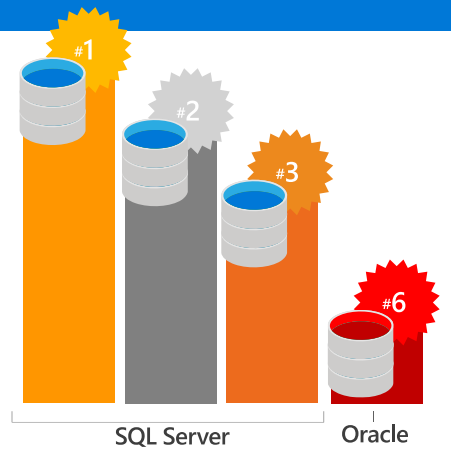
Data Warehouse

Advanced Analytics

Most secure database 6 years in a row

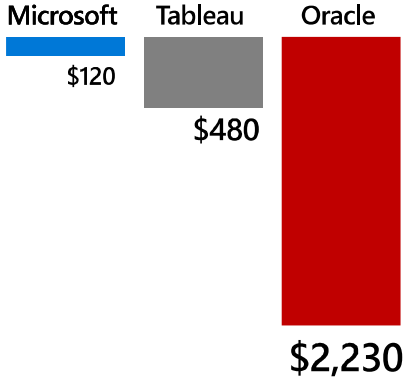


Highest performing data warehouse



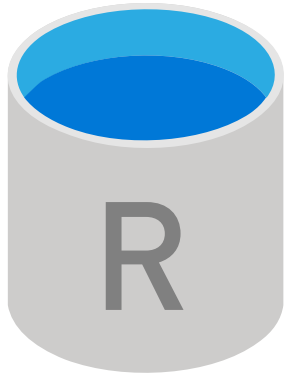
TPC-H non-clustered 10TB

End-to-end mobile BI on any device



Self-service BI per User

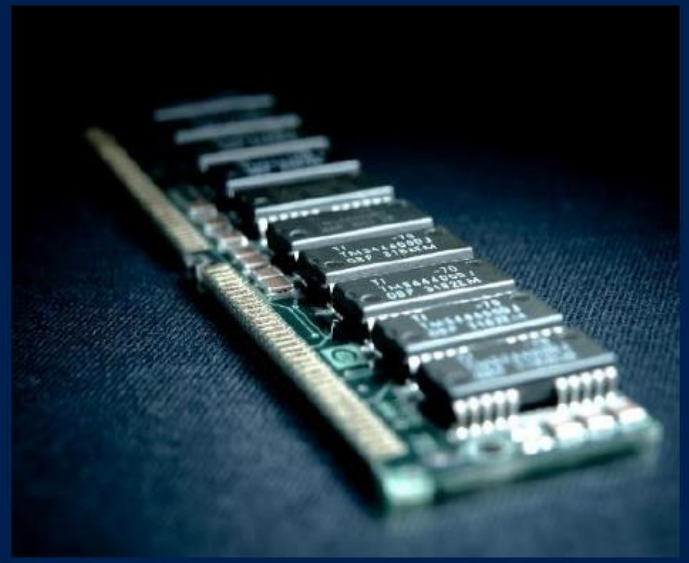
In-database Advanced Analytics



R + in-memory at massive scale

On Premises ⟷ Cloud

Enterprise use of open source R



R needs data in memory to start a computation*

R is mostly single threaded

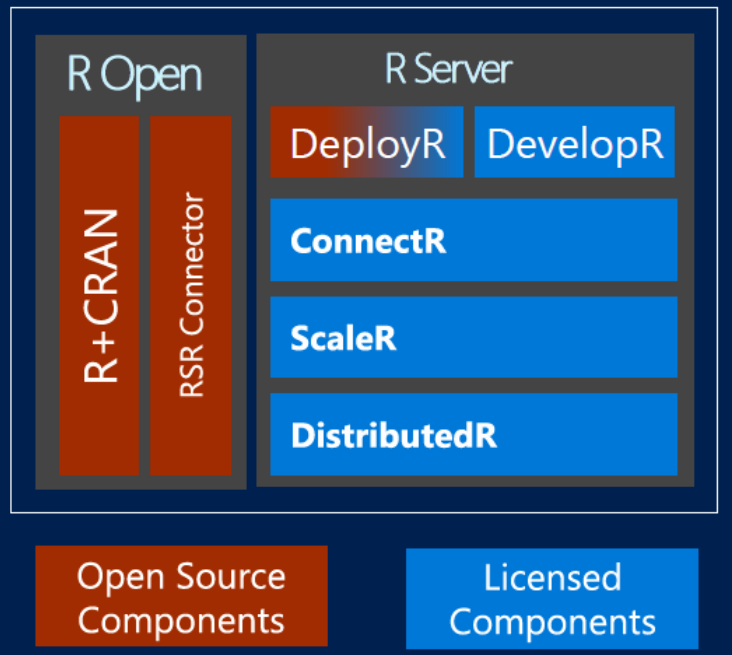
R requires skilled resource to scale out computations across a cluster and needs re-coding for R map-reduce in Hadoop

Open source R is supported by the community

Microsoft R Server

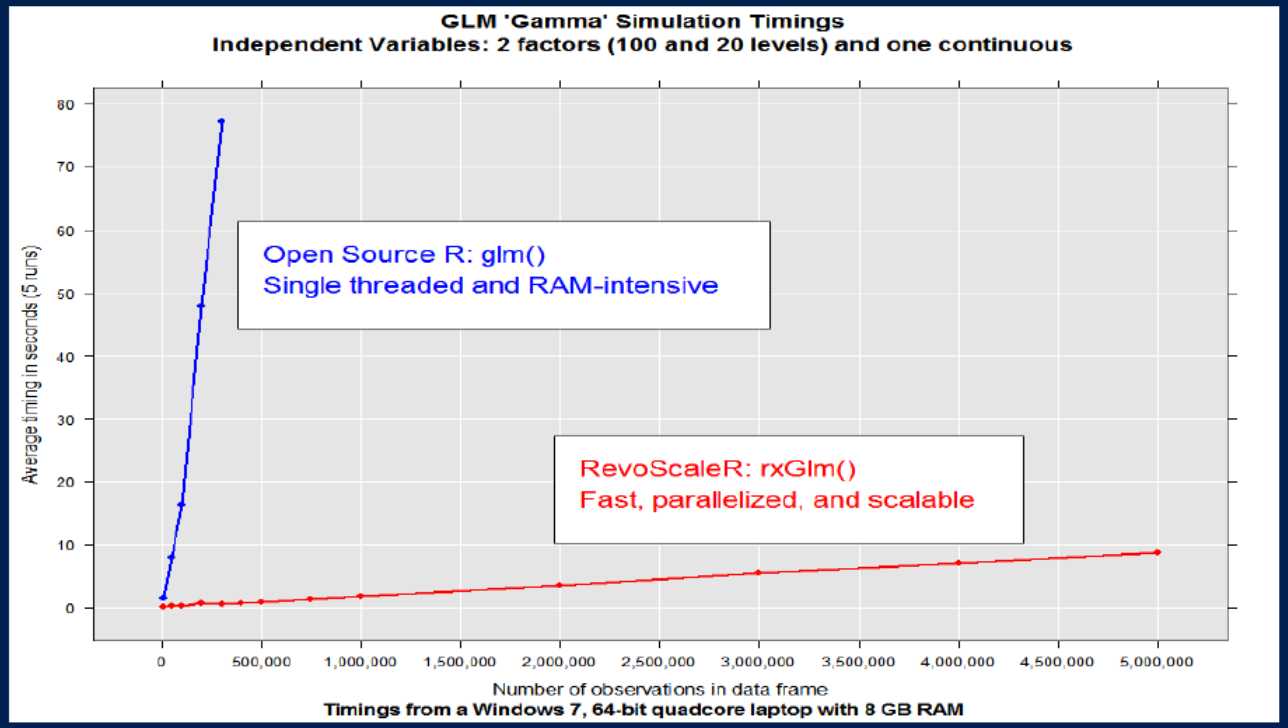
High-performance open source R plus:

- Data source connectivity to big-data objects
- Big-data advanced analytics
- Multi-platform environment support
- In-Hadoop and in-Teradata predictive modeling
- Development and production environment support
  - IDE for data scientist developers
  - Secure, Scalable R Deployment
- Technical support, training and services



ScaleR - Performance comparison

Microsoft R Server has no data size limits in relation to size of available RAM. When open source R operates on data sets that exceed RAM it will fail. In contrast Microsoft R Server scales linearly well beyond RAM limits and parallel algorithms are much faster.



File Name	Compressed File Size (MB)	No. Rows	Open Source R (secs)	Revolution R (secs)
Tiny	0.3	1,235	0.00	0.05
V. Small	0.4	12,353	0.21	0.05
Small	1.3	123,534	0.03	0.03
Medium	10.7	1,235,349	1.94	0.08
Large	104.5	12,353,496	60.69	0.42
Big (full)	12,960.0	123,534,969	Memory!	4.89
V.Big	25,919.7	247,069,938	Memory!	9.49
Huge	51,840.2	494,139,876	Memory!	18.92

- US flight data for 20 years
- Linear Regression on Arrival Delay
- Run on 4 core laptop, 16GB RAM and 500GB SSD

SQL Server 2016 – In T-SQL Stored Proc

```
create procedure TrainModelR
as
begin
   exec sp_execute_external_script
   @language = N'R',
   @script = N'

          ## Create model
          InputDataSet$Tag <- factor(InputDataSet$Tag)
          InputDataSet$Age <- factor(InputDataSet$Age)
          InputDataSet$Address <- factor(InputDataSet$Address)
          logitObj <- glm(Tag ~ ., family = binomial, data = InputDataSet)
          summary(logitObj)

          ## Serialize model and put it in data frame
          trained_model <- data.frame(model=as.raw(serialize(logitObj, connection=NULL)));'

          ,@input_data_1 = N'select * from features'
          ,@output_data_1_name = N'trained_model';
end
go
```

```
execute TrainModelR
```

Microsoft

Data Science with Microsoft SQL Server 2016

Download

Buck Woody, Danielle Dean, Debraj GuhaThakurta
Gagan Bansal, Matt Conners, Wee-Hyong Tok

Thank you.

Microsoft