SQL Server 2016 R Services
Bringing Advanced Analytics to where the data is

Marcello Benati - Advanced Analytics Product Manager
November 2015
Accelerating the speed of business

Reactive to Proactive

- Historical data to Predictions
- People and processes to Decision automation
From data to decisions and actions
Microsoft Advanced Analytics offerings

- On-premises
- SQL Server
- Cortana Analytics Suite
- In the cloud
<table>
<thead>
<tr>
<th>Expensive</th>
<th>Huge set-up costs of tools, expertise, and compute/storage capacity create unnecessary barriers to entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siloed data</td>
<td>Siloed and cumbersome data management restricts access to data</td>
</tr>
<tr>
<td>Disconnected tools</td>
<td>Complex and fragmented tools limit participation in exploring data and building models</td>
</tr>
<tr>
<td>Deployment complexity</td>
<td>Many models never achieve business value due to difficulties with deploying to production</td>
</tr>
</tbody>
</table>
First, a few foundational facts about the Advanced Analytics lifecycle and R
Typical analytical lifecycle

Model
- Ingest
- Transform
- Explore
- Model
- Deploy

Score
- Measure
- Act
- Score

SQL 2016

SQL 2016 or Cloud

$f(x)$
R—What Is It?

A language platform
- A procedural language optimized for statistics and data science
- A data visualization framework
- Provided as open source

A community
- 2.5M+ statistical analysis and machine learning users
- Taught in most university statistics and analytics programs
- Active user groups across the world

An ecosystem
- CRAN: 6500+ freely available algorithms, test data and evaluations
- Many applicable to big data if scaled
SQL Server and the R language

Motivation: marry SQL Server with the most popular language for advanced analytics

<table>
<thead>
<tr>
<th>Natural Language Processing</th>
<th>Scientific and Clinical Computing</th>
<th>Bayesian Inference</th>
<th>Chemosometrics and Computational Physics</th>
<th>Clinical Trial Design, Monitoring, and Analysis</th>
<th>Cluster Analysis &amp; Finite Mixture Models</th>
<th>Probability Distributions</th>
<th>Computational Econometrics</th>
</tr>
</thead>
</table>

- Bayesian Inference
- Chemosometrics and Computational Physics
- Clinical Trial Design, Monitoring, and Analysis
- Cluster Analysis & Finite Mixture Models
- Probability Distributions
- Computational Econometrics
- Design of Experiments (DoE) & Analysis of Experimental Data
- Empirical Finance
- Statistical Genetics
- High-Performance and Parallel Computing with R
- Medical Image Analysis
- Psychometric Models and Methods
- Natural Language Processing
- Analysis of Pharmacokinetic Data
- Official Statistics & Survey Methodology
- Phylogenetics, Especially Comparative Methods
- Multivariate Statistics
- Optimization and Mathematical Programming
- Machine Learning & Statistical Learning
- Graphic Displays & Dynamic Graphics & Graphical Devices & Visualization
- Reproducible Research
- SQL Server and the R language
Performance/Scale: PEMA

- Unique PEMAs: Parallel, external-memory algorithms
- High-performance, scalable replacements for R/SAS analytic functions
- Parallel/distributed processing eliminates CPU bottleneck
- Data streaming eliminates memory size limitations

GLM ‘Gamma’ Simulation Timings
Independent Variables: 2 factors (100 and 20 levels) and one continuous

Number of observations in data frame
Timings from a windows 7,64-bit quad core laptop with 8 GB RAM

SQL Server 2016 comes with Revolution Analytics built-in
## Parallelized 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)
- Multiple Linear Regression
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

### Variable Selection
- Stepwise Regression Linear, Logistic and GLM

### Simulation
- Monte Carlo
- Parallel Random Number Generation

### Cluster Analysis
- K-Means

### Classification
- Decision Trees
- Decision Forests
- Stochastic Gradient Boosted Decision Trees

### Combination
- Using Revolution rxDataStep and rxExec functions to combine open source R with Revolution R
- PEMA API
### History of Analytics in SQL Server

#### SQL Server Releases

<table>
<thead>
<tr>
<th>Year</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>ETL</td>
</tr>
<tr>
<td>2005</td>
<td>Operational Reports</td>
</tr>
<tr>
<td>2008 R2</td>
<td>Self Service BI</td>
</tr>
<tr>
<td>2012</td>
<td>Data Discovery</td>
</tr>
<tr>
<td>2016</td>
<td>Advanced Analytics</td>
</tr>
</tbody>
</table>

#### OLAP Services

- OLAP Services
- Analysis Services
- Data Transformation Services (DTS)

#### ETL Services

- Integration Services
- Master Data Services

#### Reporting Services

- Reporting Services

#### Power BI Services

- Power View
- Data Quality Svcs
- ColumnStore DW

#### Mobile BI

- Mobile BI

#### Data Discovery

- Revolution R in SQL
- In-DB Analytics
What Revolution R Enterprise brings to SQL 2016

Big data analytics platform
✓ Based on open source R

High-performance, scalable, full-featured
✓ Performant, scalable, distributable statistical and machine learning algorithms

Write once, deploy anywhere
✓ Scripts and models can be executed on a variety of platforms, including non-Microsoft (e.g. Hadoop, Teradata in-db)

Integration with the R Ecosystem
✓ Analytic algorithms accessed via R function with similar syntax for R users
✓ Arbitrary R functions/packages can be used in conjunction
In-database Advanced Analytics at scale

Benefits
- Faster deployment of ML models
- Less data movement, faster insights
- Work with large datasets: mitigate R memory and scalability limitations

Example Solutions
- Fraud detection
- Sales forecasting
- Warehouse efficiency
- Predictive maintenance

How is it Integrated?
- T-SQL calls a Stored Procedure
- Script is run in SQL through extensibility model
- Result sets sent through Web API to database or applications

Data Scientist
Interacts directly with data
Creates models and experiments

Data Developer/DBA
Manages data and analytics together

Analytic Library
Open Source R
Revolution PEMA
T-SQL Interface
Relational Data

Extensibility
R Integration
R

Benefits
• Faster deployment of ML models
• Less data movement, faster insights
• Work with large datasets: mitigate R memory and scalability limitations

Example Solutions
• Fraud detection
• Sales forecasting
• Warehouse efficiency
• Predictive maintenance

How is it Integrated?
• T-SQL calls a Stored Procedure
• Script is run in SQL through extensibility model
• Result sets sent through Web API to database or applications
SQL Server 2016
R Services

- Flexibility & Agility
- Performance & Scale
- Cost Effective
Flexibility & Agility

Write once deploy anywhere
- No model re-writes across platforms
- No re-writes from modeling to scoring

Hybrid modeling & scoring
- Model on premises, score on premises
- Model on premises, score in the cloud
- Model on cloud, score on premises
Deploying Analytics On-Prem in SQL 2016

Model & Deploy In SQL16:
- Complete Lifecycle
- Deploy via BI Tools or Applications

Advantages:
- No Data Movement
- Reduced Security Exposure
- SQL Skill Reuse
- Operational Stability
- SMP Parallel Performance
Hybrid Cloud Deployment:

- Data Lake in SQL
- Deploy Models to Azure
- Expose Services via Azure

Advantages:

- Score “Born In The Cloud” Data
- Deploy Globally via Cloud Services
- Secure Historic Data Locally
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
- Cloud Economics & Scale for Big Data
- SQL Server Stability, Privacy for Deploy
Performance & Scale

Reduced model execution time
- In-DB saves time and resources

Scale-out for large data models
- Not limited to box memory and drives

Hybrid memory & disk scalability
- Execute models in memory

<table>
<thead>
<tr>
<th>Example</th>
<th>Other Tools</th>
<th>R Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Scoring- Marketing and sales</td>
<td>11 hours</td>
<td>5 to 8 min</td>
</tr>
<tr>
<td>Regression on 10 million records using 4000 Variables</td>
<td>4 to 5 days</td>
<td>1 hour</td>
</tr>
<tr>
<td>Stress Test- 140,000 regressions with 20 random variables</td>
<td>1 hour</td>
<td>45 seconds</td>
</tr>
<tr>
<td>Financial- VaR calculations 2.5 B simulations</td>
<td>3.5 hours</td>
<td>42 seconds</td>
</tr>
<tr>
<td>Healthcare Scoring</td>
<td>5 weeks-35 processes</td>
<td>6 min, 1 process</td>
</tr>
</tbody>
</table>
Portable and Parallelized Analytics
SQL R is 10’s to 100’s of Time Faster Than Open Source

No RAM Limits

✓ Open Source R Exceeds RAM and Fails
✓ RRE Scales Linearly Well Beyond RAM Limits

Faster Algorithms

✓ As data grows ScaleR optimization becomes apparent

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

Best Advanced Analytics Value

R Services are part of SQL Server

✓ Built in to SQL Server 2016 Enterprise Edition

In DB analytics shrinks analysis cost and time

✓ No data movement reduces costs

No Proprietary Hardware Requirement

✓ Can be installed in commodity hardware

Leverage existing SQL investments

SQL SERVER 2016
$ 648 K
+ $120 Per user for PowerBI

Oracle
$2.2 M
+$2,230 for BI

Advanced Analytics
Business Intelligence
Data Warehousing
OLTP

Costs based on a Server with 2 proc/ 8 Cores
Best of Both Worlds

Use SQL for fast querying & in-memory ColumnStore indexes
Use R for data exploration, predictive modeling, scoring and visualization
R in the enterprise
SCENARIOS
The Data Scientist

I work from my RIDE.

I want to access and analyze large amounts of data and get quick results.
Large International Bank
Hadoop-based Analytics

Key Challenge Addressed
✓ LOB’s need to offset SAS and other legacy costs with new technology, new talent, legacy model conversion, and usage of bigger, wider data from more diverse sources

Key Technology
Revolution R Enterprise using ScaleR for “Big Data”

How is Revolution R being used
In early stages of deployment - IT, Consumer Bank, Decision Management
Consumer credit scoring, risk modeling, and acquisition modeling across LOB’s
RRE to be deployed on several large Hadoop clusters

Outcomes
Successfully trained 75 analysts (more to come) on RRE and ScaleR.
Successful installations on CDH Hadoop.

Bottom line
Significant technology and human capital savings while introducing new innovation for competitive advantage
The SQL Developer

I want use SQL to generate both predictions and visualizations without moving data outside of the database.
Large Insurance Group
Architecture Update / Risk modeling

Key Challenge
Replace SAS in the current Analytical architecture

Key Technology
Revolution R Enterprise using ScaleR Big Data Analytics

How is Revolution R being used
Data Science team is currently focused on Risk Management projects
Looking to move into actuarial in the future

Outcome
This is part of a much larger architectural change so no distinct outcomes recorded as of yet
Currently evaluating Revolution on Hortonworks
A Large Financial Institution: Customer Service

Key Challenge Addressed
LOB’s have goals of offsetting SAS and other legacy costs with new technology, new talent, legacy model conversion, and usage of bigger, wider data from more diverse sources.

Key Technology
Revolution R Enterprise

How Revolution is being used
DeployR to manage a pool of models accessed by brokers on demand to support broker recommendations to customers.
Models developed by Quantitative Equity Analysts
DeployR to manage versions, access and model interface to UI

Interfaced to SiteMinder single sign on capabilities

Outcome
Provides real-time scenario evaluation and projections enabling brokers to provide detail responses to customers while on the phone with them.

Bottom line
Improved customer response, retention and transaction volume.
The DBA

I want to help my Data Science team be effective without losing control of the database's performance and stability.
A large group of mutual insurers based in OH.
Provides lines of insurance for P&C, Corporate, Life, retirement, SMB, Farm, etc.
Parent spun off a subsidiary previously to enable company to go public.
Company’s 2014 revenues were $25B

**Key Technology**

Revolution R Enterprise

**How is Revolution R being used**

Replacing OSR and SAS models
IT Project Optimization
Marketing “in-flight” campaign analysis

**Outcome**

Analysis and continuance of high ROI projects while establishing earlier detection of likely IT project failure

Analyze and adjust campaigns in real-time to improve response rates
Up to 50X faster R processes with RRE vs. Open R

**Bottom line**

Excellent client partnership with multiple successful projects and upside potential.
Credit Card Company- General Analytics Use

Key Technology
- Revolution R Enterprise on 100 MapR Hadoop nodes
- Purchased in December 2014
- Installation / start up is underway

How is Revolution R being used
- Plans are to use RRE for various applications around credit card business – Fraud, retention, promotion, and scoring

Outcome
- This is part of a larger project that has only recently started so no distinct outcomes recorded as of yet
Appendix
Key Value from Microsoft R Server

Open Source R Delivers:

- **Capability**
  - 6500+ Algorithm & Connector Packages Available for Free in CRAN

- **Simplicity**
  - R Skills Transfer / Lower cost of Talent
  - Ease of Integration with Other Analytics Packages & Data
  - Access to Huge Libraries of R Analytical Algorithms

- **Speed**
  - Intel-Optimized Computation

Microsoft R Server\(^1\) Delivers:

- **Speed**
  - Distributed Computation using Parallelized Algorithms
  - In-Hadoop & In Teradata Analysis

- **Scale**
  - No In-Memory Limitations
  - Efficient Data Storage Formats

- **Stability**
  - Commercial Support & Services
  - Platform Longevity

- **Time-to-Results**
  - Powerful IDE & Strong Integration
  - Multi-Platform Scoring

- **Compatibility**
  - Web Services Based Integration Platform

\(^1\) Previously Revolution R Enterprise
Accelerating the speed of business
Write once. Deploy anywhere.

R+CRAN
Revolution R
DistributedR
ConnectR
ScaleR
DeployR
DevelopR

Code Portability Across Platforms

- Hadoop
- EDW
- Clustered Systems
- Workstations & Servers
- In the Cloud
- Hortonworks
- HDInsights
- Cloudera
- SQL Server
- Teradata
- Linux Clusters (LSF For Now)
- Microsoft HPC
- Linux
- Windows
- Azure
R integration into SQL Server 2016 architecture

1. **Sqlsrvr.exe** (Service Account) → **Launch External Process** → **Named Pipe** → **Rlauncher.dll**
2. **RLauncher.dll** → **Launch** → **R.dll**
3. **R.dll** → **RxLink.dll**
4. **RxLink.dll** → **Named Pipe** → **SqlSatellite.dll**
5. **SqlSatellite.dll** ← **TCP Data Channel**

**Additional Components**
- **SqlSatellite.dll**: SQL/R Reader, Writer, Converter
- **RRE**: RRE (MSLP$SQL16)
- **BxlServer.exe**: BxlServer.exe (MSLP$SQL16)
- **RTerm.exe**: RTerm.exe (MSLP$SQL16)
- **Sqlsrvr.exe** (Service Account)
Advanced Analytics in SQL Server 2016 Roadmap

2015

• RRE VMs on Azure (Preview)
• RRE for Teradata and HPC Pack 2008/2012 (remediation)

• RRE VMs on Azure (GA)
• SQL Server CTP3
• RRE on HDI/Win (Preview)
• Support for Teradata Hadoop, HortonWorks Hadoop on SUSE, updated Hadoop versions
• Azure Data Lake Read/Write
• ODBC Write

• RRE on HDI/Linux (Preview)
• RRE in AzureML (Preview)
• Hadoop Spark (Preview)
• RevoPemaR Hadoop
• Additional ML Algorithms
• Native Hadoop file readers

2016

• RRE on HDI/Linux (GA)
• RRE on HDI/Win (GA)
• RRE in AzureML (GA)
• Additional ML Algorithms
Additional motivators

Data Limitations
Size, Sources, Formats, Quality

Modeling Pace
Re-coding, Validation Delays

Talent Cost
Seniority, Longevity, Utilization

Data Lake Delays
Difficulty with Hadoop Spin-Up

Platform Instability
Rapid Evolution of Hadoop

Capabilities
Enabling Self Service, More Algo’s

Microsoft Proprietary and Confidential. Not to be disclosed outside Microsoft.