Microsoft

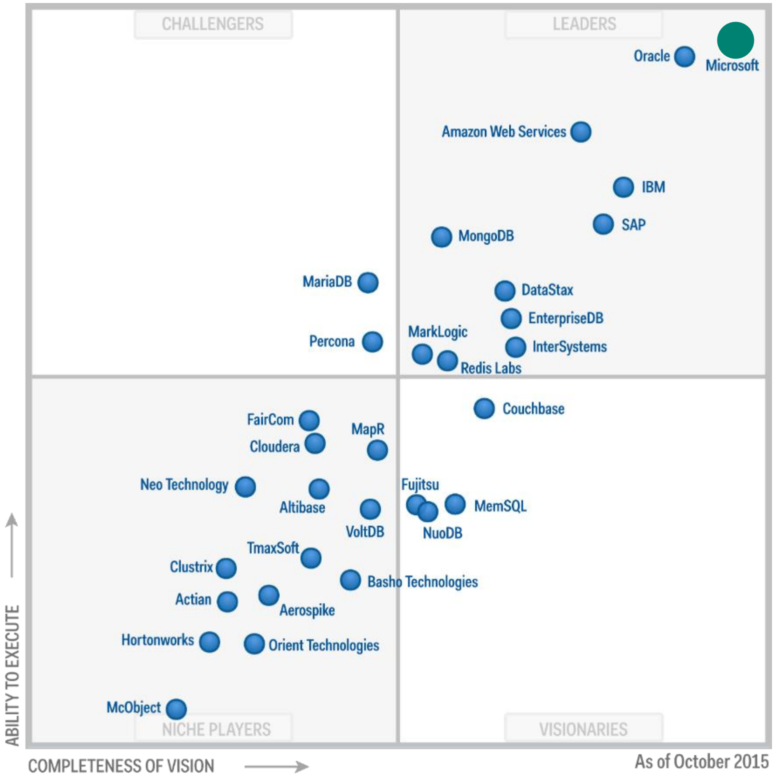
Data Warehousing and Big Data

Technology Deck

Microsoft is a leader for…

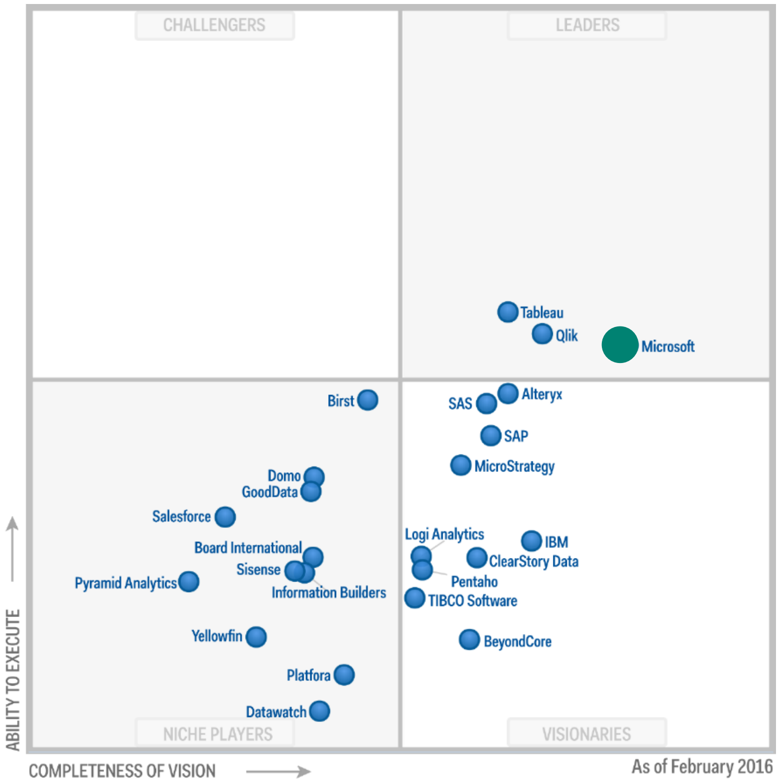
Magic Quadrant for Operational Database Management Systems[1]

Furthest in vision and ability to execute



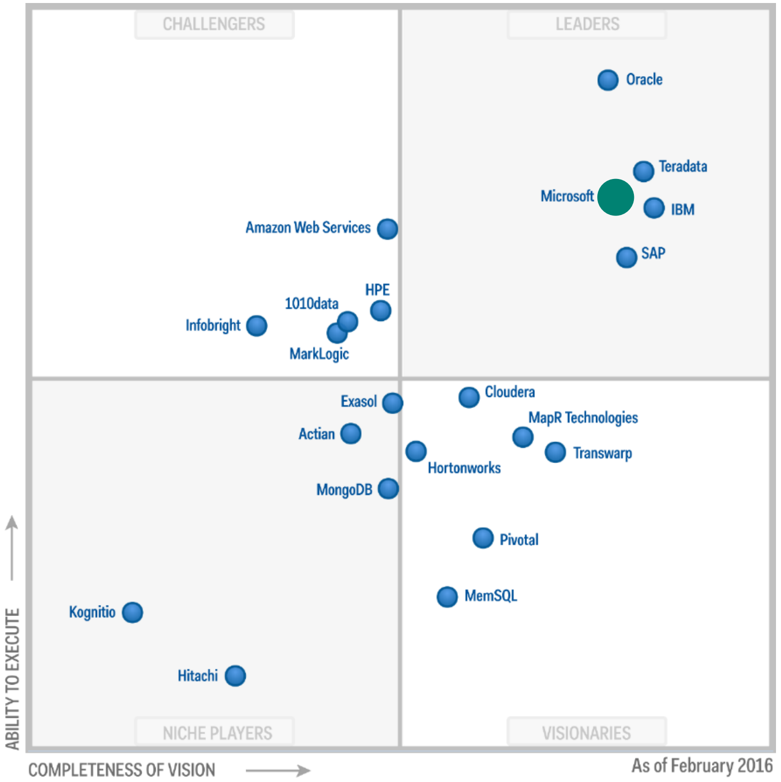
Magic Quadrant for Business Intelligence and Analytics Platforms[2]

Furthest in vision; leader 9 years running



Magic Quadrant for Data Warehouse Database Management Solutions[3]

A leader for the fifth consecutive year



[1] *Gartner "Magic Quadrant for Operational Database Management Systems," by Donald Feinberg , Merv Adrian , Nick Heudecker, Adam Ronthal, October 2015
[2] *Gartner "Magic Quadrant for Business Intelligence and Analytics Platforms," by Josh Parenteau, Rita L. Sallam, Cindi Howson, Joao Tapadinhas, Kurt Schlegel, Thomas W. Oestreich, February 4, 2016
[3] *Gartner "Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics," by Roxane Edjlali and Mark Beyer, February 25, 2016

This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from Microsoft. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

Contents

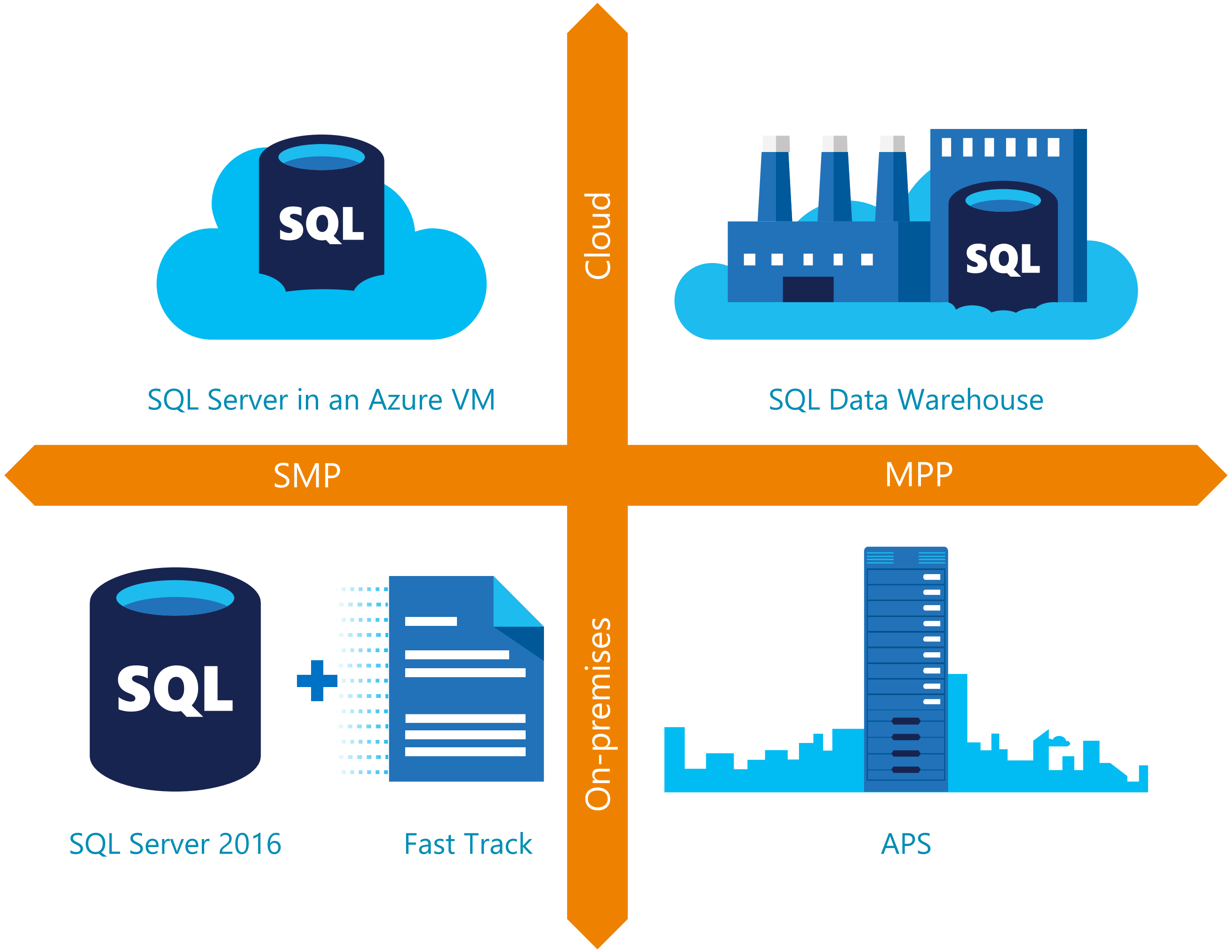
SQL Server Data Warehouse Family

- SQL Server 2016
- APS Appliance
- SQL Data Warehouse

Microsoft Big Data Solutions

- Azure Data Lake
- HD Insight
- Azure Blob Storage

SQL Server Data Warehousing solutions



Symmetric multi-processing (SMP)

On-premises: SQL Server 2016 or SQL Server Fast Track Data Warehouse

Cloud: SQL Server in an Azure VM

Massively parallel processing (MPP)

On-premise: Analytics Platform System (APS)

Cloud: Azure SQL Data Warehouse

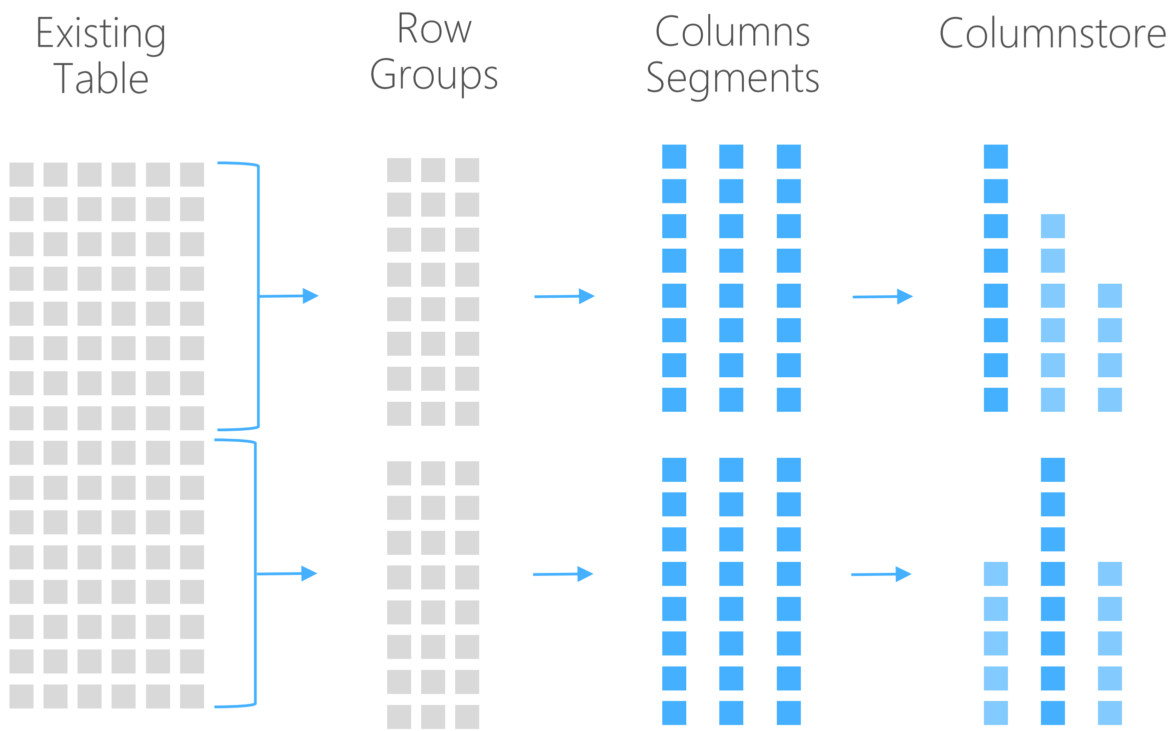
SQL Server 2016

Mission critical performance

Deeper insights across data

Hyperscale cloud

Columnstore: Query performance and data compression



Compressed column segments are added to the columnstore.

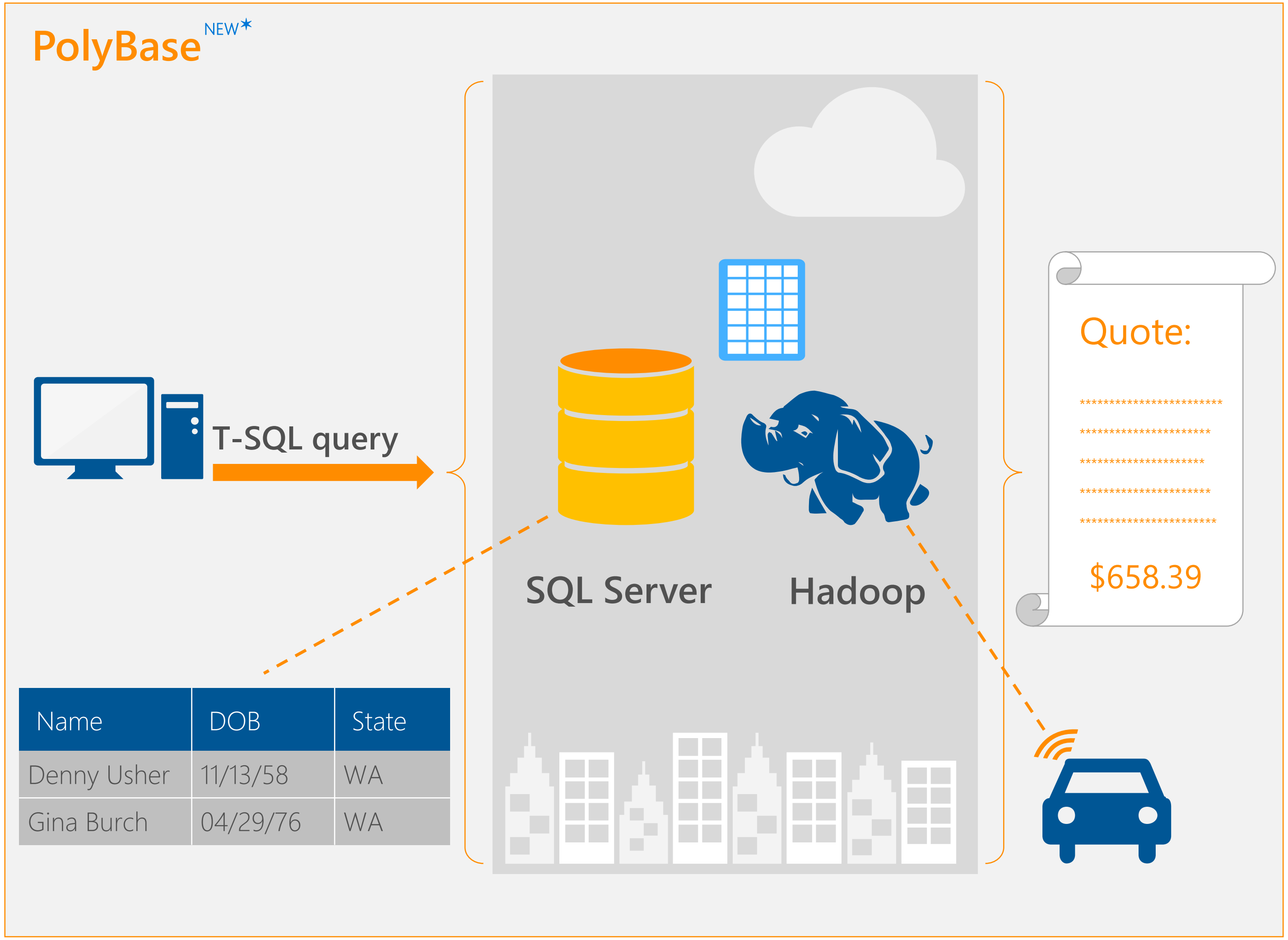
Columnstores are data structures organized in a column-based manner (as opposed to a row-based, traditional table)

Effective in scenarios where indexed columns have several repeated values

Appropriately designed columnstore indexes yield up to 100x the query performance and 10x the data compression of a traditional rowstore (table)

Remove the complexity of big data

T-SQL over Hadoop



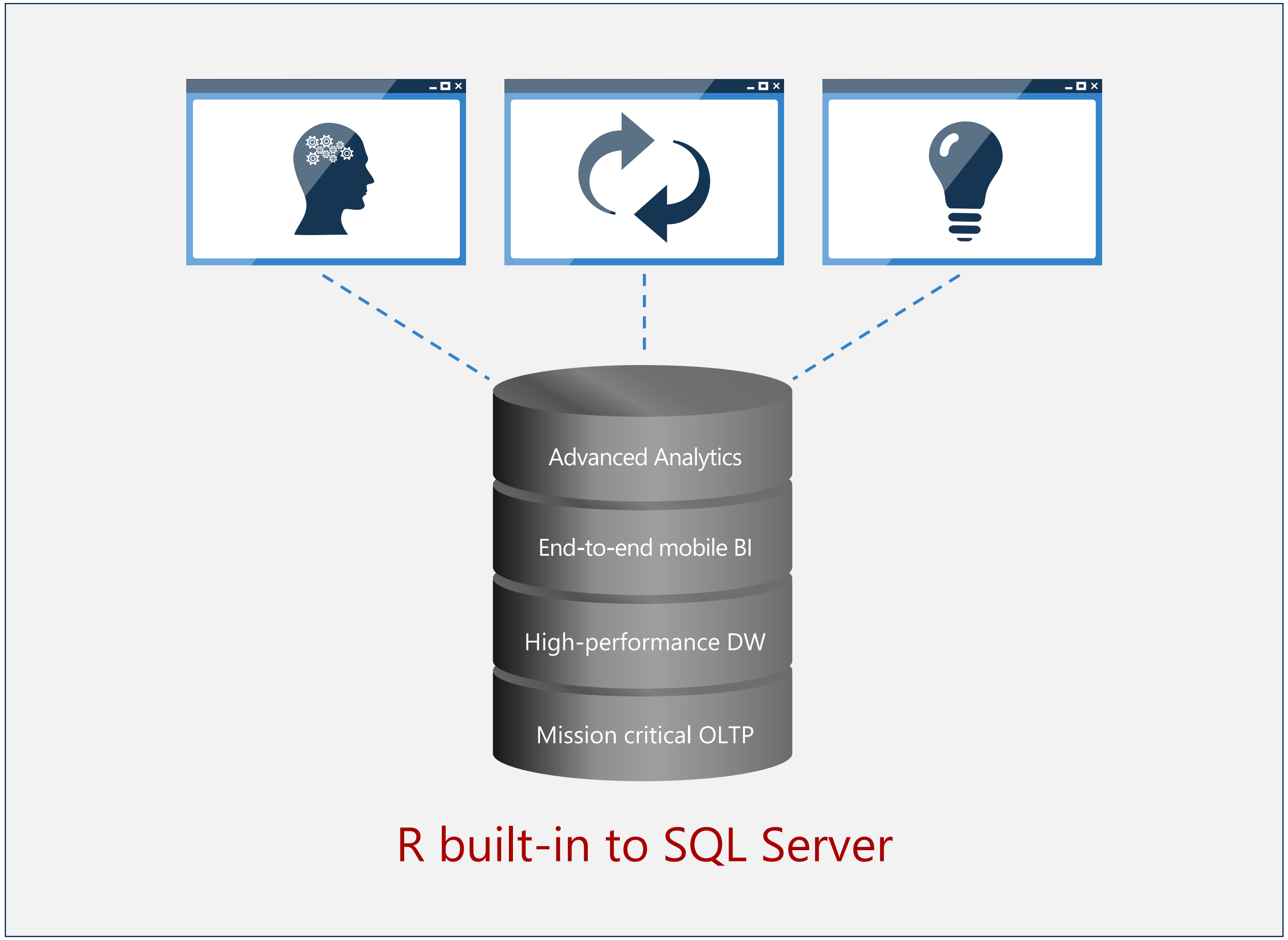
Name	DOB	State
Denny Usher	11/13/58	WA
Gina Burch	04/29/76	WA

Manage structured & unstructured data

- **Simple T-SQL** NEW✶
  to query Hadoop data (HDFS)
- **JSON support** NEW✶

In-database Advanced Analytics

Build intelligent applications with SQL Server R Services



R built-in to SQL Server

- **R built-in to your T-SQL** NEW*
- **Real-time operational analytics** without moving the data NEW*
- **Open Source R with in-memory & massive scale** - multi-threading and massive parallel processing NEW*

SQL Server 2016 (SMP) Reference Architectures

Azure Virtual Machine Image for SQL Server Data Warehouse

SQL Server 2016 pre-built VM image in the Azure gallery

Disk Configuration for Data Warehousing

Developer Edition, BYOL, or per-hour Billing

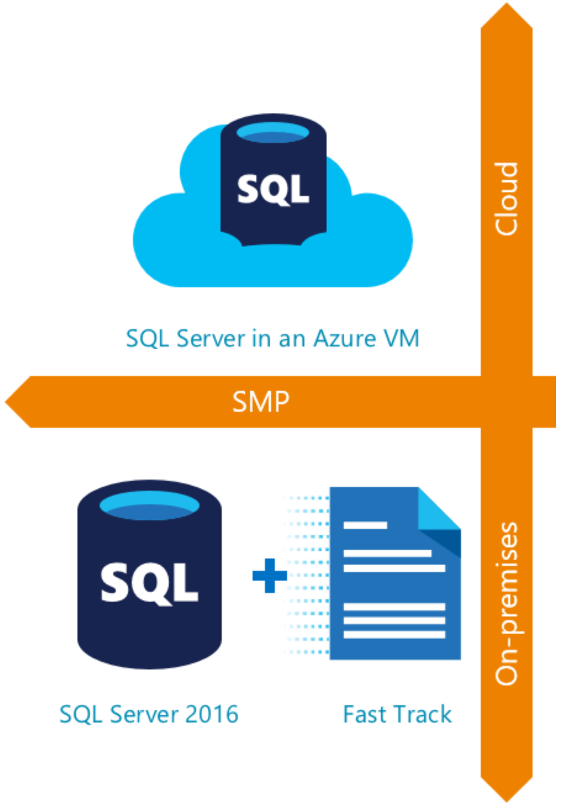
Bottomless storage with Azure Blob Storage of Database files or Polybase

Data Warehouse Fast Track

On-Prem Reference Architecture Implementations

HP, Dell, Lenovo, and other vendors

Tested Configurations from 5TB to 200TB



SQL Server 2016 MPP Solutions

SQL Data Warehouse

Data Warehouse-as-a-service

Elastic Scale in the Cloud

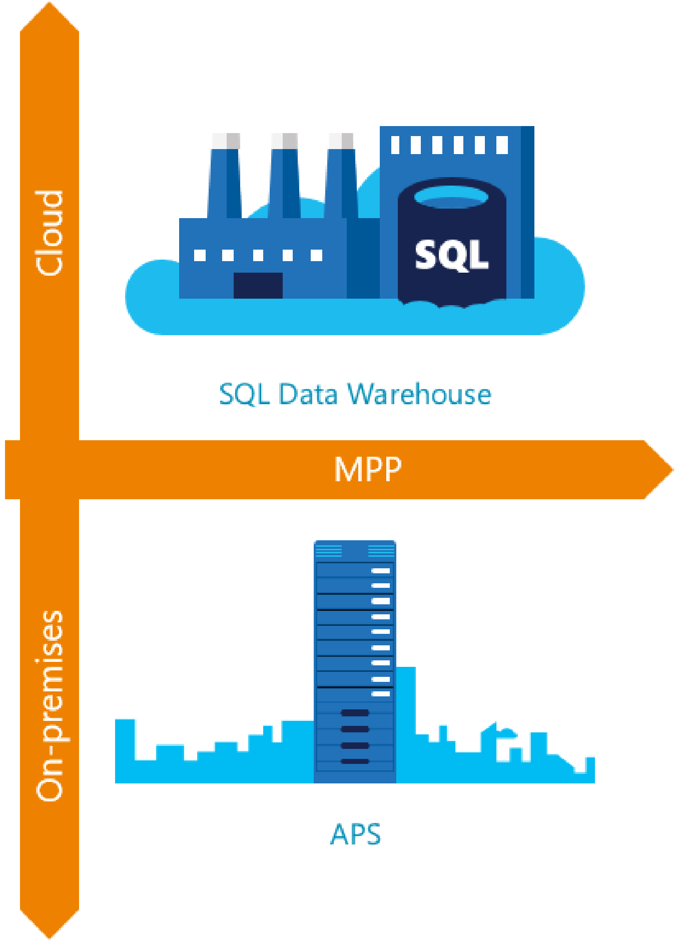
Polybase Connectivity to Azure Blob Storage

Microsoft APS

On-Prem Data Warehouse Appliance

Partial-rack to multi-rack configurations

Polybase Connectivity to Azure Blob Storage and Hadoop



Scaling out your data to petabytes

Scale-out technologies in Analytics Platform System

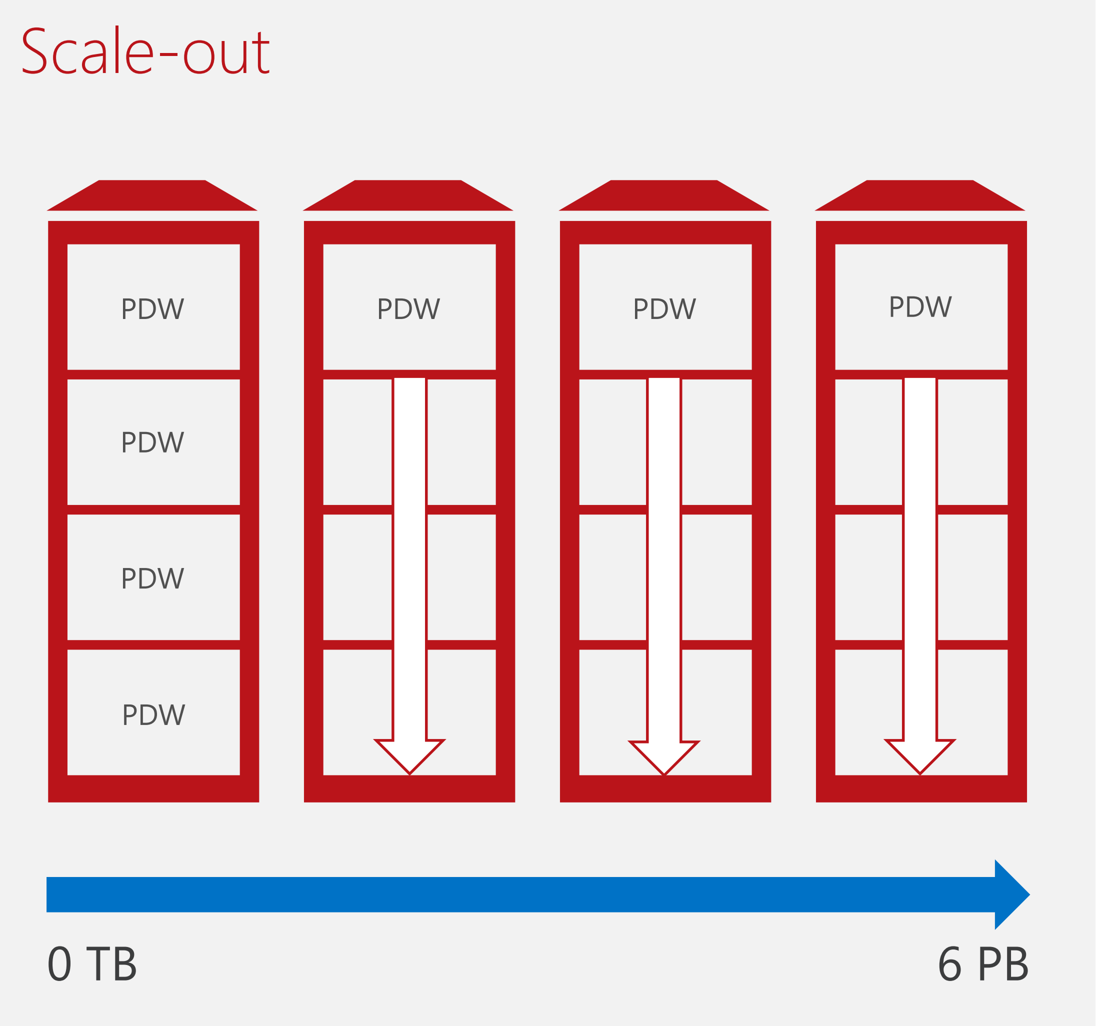
Multiple nodes with dedicated CPU, memory, and storage

Ability to incrementally add hardware for near-linear scale to multiple petabytes

Ability to handle query complexity and concurrency at scale

No “forklift” of prior warehouse to increase capacity

Ability to scale out PDW or Azure Blob Storage



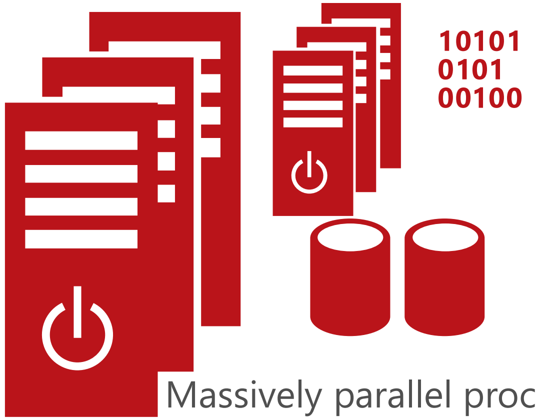
Azure SQL Data Warehouse

A relational data warehouse as a service, fully managed by Microsoft

Industry's first elastic cloud data warehouse with enterprise-grade capabilities

Support for your smallest to largest data storage needs while handling queries up to 100x faster

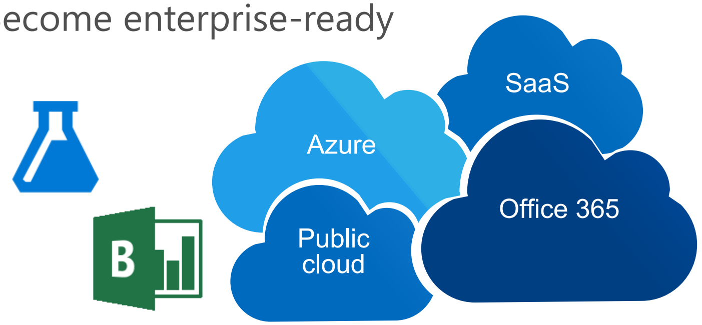
Elastic scale & performance



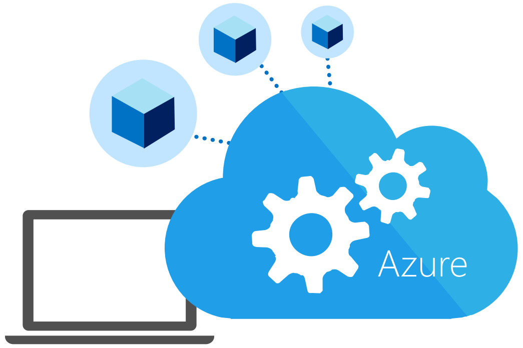
- Massively parallel processing
- Scale to petabytes of data
- Instant-on compute scales in seconds
- Query relational/non-relational

Powered by the cloud

- Get started in minutes
- Integrate with Azure ML, Power BI, and ADF
- Become enterprise-ready



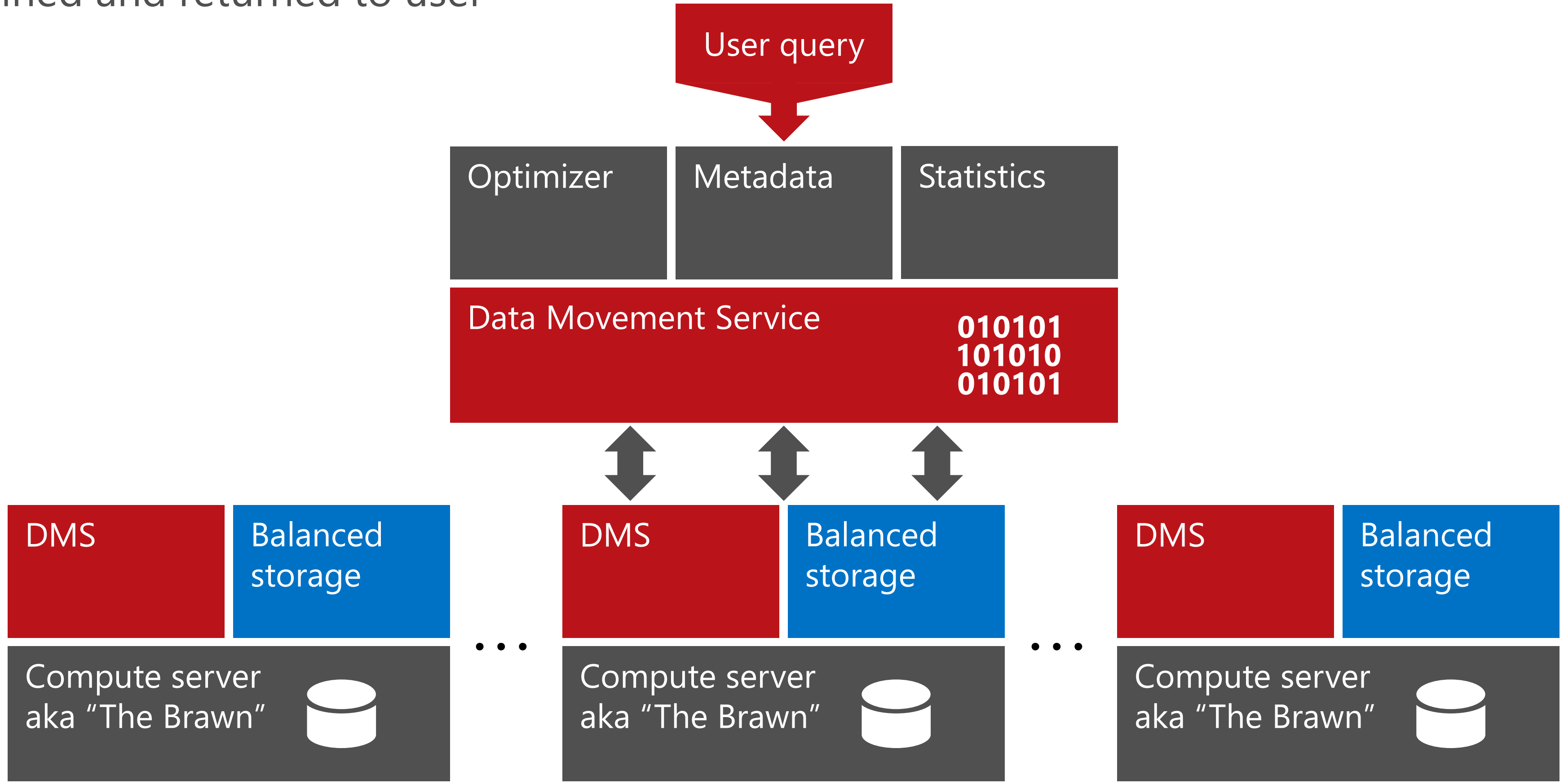
Market-leading price & performance



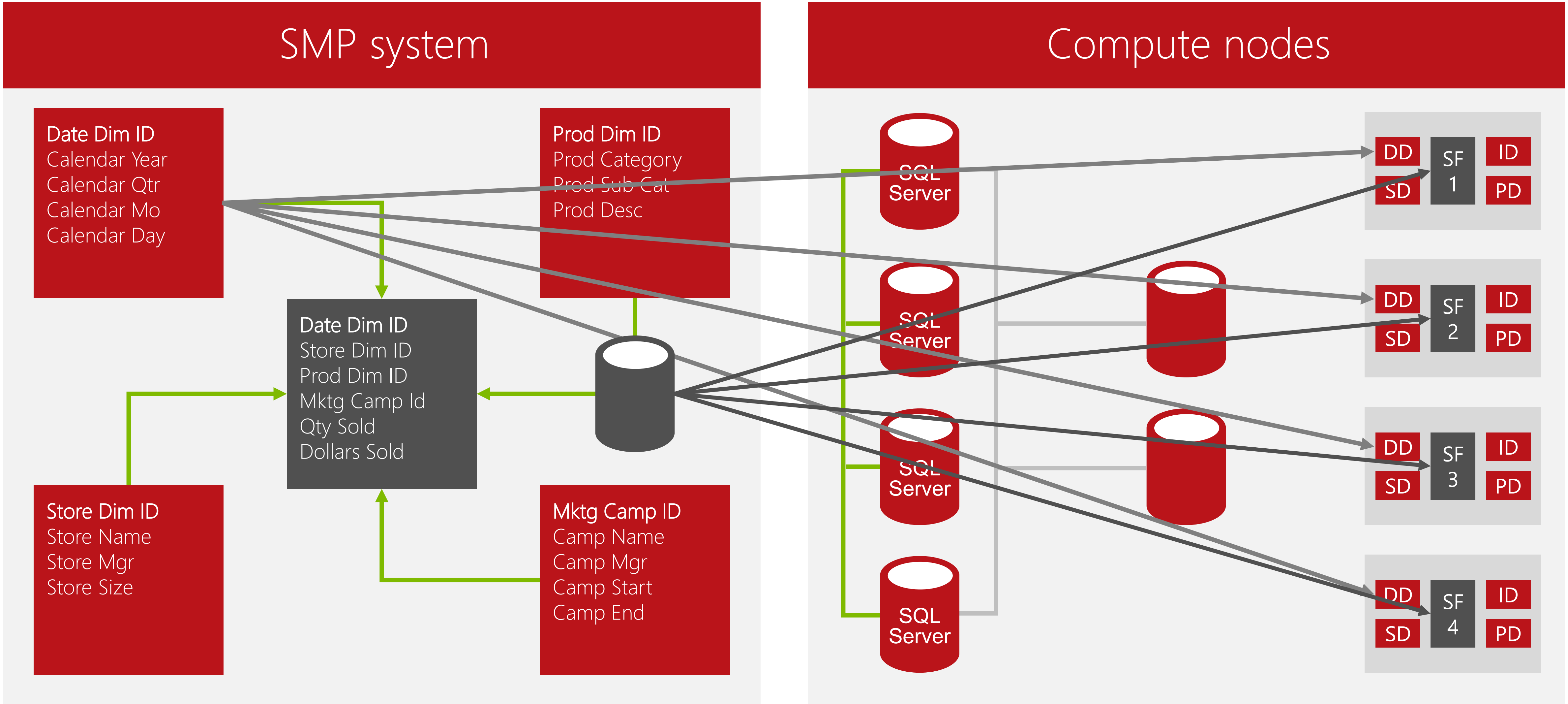
- Use simple billing compute and storage
- Pay for what you need, when you need it with dynamic pause
- Bring DW to the cloud without rewriting

Logical architecture

1. Optimizer creates parallel query plan
2. Each compute server runs portion of query in parallel
3. Data is combined and returned to user



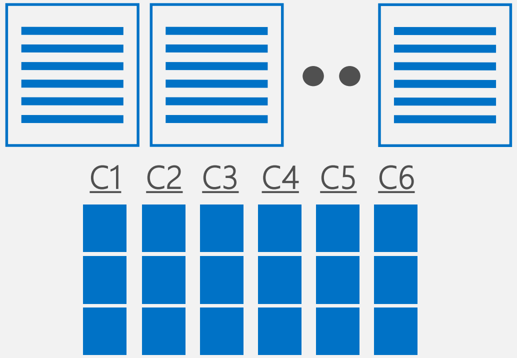
MPP SQL table geometries



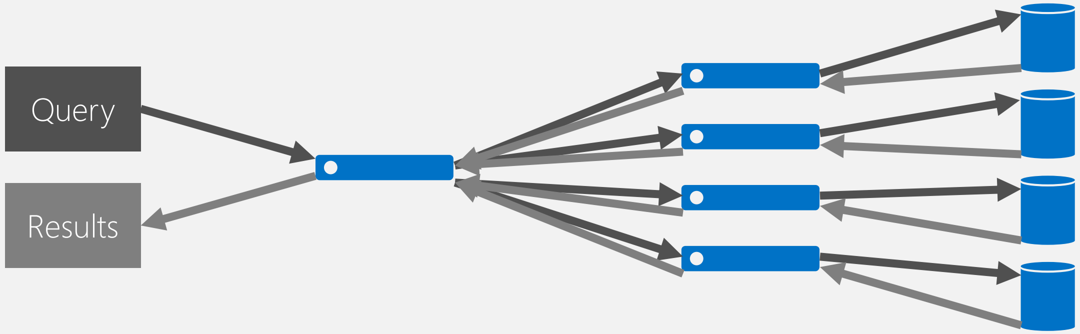
Blazing-fast performance

MPP and in-memory columnstore for next-generation performance

Columnstore index representation



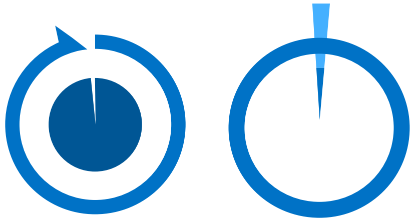
Parallel query execution

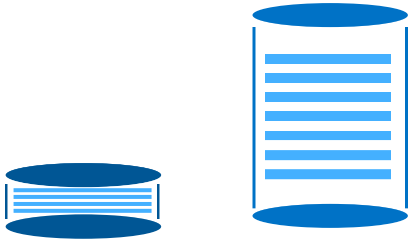


Updateable clustered columnstore vs. table with customary indexing

Up to 100x faster queries

Up to 15x more compression





Data storage in columnar format for massive compression

Data loading into or out of memory for next-generation performance, with up to 60% improvement in data loading speed

Updateable and clustered for real-time trickle loading

Better together: Azure SQL DW Service and APS

Test/dev

Test new ideas in SQL Data Warehouse before rolling out to production in APS

Age data

Age data to SQL Data Warehouse, but maintain full MPP power

Company policy restrictions

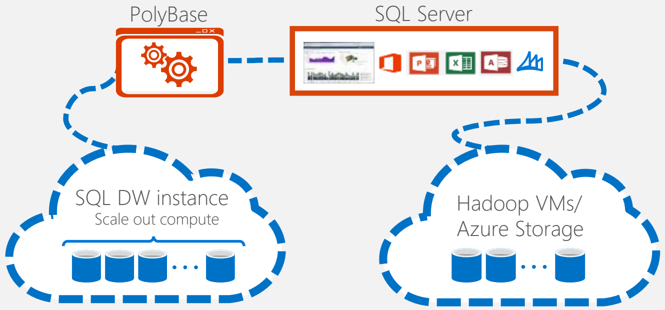
Store data in APS that company policy prohibits from being in the cloud

Disaster recovery

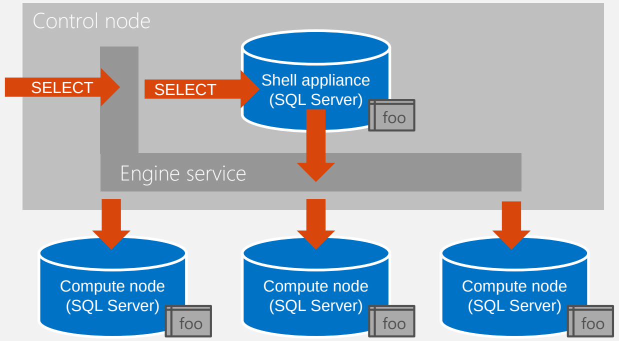
Use SQL Data Warehouse or APS as disaster recovery solution with dual load

Conclusion

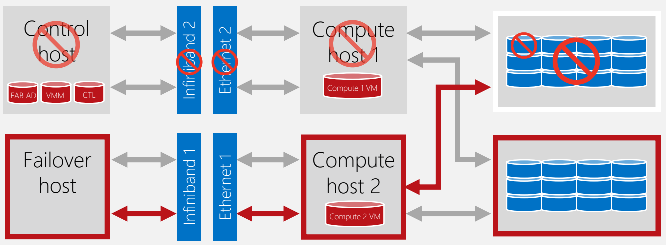
PolyBase



Massively parallel processing



High availability



Microsoft APS

The Microsoft Analytics Platform System can meet the demands of your evolving data warehouse environment with its scale-out, massively parallel processing integrated system supporting hybrid data warehouse scenarios. It provides the ability to query across relational and non-relational data by leveraging Microsoft PolyBase and industry-leading big data technologies.

Azure SQL Data Warehouse enables APS customers with different workloads to leverage a cloud-based MPP engine and cloud-based analytics by supporting a hybrid architecture or eco-system with APS + Azure SQL Data Warehouse.

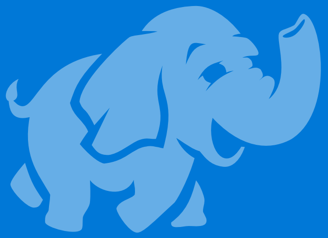
Azure SQL Data Warehouse

Big Data from Microsoft

Microsoft

Azure HDInsight

A Cloud Spark and Hadoop service for the Enterprise



Reliable with an **industry leading SLA**

Enterprise-grade security and **monitoring**

Productive platform for **developers** and **scientists**

Cost effective cloud scale

Integration with leading **ISV applications**

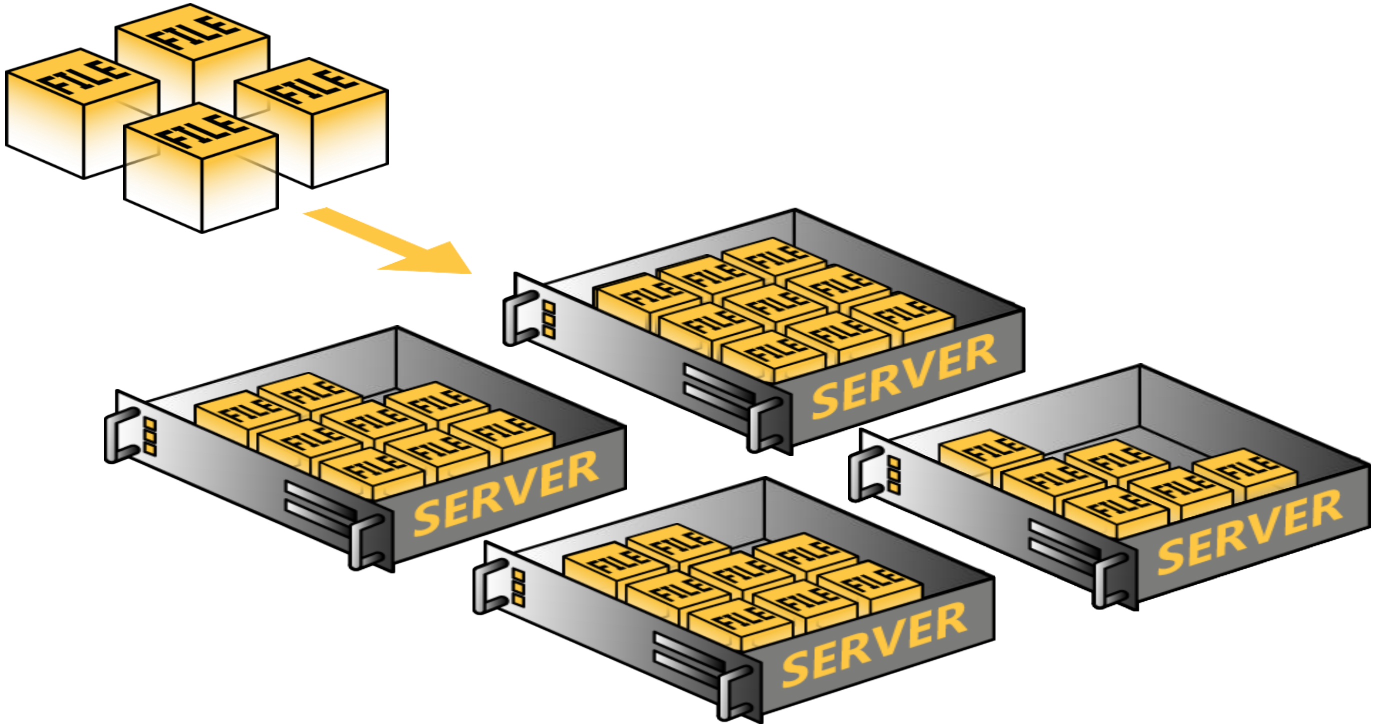
Easy for administrators to **manage**

63% lower TCO than deploy your own Hadoop on-premises*

*IDC study "The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight"

So how does it work?

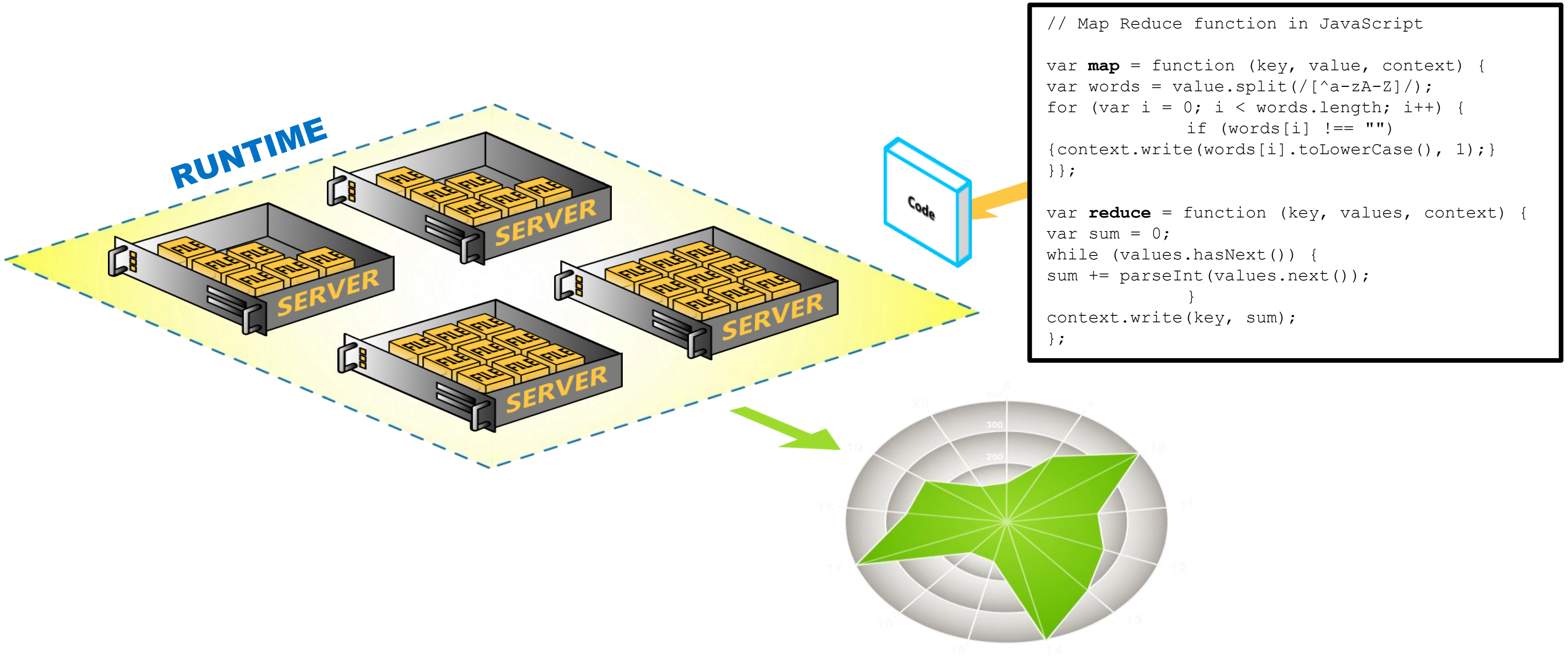
First, store the data



Microsoft Technology Centers

So how does it work?

Second, take the processing to the data

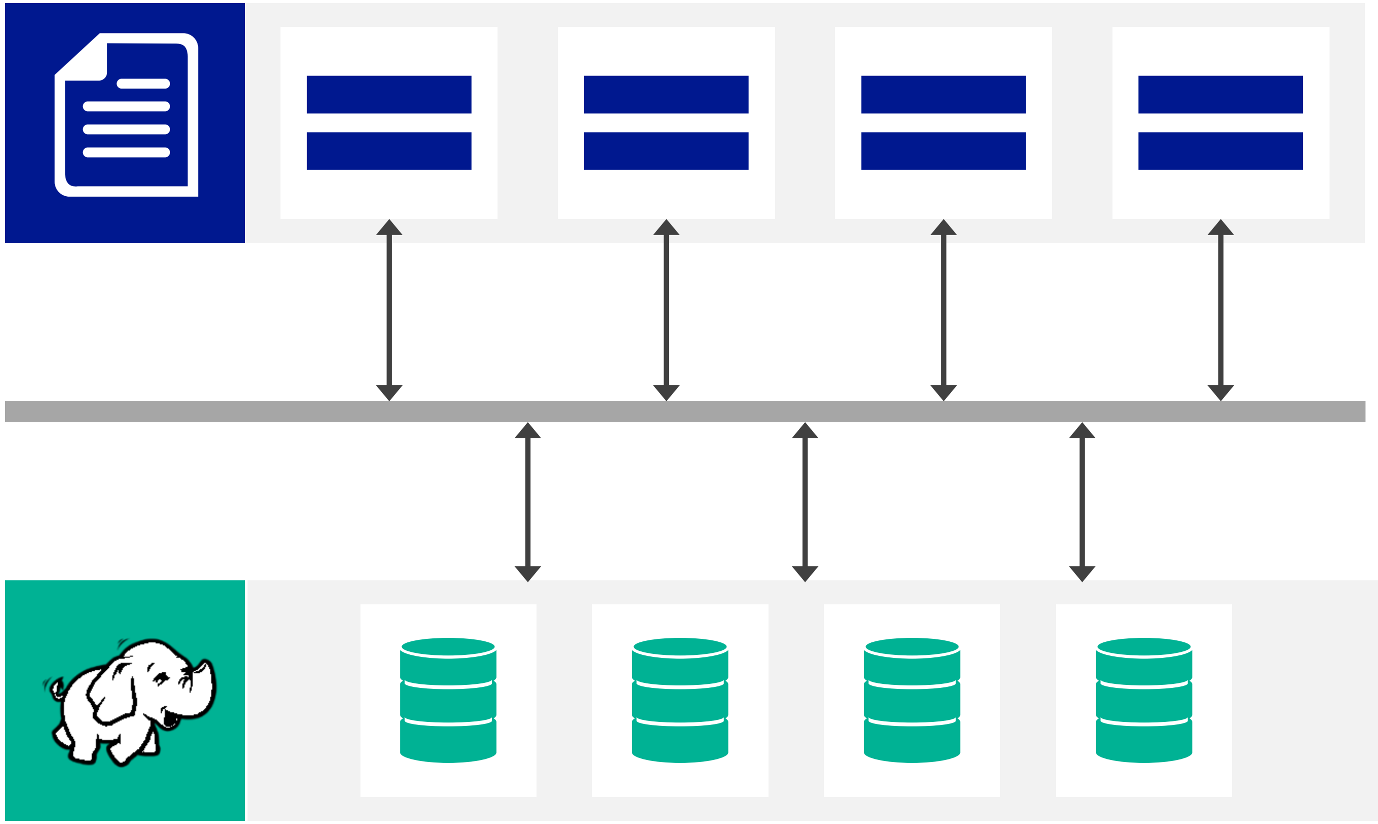


```
// Map Reduce function in JavaScript

var map = function (key, value, context) {
var words = value.split(/[^a-zA-Z]/);
for (var i = 0; i < words.length; i++) {
              if (words[i] !== "")
{context.write(words[i].toLowerCase(), 1);}
}};

var reduce = function (key, values, context) {
var sum = 0;
while (values.hasNext()) {
sum += parseInt(values.next());
              }
context.write(key, sum);
};
```

HDInsight Storage Infrastructure



Azure Blob Storage

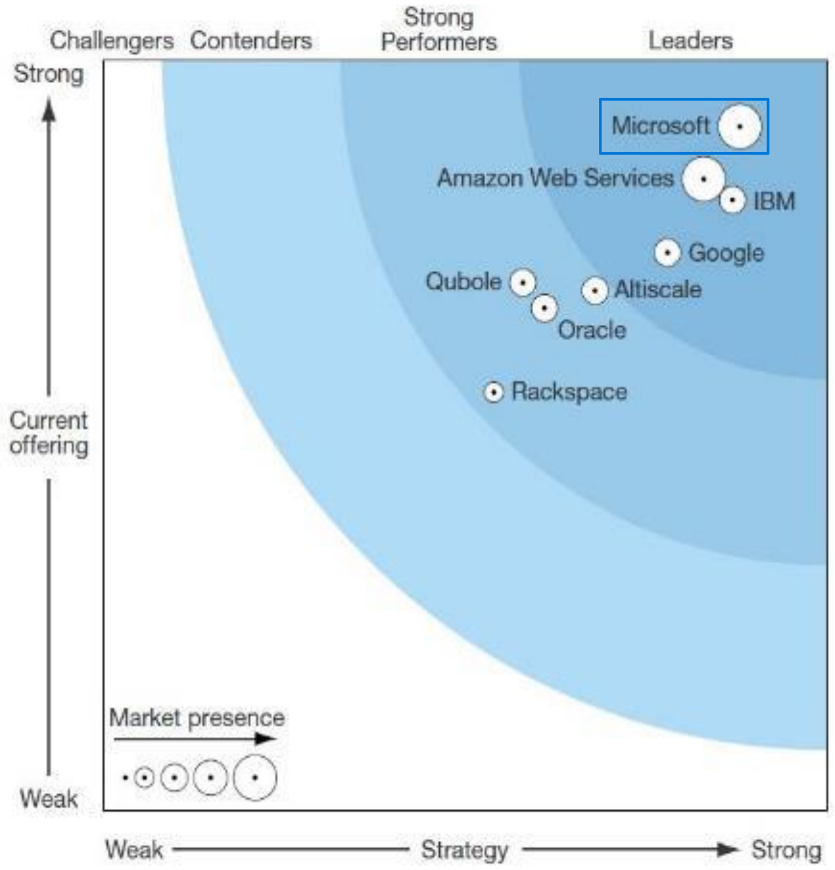
Azure Flat Network Storage

HDInsight Compute Nodes
(Large VMs)

http://dennyglee.com/2013/03/18/why-use-blob-storage-with-hdinsight-on-azure/

Recognized by top analysts

FORRESTER®



Forrester Wave for Big Data Hadoop Cloud

- Named industry leader by Forrester with the most comprehensive, scalable, and integrated platforms*
- Recognized for its cloud-first strategy that is paying off*

*The Forrester WaveTM: Big Data Hadoop Cloud Solutions, Q2 2016.

Lower total cost of ownership

- No hardware
- Hadoop support included with Azure support
- Pay only for what you use
- Independently scale storage and compute
- No need to hire specialized operations team
- 63% lower total cost of ownership than on-premises*

*IDC study "The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight"

Azure
Data Lake Store

A No limits Data Lake that powers Big Data Analytics

Petabyte size files and **Trillions of objects**

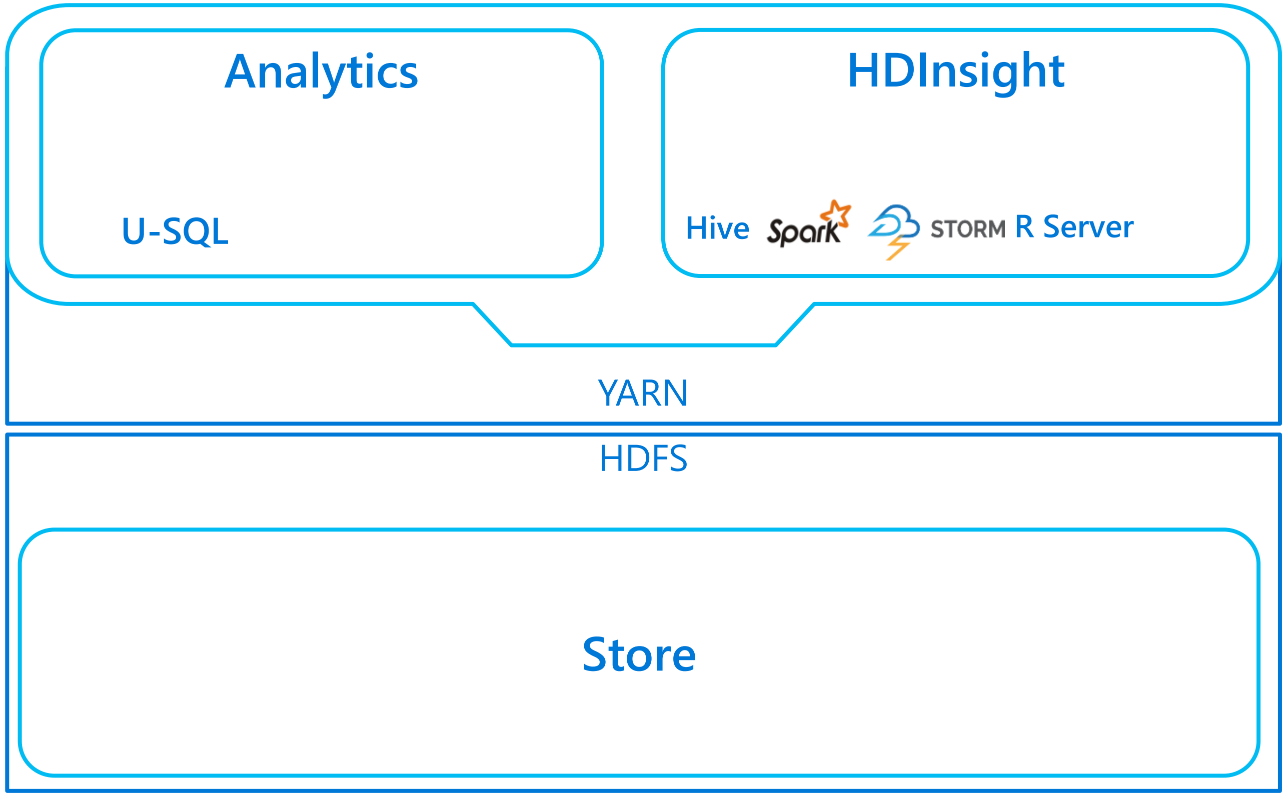
Scalable throughput for **massively parallel analytics**

HDFS for the cloud

Always encrypted, role-based security & auditing

Enterprise-grade support

Azure Data Lake



Store and analyze data of any kind and size

Develop faster, debug and optimize smarter

Interactively explore patterns in your data

No learning curve

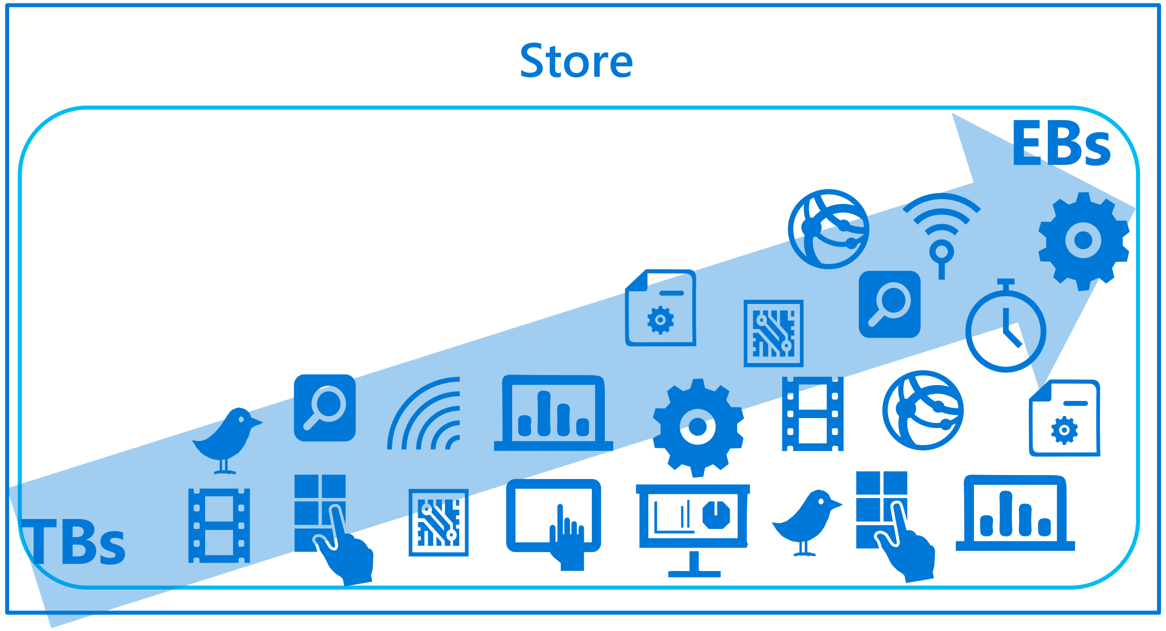
Managed and supported

Dynamically scales to match your business priorities

Enterprise-grade security

Built on YARN, designed for the cloud

Petabyte size files and Trillions of objects

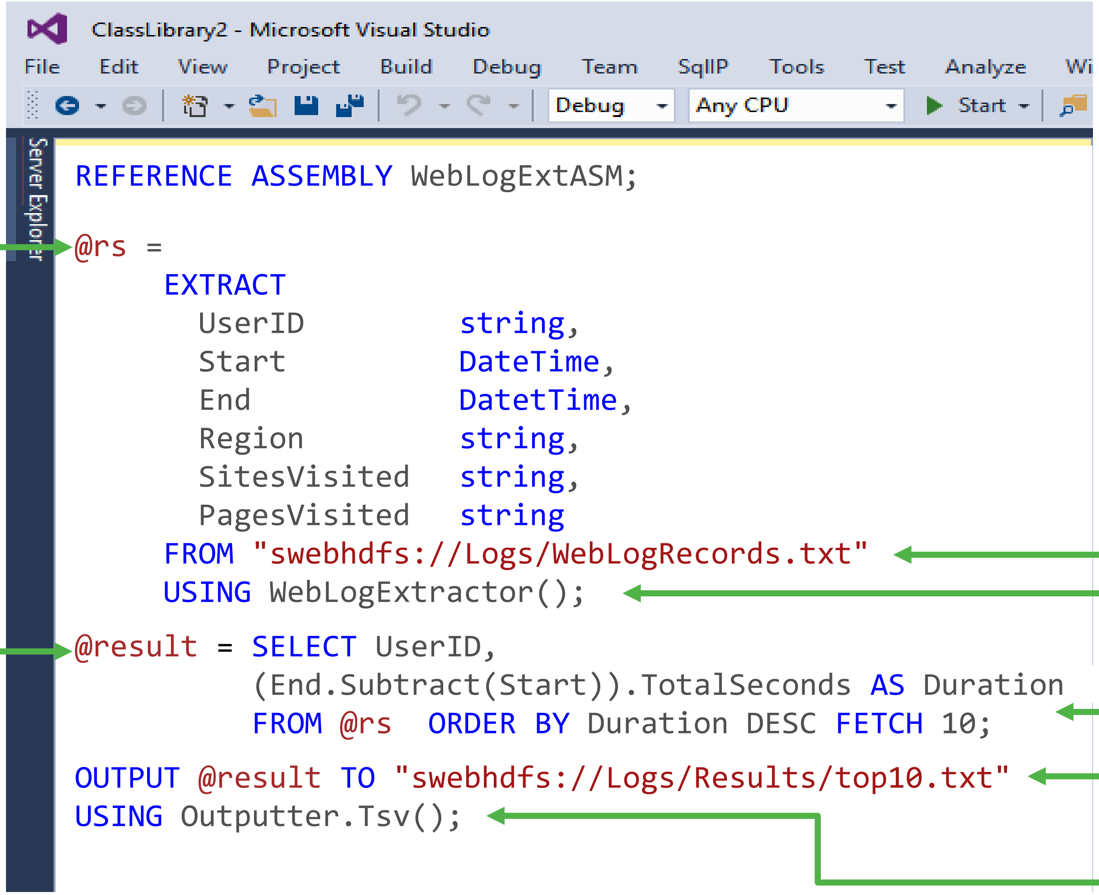


- Store data in it's native format
- PB sized files, 200x larger than anyone else
- Scalable throughput for massively parallel analytics
- No need to redesign application or reparation data at higher scale

Anatomy of a U-SQL query

Query 1

10 log records by Duration (End time minus Start time). Sort rows in descending order of Duration.



```
REFERENCE ASSEMBLY WebLogExtASM;

@rs =
     EXTRACT
       UserID          string,
       Start           DateTime,
       End             DatetTime,
       Region          string,
       SitesVisited    string,
       PagesVisited    string
     FROM "swebhdfs://Logs/WebLogRecords.txt"
     USING WebLogExtractor();

@result = SELECT UserID,
          (End.Subtract(Start)).TotalSeconds AS Duration
          FROM @rs  ORDER BY Duration DESC FETCH 10;

OUTPUT @result TO "swebhdfs://Logs/Results/top10.txt"
USING Outputter.Tsv();
```

Rowset: Conceptually is like an intermediate table… is how U-SQL passes data between statements

- U-SQL types are the same as C# types
- The structure (schema) is first imposed when the data is first extracted/read from the file (schema-on-read)

Input is read from this file in ADL

Custom function to read from input file

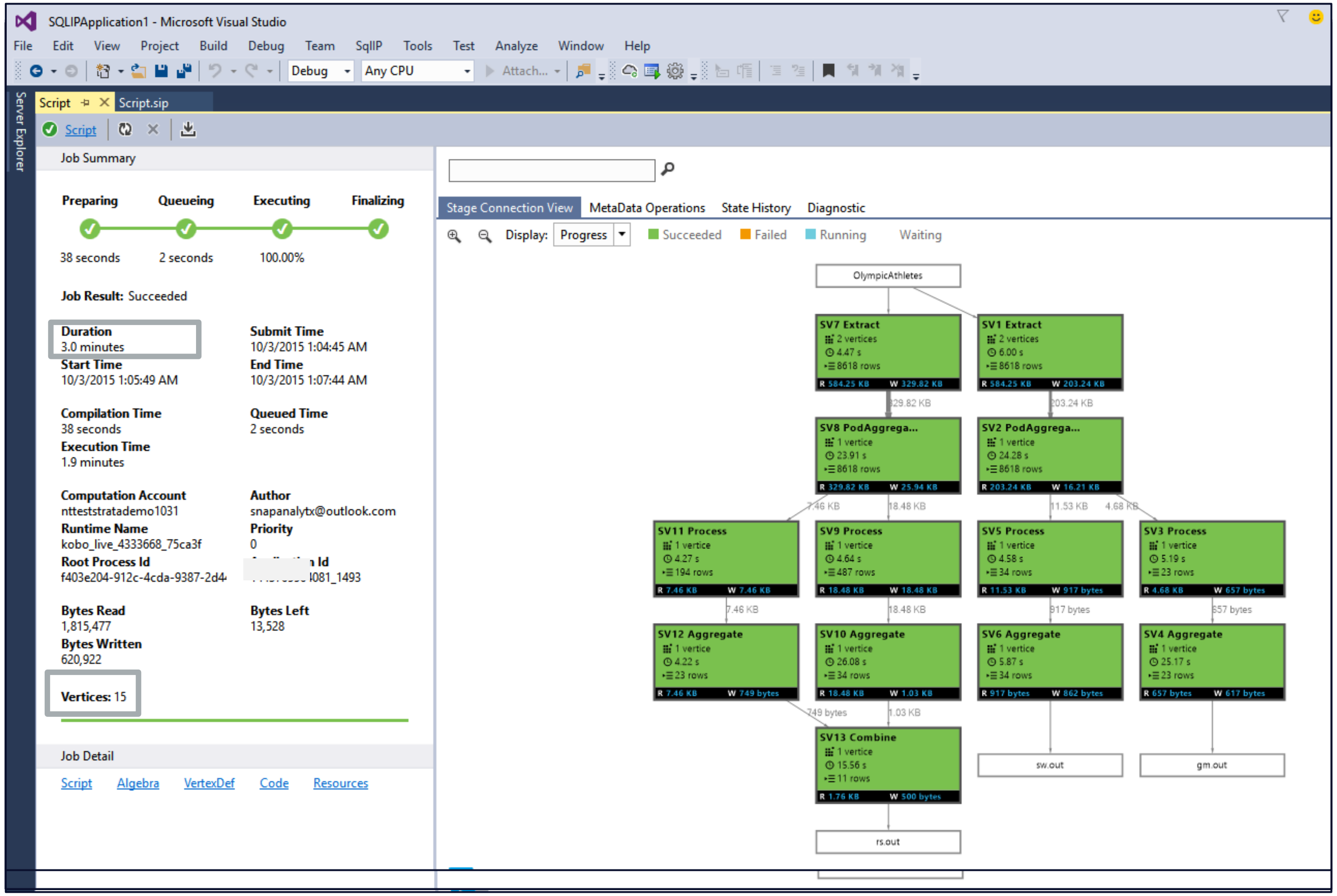
C# Expression

Output is stored in this file in ADL

Built-in function that writes the output in TSV format

Job execution graph

- After a job is submitted the progress of the execution of the job as it goes through the different stages is shown and updated continuously
- Important stats about the job are also displayed and updated continuously



Job Summary

Preparing	Queueing	Executing	Finalizing
38 seconds	2 seconds	100.00%	

Job Result: Succeeded

Duration
3.0 minutes
Start Time
10/3/2015 1:05:49 AM

Submit Time
10/3/2015 1:04:45 AM
End Time
10/3/2015 1:07:44 AM

Compilation Time
38 seconds
Execution Time
1.9 minutes

Queued Time
2 seconds

Computation Account
ntteststratademo1031
Runtime Name
kobo_live_4333668_75ca3f
Root Process Id
f403e204-912c-4cda-9387-2d4

Author
snapanalytx@outlook.com
Priority
0
[redacted]

Bytes Read
1,815,477
Bytes Written
620,922

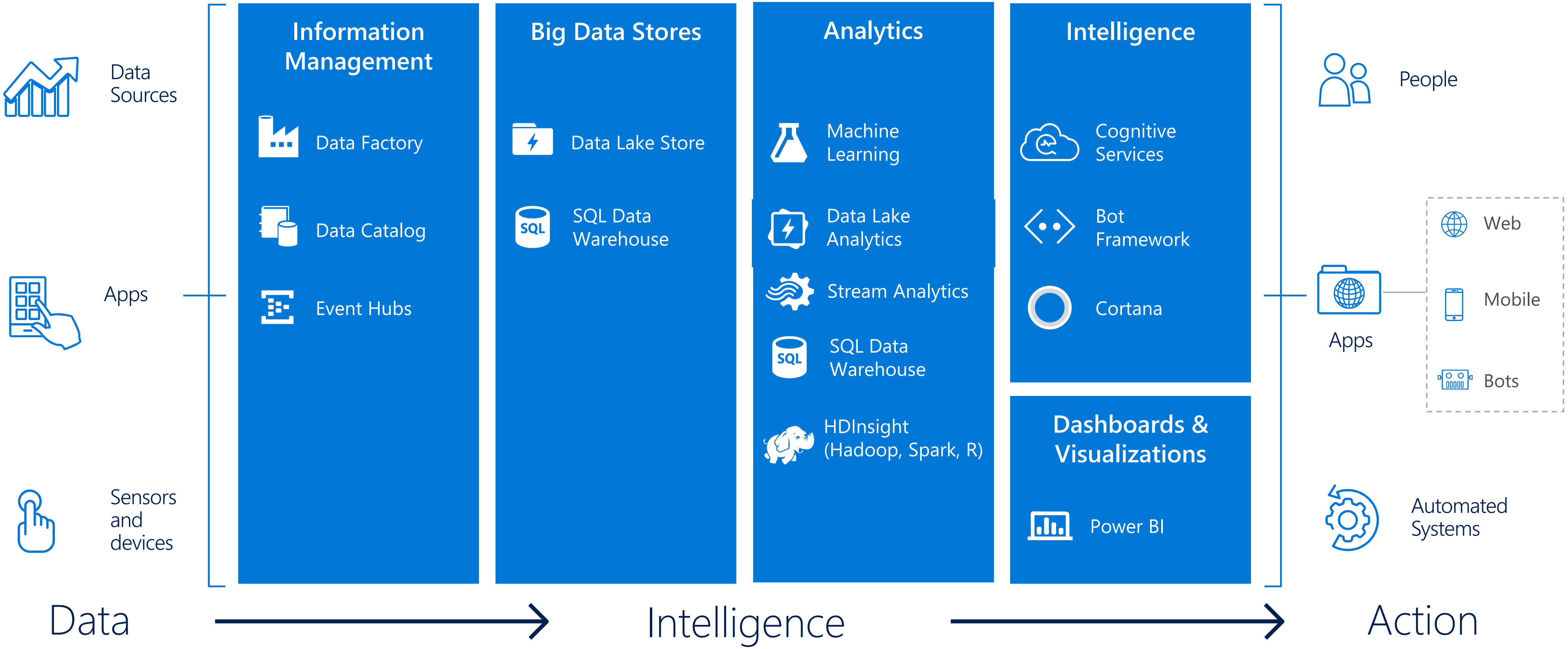
Bytes Left
13,528

Vertices: 15

Job Detail

Script Algebra VertexDef Code Resources

Putting it all together



Data Sources

Apps

Sensors and devices

Information Management

- Data Factory
- Data Catalog
- Event Hubs

Big Data Stores

- Data Lake Store
- SQL Data Warehouse

Analytics

- Machine Learning
- Data Lake Analytics
- Stream Analytics
- SQL Data Warehouse
- HDInsight (Hadoop, Spark, R)

Intelligence

- Cognitive Services
- Bot Framework
- Cortana

Dashboards & Visualizations

- Power BI

People

Apps

- Web
- Mobile
- Bots

Automated Systems

Data → Intelligence → Action

Microsoft

The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.