Data Warehousing and Big Data

Technology Deck
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
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

PolyBase

Simple T-SQL to query Hadoop data (HDFS)

JSON support

Quote: $658.39

- Name: Denny Usher, DOB: 11/13/58, State: WA
- Name: Gina Burch, DOB: 04/29/76, State: WA
In-database Advanced Analytics
Build intelligent applications with SQL Server R Services

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

R built-in to SQL Server
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

**SMP system**

- **Date Dim ID**
  - Calendar Year
  - Calendar Qtr
  - Calendar Mo
  - Calendar Day

- **Store Dim ID**
  - Store Name
  - Store Mgr
  - Store Size

- **Prod Dim ID**
  - Prod Category
  - Prod Sub Cat
  - Prod Desc

- **Mktg Camp ID**
  - Camp Name
  - Camp Mgr
  - Camp Start
  - Camp End

**Compute nodes**

- SQL Server
  - DD
  - SF
  - ID
  - PD

- SQL Server
  - DD
  - SF
  - ID
  - PD

- SQL Server
  - DD
  - SF
  - ID
  - PD

- SQL Server
  - DD
  - SF
  - ID
  - PD
Blazing-fast performance

MPP and in-memory columnstore for next-generation performance

**Updateable clustered columnstore vs. table with customary indexing**

- **Up to 100x** faster queries
- **Up to 15x** more compression

**Columnstore index representation**

- 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

**Parallel query execution**
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
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 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
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://dennyslee.com/2013/03/18/why-use-blob-storage-with-hdinsight-on-azure/
Recognized by top analysts

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 its 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.

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.
Putting it all together

Information Management
- Data Sources
  - 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

Web

Mobile

Bots

Automated Systems

Apps

Data

Intelligence

Action

Data Sources

Apps

Sensors and devices