Azure Thursdays 5/18/2017: Demystifying Data Platform Services in Azure

Kyle Wilson, Azure TSP, MAD
Daniel Sheehan, Azure DSA, NYM
Chris Santaniello, Azure TSP, NYM

Huge Credit to: James Serra, http://www.jamesserra.com/, @JamesSerra
Azure News

- Low Priority VMs for Batch Computing
- Azure Functions Runtime
- Azure Cloud Shell
- Azure Mobile App
- MySQL and Postgres managed DBs
- Cosmos DB
- Great List of Key Azure announcements from Build 2017
New Family Members

• Azure Cosmos DB
• Azure PostgresDB Service
• Azure MySQL DB Service
Boiling it Down

• Databases (inclusive of BigData & NoSQL) & Data Warehouses
• Data Analytics & AI/Machine Learning
• Data Manipulation Services/Tools
Database* Solutions in Azure

- SQL Server (IaaS)
- SQL Server Stretch DB (PaaS)
- Azure SQL (PaaS)
- Azure PostgreSQL (PaaS)
- Azure MySQL (PaaS)
- Cosmos DB (Paas)
- Redis Cache (Paas)
- HDInsight (PaaS)
- Azure Table Storage (PaaS)

- SAP Hana (IaaS)
- Oracle (IaaS)
- MySQL (IaaS)
- PostgreSQL (IaaS)
- Cassandra (IaaS)
- MongoDB (IaaS)
- Redis (IaaS)
- HortonWorks Hadoop (IaaS)
- Cloudera Enterprise (IaaS)

- And many more
Data Warehouse, Analytics & Machine Learning Solutions

- Azure SQL DW (PaaS, MPP)
- Azure Analysis Services (PaaS, SMP)
- Data Lake Analytics (PaaS)
- Stream Analytics (PaaS)
- Time Series Insights (PaaS)
- R Server for HDInsight (PaaS)
- Apache Spark for HDInsight (PaaS)
- Azure Machine Learning (PaaS)
- R Server (IaaS)
- Cognitive Services (PaaS):
  - Vision Services
  - Language Services

- Speech Services
- Search Services
- Knowledge Services
- Labs Services

MARKETPLACE

- Apache Spark (IaaS)
- Apache Storm (IaaS)
- Pyramid Analytics (IaaS)
- HPE Vertica (IaaS)
- Alteryx (IaaS)
- And many more…
Data Processing Services & Tools

- Azure Data Catalog (PaaS)
- Azure Data Factory (PaaS)
- Azure IOT HUB (PaaS)
- Azure Event Hubs (PaaS)
- Power BI Embedded (PaaS)
- Azure Bot Services (PaaS)
- BizTalk Services (PaaS)
- SQL Server Integration Services (IaaS)
- Logic Apps (PaaS)
- And many more in the Marketplace
Processing Factor - SMP vs MPP

SMP - Symmetric Multiprocessing
- Multiple CPUs used to complete individual processes simultaneously
- All CPUs share the same memory, disks, and network controllers (scale-up)
- All SQL Server implementations up until now have been SMP
- Mostly, the solution is housed on a shared SAN

MPP - Massively Parallel Processing
- Uses many separate CPUs running in parallel to execute a single program
- Shared Nothing: Each CPU has its own memory and disk (scale-out)
- Segments communicate using high-speed network between nodes
Data Factor - Considering Data Types

- **Unstructured**: Audio, video, images. Meaningless without adding some structure
- **Semi-Structured**: JSON, XML, sensor data, social media, device data, web logs. Flexible data model structure
- **Structured**: CSV, Columnar Storage (Parquet, ORC). Strict data model structure
Volume & Location Factors

• Streaming Data
• Geographic Distribution
• Batch Processing
• Small Records OR Huge Row sets
What is Big Data?

Big Data = All Data!

- Variety: It can be structured, semi-structured, or unstructured
- Velocity: It can be streaming, near real-time or batch
- Volume: It can be 1GB or 1PB
- Big data is the new currency
New Approaches

Data Lake Transformation (ELT not ETL)

All data sources are considered

Leverages the power of on-prem technologies and the cloud for storage and capture

Native formats, streaming data, big data

Extract and load, no/minimal transform

Storage of data in near-native format

Orchestration becomes possible

Streaming data accommodation becomes possible

Refineries transform data on read

Produce curated data sets to integrate with traditional warehouses

Users discover published data sets/services using familiar tools
Big Data Solutions Decision Tree

Thanks to Ivan Kosyakov: https://biz-excellence.com/2016/08/30/big-data-dt/
Machine Learning Solutions Decision Tree

Thanks to Ivan Kosyakov: https://biz-excellence.com/2016/09/13/machine-learning-dt/
The **Data Management Platform** for Analytics

- **Relational data**
  - OLTP
  - ERP
  - CRM
  - LOB

- **Non-relational data**
  - Web
  - Media
  - Social media
  - Devices

- **Data warehousing**
- **Big Data processing**

- **Data virtualization**

- **Any BI tool**
  - Dashboards
  - Reporting
  - Mobile BI
  - Cubes

- **Advanced Analytics**
  - Machine Learning
  - Stream analytics
  - Cognitive
  - AI

- **Any language**
  - .NET
  - Java
  - R
  - Python
  - Ruby
  - PHP
  - Scala

- **On-premises**
- **Cloud**
<table>
<thead>
<tr>
<th>Volume Per Day</th>
<th>Real-world Transactions Per Day</th>
<th>Real-world Transactions Per Second</th>
<th>Relational DB</th>
<th>Document Store</th>
<th>Key Value or Wide Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 GB</td>
<td>8.64B</td>
<td>100,000</td>
<td>As Is</td>
<td></td>
<td></td>
</tr>
<tr>
<td>86 GB</td>
<td>86.4B</td>
<td>1M</td>
<td>Tuned*</td>
<td>As Is</td>
<td></td>
</tr>
<tr>
<td>432 GB</td>
<td>432B</td>
<td>5M</td>
<td>Appliance</td>
<td>Tuned*</td>
<td>As Is</td>
</tr>
<tr>
<td>864 GB</td>
<td>864B</td>
<td>10M</td>
<td>Clustered Appliance</td>
<td>Clustered Servers</td>
<td>Tuned*</td>
</tr>
<tr>
<td>8,640 GB</td>
<td>8.64T</td>
<td>100M</td>
<td>Many Clustered Servers</td>
<td>Clustered Servers</td>
<td></td>
</tr>
<tr>
<td>43,200 GB</td>
<td>43.2T</td>
<td>500M</td>
<td></td>
<td></td>
<td>Many Clustered Servers</td>
</tr>
</tbody>
</table>

* Tuned means tuning the model, queries, and/or hardware (more CPU, RAM, and Flash)
Microsoft has solutions covering and connecting all four quadrants – that’s why SQL Server is one of the most utilized databases in the world.
Azure Data Lake Analytics

A new distributed analytics service

Distributed analytics service built on Apache YARN

Elastic scale per query lets users focus on business goals—not configuring hardware

Includes U-SQL—a language that unifies the benefits of SQL with the expressive power of C#

Integrates with Visual Studio to develop, debug, and tune code faster

Federated query across Azure data sources

Enterprise-grade role based access control
**Query data where it lives**

Easily query data in multiple Azure data stores without moving it to a single store

**Benefits**

- Avoid moving large amounts of data across the network between stores (federated query/logical data warehouse)
- Single view of data irrespective of physical location
- Minimize data proliferation issues caused by maintaining multiple copies
- Single query language for all data
- Each data store maintains its own sovereignty
- Design choices based on the need
- Push SQL expressions to remote SQL sources
  - Filters, Joins
  - `SELECT * FROM EXTERNAL MyDataSource EXECUTE @"Select CustName from Customers WHERE ID=1";` (not pushdown)
  - `SELECT CustName FROM EXTERNAL MyDataSource WHERE ID=1 LOCATION "dbo.Customers"` (pushdown)
Azure SQL Data Warehouse
A relational **data warehouse-as-a-service**, fully managed by Microsoft.
Industries first **elastic** cloud data warehouse with **enterprise-grade** capabilities.
Support your **smallest to your largest** data storage needs while handling queries up to **100x faster**.

<table>
<thead>
<tr>
<th>Elastic scale &amp; performance</th>
<th>Powered by the Cloud</th>
<th>Market Leading Price &amp; Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scales to petabytes of data</td>
<td>Get started in minutes</td>
<td>Simple billing compute &amp; storage</td>
</tr>
<tr>
<td>Massively Parallel Processing</td>
<td>Integrated with Azure ML, PowerBI &amp; ADF</td>
<td>Pay for what you need, when you need it with dynamic pause</td>
</tr>
<tr>
<td>Instant-on compute scales in seconds</td>
<td>Enterprise Ready</td>
<td>Bring DW to the Cloud without rewriting</td>
</tr>
<tr>
<td>Query Relational / Non-Relational</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## The data warehousing portfolio from Microsoft

### Comprehensive solutions

<table>
<thead>
<tr>
<th></th>
<th>APS with HP CS300</th>
<th>Fast Track with HP</th>
<th>SQL Server 2016</th>
<th>SQL Server in an Azure VM</th>
<th>Azure SQL Database</th>
<th>Azure SQL Data Warehouse</th>
<th>Azure Data Lake Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SMP</strong></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td><strong>MPP</strong></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SUPPORTS NON-RELATIONAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
<tr>
<td><strong>CLOUD</strong></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td><strong>PRE-ENGINEERED</strong></td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Platform Diagnostic Questions

- Can you use the cloud?
- Do the developers have Hadoop skills?
- Will you use non-relational data (variety)?
- How much data do you need to store (volume)?
- Is this an OLTP or OLAP/DW solution?
- Will you use streaming data (velocity)?
- Will you use dashboards?
- How fast do the operational reports need to run?
- Will you do predictive analytics?
- Do you want to use Microsoft tools or open source?
- What are your high availability and/or disaster recovery requirements?
- Do you need to master the data (MDM)?
- Are there any security limitations with storing data in the cloud?
- Does this solution require 24/7 client access?
- How many concurrent users will be accessing the solution at peak-time and on average?
- What is the skill level of the end users?
- What is your budget and timeline?
- Is the source data cloud-born and/or on-prem born?
- How much daily data needs to be imported into the solution?
- What are your current pain points or obstacles (performance, scale, storage, concurrency, query times, etc)?
- Are you ok with using products that are in preview?
In Summary – Learn Key Determining Factors

• Volume of Data – streaming, batch loads,
• Speed of Access – real time, batch, etc.
• Size of Data Store – gigabytes, terabytes, petabytes
• Types of Data – structured, semi-structured, unstructured
• Query Processing Time –
• Consumers of Data – users, business analysts, data scientists
• Visualization of Data – Excel, structured reports, dashboards, metrics
• Record locking and Concurrency – ACID, Optimistic Concurrency, Eventual
Tool of the Week & Tip of Week

• **Tools**
    Handy reference website with clickable graphics that share a summary and detail links for services in Azure.
    This tool helps you to analyze and existing database for migration to Azure SQL

  - Self-paced Labs
  - Instructor-led Virtual Classes (some from Pluralsight)
  - Cloud Monitor Tool
  - A small amount of individual Azure Credit (for individual playing around/exploration)
  - Functionality Overview Videos
  - Access to Trials for various cloud services
  - Career Guidance
  - Community Forums
Questions Heard In The Field

• Can I block Azure Marketplace purchases? Yes
  • EA Enrollment Administrator can turn off purchasing privileges (marketplace faq)


• Can an EMC Data Domain be integrated with Azure? Yes
  • EMC Data Domain Virtual Edition can be run from Azure
  • EMC Data Domain can leverage Azure for data tiering

• Can Azure Backup protect Hyper-V 2008 servers? Yes

• I am using SharePoint Online and OneDrive so I should setup Express Route for Office 365 to improve performance? No
  • We actually don’t recommend Express Route for Office 365 except for specific situations and performance of in this case performance of SPO and OneDrive is not guaranteed to be improved