# Examining Azure Data Storage

**Gary Grudzinskas**
CLOUD ENGINEER AND AUTHOR

@garygrudzinskas

# Objectives

Know when, where and why to use various Azure data services

Azure Blog Storage

Azure Data Lake Gen2

Azure Cosmos DB

Azure SQL DB

Azure SQL Data Warehouse

Supporting Azure Data Services

# Azure Blob Storage

Designed for images and unstructured data

Cheapest way to store data in Azure

Simple design and easy to use

HDFS and blob storage REST APIs

Scale to meet data access needs

# Blob Storage Tiers

**Premium**

Performance-sensitive data

**Hot**

Frequently accessed data

**Cool**

Infrequently accessed data

**Archive**

Rarely accessed data

# Azure Blob Storage Benefits

Share and reuse data through APIs

Geo-redundancy

Object mutability

Multiple blob types: block, page and append

Low cost

# Azure Blob Storage

**Use When....**

- Only basic storage is needed
- Data is unstructured
- Data that is older or not used as much
- Information needs to be kept but not analyzed or queried any time soon
- Money is an issue

# Azure Data Lake Storage Gen2

Built on Azure Blob storage

No need to copy or transform data in order to analyze it

Hierarchical namespace with directories and subdirectories (HDFS)

File level permissions

Use to store massive amounts of data for big-data analytics

# Azure Data Lake Storage Gen2

**Use When…**

- There is a lot of data to be stored
- Data needs analysis
- Hierarchical namespaces are needed
- File level security is desired
- Most of the data is unstructured
- Need to store a wide variety of data

# Comparing SQL and NoSQL

| SQL | NoSQL |
|---|---|
| Relational | Non-Relational |
| Vertical Scaling | Horizontal Scaling |
| Row Oriented | Multi Model Oriented |
| Static Schema | Dynamic Schema |
| Tables | Collections |
| Limited for big data | Great for big data |

# Not Only SQL

**Key-Value**

Uses a simple key/value to store data

**Document**

Contains semi-structured documents

**Columnar**

Orients data according to columns

**Graph**

Interconnects data with graphs to represent data

**SQL**

Has the same structure of an SQL database

# Azure Cosmos DB

Cloud based NoSQL

Globally distributed, multi-model database service

Built for very large databases

No schema required or schema on read

Highly responsive and highly available

# Azure Cosmos DB APIs

SQL API for SQL databases

MongoDB API for document

Gremlin API for graph

Cassandra API for columns

Table API for table storage

# Azure Cosmos DB

**Use When…**

- Database that takes unstructured and non relational data
- Have various DB models to use
- Desire a fast response time
- Have a globally disperse operation
- Don’t want to deal with schema or index management

# Azure SQL Database

Cloud-based managed relational database service

Use to scale up and scale down OLTP systems on demand

Frictionless database migration

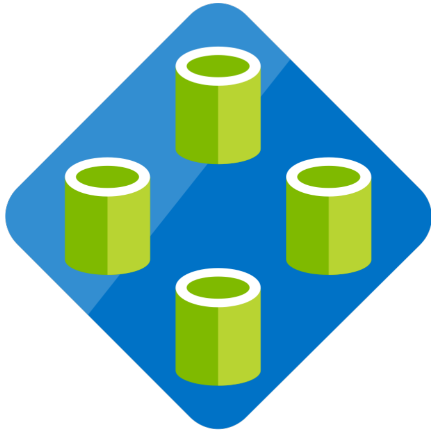
Ingest data with T-SQL, and a wide range of developer SDKs

Built-in machine learning

# Azure SQL DB Deployment Models

**Single Database**

Fully-managed
isolated database



**Elastic Pool**

Collection of single DB
with shared resources

**Managed Instance**

Contains a set of
databases

# Azure SQL DB Tiers

## General Purpose

Designed for common workloads. This is the default tier

## Business Critical/Premium

Designed for OLTP applications with high transaction rate

## HyperScale

Designed for very large OLTP database with auto-scale of storage compute

# Azure SQL Database

**Use When…**

- Scale data base in cloud vs. on-premises
- Machine learning is required
- Need an OLTP system on demand
- Have a large amounts of users

# Azure SQL Data Warehouse

Cloud-based enterprise data warehouse

Petabyte scale that is relatively easy to set up and configure

SQL Data Warehouse uses massively parallel processing (MPP)

The storage nodes are separate from the compute nodes

Coordinates and transports data between compute nodes as necessary

# Azure SQL Data Warehouse

**Use when...**

- Want to use massively parallel processing (MPP) for big data analytics
- Need to prepare loads of data that is all over the place
- Desire to release business intelligence reports in a timely fasion
- Ability to pause and resume compute
- Answer complex business questions
- Have large amounts of data and a small amount of users

# Azure Streaming Analytics

Examines data streaming in from applications, sensors, monitoring devices, and gateways

Use to respond to data events in real time or analyze large batches of data

Offers sub-second latencies

# HDInsight

Provides tools to ingest, process, and analyze big data

It includes Apache Hadoop, Spark, Kafka, HBase, Storm, and Interactive Query

Uses Hive to run ETL operations on the ingested data

# Azure Databricks

Apache Spark-based analytics platform optimized for the Microsoft Azure

Integrated with Azure to provide one-click setup, streamlined workflows

Provides an interactive workspace that enables collaboration between data scientists, data engineers, and business analysts

Globally scalable

Integrate effortlessly with a wide variety of data stores and services

# Azure Data Catalog

Enterprise-wide metadata catalog

Register, enrich, discover, understand, and consume data sources

Quickly find the data and then use it in whatever tool is required

Data stays in one place

Less time looking for data and more time using it

# Summary

**There is a lot of data being produced and it is growing exponentially**

**Data takes various forms, but it is mostly unstructured**

**Data engineers wrangle data**

**Determining factors in what Azure storage service is appropriate…**

- The value of the data
- How often the data is accessed
- What will be done to the data
- How the data is structured
- How new the data is