Microsoft

DP-201T00: Designing an Azure Data Solution

THE DP-200 EXAM

Implementation

THE DP-201 EXAM

Design

THE DP-201 EXAM

Designing data storage solutions

Designing data processing solutions

Designing for data security and compliance

Designing data storage solutions

You need to know which Azure services to recommend to meet business requirements

Relational data stores

Non-relational data stores

Designing data storage solutions

Relational data stores

Azure SQL Database
Azure Synapse Analytics

Non-relational data stores

Cosmos DB
Data Lake Storage Gen2
Blob Storage

Designing data storage solutions

Relational data stores

Azure SQL Database

Azure Synapse Analytics

Non-relational data stores

Cosmos DB

Data Lake Storage Gen2

Blob Storage

For all of the above services, you need to know how to design:

- Data distribution and partitions
- High scalability
- Disaster recovery
- High availability

Designing data storage solutions

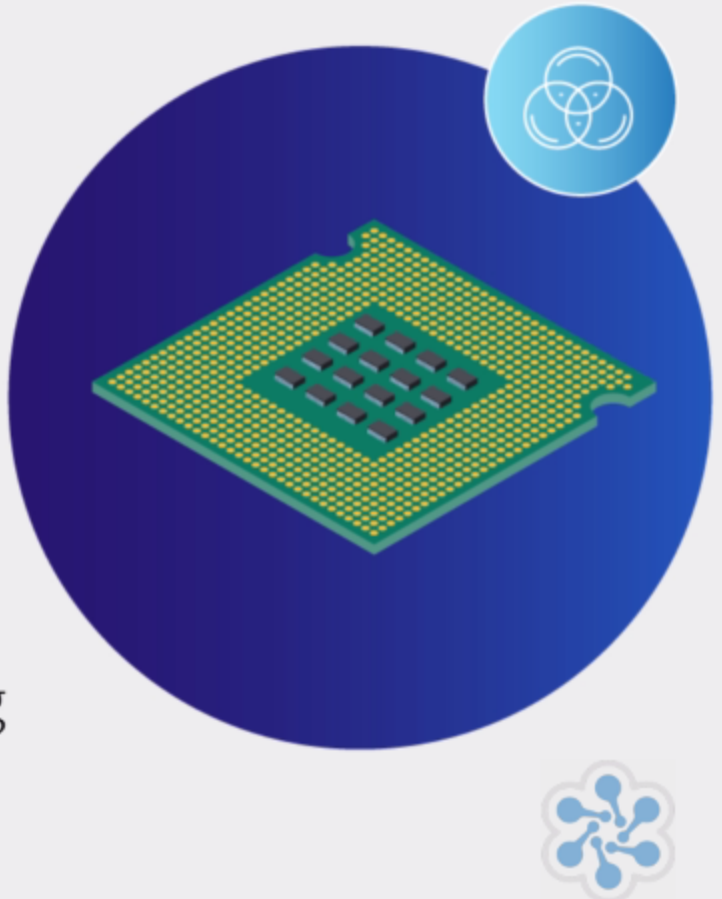
It's divided into **batch processing** and **stream processing**

Batch processing

You need to know how to design solutions using Azure Data Factory and Azure Databricks

Stream processing

You need to know how to design solutions using Stream Analytics and Azure Databricks



Designing data storage solutions

Azure Databricks is a very important service for data processing since it's used for both batch and stream processing

You also need to know how to ingest data from other Azure services and how to output the results to other services

Designing data storage solutions

You need to know how to secure your data stores

The most important decision is what authentication method to use for various use cases

Designing data storage solutions

The second part of this section deals with designing security for data policies and standards

Some of the topics include:

- Encryption, such as Transparent Data Encryption
- Data auditing
- Data masking
- Data privacy and data classification
- Data retention
- Archiving
- Purging

Azure Storage

Azure Storage

Durable and highly available

Secure

Scalable

Managed service

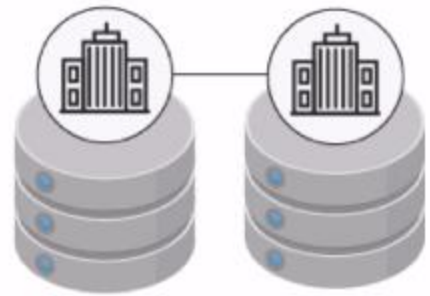
Redundancy Options

Locally-redundant storage (LRS)

Replicated across racks in the same data center

Zone-redundant storage (ZRS)

Replicated across three zones within one region



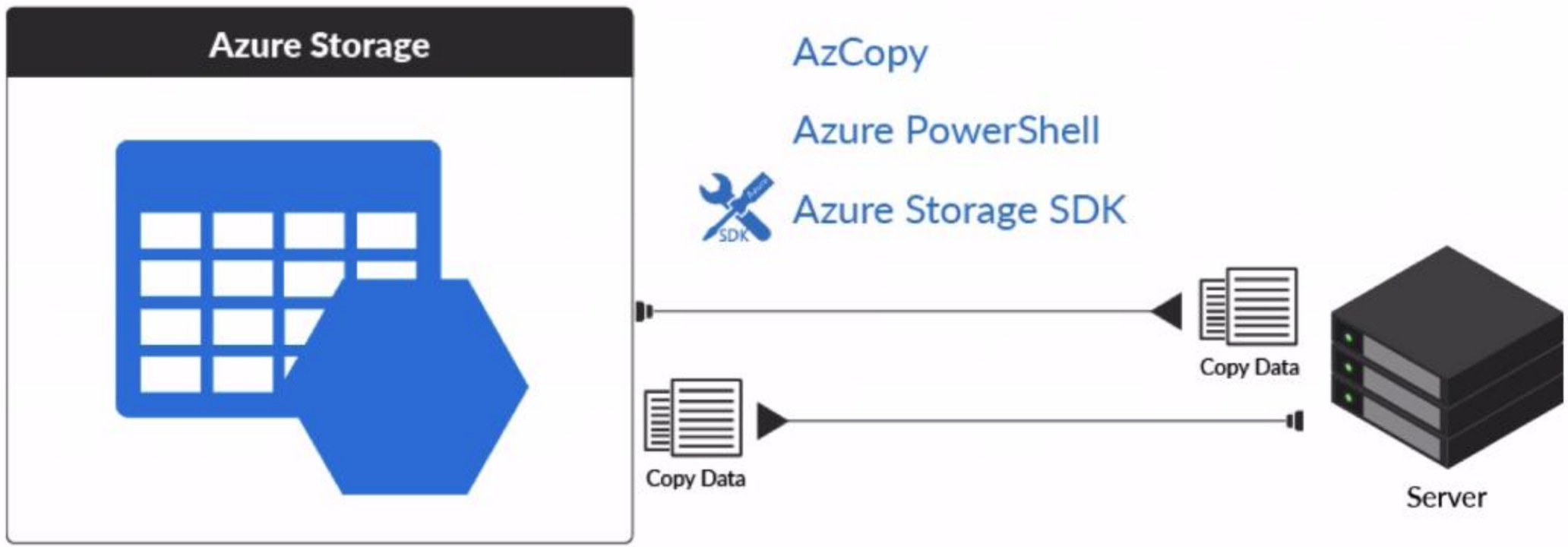
Geo-redundant storage (GRS)

Replicated across two regions

Read-access geo-redundant storage (RA-GRS)

Active read replica in secondary region

Tools for Copying Data



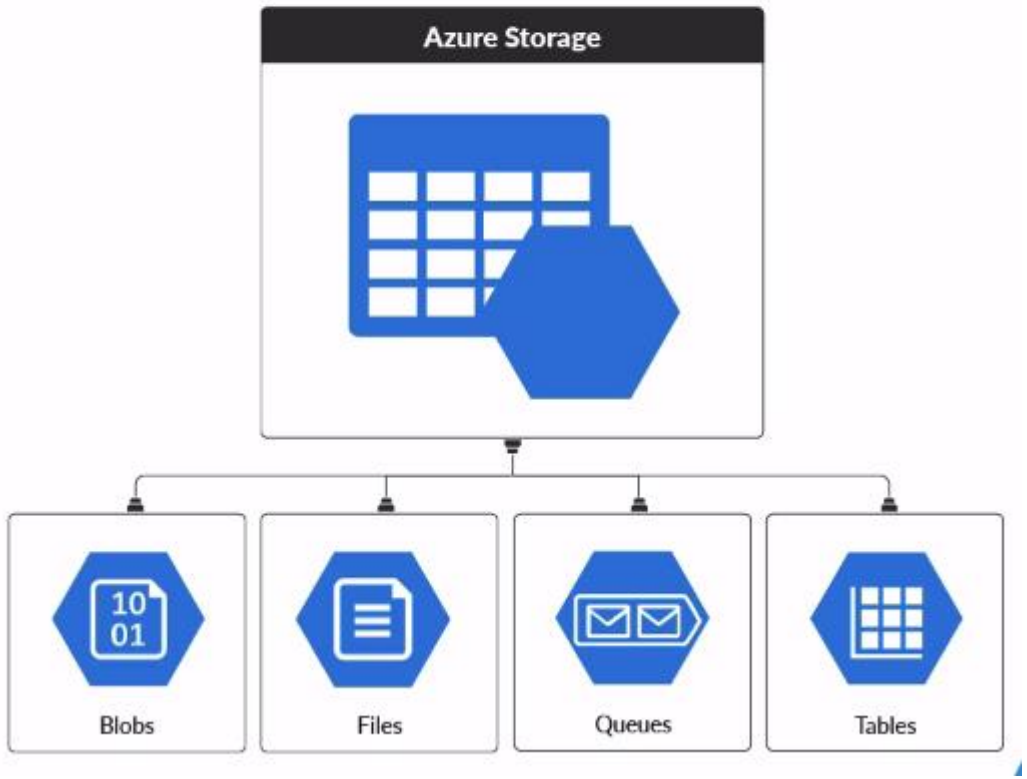
Data Types

Blobs

- Binary large object
- No organizational structure

Files

- Filesystem structure
- SMB-compliant
- Move on-premises file share to Azure
- Accessible over the web - need a shared access signature token
- Significantly more expensive than Blob storage



Blob Storage Tiers

Hot

- For data that gets accessed frequently

Cool

- For data that doesn't get accessed frequently
- Data gets retrieved immediately
- Lower storage cost, higher cost for reads and writes
- 30-day minimum

Blob Storage Tiers

Hot

- For data that gets accessed frequently

Cool

- For data that doesn't get accessed frequently
- Data gets retrieved immediately
- Lower storage cost, higher cost for reads and writes
- 30-day minimum

Archive

- Takes up to 15 hours to access when requested
- 5 times cheaper than cool tier, but far more expensive for reads
- 180-day minimum

Moving data from cool or archive tiers before minimum duration incurs an early deletion fee

Data Types

Queues

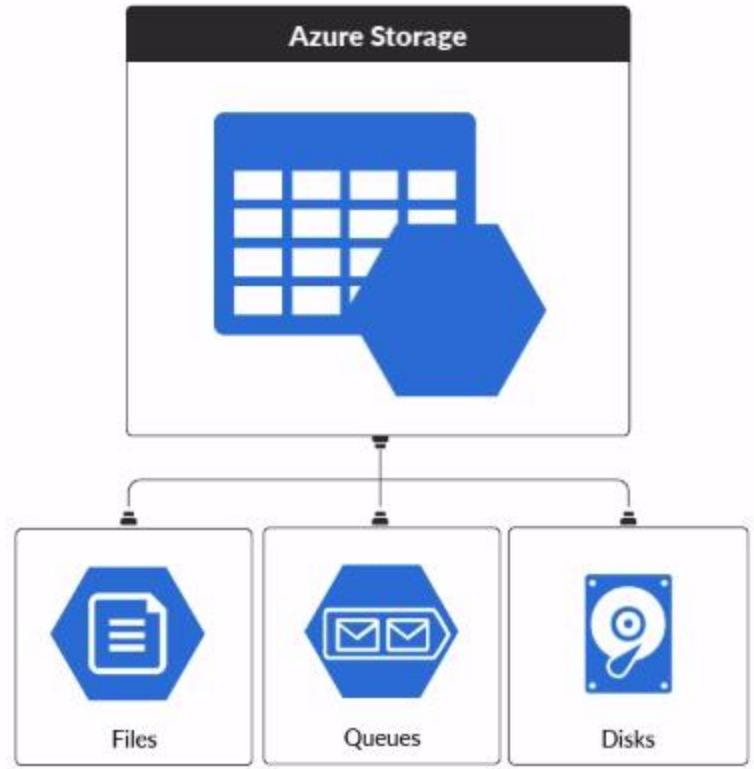
- For passing messages
- One application pushes messages onto queue, another retrieves them

Tables

- NoSQL datastore
- Storage costs about the same as File storage, but much lower transaction costs
- Premium version is part of CosmosDB service

Disks

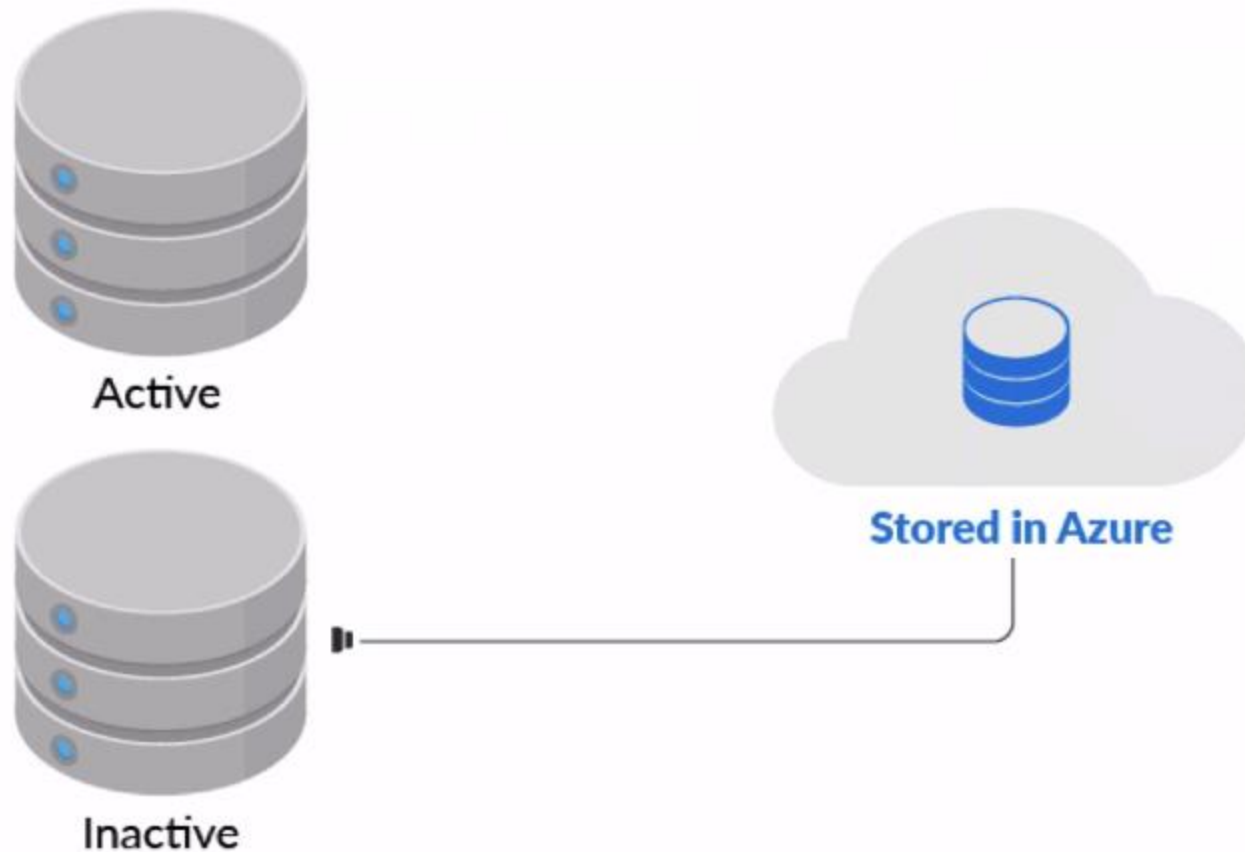
- Attached to virtual machines



StorSimple

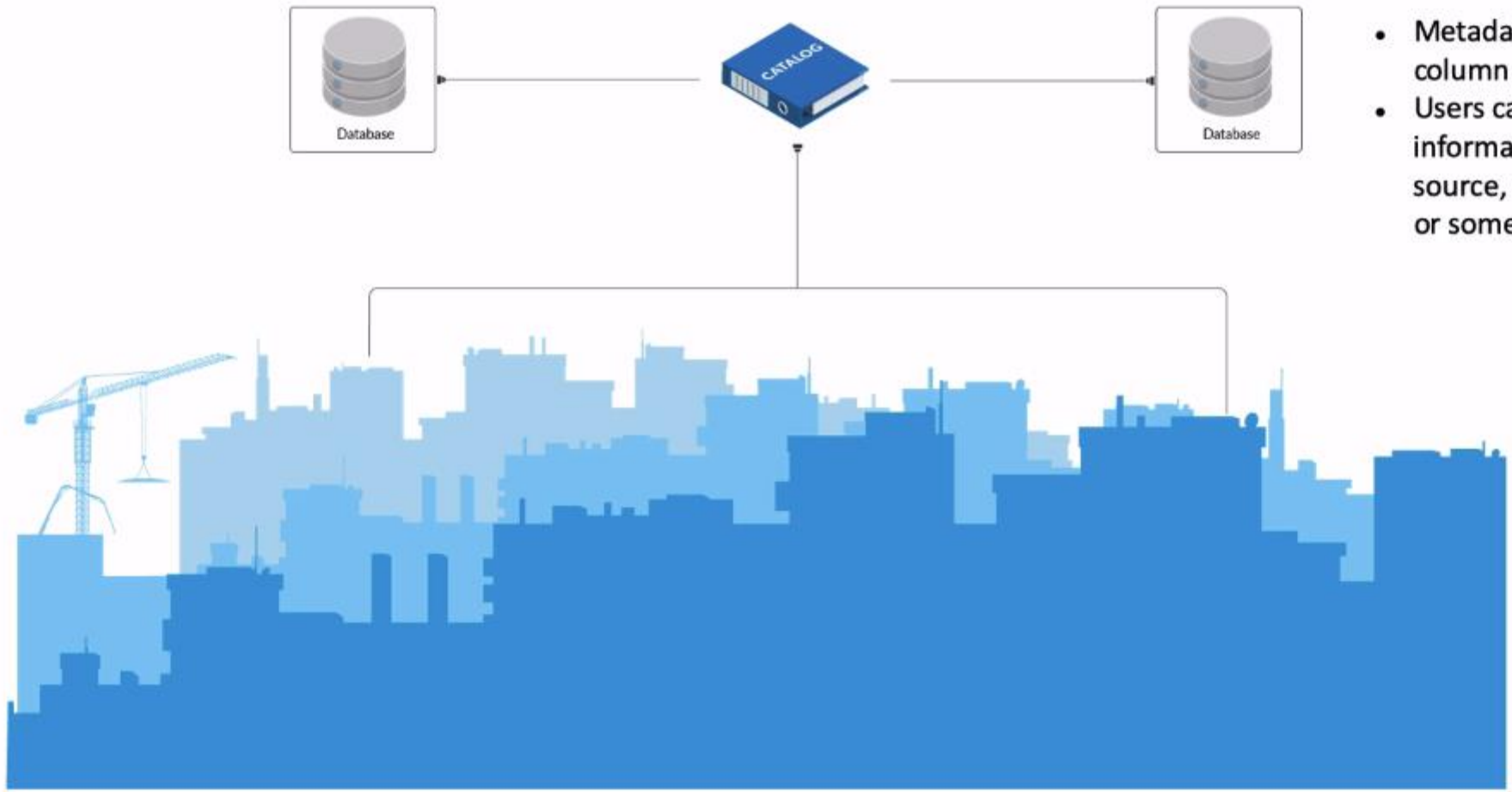
- Virtual array installed on-premises
- Backup
- Recovery
- Storage tiering

StorSimple



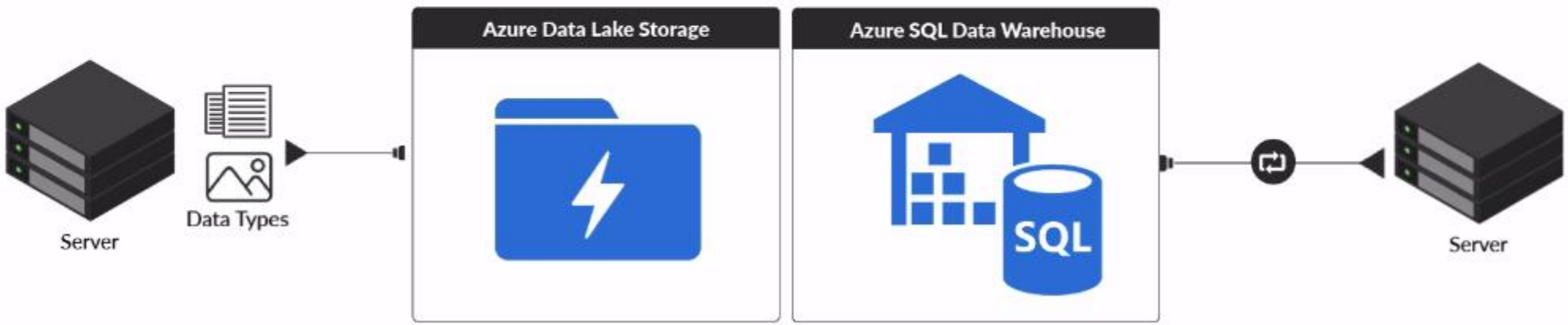
Azure Data Services

Azure Data Catalog



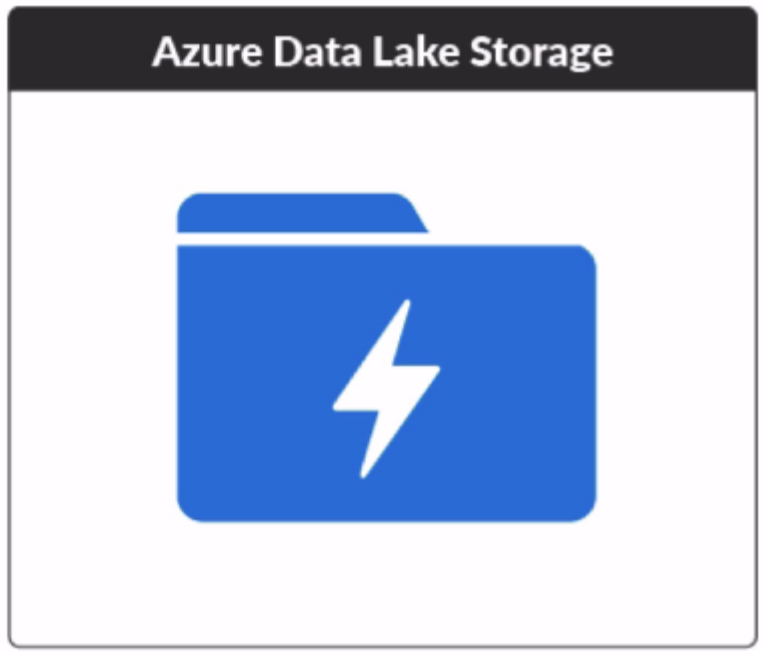
- Metadata includes things like column names and data types
- Users can also add additional information about a data source, such as a description or some tags

Azure Data Lake and Data Warehouse



If you have raw data that's not in a nicely structured format, then you'll probably need to process it before you store it.

Azure Data Lake and Data Warehouse



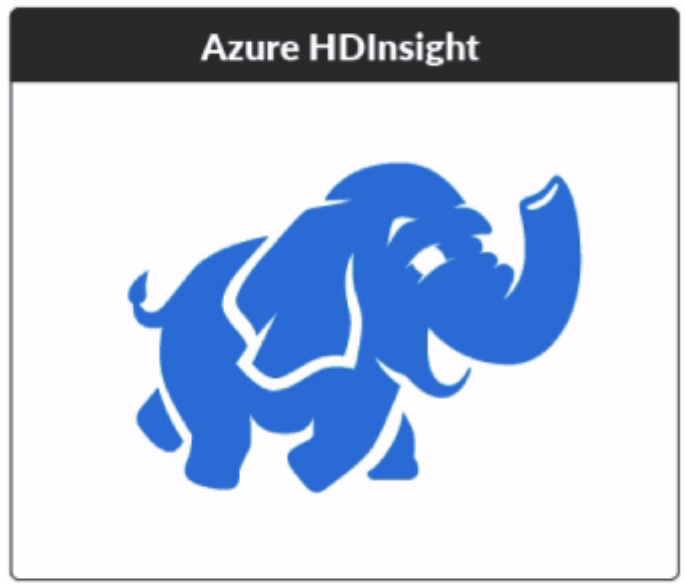
Built to work with Hadoop
No regulatory compliance
Write queries using U-SQL

Built on SQL Server
Certified for compliance
Write queries using T-SQL

Azure HDInsight

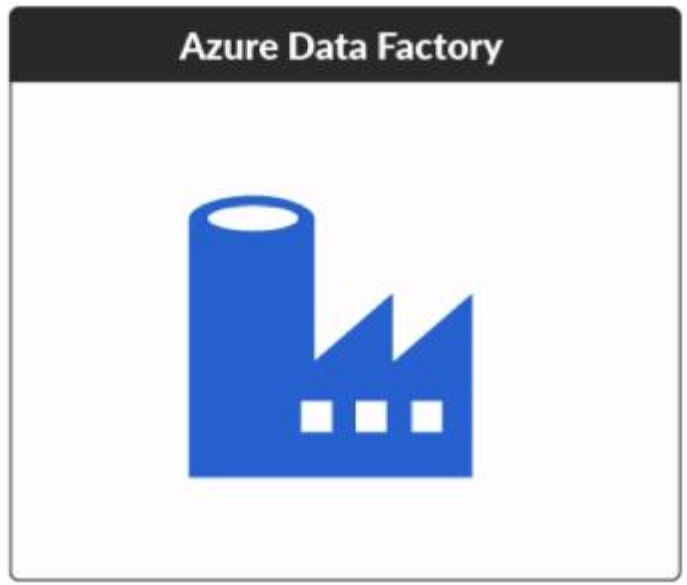
Supports a wide variety of open-source big data frameworks, including:

- Hadoop
- Spark
- Hive
- Storm
- Many others



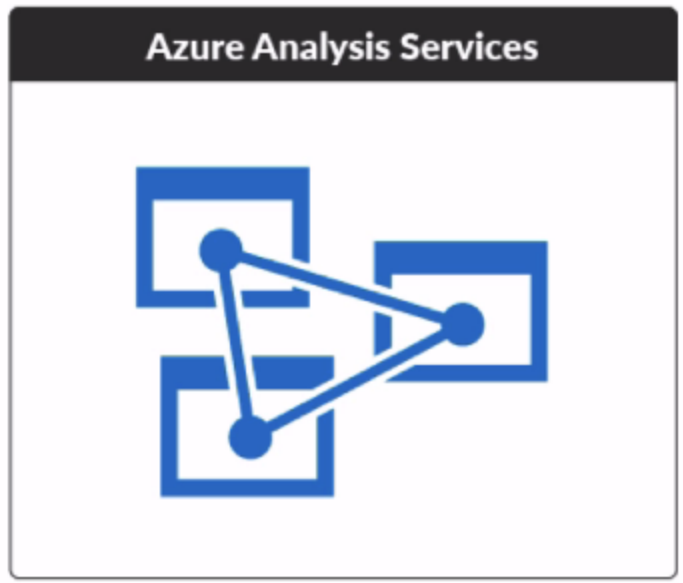
Azure Data Factory

- Automates data movement and data transformation
- Spins up and down HDInsight clusters as needed
- Creates data processing pipelines
- Automates Data Lake Analytics queries and machine learning



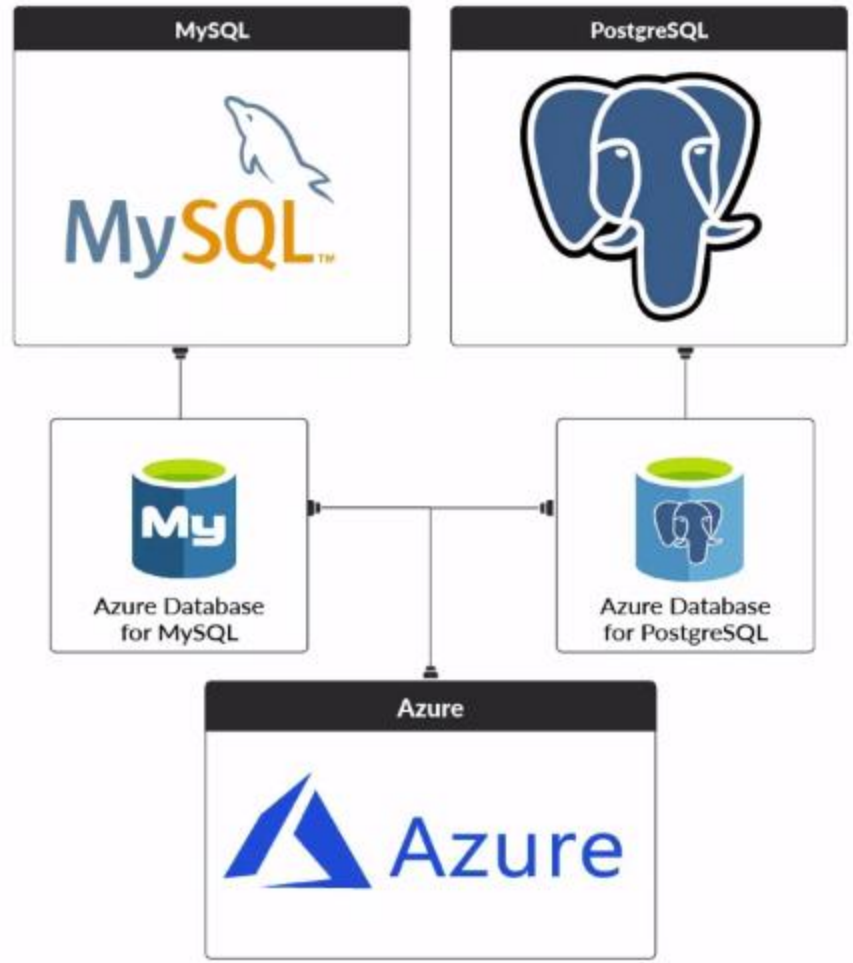
Azure Analysis Services

- Lets you create data models of existing data
- Uses in-memory caching
- Accessed through supported client tools such as Power BI, Tableau or Excel



Relational Database Storage

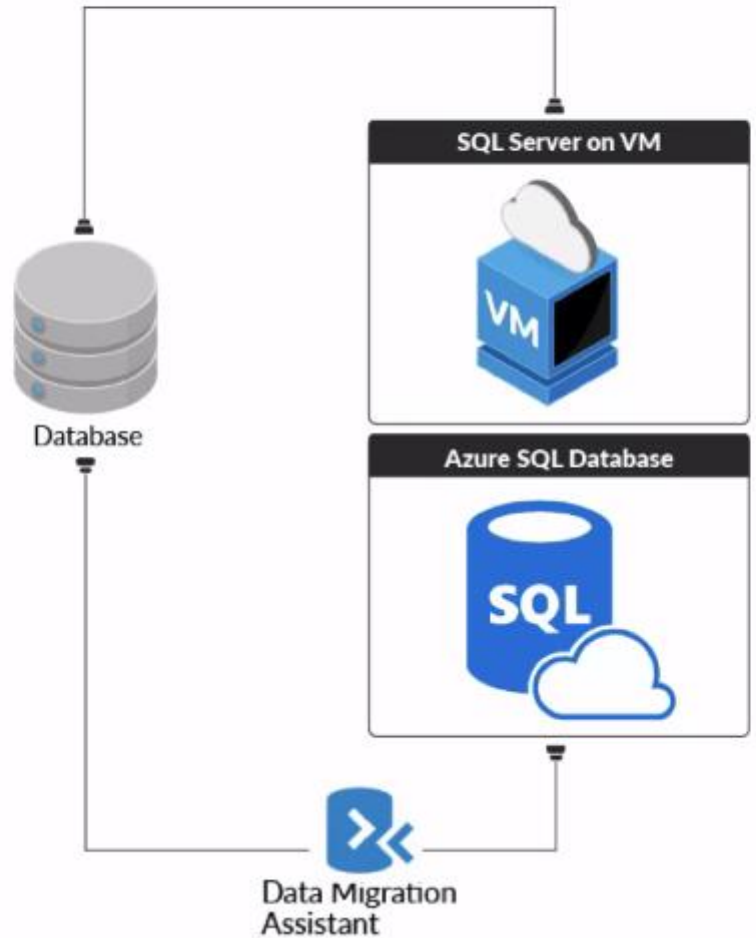
Relational Databases



SQL Server Stretch Database

- Migrates cold table rows to Azure
- You can still query migrated data
- Avoid buying more storage
- Shorten backup times
- More expensive than storing data offline
- Data Migration Assistant
  - Tells you which tables would be good candidates for Stretch Database
  - Indicates potential blocking issues

Moving the Entire Database to Azure



SQL Database Managed Instance

Nearly **100%** compatible with SQL Server

Azure SQL Database Service Tiers

General Purpose

Hyperscale

Business Critical

Azure SQL Database Service Tiers

General Purpose

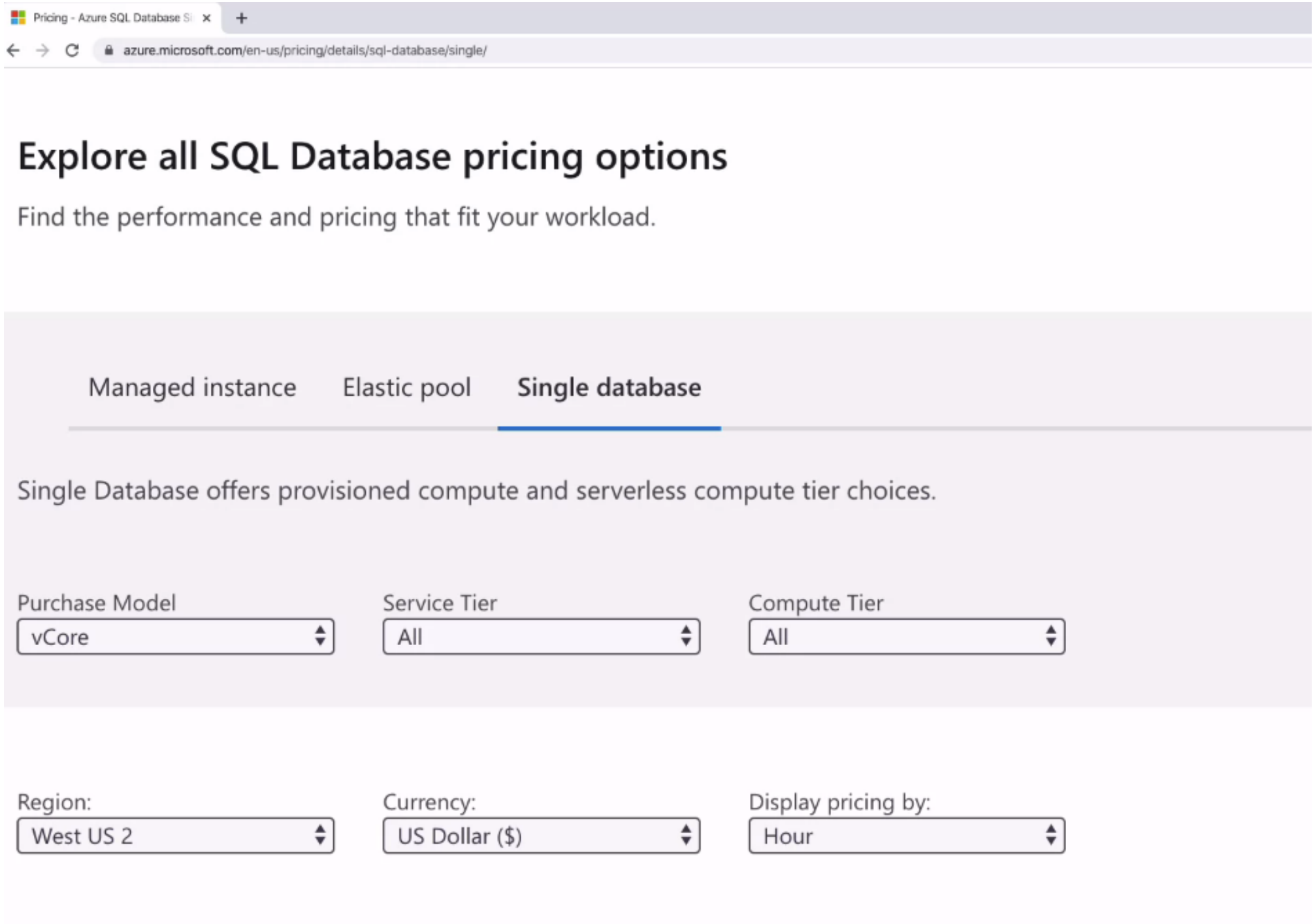
- Least expensive
- Latency: 5 - 10 milliseconds
- Availability: 99.99%
- Max size: 4TB (8TB for Managed Instance)

Hyperscale

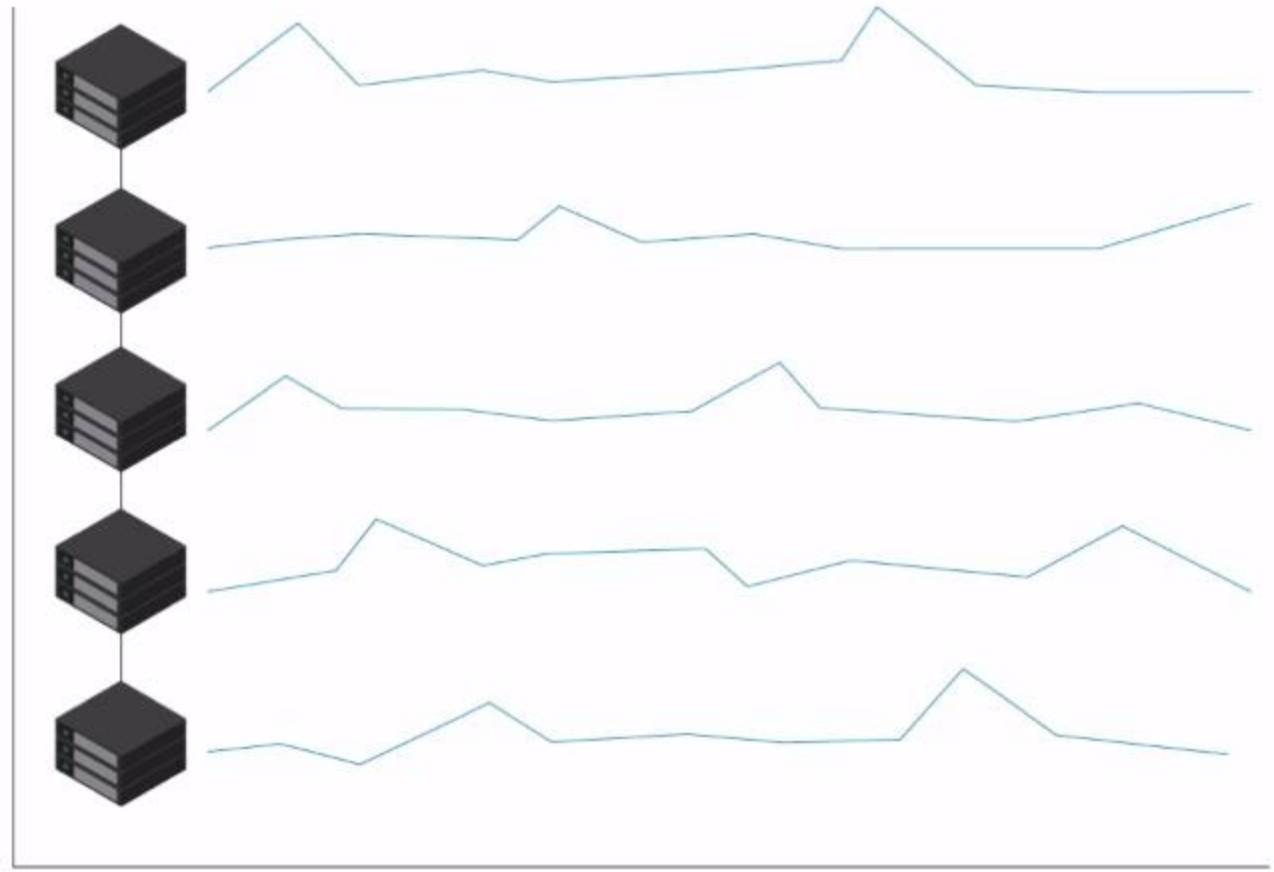
- Max size: 100TB
- Scales compute resources up and down very quickly
- Instant backups and fast database restores

Business Critical

- Latency: 1 - 2 milliseconds
- Local SSDs on 4-node cluster
- Most expensive
- Availability: 99.995% (with zone-redundant option)
- Max size: 4TB


Pricing - Azure SQL Database Si
azure.microsoft.com/en-us/pricing/details/sql-database/single/
Explore all SQL Database pricing options
Find the performance and pricing that fit your workload.
Managed instance
Elastic pool
Single database
Single Database offers provisioned compute and serverless compute tier choices.
Purchase Model
vCore
Service Tier
All
Compute Tier
All
Region:
West US 2
Currency:
US Dollar ($)
Display pricing by:
Hour

Elastic Pool Model



SQL Database Managed Instance

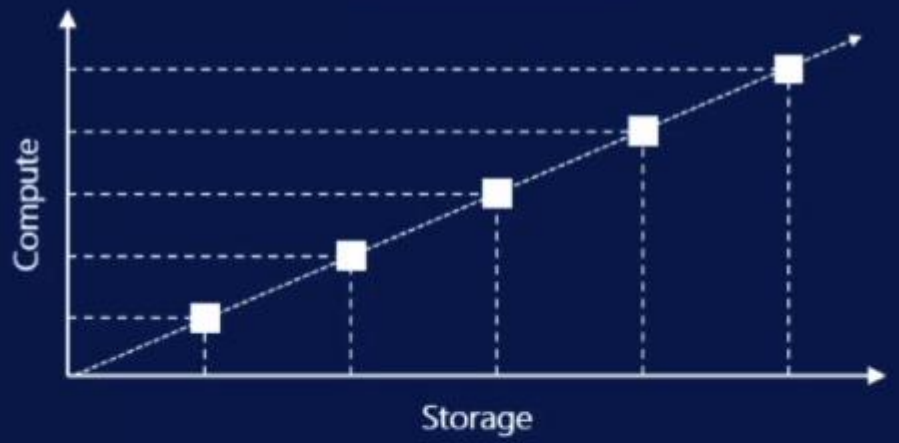
Nearly 100% compatible with SQL Server

Doesn't work with the Hyperscale tier

(Elastic pools don't work with the Hyperscale tier either)

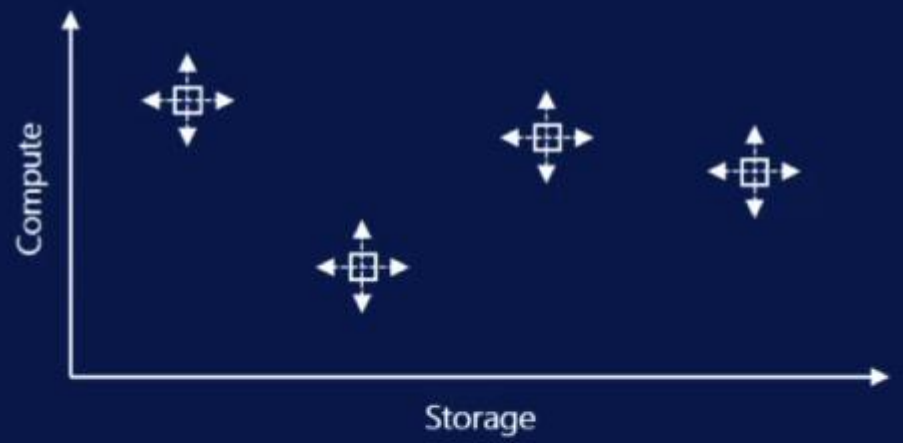
Purchasing Models

DTU model
Simple,
Preconfigured



OR

vCore model
Independent scalability



Database Transaction Unit (DTU)-based model

- Bundled measure of compute, storage and IO resources
- Best for customers who want simple, pre-configured resource options.

vCore-based model

- Independent scaling of compute, storage and IO resources
- Best for customers who value flexibility, control and transparency
- Use with Azure Hybrid Benefit for SQL Server to gain cost savings

vCore Compute Tiers

Provisioned

- Provisioned with exact resources requested
- You get charged for the database as long it's running
- To scale manually, change the number of vCores—there will be a brief loss of connectivity (<4 sec.)

Serverless

- Minimum and maximum vCores
- Autoscales based on workload demand
- If no activity, it pauses database and halts compute charges
- Cost per vCore is higher

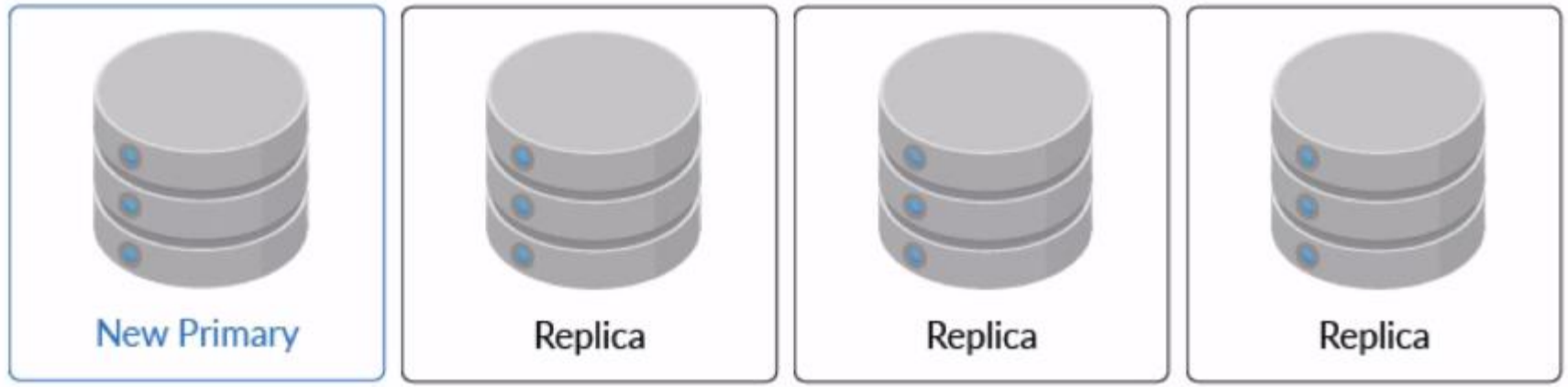
High Availability within a Region

- SQL Database uses Always ON Availability Groups technology from SQL Server to provide HA
- Failover is automatic
- It may take up to **30 seconds** to recover

Active Geo-Replication

Active Geo-Replication

Active Geo-Replication



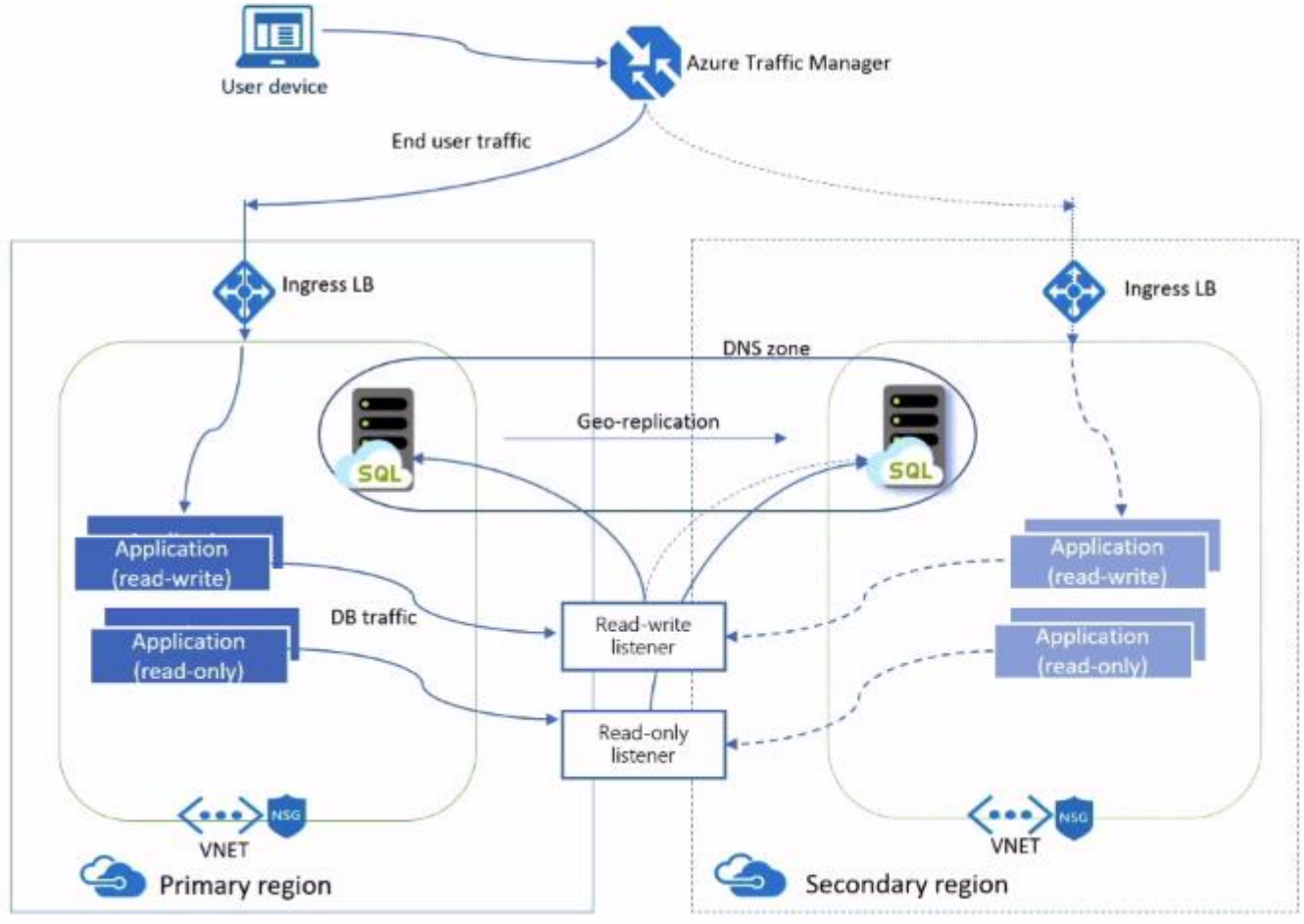
- Ensure secondaries have the same user authentication configuration as the primary
- You should use the same firewall rules for secondaries as you do for the primary
- Active geo-replication allows you to use the secondary databases to make queries faster for users in other regions

Filter by title
SQL Database Documentation
Overview
Quickstarts
SQL databases
SQL managed instances
Concepts
Common features documentation
Feature comparison
How-to guide
Security
Connect and query
Backup, restore, high availability (BCDR)
Business continuity
High availability
Backups
Failover groups and geo-replication
Active geo-replication
Auto-failover groups
Configure security for replicas
Outage recovery guidance
Recovery drills
Configure failover group
Monitor and tune
Scalability
Database features
How to
SQL databases
Download PDF

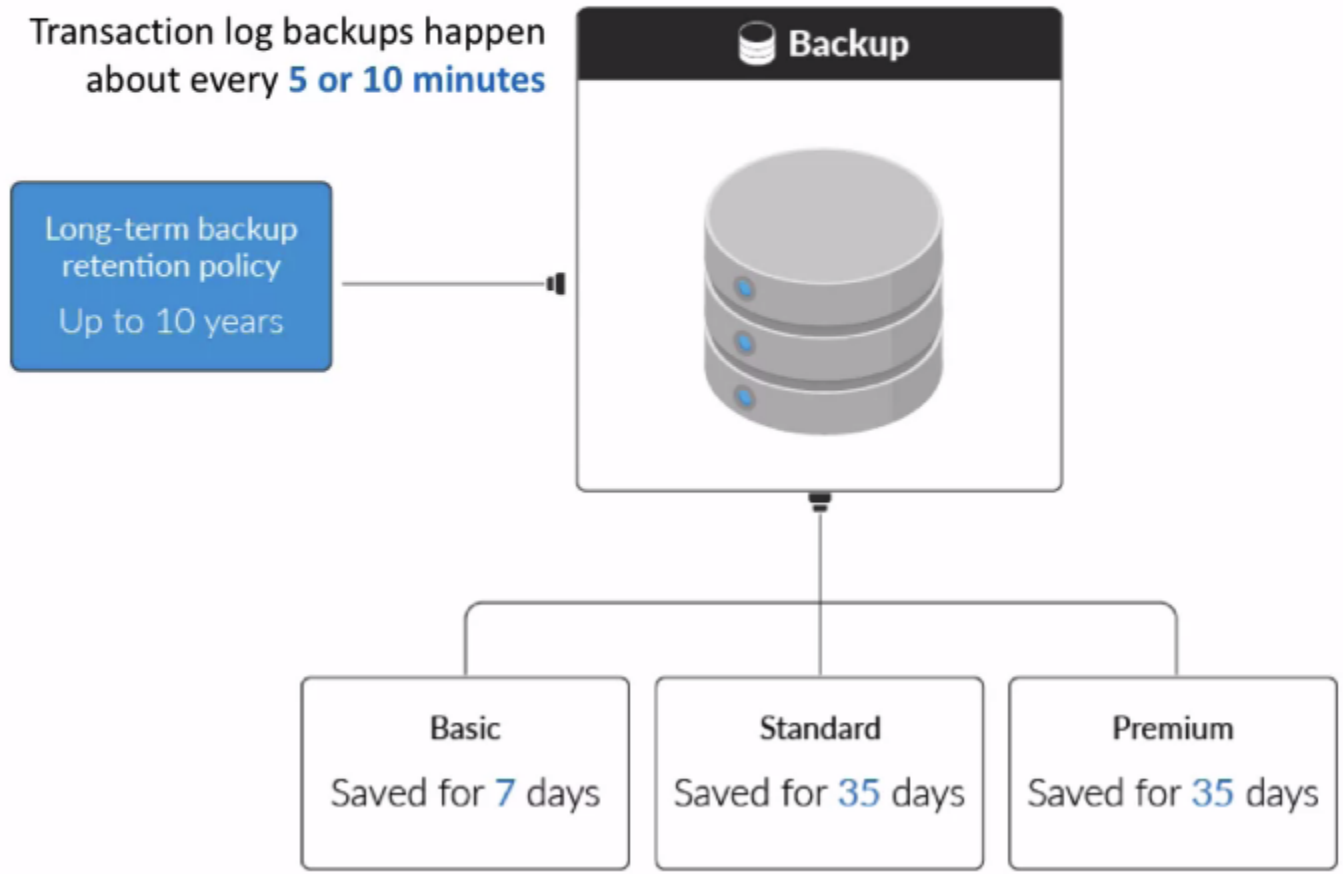
Best practices of using failover groups with managed instances

The auto-failover group must be configured on the primary instance and will connect it to the secondary instance in a different Azure region. All databases in the instance will be replicated to the secondary instance.

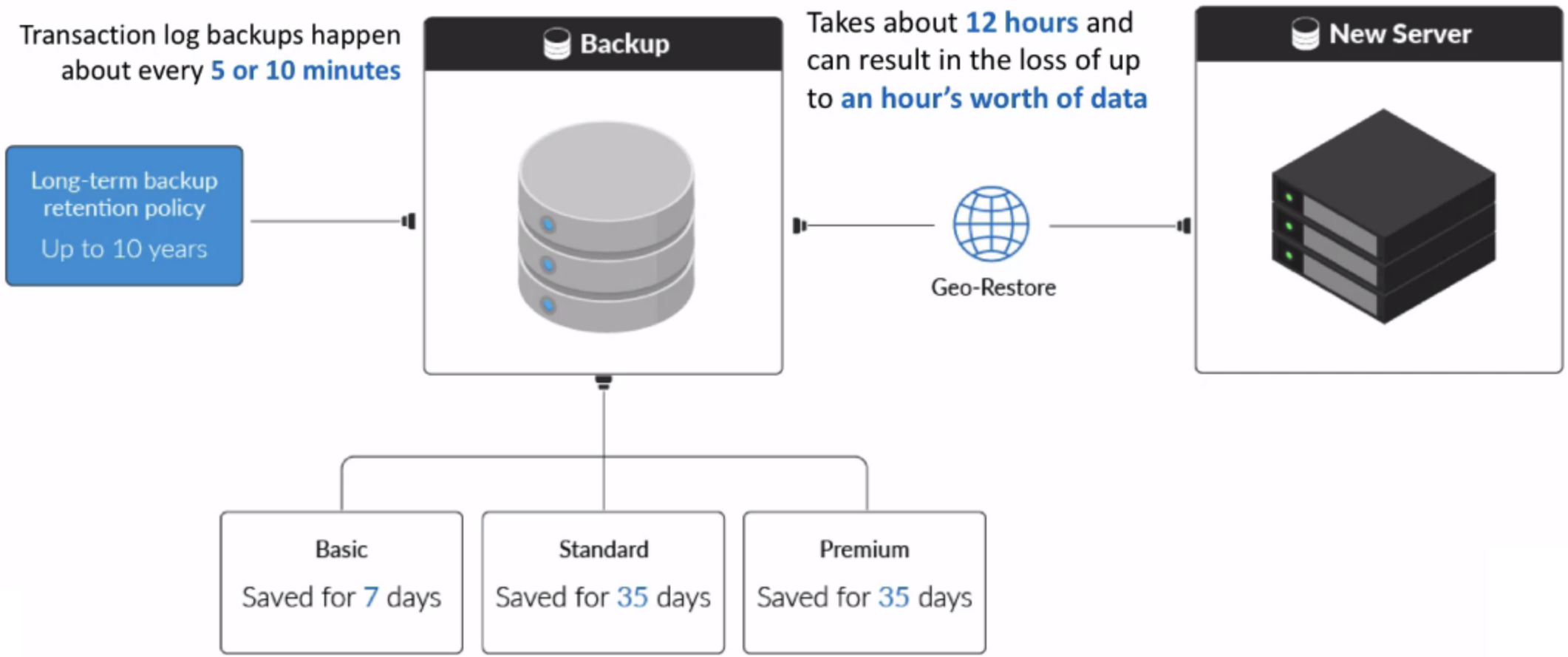
The following diagram illustrates a typical configuration of a geo-redundant cloud application using managed instance and auto-failover group.



Backups


Transaction log backups happen about every 5 or 10 minutes
Backup
Long-term backup retention policy
Up to 10 years
Basic
Saved for 7 days
Standard
Saved for 35 days
Premium
Saved for 35 days

Backups


Transaction log backups happen about every 5 or 10 minutes
Long-term backup retention policy
Up to 10 years
Backup
Takes about 12 hours and can result in the loss of up to an hour's worth of data
Geo-Restore
New Server
Basic
Saved for 7 days
Standard
Saved for 35 days
Premium
Saved for 35 days

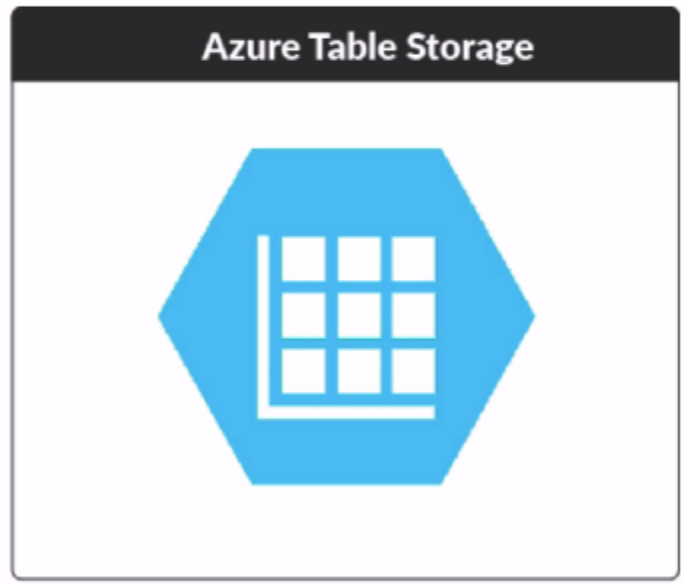
NoSQL Storage

NoSQL Datastores

- Scale better
- Satisfy fewer requirements than relational databases

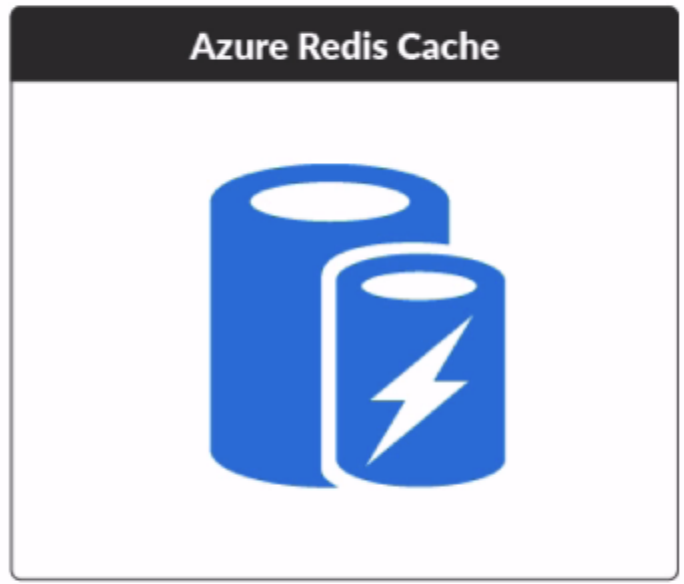
Azure Table Storage

- Intended for simple structured data
- E.g. address books and user profiles
- Schemaless design
- Indexes records
- For secondary indexes or global distribution, use Cosmos DB version instead
- For complex joins, foreign keys, or stored procedures, use relational database instead



Azure Redis Cache

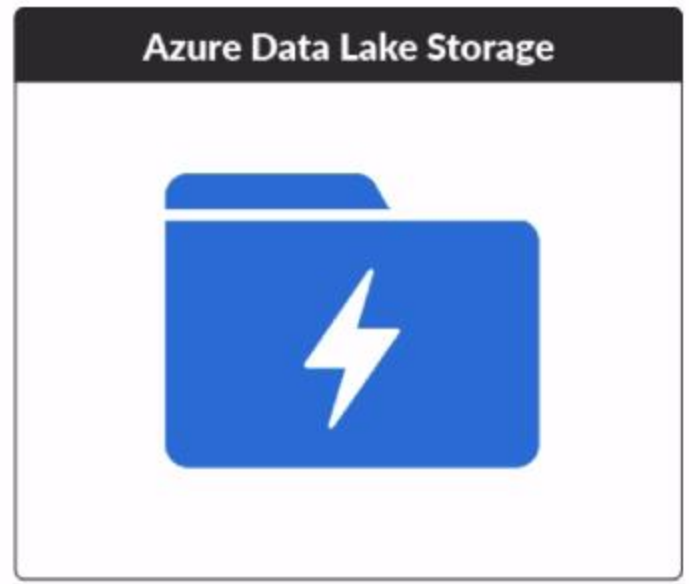
- Intended to speed up data retrieval in applications
- Managed service for Redis
- Data resides in memory
- It stores key/value pairs



Basic: Should only be used for testing and development	**Standard:** Provides a replicated, high availability cache	**Premium:** Better performance and can handle bigger workloads, disaster recovery, and more

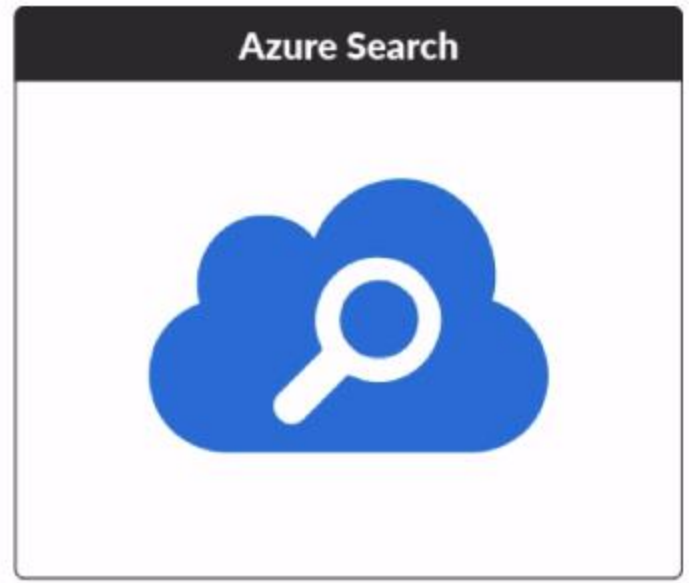
Azure Data Lake Storage

- Intended to hold large quantities of any kind of data
- Data warehouse for unstructured data
- Main purpose is data analytics



Azure Search

- Creates an index of text data
- You can embed search functionality into web, mobile, and enterprise applications
- Offers features such as:
  - Search suggestions
  - Language analyzers
  - Fuzzy searches



Time Series Insights (TSI)

- Collects time-stamped data
- Integrates with Azure IoT Hub and Azure Events Hubs
- Run queries on billions of events and get a response in seconds
- See visualizations of the data with TSI Explorer

Cosmos DB

Introduction to Cosmos DB

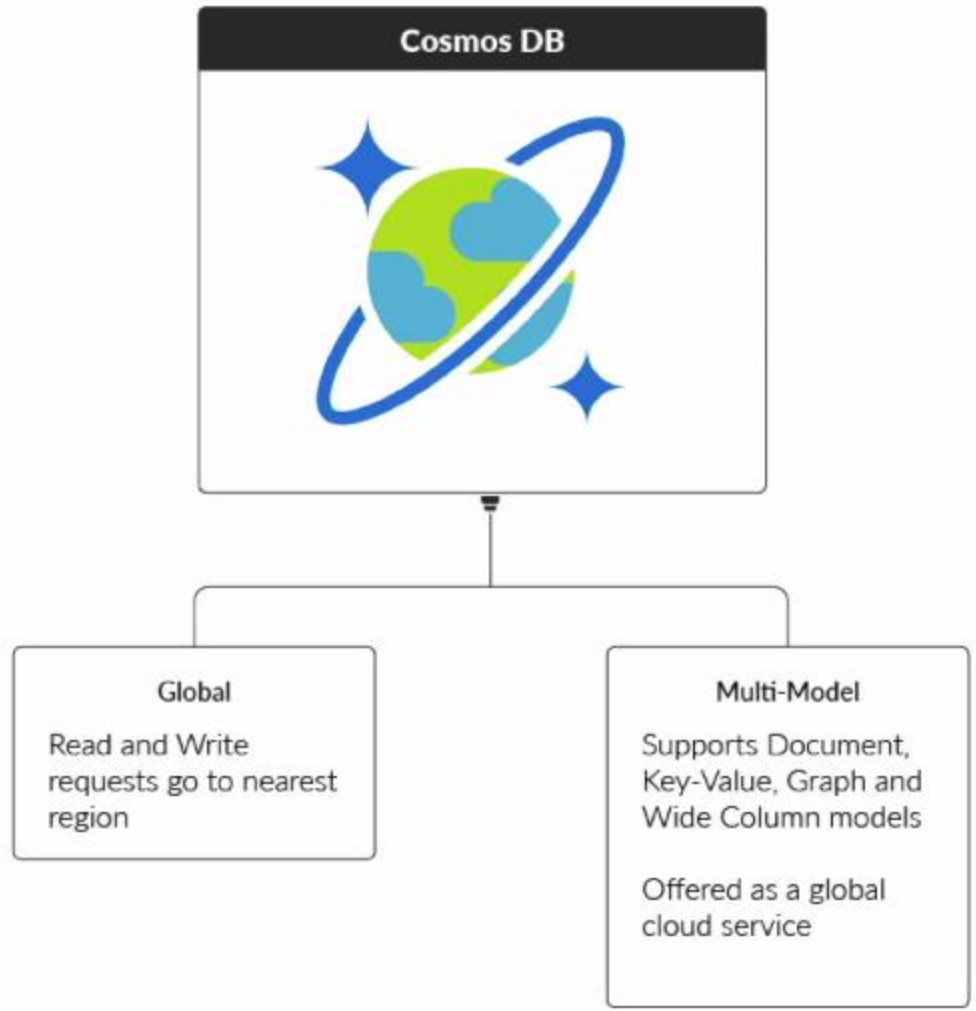
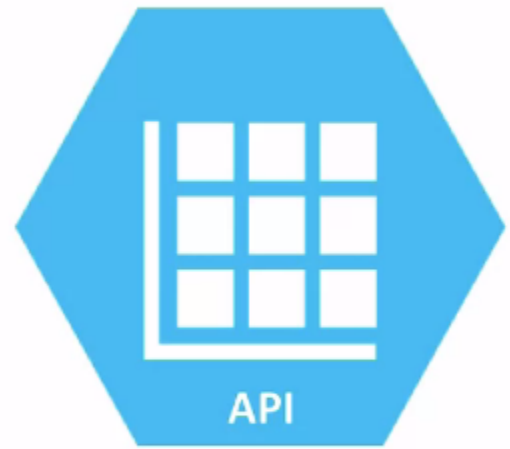


Table API

Cosmos DB's Table API is built on Table storage, but offers these additional features:

- Global distribution
- Dedicated throughput worldwide
- Single-digit millisecond latencies at the 99th percentile
- Guaranteed high availability
- Automatic secondary indexing



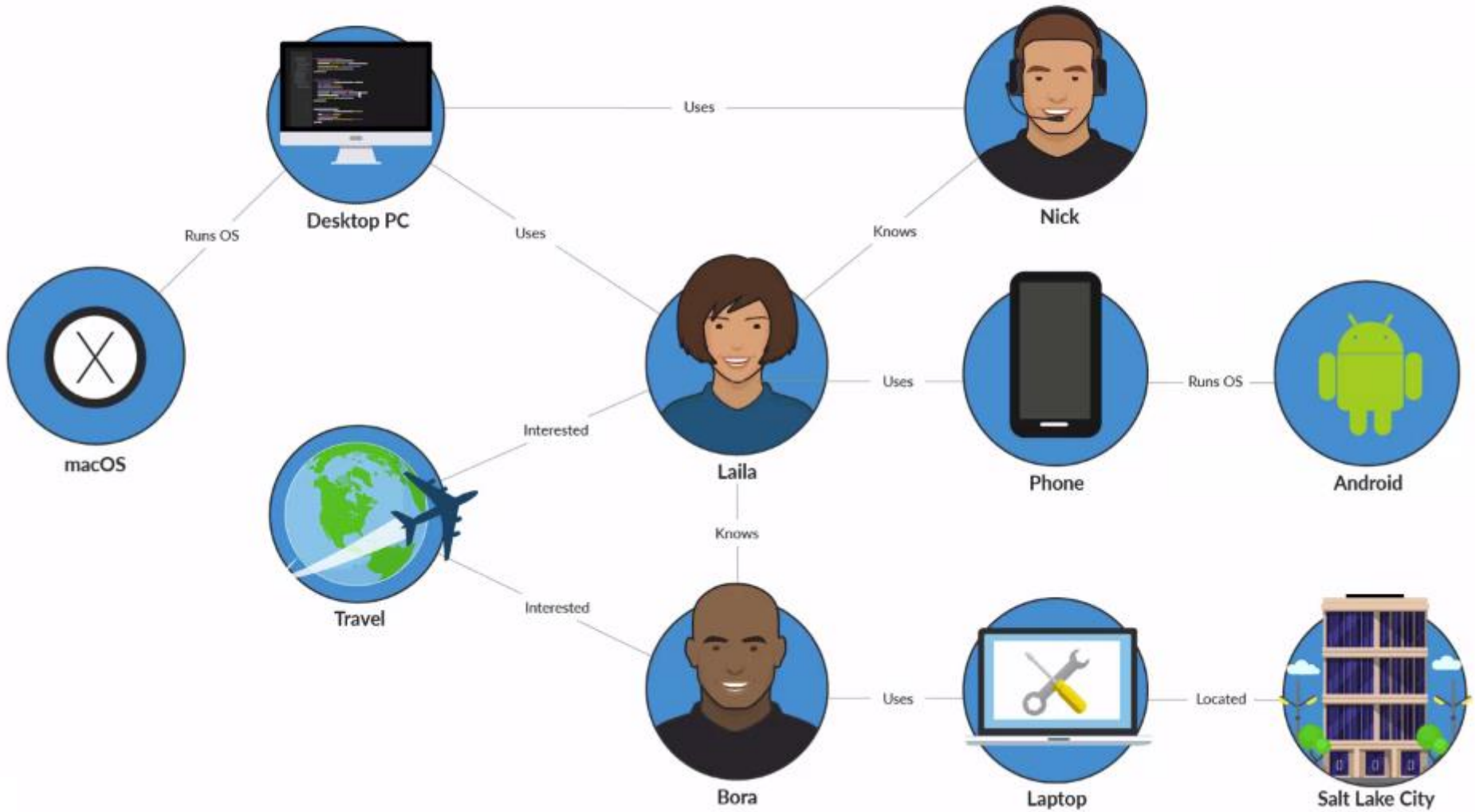
SQL API

SQL API

Lets you use a SQL-like language to query JSON documents (which is how Cosmos DB stores the data)

Easier for SQL users than the MongoDB API

Graph Databases



Wide Column Model

- Used by Apache Cassandra
- Cosmos DB provides the Cassandra API for applications that are written to use a Cassandra database