Describe how to work with relational data on Azure

13. Creating SELECT statements

- There are 6 principal clauses to the SELECT statement:
- SELECT
  - Which columns/fields do you want?
- FROM
  - Which table/query do you want it from?
- WHERE
  - What rows do you want? What is the criteria?
- GROUP BY - not used in Core (SQL)
  - When you use an aggregation (such as COUNT, SUM) you must have everything else in the GROUP BY. It allows the summarization of those fields.
- HAVING (not in Core (SQL))
  - A "WHERE" clause, but only used after the GROUP BY stage
- ORDER BY
  - What order do you want the rows.
- You can remember the order looking at a British or Spanish keyboard.

13. Variants of SQL - SELECT statements

- Some variants require a semicolon at the end; for others (e.g. T-SQL), it is optional.
- SELECT * returns all columns/fields.
- There are variants of SQL which have slight differences for these 6 clauses:
- String literals in some variants are enclosed with single quotation marks – in others, it is within speech marks.
- SELECT TOP(10) returns the first 10 rows in T-SQL, but
  - In MySQL you would end the statement with LIMIT 10

13. Query relational data in Azure SQL Database

- Using the Azure portal (previously done)
- Using the SQLCMD

  To connect, use

  ```
  sqlcmd -S <server>.database.windows.net -d <database> -U <username> -P <password>
  ```

  Then type your SQL commands.
- Using Azure Data Studio – New Query after connecting
- Using SQL Server Management Studio – New Query after connecting
- Using SQL Server Data Tools in Visual Studio – Tools – SQL Server – New Query to connect

SQL Server Management Studio (SSMS)

- Integrated environment for managing SQL.
- Available for Windows 8.1 or above.
- Not available for Linux or macOS.
- Connect to:
  - SQL Server (on-prem),
  - SQL Server (in cloud),
  - Azure SQL Database,
  - Azure Synapse Analytics.

Azure Data Studio

- Cross-platform editor for on-prem and cloud data platform.
- Available for Windows 7 or above, macOS 10.12 or above, or Linux.
- Connect to:
  - SQL Server (on-prem),
  - Azure SQL Database,
  - Azure Synapse Analytics,
  - Azure SQL Data Warehouse,
  - SQL Server Big Data Clusters
  - PostgreSQL.

SQL Server Management Studio versus Azure Data Studio

SQL Server Management Studio	Azure Data Studio
Windows only	Windows, macOS, Linux
Connect to SQL Server, SQL Database or Azure Synapse Analytics	As SSMS, plus connect to SQL Server 2019 big data cluster (preview) or PostgreSQL.
	Allows for data engineering
More advanced SQL features	
Allows Deep administrative configuration or security management	Limited deep administrative configuration or security management
Free, but not open source	Free and open source
Flagship tool for platform management tasks, with broad admin functions.	Basically for editing/running queries (most heavily used capability in SSMS)
Export to File or Text	Export to CSV, JSON, XLSX

Describe how to work with relational data on Azure

Identify the right data offering for a relational workload

Data Offering	Relational Workload	Features
Azure SQL Database	Build modern cloud applications with an always up-to-date relational database service	Serverless compute, hyperscale storage and AI-powered and automated features to optimise performance and durability
Azure SQL Managed Instance	Migrate your SQL workloads to Azure while maintaining high SQL Server compatibility	All the benefits of a fully managed and evergreen platform as a service
SQL Server on Virtual Machines	Migrate your SQL workloads to Azure while maintaining complete SQL Server compatibility	Allows for operating system-level access
Azure Database for PostgreSQL	Build scalable, secure and fully managed enterprise-ready apps on open-source PostgreSQL	Scale out single-node PostgreSQL with high performance or migrate PostgreSQL and Oracle workloads to the cloud
Azure Data for MySQL	A managed community MySQL database service or migrate MySQL workloads to the cloud	Deliver high availability and elastic scaling to open-source mobile and web apps
Azure Database for MariaDB	A managed community MariaDB database service	Deliver high availability and elastic scaling to open-source mobile and web apps

1, 11, 12. Describe the characteristics of relational data

- Tables (entity) – a topic/subject with data
    - Rows – a single instance of an entity.
    - Columns – properties of the entity.
- All rows have the same columns (but columns can be NULLable).
- Tables can contain any number of rows.
- Columns are given a specific data type, e.g. datetime.
    - Describe the characteristics of relational data

1, 11, 12. Normalization

- Each column contains a single type of data
- Reduces duplicate data
- Reduces chance of inconsistent data
- Simplify queries

- 1st Normal Form:
  - Requirements
    - The Values in each column must be atomic (indivisible),
    - Each value contains only a single value.
  - Actions
    - Eliminate repeating groups in individual tables
    - Create a separate table for each set of related data
    - Identify each set of related data with a primary key
- 2nd Normal Form:
  - Requirements
    - Is in 1st Normal Form
    - Reduce repeating information
  - Actions
    - Create separate tables for values that apply to multiple records.
    - Relate tables with a foreign key
- 3rd Normal Form:
  - Requirements
    - Is in 2nd Normal Form
    - Values that are not part of a record's key are to be removed from the table.
  - Action
    - Remove fields that are not dependent on the key

1, 11, 12. Describe the characteristics of relational data

- Keys
  - Primary Keys (a value which uniquely identifies a particular row)
  - Foreign Keys (a value which links to a primary key)

14. Describe Indexes

- Clustered Index (re-orders rows in table – only one allowed per table)
- Non-clustered Index
  - does not re-orders rows in table

  - Multiple non-clustered indexes allowed in the one table.
- Clustered indexes best used for Primary Keys
  - Primary Keys are unique and relate to Foreign Keys.
  - Only one per table.
- Non-clustered indexes best used for searching data.
  - Database uses index for seeking relevant data (in a WHERE clause),
  - Instead of having to scan through the entire table.
- Indexes need to be maintained by database.
  - Takes space.
  - Takes times to update index if data is inserted, updated or deleted.
  - Too many indexes can slow your table down.

14. Describe views

- Views are encapsulated (saved) SELECT queries.
- For example:

  *SELECT * FROM tblTable*
- can be converted into a View by using:

  CREATE VIEW vw_View AS

  *SELECT * FROM tblTable*
- You can then query the View as any other SELECT statement:

  *SELECT * FROM vw_View WHERE CustomerID = 1*

Describe how to work with relational data on Azure

14. Describe relational data structures (e.g., tables, index, views)

- Tables
    - Columns
    - Rows
    - Keys
        - Primary Keys
        - Foreign Keys
    - Constraints
        - Unique
        - Check
        - Default
        - Not Null
    - Indexes
        - Clustered Index (orders rows in table – one per table)
        - Non-clustered index
- Views
- Functions
- Stored Procedures

Describe and compare PaaS, IaaS, and SaaS solutions

On premises ("On prem")	The cloud (IaaS, PaaS, SaaS)
You know where your data is.	You have to trust the location of your data.
The physical location of your data is limited to places you own.	The location of your data can be worldwide.
You are in full control of security.	You have to trust your data's security, to an extent.
You are responsible for paying for the physical server boxes.	Your cloud provider pays for the physical server box, and you pay "rent" for the server.
To upgrade your memory, cores, hard drive space requires planning and purchase of equipment and probably several days/weeks.	To upgrade your memory, cores, hard drive space requires a click on a few buttons and a few minutes.
Capital expenditure	Operational expenses
You are responsible for doing hardware maintenance or upgrades.	Your cloud provider applies any hardware maintenance or upgrades.
You are responsible for doing any software updates.	Maybe you, maybe your cloud provider, are responsible for doing any software upgrades.

	Infrastructure as a Service	Platform as a Service	Software as a Service
On prem	IaaS	PaaS	SaaS
Physical hardware			
Buying Operating Systems			
Maintaining Operating systems (Windows, Linux)	Maintaining Operating systems (Windows, Linux)		
Database server software	Database server software		
Adding data	Adding data	Adding data	
Other applications	Other applications	Other applications	

Your responsibility

Higher administration effort
Higher capital expenditure cost
More features and control

Lower administration effort
No capital expenditure cost
Fewer features and control

Describe and compare PaaS, IaaS, and SaaS solutions

On prem	IaaS	PaaS	SaaS
Traditional servers	Virtual machines	Virtual databases	Email (Gmail), office applications, DropBox

IaaS	PaaS
SQL Server on Azure Virtual Machine	Azure SQL Database
	Azure Database for PostgreSQL
	Azure Data for MySQL
	Azure Database for MariaDB

15. Describe database service: SQL Server on Azure Virtual Machine

- A full version of Windows running a full version of SQL Server.
- IaaS – 100% compatible with SQL Server on prem.
- However, same administration requirements as on prem, e.g. backups.
- You can "lift-and-shift" from on prem to the cloud – rapid deployment.
- Hybrid development also available:
  - SQL Server in cloud.
  - Other resources on prem.
- Quickly resize Virtual Machine.
- Access database using SQL Server Management Studio (SSMS).

15. Describe database service: Azure SQL Database

- PaaS.
- Choose Single Database or Elastic Pool:
  - Single Database can be provisioned, or serverless on General Purpose tier only.
  - Elastic Pool – shared performance resources
  - You can move databases in and out of elastic pools.
- Quickly rescale resources for database/pool – no need to restart the database: 1-80 vCores, 32 Gb – 4 Tb (up to 1 Tb in China and Germany).
- Three service tiers:
  - General Purposes/Standard (the general service tier).
  - Business Critical/Premium.
  - Hyperscale – up to 100 Tb.
- Mostly compatible with on prem version of SQL Server.
- Provides:
  - Automatic patches and backups
    - Full backup every week, differential 12-14 hours, transaction log 5-10 minutes (not adjustable).
    - Long-term retention: weekly, monthly and/or yearly full backups for up to 10 years in Blob storage.
  - Point-in-time restores for up to 7 days (adjustable to 1 to 35 days – 1 to 7 for Basic).
  - Active geo-replication (up to 4 readable secondary databases),
  - Auto-failover groups,
  - Advanced threat protection,
  - Encryption:
    - Protect data in motion, TLS - Transport Layer Security.
    - Encrypt data at rest, TDE - Transparent Data Encryption (on by default).
    - Limit access to Data in use, Always Encrypted (encrypt some plain text columns).
    - Hide parts of data (e.g. credit cards), Dynamic data masking.
  - Zone-redundant databases (availability zones).

  - 99.99% high availability guarantee.

15. Describe database service: Azure SQL Database

- Does not support:
  - Linked servers,
  - Service Broker,
  - Database Mail.
- Access database using:
  - SQL Server Management Studio (SSMS),
  - Azure Data Studio, or
  - Visual Studio (including SQL Server Data Tools).
- Manage database using Azure Portal (web):
  - Adjust data storage size, or
  - Number of available cores.

15. Describe database service: Azure SQL Managed Instance

- SQL Server in the cloud – not just one database.
- Nearly 100% compatible with on prem version of SQL Server Enterprise Edition, including linked servers.
- Lift-and-shift supported.
- Provides:
  - Operating system and Database Installation and Patching,
  - Dynamic Resizing,
  - Data replication,
  - High availability configuration,
  - Access using Azure Active Directory credentials, tied to your current computer sign-in.
  - Backups and point-in-time restores – same as Azure SQL Database.
    - But only to SQL Managed Instance – not to SQL Server instance or Azure SQL Database.
- Access database using SQL Server Management Studio (SSMS) or Azure Data Studio.

16. Describe Azure Database for PostgreSQL

- PaaS – Open source.
- Uses variant of SQL called pgsql.
- Not 100% compatible with on-prem PostgreSQL.
- Option to hyperscale (Citus) to create a cluster.
    - Multiple servers ("nodes") co-ordinating with each other.
- Access database using:
    - pgAdmin (PostgreSQL GUI database manager),
    - psql (a command-line interface – CLI), or
    - Azure Data Studio.
- Migrate data from on prem using the Azure Database Migration Service.

16. Describe Azure Database for MariaDB

- Compatible with Oracle Database (which uses the SQL dialect PL/SQL).
- Community edition fully managed and controlled by Azure.
- Provides:
    - High Availability – no additional cost.
    - Scalability.
    - Automatic backups, with point-in-time restore.
    - Secure data, both in motion and at rest.
    - Built-in support for temporal data.
- Access database using MySQL Workbench or mysql command-line client.
    - Most mySQL tools with work with MariaDB.
- Migrate data from on prem using the Azure Database Migration Service.

16. Describe Azure Database for MySQL

- Open-source database.
- Includes:
    - High availability – no additional cost.
    - Scalability.
    - Automatic backups, with point-in-time restore.
    - Secure data, both in motion and at rest.
- Migrate data from on prem using the Azure Database Migration Service.
- Access database using MySQL Workbench or mysql command-line client.
    - Cannot use Azure Data Studio

Roles and responsibilities for data workloads

8. Describe responsibilities for database administrators

- DBA's responsibilities include:
    - capacity planning,
        - what processors, processor speed, memory, hard disk storage, and network interface
    - installation and configuration,
        - getting the database server on the server, and so that it works in the way you want it to.
    - database design,
        - creating the appropriate table structure
    - migration,
        - moving from one server to another
    - performance monitoring,
        - making sure your queries run as best as possible
    - security,
        - so that only the right people see the right data
    - troubleshooting
        - if users have any problems, how can these be cured
    - backup and data recovery
        - what if someone deletes the wrong data. Can you get it back?

9. Describe responsibilities for data engineers

- Data Engineers' responsibilities include:
    - Creating Extract-Transform-Load pipelines,
        - also known as data ingestion pipelines
        - they need to ensure that the data is still secure and private.
    - Translate huge amounts of data into insights
        - reporting and analytics,
        - online analytical processing,

        - multi-dimensional cubes or data warehousing for reporting
    - dashboard development,
    - data mining, process mining, and text mining,
        - extracting and discovering patterns from data, event logs, and text
    - complex event processing,
        - processing real-time things which have happened (events)
    - business performance management,
        - allows management to see how close they are to their goals,
    - benchmarking,
        - comparing time and cost of business processes against competitors
    - predictive analytics,
        - look at past facts to make predictions about the future
    - prescriptive analytics
        - not only what will happen, but why it will happen

10. Describe responsibilities for data analysts

- Data Analysts' responsibilities include:
    - inspecting,
    - cleansing
        - Removing duplicate data, data with errors, or data that is incomplete.
    - transforming
        - converting data to a preferred structure.
    - modelling data
        - connecting datasets together, using primary or foreign keys, and constraints.
- So that they can:
    - turn data into valuable information,
    - work out what it means, and
    - lead to conclusions for your company

Describe how to work with non-relational data on Azure

2, 3. Describe the characteristics of non-relational data

- Semi-structured Data structure
  - Sometimes called “NoSQL”.
  - Entities stored in collections or containers
  - “Tables” are not related to each other.
  - Flexible structure
    - No fixed schema
    - Each object self-identifies schema
  - Can be indexed (for swift retrieval of data)
- Non-structured Data
  - Files, videos, pictures
  - Hard drive images

2, 3. Determine when to use non-relational data

- Semi-structured data:
  - Receiving data very quickly, e.g. Internet browsing data
  - Internet devices
  - Social media
- Non-structured data:
  - File storage
  - Hard drive storage

2, 3. Describe the characteristics of non-relational data

- Semi-structured data Formats
  - JSON format (JavaScript Object Notation)
  - Key-value stores
  - Graph databases
  - Avro
    - Schema in JSON
    - Data in Binary
  - ORC (Optimized Row Columnar format)
    - Column-based format
    - Used in Apache Hive.
  - Parquet
    - Twitter format.
    - Contains groups of rows ("chunks") with similar schema

20, 21. Example of JSON data

```
{
        "customerID": "1",
        "customername":
      [
                {"firstname": "Phillip"},
                {"lastname": "Burton"}
        ],
        "customeraddress":
        [
                {"number": "4401"},
                {"street 1": "Floridian Way"},
                {"city": "Golden Oak"},
                {"state": "Florida"}
        ]
}
```

- Used in Document Databases

- Used in Azure Cosmos DB – Core (SQL) API, and a variant is used in MongoDB API.

20, 21. Example of Column Family data

- Column Families are stored separately
- Examples includes ORC and Parquet (used by Twitter)
- Good when you only need one Column Family instead of the whole row with Joins.
- Each row contains a unique key, and you retrieve data using the key.
- Most notable used in Apache Cassandra.
- Implemented in Azure in Azure Cosmos DB – Cassandra API.

Customer Detail			Address Detail		
1:	CustomerDetail:FirstName: CustomerDetail:LastName:	Phillip Burton	1:	AddressDetail:Number: AddressDetail:Street1: AddressDetail:City: AddressDetail:State:	4401 Floridian Way Golden Oak Florida
2:	CustomerDetail:FirstName: CustomerDetail:LastName:	Alexander Hamilton	2:	AddressDetail:Number: AddressDetail:Street1: AddressDetail:City: AddressDetail:State:	14 The Crescent Tampa Florida

20, 21. Example of NoSQL Key-value data

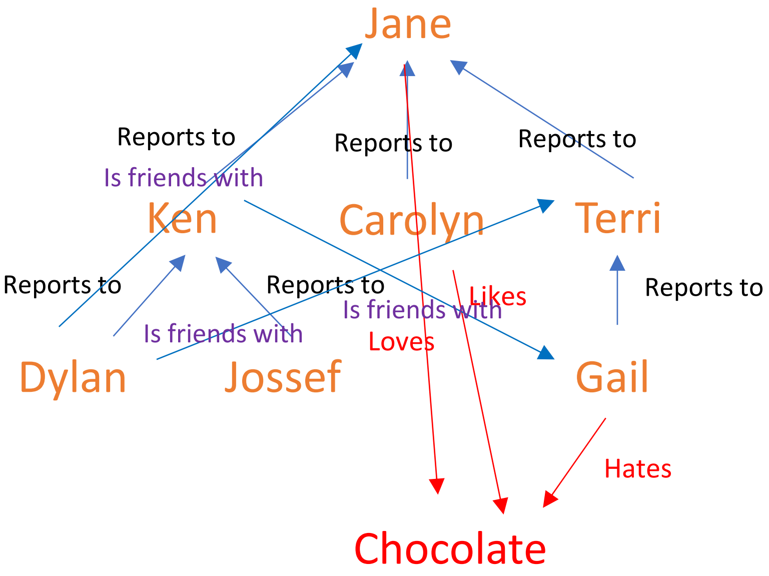
- Key is unique and is searchable.
- Data is stored in key order.
- You cannot search the values.
- Good for quick read and write.
- Can insert and delete items.
- Updates are essentially delete then insert.
- Great for loading stream data (data ingestion), e.g. data from users in a website.
- Used in Azure Table storage.
- Used in Cosmos DB - Table API.

Key	Value
1	Apples

2	Bananas	Yellow	Ecuador
3	Cherries	Large	

20, 21. Example of Graph database

- Focuses on relationships between entities.
- Entities are called “nodes” (in green).
    - In Azure, they are called “Vertices”.
- Relationship are called “edges” relationships (arrows and black words).
    - In Azure Gremlin API, edges can be assigned using “source” and “target”.
- Answer questions like:
    - Who reports to Ken?
    - Who is friends with Terri?
    - Who likes chocolate:
- Used in Cosmos DB - Gremlin API



20, 21. Describe Azure Cosmos DB APIs

Data format	Azure Cosmos DB API
JSON data (documents)	Core (SQL) API (preferred) MongoDB API
Key-value data	Table API (Also Azure Table storage)
Column store storage	Cassandra API
Graph (relationship)	Gremlin API

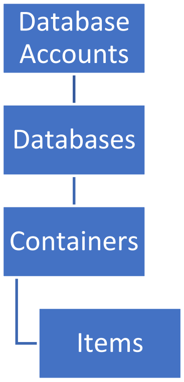
- Allows for:
  - High availability and scalability
    - 99.99% availability for single-region accounts
    - 99.999% availability for multi-region accounts.
  - Low latency (milliseconds)
  - Various consistency levels
  - Designed for huge traffic (writing and reading).
- Use for:
  - Streamed data from Internet devices.
  - Streamed data from gaming.
  - Internet shopping, web and mobile apps.

20, 21. Identify Azure data services for non-relational workloads

- Azure Cosmos DB APIs for semi-structural workloads
    - Use Core (SQL) for any new databases, unless
    - You want to query complex relationships – use Gremlin API, or
    - You are migrating Table, Cassandra or MongoDB.
    - Azure Table storage can be used for NoSQL key-value data. However, Core (SQL) is generally preferred for new databases for key-value data.
    - Core (SQL) can be serverless.
    - You can have one "free tier" Azure Cosmos DB account per subscription (not serverless).
    - You can access Cosmos DB using, among others, Azure Cosmos Explorer:
        - Query results and provide temporary/permanent read-only/read-write access.
- Azure Blob storage for individual non-structured items.
- Azure File storage for files on a computer.

20, 21. Azure Cosmos DB resource model

- PaaS
- Database Accounts
  - Configure: API, Consistency Policy, Regions, and Enable Write in multiple Region
  - You need one Cosmos DB account per different API.
- They can contain one or more Databases, a unit of management
  - Called "Keyspaces" in Cassandra API.
  - With Table API, a default "database" is automatically created.
  - Configure: Provisioned throughput (Standard or autoscale)
- Containers, the fundamental unit of scalability
  - Called "Tables" in Cassandra and Table API.
  - Called "Graph" in Gremlin API.
  - Configure: Request Units/sec, Provisioned throughput (Standard or autoscale), Serverless
- Containers contain "Items".
  - Called "Rows" in Cassandra API.
  - Called "Documents" in MongdoDB API.
  - Called "Nodes" or "edges" in Gremlin API



21. Describe Azure Cosmos DB APIs

- SQL (Core) API – for a variant of SQL queries over documents.
    - Generally use, unless there is a reason to use one of the others.
    - Uses JSON documents.
- Gremlin API – graph database interface.
    - Use for relationships (edges) between nodes (entities).
- Table API – store and retrieve documents.
    - NoSQL Key-value data. Use for switching from Table Storage on prem.
- MongoDB API.
    - Use for migrating from MongoDB on prem.
    - Uses JSON-like documents.
- Cassandra API.
    - Use for migrating from Cassandra on prem.
    - Uses column store storage.

Azure storage

17-19. Storage accounts

- General-purpose v2 accounts allows for blobs, files, queues and tables.
  - V1 account is for legacy purposes only.
- Premium performance tier
  - Store unmanaged virtual machine disks.
  - Microsoft recommends using managed disks with Azure virtual machines.
- Standard performance tier
  - Store blobs, files, tables, queues and Azure virtual machine disks.
- Hierarchy:
  - Resource group
  - Storage Account
  - Container
  - Items (e.g. Directory, Database)
- So a single GPv2 storage account may store multiple different types of containers.
- Zone replication (ZRS, GZRS, RA-GZRS) only allowed on Standard.
- Geo-redundant storage (GRS, RA-GRS) and LRS allowed on both Standard and Premium

17-19. Replication options

- LRS (locally redundant storage – copied 3 times in primary region)
  - [Premium and Standard]
- GRS and RA-GRS (Geo-redundant storage, and read-only GRS)
  - Copied 3 times in primary region, and copied to secondary region.
  - [Premium and Standard]
- ZRS (HA – zone-redundant storage. Copied across 3 availability zones in primary region)
  - [Standard only]
- GZRS and RA-GZRS (HA and Durability) – both ZRS and GRS
  - [Standard only]

4, 5. Recommend the correct data store

- Unstructured data – doesn't contain any fields

Azure storage

- You cannot search items for specific properties/columns.
- Video and audio data
  - stored as blobs
  - in Azure Storage Account
- Files
  - Main content is not structured.

19. Describe Azure Table storage

- Azure Table Storage is Azure's implementation of the NoSQL key-value model.
- Items are rows, fields are columns.
- Rows can be up to 1 Mb. Tables can be hundreds of Tbs.
- Columns may vary per row, up to 252 per row (excluding the keys).
- There are no relationships, foreign keys, stored procedures – just primary key.
- It is split into partitions, based on a partition key. This allows for quick retrieval of a single row or a range of rows (based on a range of Row Key values).
- The partition key and row key make a clustered index.
- High availability:
    - Data is replicated in an Azure region three times.
    - For additional cost, data can be replicated three times in another region (geo-redundant storage: RA-GRS or RA-GZRS). However, it is read-only unless there was a failover in the primary region.
- Advantages
    - Simple to scale.
    - No need to create relationships.
    - Adding rows is quick.
    - Retrieving rows based on keys are quick.
- Disadvantages
    - How do you filter or sort on value data?
- Ideal for:
    - Catalogues on the web. Partition = category, row = product ID.
    - Capturing Internet data.
    - Capturing logging and performance data.

19. Azure Table storage versus Azure Cosmos DB Table API

Azure Table storage	**Azure Cosmos DB Table API**
Fast	Very fast (milliseconds)
Up to 20,000 operations per second	Scalable – supports more than 10 million operations per second per table
Single region, with optional readable secondary read-only (HA)	One to a number of regions. Support for automatic/manual failovers. Multiple write regions.
No secondary index – just one primary key on PartitionKey and RowKey.	Complete indexing on all properties by default.
Uses index for primary key; scans for others	Automatic indexing of properties
Strong consistency within primary region; eventual in secondary	Your choice of five well-defined consistency levels
SLA of 99.9%-99.99%	99.999% read availability, and 99.99% (single-region)-99.999% (multiple) write.

17. Describe Azure Blob storage

- Blob = Binary Large Object
- Block blobs – a collection of blocks.
    - Each blob is a large single binary object – e.g. files, images, videos (unstructured data).
    - A blob can consist of up to 50,000 blocks, each of up to 100 Mb, total over 4.7Tb.
    - Used when individual blobs rarely change. Use BlockBlobStorage.
- Page blobs – a series of 512-byte page, up to 8 Tb.
    - Can be used for random read/write operations. Imagine a hard drive.
    - Used in Azure for disk storage for Virtual Machines. Use StorageV2.
- Append blobs – up to around 195 Gb.
    - Each block is up to 4 Mb.
    - Read and write only – no deletions or updating.
    - Used for logs or archiving. Use BlockBlobStorage.
- Blobs are stored in containers.
    - Think of a container as a folder, with the blobs as files.
- Containers can be held in a hierarchy of folders.
- Access tiers:
    - Hot – frequent use.
        - Higher storage costs, lowest access costs
    - Cool – less frequent use, stored for 30+ days. Higher access cost.
        - Lower storage costs, higher access costs.
        - If deleted before 30 days, early deletion charge applies.
    - Archive – rarely used, stored for 180+ days. Needs to be rehydrated for use (may take hours, but smaller items rehydrate more quickly).
        - Cannot be read, overwritten or modified until rehydrated. Can be listed (catalogued).
        - If deleted before 180 days, early deletion charge applies.
- Useful for:
    - Images or documents, maybe to serve a website.
    - Access to files.
    - Streaming video or audio
    - Backup or archiving data.
    - Data for Azure or on prem service.
- In addition to redundancy (or geo-redundancy at extra cost):
    - Versioning / snapshots – restore earlier versions of blobs.
    - Soft delete – recover deleted/overwritten blobs.

18. Describe Azure File storage

- Store files on the cloud.
- Create in a storage account with up to 100 Tb.
  - Add file shares and grant access to other users.
  - Up to 2,000 connections per file at the same time (for reading).
  - However, probably only 1 connection can write at any one time.
- You can use the following Storage accounts:
  - General purpose version 2 Storage Accounts (Standard only – using hard disks), and
  - Azure FileStorage storage accounts (Premium – using solid-state disks (SSDs), with greater throughput but higher cost). This can only store files (not blobs etc.).
- Upload using
  - Azure File Storage,
  - AzCopy utility,
  - Azure File Sync to have local versions of shared files.
- Use for:
  - Moving existing data to Azure.
  - Share data on prem, cloud and in apps.
    - Including server data like logs, data from Internet monitors, and backups.
  - Writing High Availability backups.
  - Mounting shares (e.g. to a drive letter) from anywhere using SMB 3.0.
- Includes:
  - Encrypted at rest, and you can enable encrypted in transit, using SMB 3.0.
  - Mount Azure File Storage on any computer or in an app.
  - Data is replicated in an Azure region (LRS).
  - For additional cost, data can be zone-redundant (ZRS).
  - For additional cost, standard performance tier only can be either:
    - GRS – only readable in the event of a problem in the original region,
    - RA-GRS – always readable.
    - It can also be GZRS.

Describe an analytics workload on Azure

6. Transactional systems

- Transactions are a small, complete, separate part of work.
- Transactions can be high volume.
- Online Transactional Processing (OLTP)
- Data divided into small tables ("normalization"), so unnecessary data not needed to be written.
- Supports fast writing.
- Querying can be slower, due to normalization.

6. Describe transactional workloads

- Transactions
  - Begin
  - Commit or Rollback
- ACID
  - Atomicity – single unit
  - Consistency – always valid
  - Isolation – treated separately
  - Durability – remains committed
- Locks
- Distributed Databases

25. Describe batch data

- A batch is a group of new data.
- This data can be processed:
    - At a particular time
    - After a particular interval,
    - After a limit to the new data has been reached,
    - Some other event
- Examples
    - Data received over the day to be processed at night.
    - Lots of records received from a data source by email/CD.
- Advantages
    - Can be processed when convenient, including overnight.
    - Potentially huge data set processed at once.
- Disadvantages
    - Time delay
    - Batch needs to be complete before processing begins.

25. Describe streaming data

- Streaming data is data in real time.
- Examples:
    - Stock market data in real time.
    - Social media data.
    - Online gaming data.
    - Your location.
- Advantages
    - Can provide near instant response (trigger or action).
- Other characteristics:
    - Only small number (or even one) record processed at once.
    - If more data arrives at once, multiple processing actions may be required.

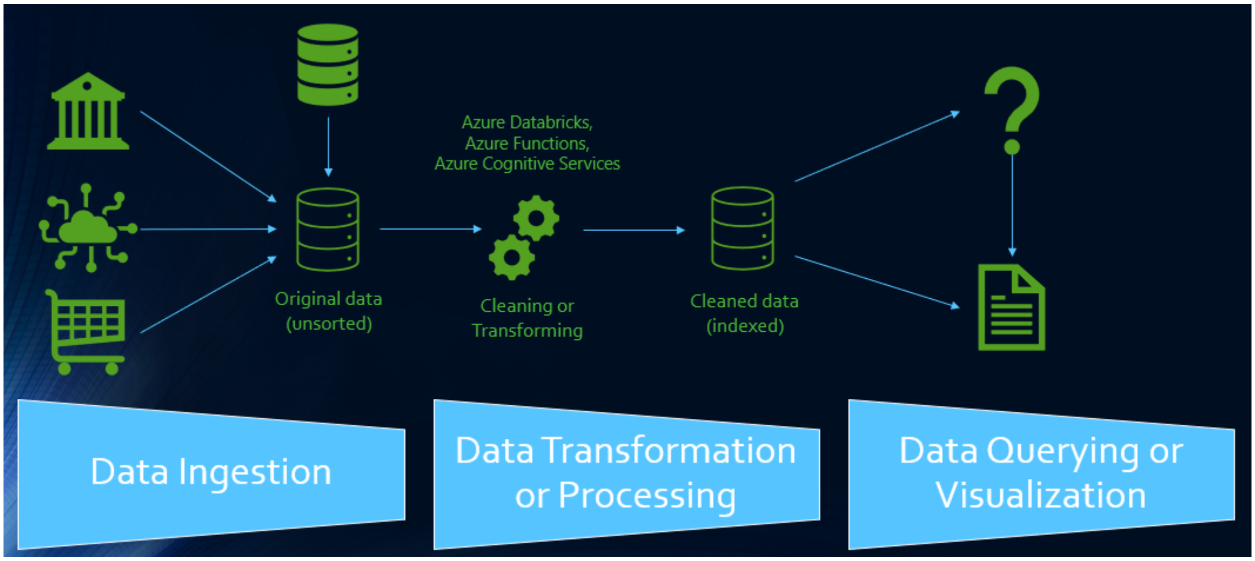
25. Describe the difference between batch and streaming data

Batch data	Streaming data
Can be processed when convenient	Needs to be captured when available.
Can be used for high quantity of data.	Data generally is small in size.
Processes all data in batch.	Only process recent data.
Delay until processing complete.	Processing generally occurs immediately.
Complex data	Simple data

Describe the concepts of data processing

- Transform/process data into a suitable form:
  - For querying, or
  - For visualizing.
- For example:
  - From one type (document database) to another (relational database)
  - Changing string data that looks like dates into date fields.
  - Adding additional data from other sources.
  - Creating summaries.
- You can use:
  - Azure Databricks
  - Azure Functions
  - Azure Cognitive Services

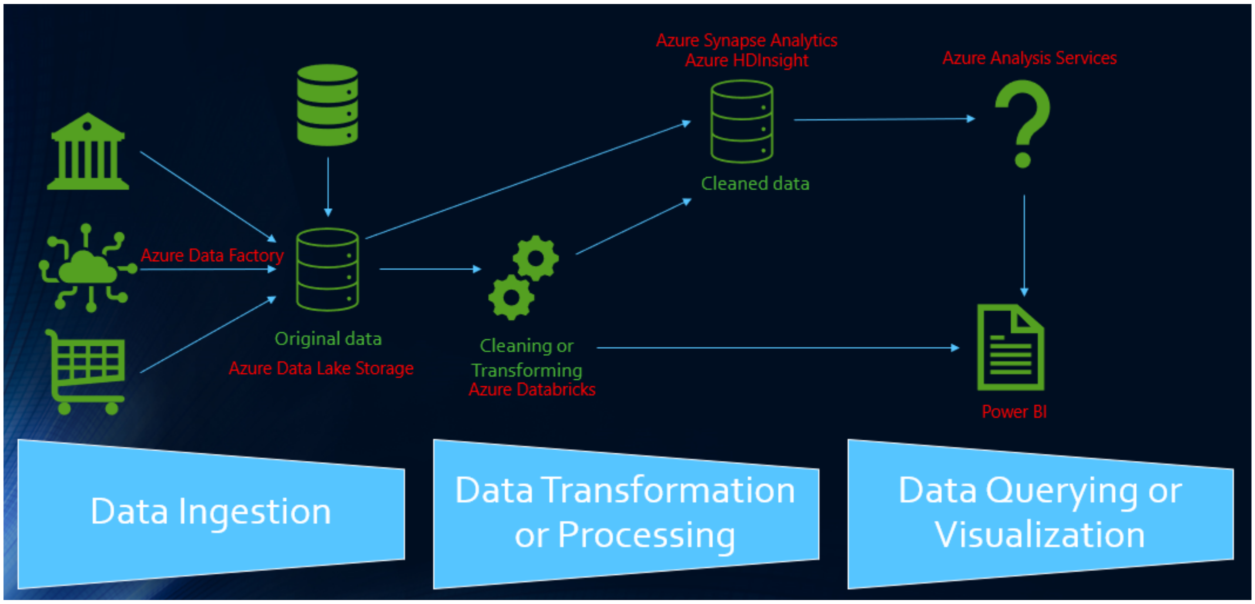
22. Analytical systems



7. Describe the difference between a transactional and an analytics workload

Transaction systems	Analytics systems
Transactional systems are to process transactions, e.g. online shopping and inventory.	Analytics systems are to query data (including one off queries and Big Data analysis) and generate insights.
Read-write, with an emphasis on write	Read only
Capturing data	Ingest – Transform, so it can be queried
Current data	Snapshot (or series of snapshots)
Online Transactional Processing (OLTP)	Online Analytical Processing (OLAP)

Describe data warehousing workloads



Azure Data Factory

- Data Integration Service.
- Use in ETL or ELT projects.
- Input from multiple data sources, both cloud and on prem.
  - Data may be structured or semi-structured, static or streaming.
  - Remove unwanted data ("noise").
- Azure Data Factory "orchestrates" services:
  - Directs, controls and connects them, for automated sequences of complex operations.
  - Uses "linked services", with what the service is, and the AuthN.

- Datasets are data you want to ingest (input to Data Factory) or store (output).
- Work is designed in a "pipeline":
    - Created in a Microsoft GUI, or
    - Creating in your own code.
    - Can use other Azure services - Azure Functions, or Azure Databricks Notebooks etc.
    - Pipelines can include branching (If), looping (ForEach) or remapping column names.
- Convert remaining ("interesting") data, e.g. formatting data, to create uniform output
        - What if data contained British dates in a string format?
        - What if single strings contained multiple fields?
        - Simple conversions for streaming data.
        - Do you need to deduplicate, filter, or re-map columns from source to target?
- Pipeline runs:
    - Manually,
    - On a trigger (at a specified schedule or repeated intervals), or
    - On an event (new data).
- Data is then "ingested" into a data store,
    - Most notably for a Data Warehouse: Azure Data Lake Storage and Azure Synapse Analytics
    - Other destinations could also include:
        - Azure Blob storage,
        - Azure Cosmos DB (SQL API)
        - Azure Database for mySQL or PostgreSQL,
        - Azure SQL Database,
        - Azure SQL Managed Instance.

PolyBase

- Parallel-processing engine enables you to run T-SQL against external data sources.
- Run data managed by:
    - Hadoop,
    - Spark,
    - Azure Blob Storage,
    - Databases like Cosmos DB, Oracle, Teradata, MongoDB.
- Transfer data into a table into SQL Server, or into Azure Synapse Analytics.

- Join tables from an SQL database with external data.
- Not supported by Azure SQL Database.

SQL Server Integration Services (SSIS)

- On prem ETL platform.
- Use GUI.
- Why are we talking about an On prem platform?
    - Azure Data Factory can run your existing SSIS packages in its pipeline.
- SSIS Feature Pack for Azure:
    - Connects to Azure services,
    - Transfer data between on prem and Azure data sources,
    - Process data in Azure.

Azure Data Lake

- Stores raw data:
    - Very fast to load/upload
    - Not easy to analyse.
- Staging storage.
- A version of Azure Blob storage, with:
    - Allows RBAC (Role-based Access Control) on your data, and
    - Compatible with HDFS (Apache's Hadoop Distributed File System).
    - Run Hadoop using Azure HDInsight.
- Also, can be organized into Directories and subdirectories
    - You need to enable “hierarchical namespace” on the Advanced tab.
    - With a lot of data, this can lead to better performance.
    - Without this enabled, it has a flat namespace (like Azure Blob storage).
- Can be used by:
    - Azure Data Factory,
    - Azure Databricks,
    - Azure HDInsight,
    - Azure Data Lake Analytics,
    - Azure Stream Analytics.
        - Uses U-SQL, hybrid SQL and C# language.

24b. Azure Databricks

- Provides "big data" modelling, processing, streaming, and machine learning.
- Also supports exploration and data visualization.
- Apache Spark environment on Azure:
  - Parallel processing engine.
- Can also be used for ingesting.
- Can use drivers to import data from:
  - Azure Blob Storage,
  - Azure Data Lake Store,
  - Hadoop storage,
  - Flat files,
  - Databases (including Azure SQL Database and Azure Cosmos DB),
  - Data warehouses, and
  - Streaming data.
- However, data is written to Azure Blob Storage or Data Lake Storage before processing.
- Clusters are computation resources and configurations
- Spark code is created in "notebooks” inside clusters.
  - "Notebooks" contains "cells", a separate block of code (a series of steps).
  - Cells can read and process data from multiple data sources, and write results to a data store.
- Includes GUI with Spark code, and you can use query data with:
  - R,
  - Python,
  - Scala,
  - Java and
  - SQL.
- Supports structured stream processing.
  - Calculations can be updated when new data arrives.
- Jobs can run notebooks:
  - Immediately, or
  - When scheduled.
  - Email alerts can be set up in case of job start, success or failure.
  - Jobs create automated job clusters when run, and terminates the cluster when complete.
  - Automated job clusters cannot be restarted.
- Other clusters are all-purpose clusters:
  - Can be manually terminated and restarted.
  - Multiple users can share all-purpose clusters.

26b. Azure Synapse Data Explorer

- Optimised for log analytics. Query telemetry and log data to get insights.
- Features:
  - Easy ingestion,
  - No need to build complex data models or to transform data,
  - No need for you to maintain indexes,
    - Data Explorer structures semi- and un-structured data, such as JSON and standard string text.
  - Allows you to use Kusto Query Language (KQL) to investigate telemetry and time series data,
    - Easy-to-read, but allowing composition of complex data processing queries.
  - Can work with a massive amount of data (gigabytes or petabytes),
  - Integrated between Data Explorer, Apache Spark, and databases.
- Use it for:
  - analysing your log and event data, from on-premises, cloud, and other data sources,
  - building Internet of Things (IoT) analyses,
  - building Software as a Service (SaaS) solutions for yourselves and your customers.

24a. Azure Synapse Analytics

- Formally Azure SQL Data Warehouse.
- Analytics ELT engine, processes huge data quickly.
  - Ingest from Azure Data Lake or other sources
  - Transform/aggregate data.
  - Run complex queries.
- Massively Parallel Processing (MPP)
  - (The opposite of MPP is Symmetric Multiprocessing – SMP)
  - Control Node interacts with outside applications,
  - Then optimizes requests and controls the Compute nodes.
  - When Compute nodes are finished, results sent back to Control Node.
  - Uses pipelines – group of activities.

Describe an analytics workload on Azure

- Interact using Synapse Studio – web GUI.
- Query using Azure Portal's query editor, SSMS, Azure Data Studio, Visual Studio
- Two computational models
- SQL Pool:
    - Compute nodes use Azure SQL Database and Azure Storage, using T-SQL.
    - Control and compute nodes move data using Data Movement Service (DMS) in chunks called "distributions".
- Can use data received using PolyBase:
    - Makes external data look like SQL tables.
    - Retrieves data from multiple data types, e.g. text files, Azure Blob Storage, Azure Data Lake Storage
- You specify number of nodes.
    - Can be changed when not running a T-SQL query.
- When not in use, pause the service – releases resources and reduces costs:
    - Resuming pools take a few minutes.
- Spark pools:
    - Nodes are Apache Spark clusters.
    - Code written in Notebooks.
        - C#, Python, Scala, Spark SQL.
        - Supports Azure Machine Learning.
    - Spark cluster converts work into parallel tasks.
    - Data can be saved in Azure Storage or Data Lake Storage.
- Spark is optimized for in-memory processing:
    - Faster than disk-based, but more memory resources required.
- You specify number of nodes:
    - Number of nodes can be autoscaled.
    - Nodes can be altered when running queries.
- When not in use, pause the service – releases resources and reduces costs.

- Use SQL Pools for:
    - Complex reporting, and
    - Data ingestion via PolyBase.
- Use Spark Pools for:
    - Data Engineering/Data Preparation, and
    - Machine Learning.
- These pools can co-exist.

26c. Spark structured streaming

- Structured Streaming allows you to do calculations for a small number of rows and updates the results.
    - The main way for processing streaming (real-time) datasets in Apache Spark. You can use Azure Databricks, often with Python or Scala.
    - Use it for large-scale, real-time data, such as Internet of Things (IoT) devices, social media, and online transactions
- It uses triggers to specify how often data should be processed.
- It needs to have robust failure handling
    - So if a failure happens, it can then restart, potentially using a new cluster.

Azure Analysis Services

- Build tabular models to support OLAP (analytical) processing.
- Data sources include:
    - Azure SQL Database or Cosmos DB,
    - Azure Synapse Analytics, and
    - Azure Data Lake Store.
- Includes a GUI for connecting data sources together and creating queries.
    - Explore data from Analysis Services, or
    - Use Power BI for visualization.

24a. Azure Synapse Analytics versus Azure Analysis Services

Azure Synapse Analytics	**Azure Analysis Services**
Multi Tb data or bigger.	Small Tb data.
Fewer than 128 users	Thousands or users.

Very complex queries/aggregations.	
Data mining and exploration.	
Complex ETL operations.	
	Multiple correlated sources (model)
	Power BI analysis.
	Rapid development.
You can summarise big data here …	… and analyse the reduced data here, and then use Power BI

24c. Azure HDInsight

- Big data processing service.
- Similar to Azure Synapse Analytics using:
  - Spark nodes and Spark clusters,
  - Apache Kafka, and
  - Apache Hadoop processing model.
- Analyze data from:
  - Apache Spark - splits data into smaller parallel tasks,
    - Hadoop Map/Reduce – similar – largely replaced by Apache Spark,
  - Apache Hive – an SQL dialect for use with HDInsight cluster running MPP (massively parallel processing) queries,
  - Apache Kafka – clustered streaming service for ingesting streaming data,
  - Apache Storm – real-time data processing, and
  - R.
- Stores data using Azure Data Lake storage.
- Can be used with, or instead of, Azure Synapse Analytics.

Data visualization in Microsoft Power BI

27. What is Power BI?

- Visualization platform.
- Allows you to create dashboards and reports using datasets.
- Easily connect to your data sources.
    - Get and Transform the data.
    - Create visualizations.
    - Model your data (connect different tables together, and add calculated columns and measures).
- Share your analysis with others in your company.
- Power BI Desktop is FREE.
- The Power BI Service has a FREE version, and a $10/month/user version.

Data visualization in Microsoft Power BI

29. Describe data visualization (e.g. visualization, reporting, business intelligence)

- Reporting is showing the results of queries in easy-to-read form.
- Visualization is the showing of a graph, chart, or other graphic to help communicate reporting.
- Business intelligence combines:
  - Data gathering
  - Data storage
  - Making best use of management

27. Describe the role of paginated reporting

- Designed to be printed or shared, and fit on a page.
- Can flow onto multiple pages if space is needed – able to print all tabular data, regardless of length.
  - Non-paginated reports will print what you can see – beware of scroll bars!
- A collection of visualizations.
- One dataset from multiple data sources.
- You can "pixel perfect" the locations of the visualizations.
- Need Power BI Premium to publish from Power BI Report Builder to Power BI Service.
  - But you can print directly from Power BI Report Builder.

27. Describe the role of interactive reports

- Create self-service analytics
  - Save your time recreating almost the same report.
  - Can subscribe (be alerted when the data changes).
- Explore your data.
- Filter on locations or time
- Slicer – reduce data shown in visualizations
- Remove unnecessary detail using:
  - Tooltips
  - Drill-down
    - Move to a different level of granularity.
  - Drill-through to another report

27. Describe the role of dashboards

- Used for “company at a glance” viewpoint.
- A collection of visualizations from reports.
  - These are called ”tiles”, and are “pinned” from reports.
- Or stand-alone elements
  - Images, videos, text boxes, streaming data and web content.
- Only one page per dashboard.
- You can drill-through the tile into the underlying report.
- Unlike reports, you cannot filter or slice in a dashboard.
- You can export the data to Excel.
- Multiple datasets.
- Multiple reports.
- One dashboard can be your “featured dashboard”
  - The initial dashboard shown when you open the Power BI Service.

Additional videos

The following videos were required in the DP-900 exam, but as of April 2022 are no longer required.

However, you may find them useful in understanding more about Azure data services generally.

sqlcmd utility

- Allows you to run queries:
  - At the command prompt
  - In Query Editor
  - In a Windows script file (.bat)
  - In a SQL Server Agent job using an operating system job step.
- Supports:
  - Azure AD authentication,
  - Azure Synapse Analytics, and
  - Always Encrypted.
- Downloadable from the Internet.
- Also built into Azure Cloud Shell.

Query relational data - Other query tools

- Excel
  - For Windows or macOS
  - Go to Data – Get Data – from Database – From SQL Server Database on Windows
- Visual Studio
  - For Windows, macOS or Linux
  - Go to Tools – SQL Server – New Query to connect
- Incorporate into Visual Studio code
  - For Windows, macOS or Linux

Compare Data Definition Language (DDL) versus Data Manipulation Language (DML)

Data Definition Language (DDL)	Data Manipulation Language (DML)
	SELECT – runs a query
CREATE – adds new objects	INSERT – adds rows into a table
ALTER – modifies existing objects	UPDATE – updates existing rows in a table

DROP – deletes existing objects	DELETE – deletes existing rows in a table

Describe provisioning and deployment of relational data services

- You specify:
  - Type of Database service,
  - Storage, including Backups and location,
  - Memory, and Compute power,
  - Other things, such as networking and tags.
- You can specify it in one of four main ways:
  - Azure portal,
  - Azure command-line interface (CLI),
  - Azure PowerShell, and
  - Azure Resource Manager templates.
- Azure then creates your service behind the scenes, without any further input.
- When creating PostgreSQL server
  - the default database is called “postgres”.
  - You can create read-only replicas (up to 5 replicas).
- When creating MySQL database:
  - You can create read-only replicas (up to 5 replicas).

Identify data security component and basic connectivity issues)

- Virtual Networks (VNets) – a series of Azure services connected together, but isolated from other VNets.
  - If you want other VNets to access resources in your VNet, you will need to add them to the “Allowed networks”
- Firewalls – the way that Azure stops anyone from accessing your VNet by default.
  - Add IP address for any on-prem computers or Internet computers.
  - Firewall rule of 0.0.0.0 allows all Azure services to bypass firewall.
  - Azure SQL Database uses port 1433.
  - Azure Database for MySQL uses port 3306.
  - Azure Database for PostgreSQL uses port 5432.
- Private Endpoints allow you to connect the database to the VNet.
- Public Endpoints allow you to connect the database to devices outside the VNet.
- Point-to-site VPN and Private Endpoints – connect mobile devices to databases.
- Security – Advanced Data Security shows assessments and threat protection.

Identify data security component and basic connectivity issues)

- Authentication (AuthN) – who are you?
  - Uses Azure Active Directory (Azure AD or AAD).
  - Alternative to SQL Server Authentication for Azure SQL Database and Azure SQL Managed Instance.
  - Can also use Multi-factor Authentication (MFA) with Azure AD.
- Authorization (AuthZ) – what do you have access to?
  - Uses Role-Based Access Control (RBAC).
- Role assignments are added in the Access Control (IAM) page, and consist of:
  - Security Principal – who (or what object) are you?
  - Scope - what do you want access to?
  - Role definition (also known as “role”) – how much access:
    - Owner – Full access, including delegating access.
    - Contributor – Full access, but not delegating access.
    - Reader – View access.

Identify management tools for non-relational data

- Uploading data into Cosmos DB
  - Azure Portal (but only one document at a time – you can also run ad hoc queries),
  - Cosmos DB Data Migration Tool (downloadable from GitHub), importing data from:
    - JSON files (you can also export to JSON, either locally or Azure Blob storage)
    - MongoDB
    - SQL Server
    - CSV files
    - Azure Table storage
    - Amazon DynamoDB
    - Hbase
    - Azure Cosmos containers
  - Azure Data Factory
  - Application using Cosmos DB BulkExecutor (library).
  - Application using Cosmos DB SQL API client (library).
  - https://cosmos.azure.com/,

Describe provisioning and deployment of non-relational data services

- You specify:
  - Type of service,
  - Storage, including Backups and location,
  - Memory, and Compute power (if appropriate),
  - Amount of throughput (in Request Units per second – RU/s), at a database or container level,
    - Whether throughput is shared throughput database or dedicated throughput.
  - Other things, such as networking and tags.
- You can specify it in one of four main ways:
  - Azure portal,
  - Azure command-line interface (CLI),
  - Azure PowerShell, and
  - Azure Resource Manager templates.
- Azure then creates your service behind the scenes, without any further input.

- Replication – change in “Replicate data globally”
    - By default, it is replicated to another region at Azure’s choice.
    - Additional regions, which increase the price, are read-only replicas by default.
    - You can enable asynchronous multi-region writes, at an additional cost.
        - What if there is a conflict in the writes?
- Consistency – change in ”Default consistency”
    - Eventual – lowest latency but least consistency.
    - Consistent Prefix – changes may appear in sequence.
    - Session – read your own writes.
    - Bounded Staleness – lag between writing and able to read.
    - Strong – Data only available when written everywhere.
- Storage accounts – “Configuration”
    - Set the default Access tier (can be overridden by individual BLOBs)
    - How Replicated
    - Azure AD integration.
- Encryption
    - By default, Microsoft-managed keys.
    - Customer-managed keys can be added to Azure Key Vault.
- Shared Access Signatures
    - Limited right to Azure storage for a limited time.

Describe method for deployment using the Azure portal

Creating a storage account

Sku can be e.g. Premium_LRS, Premium_ZRS kind is BlobStorage, BlockBlobStorage, FileStorage, Storage, or StorageV2 access-tier can be Cool or Hot

Provision Blob storage in a storage account (object data store)

- Public Access can be blob, container or off (private).
- "Blob" supports anonymous read-only, and is probably the most appropriate.
- But there is no list (catalog) for unauthenticated clients.
- You can use Azure Portal, Azure CLI, Azure PowerShell and the AzCopy utility to upload/download files, including blobs.

Provision File storage in a storage account

- You can use Azure Portal and AzCopy utility to upload/download files.
- You can also use Azure Storage Explorer, downloadable from the Microsoft website.

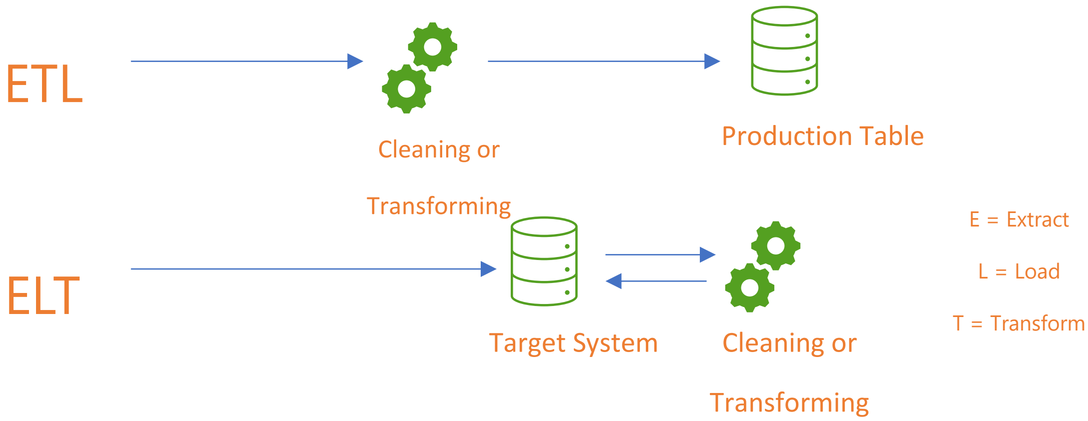
- https://azure.microsoft.com/en-us/features/storage-explorer/
- It is also included in Azure Portal – Storage Explorer.
- With Azure Storage Explorer, you can upload, download and manage:
    - Azure blobs, files, queues and tables,
    - Azure Cosmos DB and

Additional videos

Identify data security components and Identify basic connectivity issues

- Firewall
    - By default, Azure Cosmos DB and Azure Storage's access is enabled to:
        - Vnet, On prem, or Internet.
    - However, that can be changed in the "Firewalls and virtual networks" page.
    - Need to have a valid authorization (AuthZ) token to access data.
    - You can enable "Accept connections from within Azure datacenters" to configure access from, e.g. Azure Functions, which do not have a static IP address.
- Private Endpoints allow you to connect the database to the VNet.
- Public Endpoints allow you to connect the database to outside the VNet.
- To allow access from a subnet on a VNet:
    - Enable a Service endpoint on the subnet, and
    - Add it to your Azure Cosmos account ("Firewalls and Virtual Networks" setting).
- Authentication (AuthN) – who are you, by:
    - Azure Active Directory (Azure AD) preferably, or
    - Access Keys. You can use either of the two keys.
- Authorization (AuthZ) – what do you have access to?
    - Uses Role-Based Access Control (RBAC).
- Role assignments are added in the Access Control (IAM) page, and consist of:
    - Security Principal – who (or what object) are?
    - Scope - what do you want access to?
    - Role definition (also known as "role") – how much access:
        - Owner – Full access, including delegating access.
        - Contributor – Full access, but not delegating access.
        - Reader – View access.
- Advanced Threat Protection for Azure Storage is available – for a fee.

Describe ELT and ETL processing



- Processes to help with ELT and ETL:
  - SQL Server Integration Services (if using Azure SQL Database)
  - Azure Data Factory
    - Scheduled pipelines (workflows)
  - Azure HDInsight Hadoop
  - Azure Databricks
  - Azure SQL Database

Determine when a data warehouse solution is needed

- When you need quick answers from big data.
- Ideal for reading from, rather than writing.
- Data can include both:
  - Historic data (previously processed), and
  - Up-to-date real time data (streamed data).

DP-900 Microsoft Azure Data Fundamentals

April 2022 update

Additional videos

Describe modern data warehousing architecture and workload

- Snowflake schema (OLTP)
- Star schema (DW/OLAP)
  - Achieved by denormalizing
  - Results in big tables (many columns)
  - Better for reading
- Fact tables
  - Measures, and links to dimension tables
- Dimension tables

Describe analytics techniques (e.g., descriptive, diagnostic, predictive, prescriptive, cognitive)

- Descriptive analytics - What happened in the past.
- Diagnostic analytics - Why things happened in the past.
  - Drill down dashboards and hierarchies
- Predictive analytics - What will happen in the future.
  - Machine Learning Strategies and statistical algorithms
- Prescriptive analytics – What actions should be taken to achieve a future goal.
  - Machine Learning Strategies
- Cognitive analytics – Inference
  - Generate hypothesis from the past
  - Test hypothesis in the future
  - How well did it do? Repeat