Exam Preparation Guide

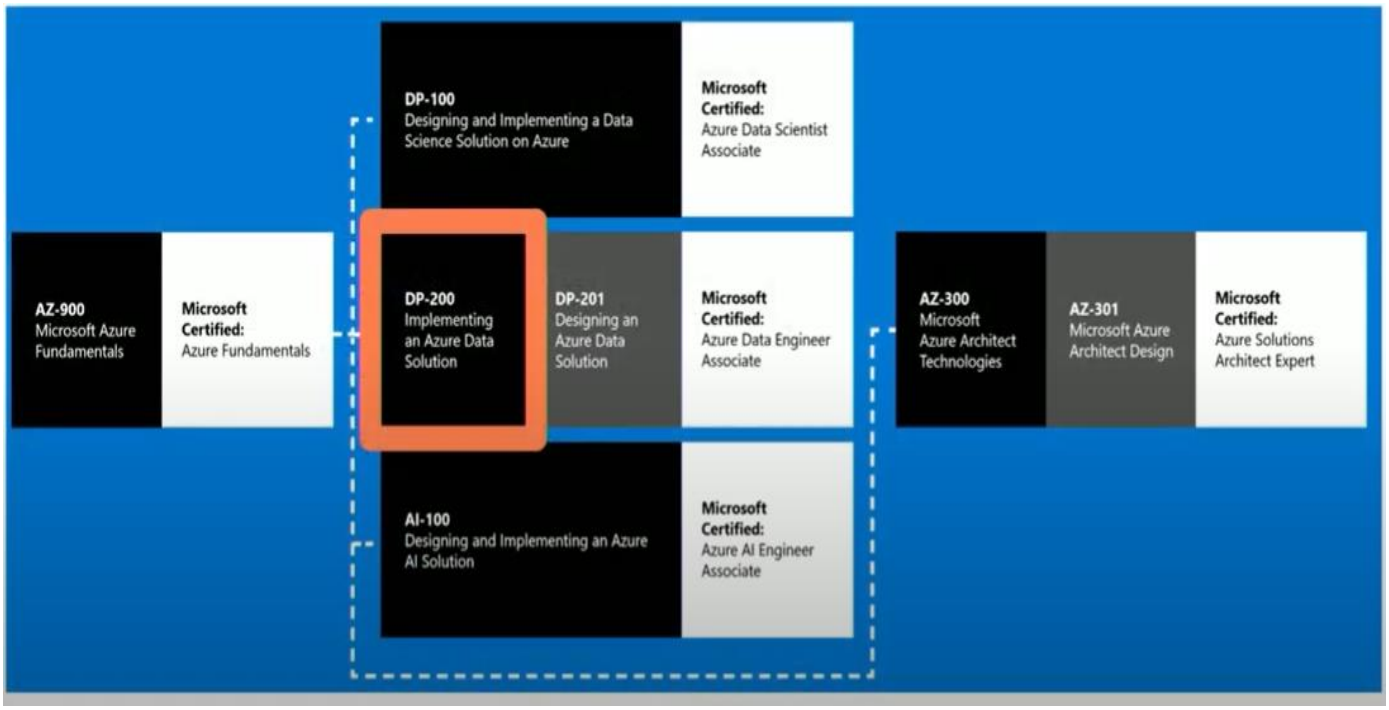
Agenda

- Certification overview
- Exam overview and topics
- How to prepare
- Tips & Tricks

There are different Roles Path available today:

- Apps and Infrastructure
- Data & AI
- Modern Workplace
- Business Applications

For people interested in Data and Artificial Intelligence:



Performance-based testing

- Requires you to prove your skills with hands-on labs.
- Real-world scenarios where you configure, manage resources and services.

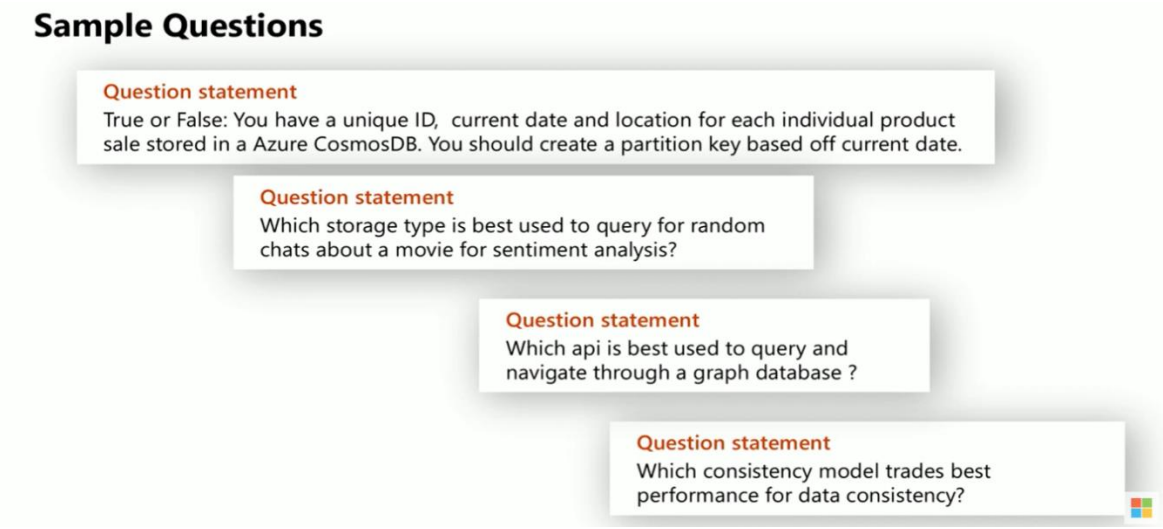
Objective Domain

- Implement data storage solutions (40-45%)
- Manage and develop data processing (25-30%)
- Monitor and optimize data solutions (30-35%)

IMPLEMENT DATA STORAGE SOLUTIONS

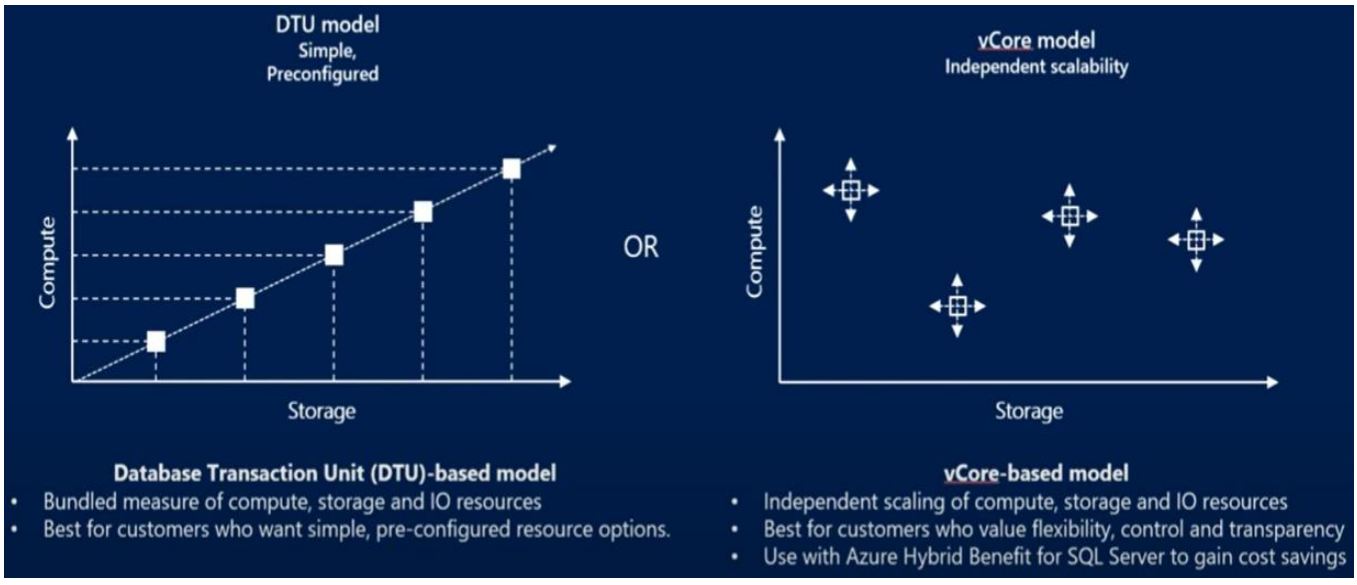
- **Implement Non-Relational Data Store**
  - Azure Data Lake Gen 2
    - Use Gen 1 only when no upgradation is allowed in the solution.
    - Perfect for Machine Learning and Data Science Engineers
    - Supports multiple data types and format
    - Used to store unprocessed raw data
  - WASB (Blob Storage)
    - Used to store raw unstructured data like image, video and audio files.
  - Cosmos DB
    - Supports relational, columnar (Cassandra), graph data types
    - Data distribution and partitioning: One common way to select the partition keys is to observe the 'where' clause in your query. The column which is used most often in your 'where' clause is generally a good candidate for a partition key.
    - Implement a consistency model: Cosmos DB supports many levels:
      1. **Strong**- Closest to RDBMS, consistent, high latency, 1 region, cost higher.
      2. **Bounded Staleness**- at most K or within T seconds.
      3. **Session**- Read your own writes.
      4. **Consistent Prefix**- In order writes, in order writes.
      5. **Eventual**- Eventually everything gets in order.
    - **Provision a non-relational data store**
    - Question on how to use PowerShell or Azure CLI. Will not be asked to write the command from scratch, but focus will be on the arguments like name, location, etc.
    - Azure supports the following APIs:
      1. SQL API
      2. Gremlin API
      3. Mongo DB using Leaf
      4. Table API
      5. Cassandra API
    - Most of the questions focus on checking if you understand the difference between these APIs.
    - Provide access to data to meet security requirement
      1. Use Firewall rules
      2. Enable authentication
         - SQL vs Azure AD (strongest)
         - Use RBAC roles with Azure AD
         - Know built-in roles for storage and SQL
      3. Enable encryption on database
      4. Use row-level security
      5. Enable auditing
      6. Enable threat detection
      7. Enable feature restrictions
    - Availability: Azure provides two different tiers:
      1. Standard/General purpose Tier
      2. Premium/Business critical service tier (collocated compute and storage)
    - Multi-master replication: Available in Premium tier, 99.999% SLA.

```
az cosmosdb create \
--name mycosmosdbaccount \
--resource-group myResourceGroup \
--kind GlobalDocumentDB \
--default-consistency-level Session \
--locations regionName=EastUS failoverPriority=0 isZoneRedundant=True \
--locations regionName=WestUS2 failoverPriority=1 isZoneRedundant=True \
--enable-multiple-write-locations
```



    - You're charged per **RU (Request Unit)** while querying data.

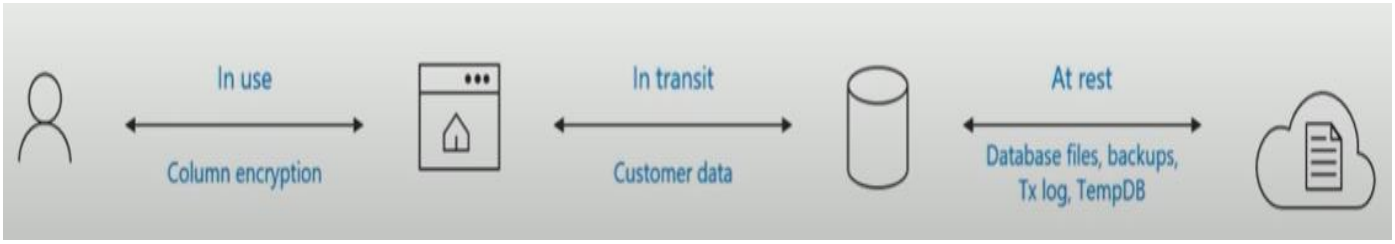
- **Implement Relational Data Store**
    - You should know-
        - Select, Insert, Update, Delete
        - Having, Group by, Where
        - Create Table, Drop
    - Azure SQL
        - Configure Elastic Pools
            1. **DTU**- Distributed Transaction Unit (Fixed combined compute, storage and IO measurement)
            2. **vCore**- Configurable compute, storage and IO performance
            3. **eDTU**- Elastic DTU; Multiple databases share DTU's on server. Best used by low average utilization with infrequent spikes.



- Configure Geo-Replication
    1. Not all regions can be geo-replicated.
    2. Understand how data recovery works when Geo-Replication is enabled.
- High Availability- 2 architecture models
    1. Standard Availability
    2. Premium Availability
        - Think about Clusters
        - Always On availability group.

- Azure Data Warehouse
    - Know about:
        1. Bottom Up architecture (Star Schema)
        2. Top Down architecture (Traditional Normalized)
        3. MPP (Massively Parallel Processing) architecture
            - Up to 60 worker nodes
        4. Sharding of Data Warehouses
            - Hash
            - Round Robin
            - Replicated
    - Implement Distribution and Partitions
    - Implement Polybase
        1. Extract data from different sources using SQL queries.
        2. Expect questions on:
            - Polybase and T-SQL
            - ADF (with Polybase)
            - SSIS
            - BCP
            - 3rd Party tools
        3. **Steps to implement Polybase**
            - Extract into text file
            - Load into BLOB, Hadoop (Databricks) or Data Lake
            - Import data into SQL DW temp table using Polybase
            - Transform
            - Insert into production tables
    - Dynamic Masking Policy
        1. Exclude AAD identities
        2. Masking Rules
        3. Masking Functions
            - Default
            - Credit Card
            - Email
            - Random Number
    - Encrypt data at rest and in motion
        1. You should know
            - Types of data encryption
            - What Always Encrypted means
            - How to protect data from high privileged users

Data encryption	Encryption technology	Customer value
In transit	Transport Layer Security (TLS) from the client to the server	Protects data between client and server against snooping and man-in-the-middle attacks *Azure SQL Database is phasing out Secure Sockets Layer (SSL) 3.0 and TLS 1.0 in favor of TLS 1.2
At rest	Transparent Data Encryption (TDE) for Azure SQL Database	Protects data on the disk Key management is done by Azure, which makes it easier to obtain compliance
In use (end-to-end)	Always Encrypted for client-side column encryption	Data is protected end-to-end, but the application is aware of encrypted columns This is used in the absence of data masking and TDE for compliance-related scenarios



2. Transparent Data Encryption
    - Use keys generated by Azure
    - Bring your own keys and store them in Azure Key Vault.
- Maria, PostgreSQL, MySQL
    - Expect basic implementation questions.

MANAGE AND DEVELOP DATA PROCESSING

- Develop batch processing solutions
    - Develop batch processing solutions by using Data Factory and Azure Databricks.
    - Azure Databricks
        - Microsoft's version of Apache Spark.
        - Complete environment for building and running ML models.
        - Deals with compute power, creating clustered nodes.
        - Provides Notebook environment.
    - Azure Data Factory
        - An orchestrator to tell other services which jobs to run
        - Doesn't actually run jobs, other services run jobs
        - Example: Schedules Azure Data Bricks to analyse and query data and send results to Cosmos DB.
        - Pipeline configured using Graphical UI and JSON.
    - Synapse Analytics: Brings it all together to form a compact development environment.
- Develop streaming solutions
    - Know your Stream Analytics Interfaces
        - Input- Event Hub, IoT Hub, BLOB Storage
        - Output- Cosmos DB, BLOB, Azure SQL, Event Hub, etc.
    - Know your Window functions
        - Tumbling (ex: last 10seconds of data)
        - Hopping (ex: last 3 records in last 10 seconds of data)
        - Sliding
        - Session

MONITOR AND OPTIMIZE DATA SOLUTIONS

- Monitor data storage
    - Go through every Monitor tab in various data resources and understand what kind of data is provided.
    - Ex: Number of incoming requests, latency, number of errors generated, network speed.
    - Cosmos DB monitoring
    - SQL DB monitoring
- Monitor data processing
    - **Azure Monitor**
- Optimize Azure data solutions
    - Manage data life cycle- Threat detection
    - Set up -> Alert -> Explore
- Troubleshooting
    - Connection Issues- Check:
        - Credentials
        - Access keys
        - Deployment failures
        - Connection timeouts
        - Verify service is running- Azure Service Dashboard
    - Performance Issues
        - Queries Slow- Cosmos- Avoid full scans, query in parallel, increase RU's
        - Queries Slow- Data Lake- Check hierarchical namespace