DP900 Final Notes – Part 2

Identify considerations for relational data on Azure (20—25%)

Describe relational concepts

• Identify features of relational data

Tables, Rows, Indexing, Views, Stored procedures

• Describe normalization and why it is used

Do not replicate data.

• Identify common structured query language (SQL) statements

DDL (Data definition language)
Create, Alter, Delete, Drop
DML (Data manipulation language)
Insert, Update, Delete (Select)
DQL (Data query language)
Select
DCL (Data control language)
Deny, Revoke

• Identify common database objects

When you create a database in Microsoft Access, you have a number of different types of object: **tables, forms, reports, queries, macros and modules**.

Describe relational Azure data services

• Describe the Azure SQL family of products including Azure SQL Database, Azure SQL Managed Instance, and SQL Server on Azure Virtual Machines

Azure SQL Database = PaaS offering. Based on Microsoft SQL Server database engine and has nearly the same capabilities as SQL Server on-premises. Elastic with automatic scaling. vCore based. DTU based.
Azure SQL Managed Instance = SaaS offering. Migration from on-premise using Azure Database Migration Service.
SQL Server on Azure Virtual Machines = IaaS offering. Runs on a vm in Azure. Require Azure storage for the vm disk, Azure Vnet for connectivity and Azure Compute service which acts as the hypervisor. Can have automated patching, automated backups, LRS/GRS and high availability.

Azure SQL is the collective name for a family of relational database solutions based on the Microsoft SQL Server database engine. Specific Azure SQL services include:
Azure SQL Database – a fully managed platform-as-a-service (PaaS) database hosted in Azure
Azure SQL Managed Instance – a hosted instance of SQL Server with automated maintenance, which allows more flexible configuration than Azure SQL DB but with more administrative responsibility for the owner.
Azure SQL VM – a virtual machine with an installation of SQL Server, allowing maximum configurability with full management responsibility.
Database administrators typically provision and manage Azure SQL database systems to support line of business (LOB) applications that need to store transactional data.
Data engineers may use Azure SQL database systems as sources for data pipelines that perform *extract*, *transform*, and *load* (ETL) operations to ingest the transactional data into an analytical system.

Data analysts may query Azure SQL databases directly to create reports, though in large organizations the data is generally combined with data from other sources in an analytical data store to support enterprise analytics.

• Identify Azure database services for open-source database systems

Azure includes managed services for popular open-source relational database systems, including:

Azure Database for MySQL - a simple-to-use open-source database management system that is commonly used in *Linux*, *Apache*, *MySQL*, and *PHP* (LAMP) stack apps.

Azure Database for MariaDB - a newer database management system, created by the original developers of MySQL. The database engine has since been rewritten and optimized to improve performance. MariaDB offers compatibility with Oracle Database (another popular commercial database management system).

Azure Database for PostgreSQL - a hybrid relational-object database. You can store data in relational tables, but a PostgreSQL database also enables you to store custom data types, with their own non-relational properties.

As with Azure SQL database systems, open-source relational databases are managed by database administrators to support transactional applications, and provide a data source for data engineers building pipelines for analytical solutions and data analysts creating reports.

Other Azure database services

Azure Cosmos DB

Azure Cosmos DB is a global-scale non-relational (*NoSQL*) database system that supports multiple application programming interfaces (APIs), enabling you to store and manage data as JSON documents, key-value pairs, column-families, and graphs.

In some organizations, Cosmos DB instances may be provisioned and managed by a database administrator; though often software developers manage NoSQL data storage as part of the overall application architecture. Data engineers often need to integrate Cosmos DB data sources into enterprise analytical solutions that support modeling and reporting by data analysts.

Azure Storage

Azure Storage is a core Azure service that enables you to store data in:

Blob containers - scalable, cost-effective storage for binary files.

File shares - network file shares such as you typically find in corporate networks.

Tables - key-value storage for applications that need to read and write data values quickly.

Data engineers use Azure Storage to host *data lakes* - blob storage with a hierarchical namespace that enables files to be organized in folders in a distributed file system.

Azure Data Factory

Azure Data Factory is an Azure service that enables you to define and schedule data pipelines to transfer and transform data. You can integrate your pipelines with other Azure services,

enabling you to ingest data from cloud data stores, process the data using cloud-based compute, and persist the results in another data store.
Azure Data Factory is used by data engineers to build *extract*, *transform*, and *load* (ETL) solutions that populate analytical data stores with data from transactional systems across the organization.

Azure Synapse Analytics

Azure Synapse Analytics is a comprehensive, unified data analytics solution that provides a single service interface for multiple analytical capabilities, including:
Pipelines - based on the same technology as Azure Data Factory.
SQL - a highly scalable SQL database engine, optimized for data warehouse workloads.
Apache Spark - an open-source distributed data processing system that supports multiple programming languages and APIs, including Java, Scala, Python, and SQL.
Azure Synapse Data Explorer - a high-performance data analytics solution that is optimized for real-time querying of log and telemetry data using Kusto Query Language (KQL).
Data engineers can use Azure Synapse Analytics to create a unified data analytics solution that combines data ingestion pipelines, data warehouse storage, and data lake storage through a single service.
Data analysts can use SQL and Spark pools through interactive notebooks to explore and analyze data, and take advantage of integration with services such as Azure Machine Learning and Microsoft Power BI to create data models and extract insights from the data.

Azure Databricks

Azure Databricks is an Azure-integrated version of the popular Databricks platform, which combines the Apache Spark data processing platform with SQL database semantics and an integrated management interface to enable large-scale data analytics.
Data engineers can use existing Databricks and Spark skills to create analytical data stores in Azure Databricks.
Data Analysts can use the native notebook support in Azure Databricks to query and visualize data in an easy to use web-based interface.

Azure HDInsight

Azure HDInsight is an Azure service that provides Azure-hosted clusters for popular Apache open-source big data processing technologies, including:
Apache Spark - a distributed data processing system that supports multiple programming languages and APIs, including Java, Scala, Python, and SQL.
Apache Hadoop - a distributed system that uses *MapReduce* jobs to process large volumes of data efficiently across multiple cluster nodes. MapReduce jobs can be written in Java or abstracted by interfaces such as Apache Hive - a SQL-based API that runs on Hadoop.
Apache HBase - an open-source system for large-scale NoSQL data storage and querying.
Apache Kafka - a message broker for data stream processing.

Data engineers can use Azure HDInsight to support big data analytics workloads that depend on multiple open-source technologies.

Azure Stream Analytics

Azure Stream Analytics is a real-time stream processing engine that captures a stream of data from an input, applies a query to extract and manipulate data from the input stream, and writes the results to an output for analysis or further processing.
Data engineers can incorporate Azure Stream Analytics into data analytics architectures that capture streaming data for ingestion into an analytical data store or for real-time visualization.

Azure Data Explorer

Azure Data Explorer is a standalone service that offers the same high-performance querying of log and telemetry data as the Azure Synapse Data Explorer runtime in Azure Synapse Analytics.
Data analysts can use Azure Data Explorer to query and analyze data that includes a timestamp attribute, such as is typically found in log files and *Internet-of-things* (IoT) telemetry data.

Microsoft Purview

Microsoft Purview provides a solution for enterprise-wide data governance and discoverability. You can use Microsoft Purview to create a map of your data and track data lineage across multiple data sources and systems, enabling you to find trustworthy data for analysis and reporting.
Data engineers can use Microsoft Purview to enforce data governance across the enterprise and ensure the integrity of data used to support analytical workloads.

Microsoft Power BI

Microsoft Power BI is a platform for analytical data modeling and reporting that data analysts can use to create and share interactive data visualizations. Power BI reports can be created by using the Power BI Desktop application, and then published and delivered through web-based reports and apps in the Power BI service, as well as in the Power BI mobile app.