# AZURE Data Foundations Certification

## Skills measured as of November 2, 2023

### Audience profile

This exam is an opportunity to demonstrate your knowledge of core data concepts and related Microsoft Azure data services. As a candidate for this exam, you should have familiarity with Exam DP-900’s self-paced or instructor-led learning material.

This exam is intended for you, if you’re a candidate beginning to work with data in the cloud.

You should be familiar with:

* The concepts of relational and non-relational data.
* Different types of data workloads such as transactional or analytical.

You can use Azure Data Fundamentals to prepare for other Azure role-based certifications like Azure Database Administrator Associate or Azure Data Engineer Associate, but it is not a prerequisite for any of them.

### Skills at a glance

* Describe core data concepts (25–30%)
* Identify considerations for relational data on Azure (20–25%)
* Describe considerations for working with non-relational data on Azure (15–20%)
* Describe an analytics workload on Azure (25–30%)

### Describe core data concepts (25–30%)

#### Describe ways to represent data

1. Describe features of structured data
   * Adheres to a fixed schema, so all of the data has the same fields or properties.
   * Most commonly, the schema for structured data is tabular with rows and columns.
   * Often stored in a database in which multiple tables can reference one another by using key values in a relational model.
2. Describe features of semi-structured
   * Has some structure, but allows for some variation between entity instances. i.e. some customers may have an email, multiple emails, or none at all.
   * One common format of semi-structured data is JSON
3. Describe features of unstructured data
   * Some examples of unstructured data are:
     + Documents
     + Images
     + Audio
     + Video
     + Binary Files

#### Identify options for data storage

* Describe common formats for data files
  + Delimited Text Files
    - Plain text format with specific field delimiters and row terminators
      * Comma-separated values (csv)
      * Tab-separated values (tsv)
      * Field width data
    - Delimited text is a good choice for structured data that needs to be accessed by a wide range of applications and services in a human-readable format.
  + JavaScript Object Notation (JSON)
    - JSON is a ubiquitous format in which a hierarchical document schema is used to define data entities (objects) that have multiple attributes.
    - JSON is a flexible format that’s good for both structured and semi-structured data.
  + Extensible Markup Language (XML)
    - Has largely been superseded by JSON, but there are still some systems that use XML to represent data.
    - Much like HTML, XML uses tags enclosed in angle brackets to represent data, defining elements and attributes.
  + Binary Large Object (BLOB)
    - Unstructured data in raw binary usually storing images, audio, video, and application specific documents.
  + Optimized file formats
    - Avro
    - ORC (Optimized Row Columnar format)
    - Parquet
* Describe types of databases

#### Describe common data workloads

* Describe features of transactional workloads
  + The primary function of business computing
  + Transactional data processing systems record transaction that encapsulate specific events that the organization wants to track.
  + There are many types of transaction but think of a transaction as a small, discreet unit of work.
  + High-volume and quick processing
  + Is often call Oline Transactional Processing (OLTP)
  + OLTP system rely on crud operation to support transactional workloads
  + In order to ensure integrity, OLTPs enforce transaction that support ACID semantics:
    - Atomicity- Each transaction is treated as single unit, which succeeds completely or fails completely
    - Consistency- transactions can only take the data from one valid state to another valid state.
    - Isolation- concurrent transactions cannot interfere with one another.
    - Durability- once a transaction has been committed it shall remain committed and persist even through power outages.
* Describe features of analytical workloads
  + Uses read-only (or read-mostly) systems that store vast amounts of historical or business metrics.
  + Can be based on a snapshot of the data at a given point of time or a series of snapshots.
  + Analytical Processing systems vary but a common architecture for enterprise-scale analytics looks like this:
    - Operational data is extracted, transformed, and loaded ETL into a data lake for analysis.
    - Data is loaded into a schema of tables- typically in Spark-based data lakehouse with tabular abstractions over files in the data lake, or a data warehouse with a fully relational SQL engine.
    - Data in a data warehouse may be aggregated and loaded in to an online analytical processing (OLAP) model, or cube. Aggregated values (measures) from fact tables are calculated for intersections of dimension from dimension tables i.e. sales revenue might be totaled by date, customer, and product.
    - The data in the data lake, data warehouse, and analytical model can be queried to produce reports, visualization, and dashboards.
    - Data lakes are common in large-scale data analytical processing scenarios, where a large volume of file-based data must be collected and analyzed.
    - Data warehouses are an established way to store data in a relational schema that is optimized for read operations- primarily queries to support reporting and visualizations. Data Lakehouses are a more recent innovation that combine the flexible and scalable storage of a data lake with the relational querying semantics of a data warehouse. The table schema may require some denormalization (s some duplication to make queries perform faster).
    - An OLAP model is an aggregated type of data storage that is optimized for analytical workloads. Data aggregations are across are across dimensions at different levels, enabling you to drill up/down to view aggregations at multiple hierarchical levels; for example, to find total sales by region, by city, or for an individual address. Because OLAP data is pre-aggregated, queries to return the summaries it contains can be quickly.

#### Identify roles and responsibilities for data workloads

* Describe responsibilities for database administrators
  + Manage Databases, assigning permissions to users, storing backup copies of data and restore data in the event of a failure.
  + Responsible for the design, implementation, maintenance, and operational aspects of on-premises and cloud-based database systems. They’re responsible for the overall availability and consistent performance and optimization of databases.
  + They work with stakeholders to implement policies, tools, and processes for backup and recovery plans to recover following a natural disaster or human-mase error.
  + Management of the security of the data in the database, granting privileges over the data, granting or denying access to users as appropriate.
* Describe responsibilities for data engineers
  + Manage infrastructure and processes for data integration across the organization, applying data cleaning routines, identifying data governance rules, and implementing pipelines to transfer and transform data between systems.
  + Collaboration with stakeholders to design and implement data workloads, including data ingestion pipelines, cleansing and transformation activities, and data stores for analytical workloads. They use a wide range of data platform technologies, including relational and non-relational databases, file stores, and data streams.
  + They are responsible for ensuring that the privacy of data is maintained within the cloud and spanning from on-premises to the cloud data stores. They own management and monitoring of data pipelines to ensure that the data loads perform as expected.
* Describe responsibilities for data analysts
  + Explore and analyze data to create visualizations and charts that enable organizations to make informed decisions.
  + Enable businesses to maximize the value of their data assets.
  + Responsible for exploring data to identify trends and relationships., designing and building analytical models, and enabling advanced analytics capabilities through report and visualizations.
  + A data analyst processes raw data into relevant insights based on identified business requirements to deliver relevant insights.
* Identify Data Services
  + Microsoft Azure is a cloud platform that powers the applications and IT infrastructure for some of the world’s largest organizations. It includes many services to support cloud solutions, including transactional and analytical data workloads.
  + Azure SQL- collective name for a family of relational database solutions based on the Microsoft SQL Server database engine.
    - 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 admins. use AzureSQL database system to support line of business (LOB) apps that need to store transactional data.
    - Data Engineers may use Azure SQL database systems as sources of data pipeline the perform, extract, transform, and load ETL operations to ingest the transactional data into an analytical system.
    - Data Analyst may query the AzureSQL databases directly to create reports, though in large organization that data is general combined with data from other sources in an analytical data store to support enterprise analytics.
  + Azure Database for open-source relational databases
    - Azure includes managed services for popular open-source relational database systems, including:
    - Azure Database for MySQL – a simple-to use open-source relational database systems, including:
    - Azure Database for MariaDB- a newer database management system that offers compatibility with Oracle Database.
    - Azure Database for PostSQL- 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.
    - Open-source relational databases are used by Database admins, engineers and analyst in similar fashion as AzureSQL database systems.
    - Azure Cosmos DB
      * a global-scale non-relational (No-SQL) 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.
      * Cosmos DB instance can be provisioned and managed by data admins or software developers a spart of the overall application architecture.
    - Azure Storage
      * A core Azure service that enables you store data in:
        + Blob containers – scalable, cost-effective storage for binary files
        + Files 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
      * An Azure service that enables you to define and schedule data pipelines to transfer and transform data.
      * Able to integrate 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 engineers to build ETL solutions that populate analytical data stores with data from transactional systems across the organization.
    - Azure Synapse Analytics
      * A comprehensive, unified Platform-as-a-Service (PaaS) solution for data analytics 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 Querying 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 Analyst 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
        + An Azure-integrated version of the popular Databricks platform, which combines 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 Analyst can use the native notebook support in Azure Databricks to query and visualize data in an easy-to-use web-based interface.
      * Azure HDInsight
        + 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- an 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 form 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 for real-time visualization.
    - Azure Data Explorer
      * A standalone service the offers the same high-performance querying of log and telemetry data as the Azure Synapse Data Explorer runtime in Azure Synapse Analytics.
      * Data analyst can use Azure Data Explore to query and analyze data that includes a timestamp attribute, such as is typically found in log files and internet-of-things (loT) telemetry data.
    - Microsoft Purview
      * Provides a solution for enterprise-wide data governance and discoverability. You can use Microsoft Purview to create a map for your 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 Fabric
        + A unified software-as-a service(SaaS) analytics platform based on open and governed lakehouse that includes functionality to support:
        + Data ingestion and ETL
        + Data lakehouse analytics
        + Data warehouse analytics
        + Data Science and machine learning
        + Realtime analytics
        + Data visualization
        + Data governance and management

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

#### Describe relational concepts

* Identify features of relational data
  + A model of collections of entities from the real world as tables.
  + A table contains rows and each row represents an instance of an entity
  + A format for structured data where each row in a table has the same columns.
  + Each column stores data of a specific datatype.
  + Standard datatypes defined by the American National Standards Institute (ANSI) that are supported by most database systems.
* Describe normalization and why it is used
  + A schema design process that minimizes data duplication and enforces data integrity.
    - 1- Separate each entity into its own table
    - 2- Separate each discrete attribute into its own column.
    - 3- Uniquely identify each entity instance (row) using a primary key.
    - 4- Use foreign key columns to link related entities.
* Identify common structured query
* language (SQL) statements
  + Used to communicate with relational databases.
  + SQL Statement Types
    - Data Definition Language (DDL)
      * Used to create, modify, and remove tables and other objects in a database (table, stored procedures, views, and so on)
      * Create- Create a new object in the database, such as a table or a view.
      * Alter- Modify the structure of an object. For instance, altering a table to add a new column.
      * Drop- Remove an object from the database and is permanent.
      * Rename- Rename an existing object.
      * Table Creation
        + Create Table Product

(

ID INT PRIMARY KEY,

Name VARCHAR (20) NOT NULL,

Price DECIMAL NULL

);

* + - Data Control Language (DCL)
      * Used to manage access to objects in a database by granting, denying, or revoking permissions to specific users or groups.
      * GRANT – grant permission to perform specific actions
      * Example:
        + GRANT SELECT, INSERT UPDATE
        + ON Product
        + TO user1;
      * DENY – Deny permission to perform specific actions
      * REVOKE – Remove a previously granted permission
    - Data Manipulation Language (DML)
      * Used to manipulate rows in tables. These statements enable you to retrieve query data, insert new rows, modify existing rows. You can also delete rows.
      * Select – read rows from a table
      * Insert- new rows
      * Update- modify existing data in rows
      * Delete- Delete existing rows
* Identify common database objects
  + In addition to tables, a relational database can contain other structures that help to optimize data organization, encapsulate programmatic actions, and improve the speed of access:
  + Views- a virtual table based on the results of a SELECT query.
    - You can think of a view as a window on specified rows in one or more underlying tables.
    - CREATE VIEW Deliveries
    - AS
    - SELECT o.OrderNo, o.OrderDate,
      * C.FirstName, c.LastName, c.Address, c.City
    - FROM Order AS o JOIN Customer AS c
    - ON o.Customer = c.ID;
  + Stored Procedures
    - Defines SQL statement that can be run on command.
    - Used to encapsulate programmatic logic in a database for actions that apps need to perform when working with data.
    - Can contain parameters to create a flexible solution for common actions that might need to be applied to data based on a specific key or criteria.
    - Create Procedure RenameProduct
      * @ProductId INT,
      * @NewName VARCHAR(20)
    - AS
    - UPDATE Product
    - SET Name = @NewName
    - WHERE ID = #ProductID;
    - (And then you can execute the stored procedure)
    - EXEC RenameProduct 201, ‘Spanner’;
  + Indexes
    - Helps you search for data in a table.
    - A specification of a column from a table with pointers to the corresponding rows
    - CREATE INDEX idx\_ProductName
    - ON Product(Name)

#### Describe relational Azure data services system

Microsoft Azure provides multiple services for relational databases. You can choose the relation database management system that’s best for your needs, and host relational data in the cloud.

Azure SQL is a collective term for a family of Microsoft SQL Server based database services in Azure.

* Describe the Azure SQL family of products including
  + Azure SQL Database
    - A fully managed, highly scalable (PaaS) that is designed for the cloud. This service includes the core database-level capabilities of on-premises SQL server, and is a good option when you need to create a new application in the cloud.
  + Azure SQL Managed Instance
    - A platform-as-a-service(PaaS) that provide near 100% compatibility with on-premises SQL Server instances while abstracting the underlying hardware and operating system. The service includes automated software updated management, backups, and other maintenance tasks, reducing the administrative burden of supporting a database server instance.
  + SQL Server on Azure Virtual Machines
    - A virtual machine running in Azure with an installation of SQL server. The use of a VM Makes this option an infrastructure-as-service (IaaS) solution that virtualizes hardware infrastructure for compute, storage, and networking in Azure; making it a great option for “lift and shift” migration of existing on premises SQL Server installations to the cloud.
    - Migration is easy
    - This is a (LaaS) approach
    - Can also incorporate extension from on-premise hardware into the cloud for extended features and capabilities.
    - Easily scalable without having to re-install.
  + Azure SQL Edge-
    - A SQL engine that is optimized for internet-of-things (loT) scenarios that need to work with streaming time-series data.
  + Identify Azure database services for open-source database s change systems
    - Azure data services are available for other popular relational database systems, including MYSQL, MariaDB, PostgreSQL. The primary reason for these services is to enable organizations that use them in on-premises apps to move to Azure quickly, without making significant changes, to their applications.
  + What are MYSQL, MariaDB, and PostgreSQL?
    - Relational Database management systems that that are tailored for different specializations.
    - MySQL
      * Leading open source database for Apache, MySQL, Linux, and PHP (LAMP)
      * Community, Standard, Enterprise
    - Azure Database for MySQL
      * A PaaS implementation of MySQL in the cloud.
      * Automatic Backups are provided with poin-in time restore.
      * Connection Security
        + Firewall rules
        + SSL connections
        + Server setting such as lock modes, max number of connections and timeouts.
      * A global database system capable of scaling up without the need to manage underlying components.
      * Certain operations concerning security and admin aren’t available as they are managed by Azure itself.
      * Benefits of Azure Database for MySQL
        + Built-in features
        + Predictable performance
        + Easy scaling
        + Secure data, both at rest and in-motion.
        + Automatic backup and point-in-time restore for the last 35 days.
        + Enterprise-level security and legislation compliance.
        + Pay-as-you go pricing
        + Monitoring functionality: add alerts, metrics views and logs.
    - MariaDB
      * Created by SQL for improved performance.
      * Compatibility w/ Oracle
      * Support for temporal data
    - Azure Database for MariaDB
      * An implementation of the MariaDB database management system adapted to run in Azure.
      * Fully managed and controlled by Azure require no additional admin besides the transfer of data.
      * Benefits Delivered
        + Built-in availability of features with no additional cost
        + Predictable performance
        + Pay-as you-go
        + Scaling as needed within seconds
        + Secured protection of sensitive data at rest and in-motion
        + Automatic backup and point-in-time restore for the last 35 days.
        + Enterprise-level security and legislation compliance.
    - PostgreSQL
      * Hybrid (Relational with non-relational capabilities)
      * You can add code modules which can be run with queries.
      * Can store and manipulate geometric data (lines, circles, and polygons)
      * Postgre has its own querying language called pgsql, an sql variant and has features that can run stored procedures.
    - Azure Database for PostgreSQL
      * Runs on A PaaS implementation of PostgreSQL
      * Provides a lot of the same features as MySQL services.
      * Some feature such as writing stored procedure in different languages and interacting directly with the operating system are not available in Azure Database for PostgreSQL
      * There is a core list of the most frequently used extensions that is supported.
    - Azure Database for PostgreSQL Flexible Server
      * A fully-managed database service
      * High level of control and server configuration customizations with cost optimization controls.
    - Benefits of Azure Database for PostgreSQL
      * Azure Database for PostgreSQL is a highly available service with built-in failure detection and failover mechanisms.
      * Features the PgAdmin tool, which can be used to manage and monitor a PostgreSQL database and to connect to Azure database for PostgreSQL.
      * However, some server-focused functionality, such as server backup and restore aren’t available because they are fully managed by Microsoft.
      * Records information about queries run against databases on the server and saves them in a database named azure\_sys. You can query the query\_store.qs\_view view to see this information and use it to monitor and fine-tune the queries that the users are running.

### Describe considerations for working with non-relational data on Azure (15–20%)

#### Describe capabilities of Azure storage

* Describe Azure Blob storage
  + **A service that enables you to store massive amounts of unstructured data as large binary objects, or blobs, in the clouds.**
  + **Apps can read and write them by using the Azure blob storage API**
  + **Blobs are stored in containers, which can be organized by a hierarchy of files similar to files in a file system on disk.**
  + **These files are purely virtual**
  + **Supports three different types of blob:**
    - **Block blobs**
      * Smallest unit of data that can be read or written
      * Each block can vary in size, up to 4000 MiB
      * Best used to store discreet, large, binary objects that change infrequently.
    - **Page blobs**
      * Organized as a collection of fixed size 512-byte pages.
      * Optimized to support random read and write operations
      * Single page fetch and store is also possible
      * A page blob can hold up to 8TB
      * Used to implement virtual disk storage for virtual machines
    - **Append blobs**
      * A type of block blob used for append operations
      * You can only add blocks to the end of an append blob while updating and deleting existing blocks is not supported
      * Each block can vary in size, up to 4mb
      * The max size for and append blob is just over 195 GB
  + **Blob Storage provides three access tiers:**
    - And only after the rehydration process is complete the Hot Tier
      * The default for frequently accessed.
      * Stored on high-performance media
    - The Cool Tier
      * Lower performance requirements and less storage charges than the hot tier
      * Migration from hot to cool or cool to hot is based on the frequency of use over time.
    - The Archive tier provides the lowest storage cost, but with increased latency.
    - Intended for historical data that mustn’t be lost but is required only rarely
    - Stored in an offline state
    - Typical reading latency for hot and cool is within milliseconds while it may take hours to retrieve data for the archive tier
    - In order to read archived data, it must be rehydrated and transferred to hot or cool and only after the rehydration process is complete.
* Explore Azure DataLake Storage Gen2
  + A separate service for hierarchical data storage for analytical data lakes.
  + Used by big data analytical solution that work with structured, semi-structured and unstructured data stored in files.
  + Gen2 is a newer version of this service integrated into Azure Storage.
  + Benefits include:
    - Scalability of Blob Storage
    - Cost control of storage tiers
    - Hierarchical fil system capabilities
    - Compatibility with major analytics systems of Azure Data Lake Store
* Describe Azure File storage
  + Azure file shares is essential a way to create cloud-based networked file shares, to make document and other files available to multiple users.
  + Benefits to hosting files shares in Azure include:
    - Elimination of hardware costs and maintenance overhead.
    - High availability and scalable cloud storage for files.
  + Azure File Storage in stored in a storage acct. with up to 100 TB of storage with a max. file size of 1 TB
  + Up to 2000 concurrent connections per shared file.
  + Ways to upload files into Azure File Storage:
    - The Azure Portal
    - The AzCopy utility
    - The Azure File Sync service
  + Two performance tiers:
    - Standard- used hard-disk-based hardware
    - Premium- uses solid state with higher throughput but at a higher rate.
  + Azure Files supports two common network file sharing protocols:
    - Server Message Block (SMB) file sharing is commonly used across multiple operating systems (Windows, Linux, macOS)
    - Network File System (NFS) shares are used my Linux and macOS versions.
      * Only works on premium tier storage with a virtual network through which access to the share can be controlled.
* Describe Azure Table storage
  + Is a NoSql data storage solution in tabular form where rows contain columns in key/value form.
  + The rows and columns in Azure Table are non-relational in nature which differs from the tables of rows and columns that you would find in relational databases.
  + The key in Azure Table storage is made up of a partition key and a row key.
  + There is a timestamp that accompanies the row every time the row is modified.
  + Allows for unstructured data meaning that data in each column can vary by row.
  + Unlike the tables in relational databases, non-relational Azure Table Storage have no concept of foreign keys, relationships, stored procedures, views, etc.
  + Denormalized
  + Organized and optimized and scalable by a process of partitioning where there is an unlimited amt. of partitions allowed per table.
  + Querying the data by partition number is also optimized by speed and storage.

#### Describe capabilities and features of Azure Cosmos DB

* Identify use cases for Azure Cosmos DB
  + For Developer creating applications that need access to data through the use of APIs.
  + Lot and telematics
    - Ingest large amounts of data w/ infrequent burst of activity
    - Used by analytics services such as Azure Machine Learning, Azure HDInsight, and PowerBI
    - You can process the data in real-time using Azure functions that are triggered as data arrives in the database.
  + Retail and Marketing
    - Used as an e-commerce platform that run as a part of Windows store and XBOX Live.
    - Also used in the retail industry for storing catalog data and for order processing and event sourcing.
  + Gaming
    - These services make gaming lit with single-millisecond latencies for reads and write to provide an engaging in-game experience.
  + Web and mobile applications
    - Modeling social interactions
    - Integrating third-party services
    - Building rich and personalized experiences
    - iOS and Android apps using Xamarin framework
* Describe Azure Cosmos DB APIs
  + Fully managed and serverless distributed database for applications of any size or scale
  + Relational or non-relational workloads

### Describe an analytics workload on Azure (25–30%)

#### Describe common elements of large-scale analytics

* Describe considerations for data ingestion and processing
* Describe options for analytical data stores
* Describe Azure services for data warehousing, including Azure Synapse Analytics, Azure Databricks, Azure HDInsight, and Azure Data Factory

#### Describe consideration for real-time data analytics

* Describe the difference between batch and streaming data
* Describe technologies for real-time analytics including Azure Stream Analytics, Azure Synapse Data Explorer, and Spark Structured Streaming

#### Describe data visualization in Microsoft Power BI

* Identify capabilities of Power BI
* Describe features of data models in Power BI
* Identify appropriate visualizations for data

## Study resources

We recommend that you train and get hands-on experience before you take the exam. We offer self-study options and classroom training as well as links to documentation, community sites, and videos.

| **Study resources** | **Links to learning and documentation** |
| --- | --- |
| Get trained | [Choose from self-paced learning paths and modules or take an instructor-led course](https://learn.microsoft.com/en-us/certifications/exams/dp-900#two-ways-to-prepare) |
| Find documentation | [Azure SQL documentation - Azure SQL](https://learn.microsoft.com/en-us/azure/azure-sql/?view=azuresql) [SQL Server technical documentation - SQL Server](https://learn.microsoft.com/en-us/sql/sql-server/?view=sql-server-ver16) [Azure Blob Storage documentation](https://learn.microsoft.com/en-us/azure/storage/blobs/) [Azure Table storage documentation](https://learn.microsoft.com/en-us/azure/storage/tables/) [Azure Storage documentation](https://learn.microsoft.com/en-us/azure/storage/) [Azure Cosmos DB](https://learn.microsoft.com/en-us/azure/cosmos-db/) [Azure Synapse Analytics](https://learn.microsoft.com/en-us/azure/synapse-analytics/) [Azure Databricks](https://learn.microsoft.com/en-us/azure/databricks/) [Data Factory](https://learn.microsoft.com/en-us/azure/data-factory/) [Power BI documentation - Power BI](https://learn.microsoft.com/en-us/power-bi/) |
| Ask a question | [Microsoft Q&A | Microsoft Docs](https://learn.microsoft.com/en-us/answers/products/) |
| Get community support | [Analytics on Azure | TechCommunity](https://techcommunity.microsoft.com/t5/analytics-on-azure/bd-p/AnalyticsonAzureDiscussion) [Azure Synapse Analytics | TechCommunity](https://techcommunity.microsoft.com/t5/azure-synapse-analytics/bd-p/AzureSynapseAnalytics) [Welcome to the SQL Server Community (microsoft.com)](https://techcommunity.microsoft.com/t5/sql-server/ct-p/SQL-Server) [Azure PaaS - Microsoft Tech Community](https://techcommunity.microsoft.com/t5/azure-paas/bd-p/AzurePaaS) |
| Follow Microsoft Learn | [Microsoft Learn - Microsoft Tech Community](https://techcommunity.microsoft.com/t5/microsoft-learn/ct-p/MicrosoftLearn) |
| Find a video | [Exam Readiness Zone](https://learn.microsoft.com/en-us/shows/exam-readiness-zone/) [Data Exposed](https://learn.microsoft.com/en-us/shows/data-exposed/) [Browse other Microsoft Learn shows](https://learn.microsoft.com/en-us/shows/browse) |

## Change log

Key to understanding the table: The topic groups (also known as functional groups) are in bold typeface followed by the objectives within each group. The table is a comparison between the two versions of the exam skills measured and the third column describes the extent of the changes.

| **Skill area prior to November 2, 2023** | **Skill area as of November 2, 2023** | **Change** |
| --- | --- | --- |
| Audience profile |  | No change |
| **Describe core data concepts** | **Describe core data concepts** | No change |
| Describe ways to represent data | Describe ways to represent data | No change |
| Identify options for data storage | Identify options for data storage | No change |
| Describe common data workloads | Describe common data workloads | No change |
| Identify roles and responsibilities for data workloads | Identify roles and responsibilities for data workloads | No change |
| **Identify considerations for relational data on Azure** | **Identify considerations for relational data on Azure** | No change |
| Describe relational concepts | Describe relational concepts | No change |
| Describe relational Azure data services | Describe relational Azure data services | Minor |
| **Describe considerations for working with non-relational data on Azure** | **Describe considerations for working with non-relational data on Azure** | No change |
| Describe capabilities of Azure storage | Describe capabilities of Azure storage | No change |
| Describe capabilities and features of Azure Cosmos DB | Describe capabilities and features of Azure Cosmos DB | No change |
| **Describe an analytics workload on Azure** | **Describe an analytics workload on Azure** | No change |
| Describe common elements of large-scale analytics | Describe common elements of large-scale analytics | No change |
| Describe consideration for real-time data analytics | Describe consideration for real-time data analytics | No change |
| Describe data visualization in Microsoft Power BI | Describe data visualization in Microsoft Power BI | No change |

## Skills measured prior to November 2, 2023

### Audience profile

This exam is an opportunity to demonstrate knowledge of core data concepts and related Microsoft Azure data services. Candidates for this exam should have familiarity with DP-900’s self-paced or instructor-led learning material.

This exam is intended for candidates beginning to work with data in the cloud.

Candidates should be familiar with the concepts of relational and non-relational data, and different types of data workloads such as transactional or analytical.

Azure Data Fundamentals can be used to prepare for other Azure role-based certifications like Azure Database Administrator Associate or Azure Data Engineer Associate, but it is not a prerequisite for any of them.

### Skills at a glance

* Describe core data concepts (25–30%)
* Identify considerations for relational data on Azure (20–25%)
* Describe considerations for working with non-relational data on Azure (15–20%)
* Describe an analytics workload on Azure (25–30%)

### Describe core data concepts (25–30%)

#### Describe ways to represent data

* Describe features of structured data
* Describe features of semi-structured
* Describe features of unstructured data

#### Identify options for data storage

* Describe common formats for data files
* Describe types of databases

#### Describe common data workloads

* Describe features of transactional workloads
* Describe features of analytical workloads

#### Identify roles and responsibilities for data workloads

* Describe responsibilities for database administrators
* Describe responsibilities for data engineers
* Describe responsibilities for data analysts

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

#### Describe relational concepts

* Identify features of relational data
* Describe normalization and why it is used
* Identify common structured query language (SQL) statements
* Identify common database objects

#### 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
* Identify Azure database services for open-source database systems

### Describe considerations for working with non-relational data on Azure (15–20%)

#### Describe capabilities of Azure storage

* Describe Azure Blob storage
* Describe Azure File storage
* Describe Azure Table storage

#### Describe capabilities and features of Azure Cosmos DB

* Identify use cases for Azure Cosmos DB
* Describe Azure Cosmos DB APIs

### Describe an analytics workload on Azure (25–30%)

#### Describe common elements of large-scale analytics

* Describe considerations for data ingestion and processing
* Describe options for analytical data stores
* Describe Azure services for data warehousing, including Azure Synapse Analytics, Azure Databricks, Azure HDInsight, and Azure Data Factory

#### Describe consideration for real-time data analytics

* Describe the difference between batch and streaming data
* Describe technologies for real-time analytics including Azure Stream Analytics, Azure Synapse Data Explorer, and Spark Structured Streaming

#### Describe data visualization in Microsoft Power BI

* Identify capabilities of Power BI
* Describe features of data models in Power BI
* Identify appropriate visualizations for data