DP-900: Microsoft Azure Data Fundamentals

Core Concepts

Data

- Data is a collection of facts such as numbers, descriptions, and observations used in decision making
  - Structured
    - Usually stored in a relational database
    - Tabular data that is represented by rows and columns
  - Semi-structured
    - Information that doesn't reside in a relational database but still has some structure to it
    - Usually stored in a Binary Large Object (Blob) storage
    - JSON
    - XML
    - HTML
    - Text
  - Unstructured
    - For example, audio, video files, binary data files...
  - Access levels
    - Read only
    - Read/write
    - Owner
- Data processing
  - Data processing is simply the conversion of raw data to meaningful information through a process.
  - Transactional System
    - Records transactions
    - File work is often referred to as Online Transactional Processing (OLTP)
    - Operations
    - Transactions
    - Each transaction is treated as a single unit.
    - A transaction ensures that the data from you called state is pipeline
    - Consistency
    - Isolation
    - Durability
    - Ensures that processing execution of transactions leaves the database in the same state it the transactions were executed sequentially
    - Once a transaction has been committed, it will remain committed
    - A transaction is a sequence of operations that are atomic
    - Most adhere to the ACID properties
  - Analytical System
    - Query data and gain a big picture view
    - The work is often referred to as Online Analytical Processing (OLAP)
    - Typically read-only systems that store vast volumes of historical data
    - Capture raw data generate insights
    - Types
      - Capturing the raw data
      - Data Ingestion
      - Digital Transformation/Modernization
      - Querying for results
      - Data Queries
      - Operate on the results
      - Actions
      - Apps that already exists
- Batch data
  - Handles data in real time
  - Always being generated
  - Advantages
    - The work is often referred to as Online Analytical Processing (OLAP)
    - Typically read-only systems that store vast volumes of historical data
    - Capture raw data generate insights
  - Disadvantages
    - Nearly arriving data elements are collected into a group
    - A batch process is then processed at a future time as a batch
    - Large volumes of data can be processed at a convenient time
    - It can be scheduled
    - No time delay between capturing the data and acting on it
    - Complex to implement and maintain
    - A batch job is time-ly to bring it into Load the processing
    - Blob files
    - Text files
    - Apps that already exists
    - Web files
- Streaming data
  - Handles data in real time
  - Always being generated
  - Sources
    - A data together service that you can use to manage system data, ensuring that each event is processed in exactly-once
    - Azure Event Hubs
    - A data ingestion service that is similar to Azure Event Hubs, but optimized specifically to process event data from Internet-of-things (IoT) devices
    - Azure IoT Hub
    - A highly scalable storage service that is often used in batch processing scenarios, but which can also be used as a source of streaming data
    - Azure Data Lake Store Gen 2
    - An open source data ingestion solution that is commonly used together with Apache Spark. You can use Azure HDInsight to create a Kafka cluster
    - Apache Kafka
    - Used to queue the processed data for further downstream processing
    - Azure Event Hubs
    - Used to persist the processed results as a file
    - Azure Data Lake Store Gen 2 or Azure Blob storage
    - Used to extract the processed results in a relational table for querying and analysis
    - Azure Synapse Analytics, or Azure Database for MySQL, Azure Database for PostgreSQL, or Azure SQL Database
    - Used to generate real-time data visualizations in reports and dashboards
    - Power BI
- Data workloads
  - Batch processing
    - Batch processing is the process of the data in the storage
    - Data Storage
    - Dataset processing typically only has access to the entire source data-storage
    - Data Ingestion
    - Batch processing is suitable for handling large datasets efficiently
    - Dataset processing is intersected for information/insights
    - Data Processing
    - Batch processing typically involves latency in its data feeds
    - Data Analytics
    - Batch processing typically involves complex algorithms
    - Data Visualization
    - The tools use batch processing for performing complex analytics
    - Reporting
    - Batch processing is used for whole reports functions
  - Difference between the two
  - Stream processing
    - Stream processing typically has access to the most current data-stream
    - Stream processing is suitable for handling large datasets efficiently
    - Stream processing is intersected for information/insights
    - Stream processing typically involves latency in its data feeds
    - Data Ingestion
    - Stream processing typically involves complex algorithms
    - Data Processing
    - The tools use batch processing for performing complex analytics
    - Data Analytics
    - Stream processing is used for whole reports functions
    - Data Visualization
    - Reporting
- Relational data
  - Characteristics
    - Tables - rows and columns
    - Tables in a query that uses the "JOIN" to combine
    - Primary Key (PK) - unique identifier for a row
    - Relationships using foreign keys
    - Multiple tables make a single table
  - Structure
    - The structure of a database
    - Entities/Tables
    - Attributes/Columns
    - Table names
    - Column names
    - Data types
    - Keys

Roles

- Database Administrators
  - Managing permissions
  - Performing backups
  - Applying data cleaning routines
  - Identifying business rules
- Data Engineers
  - Identifying business rules
  - Turning data into useful information
- Data Analysts
  - Explore and analyze data to create visualizations and charts

Chart

- Easy to interpret answering data...
  - Bar
  - Line
  - Pie
- Data visualization
  - Bar
  - Line
  - Pie

Data processing

- Data analytics is concerned with examining, transforming, and arranging data so that you can study it and extract useful information
- Analytics techniques
  - Descriptive
    - What happened?
    - Summarizes large datasets to describe outcomes to stakeholders
  - Diagnostic
    - Why did it happen?
    - Deep dive into existing data
    - Take the findings from descriptive analytics and dig deeper to find the cause
  - Predictive
    - What will likely happen?
    - Based on past trends
    - Uses historical data to identify trends and determine if they're likely to recur
  - Prescriptive
    - What should I do?
    - Something happened, you should act this...
    - Relies on machine learning algorithms to find patterns in large datasets
  - Cognitive
    - Artificial intelligence and machine learning
    - Attempt to learn what happens if circumstances change, and how you might handle these situations
    - Ask Google Maps it traffic path is different today...
    - My Self-driving car's
    - Analyzing data to come up with a model
- Stages
  - Data Ingestion
    - Collecting data
    - Possible to perform some transformations
    - Ingestion
    - Data Processing
    - At the process of obtaining and integrating data for analysis, ingest it or persist it in a location
    - Convert the raw data into one or more business-ready
    - Stores
    - ELT & ETL
    - E: Extraction
    - L: Transformation
    - L: Loading
    - Raw data is stored before being transformed
    - Extract data
    - Save into your database
    - Transform the data for your needs
    - ETL
    - Raw data is transformed before being stored
    - Extract data
    - Transformed the data for your needs
    - Save into your database
    - Data Processing
    - Data analytics is the process of examining raw data to uncover trends, and discover information used to ask and answer questions
    - It is a data analytics solution; you combine data and construct a data warehouse that you can use to ask/answer questions about your business operations.
- Types
  - Online Transaction Processing (OLTP)
    - Use the backend database to your business systems
    - Optimized for general use
    - Existing rows can be updated
    - Data can be changed by SQL statements
    - Normalized
    - Database normalization
    - Schema
    - Consistency
    - Heavy writes, medium reads
    - Updatable
    - Size from MBs to TBs
    - Azure SQL Database
    - SQL Server in a VM
    - Azure Database for MySQL
    - Azure Database for PostgreSQL
    - Online Analytics Processing (OLAP)
    - Reviews (rechecking said day-to-this business)
    - No backup of a data
    - Denormalized
    - Read only
    - Multi-dimensional indexing
    - Data 10 / 1000
    - SQL Server with Columnstore Indexes
    - Azure Analysis Services
    - Azure optimized
    - SQL Server Analysis Service

Data Analytics Workloads

- The term big data refers to data that is too large or complex for traditional database systems
- Data warehouse
  - Contains a mixture of relational and non-relational data, including files, social media streams, and Internet of Things (IoT) sensor data
  - Holds structured information
  - Characteristics
    - All data comes from somewhere else
    - Read or more data sources
    - Can be structured or semi-structured
    - When queries affect day-to-day operations
    - When data needs further processing (ELT or ETL)
    - Offline to use it
    - When you want to remove historical data from the day-to-day system (trim down)
    - When you need to integrate data from different sources
    - Central repository of data from different sources
    - Current and historical data
    - Can rename/normal columns
    - Does not affect day-to-day business
  - Types
    - Azure SQL Database
    - SQL Server in a VM
    - Azure Synapse Analytics
    - Azure Hive or HDInsight
    - Massively Parallel Processing (MPP)
    - Interactive Query (Hive LLAP) on HDInsight
    - Symmetric Multi Processing (SMP)
    - Massively Parallel Processing (MPP)
- Azure Data Factory
  - To retrieve data from one or more data sources, and convert it into a format that you process
  - Responsibilities
    - Maintained by a Data Engineer
    - Scheduling
    - Monitoring
    - Alerting
    - Data security
    - Performance
    - External data connections
    - To bring data from external resources...
    - Data Orchestration
    - Activities
    - Online activities to perform on the data
    - Pipelines
    - A logical grouping of activities to perform some task
    - Triggers
    - Pipelines can be run manually or using a trigger
    - Connections
    - Pipelines information needed for Data Factory to connect to source or destination
    - Linked services
    - Datasets
    - Represents the data that you want to ingest (input), or store (output)
    - Integration Runtime
    - A feature of SQL Server and Azure Synapse Analytics that enables you to run Transact-SQL queries that read data from external data sources.
    - Polybase
    - A platform for building enterprise-level data integration and data transformations solutions.
    - SQL Server Integration Services (SSIS)
    - Can process data held in many different types of storage, including streaming data.
    - Azure Databricks
    - Azure Data Lake Storage
    - Combines the hierarchical directory structure and file system semantics of a traditional file system
    - Characteristics
    - Organizes your flat-key descriptors and subdirectories
    - Supports the Portable Operating System Interface, X (POSIX)
    - Supports Role-Based Access Control (RBAC) on your data
    - Compatible with the Hadoop Distributed File System (HDFS)
- Azure Databricks
  - Apache Spark is a highly efficient data processing engine that can consume and process large amounts of data very quickly
  - Notebook
  - A notebook contains a series of steps (cells), each of which contains a block of code.
  - Cluster
  - Clusters are the computer resources that run your notebooks.
  - Notebook
  - Notebooks can be run manually or using a trigger
  - Azure Synapse Analytics
  - Pipelines information needed for Data Factory to connect to source or destination
  - Linked services
  - Datasets
  - Represents the data that you want to ingest (input), or store (output)
  - Integration Runtime
  - Triggers
  - Pipelines
  - Activities
  - Pipelines
  - Connections
  - Activities
  - Triggers
  - Pipelines
  - Datasets
  - Linked services
  - Integration Runtime
  - Azure HDInsight
  - Reporting
  - Techniques
  - Business Intelligence (BI)
  - Data Visualization
  - Interactive representation of information and data
  - It's a collection of software services, apps, and connections that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights
  - Power BI
  - A visual representation of data
  - Visualizations
  - A collection of data that Power BI uses to create its visualizations
  - Datasets
  - A collection of visualizations that appear together on one or more pages
  - Reports
  - A collection of visuals from a single page that you can share with others
  - Dashboards
  - A single visualization on a dashboard. It's the rectangular box that holds an individual visual
  - Tiles
  - An app is a collection of ready-made visuals, pre-arranged in dashboards and reports. You can get apps that connect to many online services from the AppSource.
  - Paginated Reports
  - Designed to be printed
  - Can span multiple pages.
  - Created with the Powerful Features with the Report Builder.
  - Interactive Reports
  - Designed to be viewed on screen
  - Drill-down
  - Report can change based on user actions
  - Dashboards
  - Mix of chart types on a single page
  - Fit the relevant data on a glance
  - Power BI Content Flow
  - 1. Connect to the data source
  - 2. Pull the data into the in-memory data model
  - 3. Edit, transform data
  - 4. Build reports using Power BI Desktop
  - 5. Share reports

Data Ingestion and Processing

Data Visualization

Relation Database Concepts

- Characteristics
  - An entity is described as a thing about which information needs to be known or held
  - Columns define the properties of the entity
  - Rows are the values of those properties of the entity
  - Primary key (PK) indicates the column that uniquely identify each row
  - Foreign key (FK) reference to primary key of another table
- Benefits
  - Normalization reduces duplication
  - Data integrity
  - Enforced constraints, fixed schema
  - Relationships show data insights
  - Stable tested
- Services
  - SQL Server in a VM
    - Enables you to use full versions of SQL Server in the Cloud without having to manage any on-premises hardware
    - SQL virtual machines are lift-and-shift ready for existing applications that require find migration to the cloud with minimal changes
    - Infrastructure as a Service
    - Guaranteed compatibility to SQL server on Premise
    - Flex manage everything
      - OS upgrades
      - Software upgrades
      - Tuning
    - No limitation
    - Supports data above 4TB
    - Ideal for server lift and shift
  - Azure SQL Database
    - Provide a managed database service in the Cloud, and then deploy your database on the platform
    - Business benefits
      - Automatically updates and patches the SQL Server software
      - Automatic updates
      - Automatic recovery
      - Advanced threat protection
      - Auditing tracks database events
      - Encryption
    - Elastic Pool
      - Provide and run a database server in the cloud, and put custom your database through his server
    - Managed Instance
      - Scalability
      - Compatibility
      - Cross-platform data
    - Deployment
      - Provisioning resources take seconds, which might be required for a number of database benefits in future those applications
      - Ability to scale up or down to meet demand
      - Pay-as-you-go billing model
      - Exclude individual resources in the pool to deploy
    - Single Database
    - Elastic Pool
    - Managed Instance
  - SQL Managed Instance
    - If you want to lift and shift an on-premises SQL Server instance and all its databases to the cloud, Azure SQL Managed Instance is the best option for running SQL Server on a virtual machine
    - Eliminates as much management overhead as possible from administering databases and servers
    - Automated backups
    - Complete control, much as you would for an on-premises server
    - Depend on other Azure services
    - Azure Storage for backups
    - Azure Event Hubs for telemetry
    - Azure Active Directory for authentication
    - Azure Key Vault for Transparent Data Encryption
  - Azure s Databases
    - The Database Migration Service enables you to migrate your SQL Server databases to Azure with minimal downtime to a database running on Azure Data Services
    - MySQL
    - It is the leading open source relational database for Linux, Apache, MySQL, and PHP (LAMP) stack apps
    - MariaDB
    - Server version of MySQL
    - PostgreSQL
    - Object-relational-object database
- Structure
  - Tables
    - Index you search for data in a table
    - Delete one type of data
    - Clustered
    - Non-clustered
    - Unique
    - Views & columns
    - Stored in RAM
    - Query table should have a Primary Key (PK)
    - Relate one have a Foreign key
    - Refers to the primary key of another table
    - Foreign key
    - Used to improve the performance of queries
    - The Primary Key is by default an index
    - Indexes
    - A unique index based on the result set of a query
  - Views
    - Azure enables you to create a virtual infrastructure in the cloud that mirrors the way an on-premises data center might work
    - Infrastructure as a Service (IaaS)
    - Flexibility
    - Control
    - Virtual clouds
    - Cost
    - Storage in Cloud
    - Virtual machines of Cloud
    - SQL Server as a VM
    - SQL Server options
    - Web
    - Standard
    - Enterprise
    - Developer
    - Express
    - Platform as a Service (PaaS)
    - Rather than creating a virtual infrastructure, and installing and managing the database software yourself, a PaaS (Platform-as a Service) solution does this for you
    - Azure Data Services
    - These services are a series of DBMSs managed by Microsoft in the cloud
    - Flexibility
    - Can limit the range of custom administration tasks you can perform
    - Cost
    - SLA available at least 99.99%
    - Security
    - Always on (not able to restart)
    - Azure SQL Server engine underneath
    - Scalability
    - Designed to work in the cloud
    - Azure SQL Database
    - Scalability
    - Automatic maintenance in cloud
    - Flexibility
    - Has options to...
    - Tuning
    - Flex types
    - Single Database
    - Isolate resource to a specific database
    - Elastic Pool
    - Choose resources to a pool of databases to share them
    - Azure Synapse Analytics
    - Big data
    - Designed for reporting and data analytics
    - Azure Database
    - PostgreSQL, MySQL, and MariaDB
    - Allows you to migrate an existing solution to the cloud
    - SaaS
    - Flash your apps
    - Flash your apps
    - The act of purchasing series of tools that a service provider, such as Azure SQL Database, performs to create and configure a service
    - Provisioning
    - Azure Portal
    - Azure command line interface (CLI)
    - The CLI provides a set of commands that you can run from the operating system command prompt or the Cloud Shell in the Azure portal
    - Azure PowerShell
    - Azure Resource Manager templates
    - A JSON file the ARM you interpret
    - Connectivity
    - Firewalls and virtual networks
    - Configure authentication with Azure AD
    - Access control
    - Role
    - Security principal
    - An object that represents a user, group, service principal, or managed identity that is requesting access to Azure resources
    - Scope
    - Full access to all resources including the right to delegate access to others
    - Resource group
    - Applies to a group of Azure resources but can't grant access to others
    - Resource
    - Applies to a single Azure resource
    - Role definition
    - Collection of permissions
    - Role assignment
    - Ties the set of resources that the access applies to
    - Configuring
    - Query tools
    - Query Editor (Portal)
    - Azure editor into the portal
    - Execute queries
    - Provides a graphical user interface for managing many different database systems
    - Cross-platform software to work with SQL Server database data
    - Charting
    - Azure Data Studio
    - Notebook
    - Write analysis
    - Source control
    - Integrated terminal
    - Benefits
    - Extensible
    - Free
    - SQL Server Management Studio (SSMS)
    - Provides a graphical interface, allowing you to query data, perform general database administration tasks, and generate scripts for automating database maintenance and support operations
    - Supports database admin tasks
    - Sqlcmd
    - Execute T-SQL statements, stored procedures, script files
    - SQL command line utility (sqlcmd)
    - Not full or limited command
    - Structured Query Language (SQL)
    - Data Definition Language (DDL)
    - Manipulation of the database structure (schema)
    - Create
    - Alter
    - Drop
    - Truncate
    - Comment
    - Rename
    - Data Manipulation Language (DML)
    - Manipulation of data
    - Select
    - Insert
    - Update
    - Delete
    - Lock table
    - Data Control Language (DCL)
    - Granting rights and permissions to authority
    - Grant
    - Revoke
    - Transaction
    - Commit the SQL statements
    - Rollback
    - Rollback versions over different tables

Manage Relation Databases

Non-relational Database Concepts

- Characteristics
  - Flexible to store data in a very flexible manner
  - Focussed on the data itself rather than how to structure it
  - Information for entities is stored in collections or containers
  - Each entry should have a unique key value
- Benefits
  - Optimized for specific data types
  - High performance
  - Not a lot of CPU power needed
  - Open source
  - Supports schema schemas
- Types
  - Also known as Not only SQL (NoSQL)
  - Usually in JSON format
    - JSON
    - ORC
    - Parquet
  - Optimized for different uses
  - Key/Value
    - Key/Value store has a unique id
    - Contains the entire data for an entry
    - Easy to the documents are transparent to the database management system
    - Does not require that all documents have the same structure
  - Document
    - Organizes data into cases and columns
    - Appears, conceptually, very similar to a relational database
    - Denormalized
    - Column-Family data
    - Table Storage
    - Key-Value Data
    - Stores entries under shared relationships
    - Graph Data
    - Nodes
    - Information about objects
    - Edges
    - Information about the relationships between objects
    - Time Series Data
    - Stores data under shared timestamps
    - Object Data / Blob storage
- Data/Types
  - Microsoft Azure
    - Cosmos DB
    - It is a multi-model NoSQL database management system
    - Meets the concept of Request Units per second (RU/s) to manage the performance and cost of resources
    - The amount of computation and I/O resources required to satisfy a single read request
    - Table
    - Gremlin
    - Mongo
    - Cassandra
    - Azure Cosmos DB
    - Cosmos account
    - Contains many data models (graph, document, key-value)
    - Database
    - Manageable with established APIs (Cassandra, MongoDB...)
    - Container
    - Size 10 GB limit
    - Cosmos container
    - Contains schemas
    - Contains schemas
    - Stores partitions that share a common partition key
    - Azure Table Storage
    - Stores key/value pairs
    - Table Storage
    - Cosmos DB Table Storage
    - Provides a table store in Azure Cosmos DB
    - Azure Cosmos DB
    - Cosmos DB Table API
    - Azure Table Storage
    - Table API
    - Cosmos DB Table API
    - Cosmos DB Table API
    - Cosmos DB Table API
    - Cosmos DB Table API
    - Azure Data Factory
    - Ability to import Azure Table Storage data into Cosmos DB
    - Cosmos DB BulkExecutor library
    - Cosmos DB Table API
    - Cosmos DB Table API
    - Cosmos DB Table API
    - Cosmos DB Table API
    - Azure
    - The SQL API is preferred for SQL-like queries against Cosmos DB databases
    - Cassandra
    - Apache Kernel Machine (NoSQL) series
    - Provides controlled, time-limited, anonymous access to sources and relational or a storage account
    - Blob
    - Azure Blob Storage
    - Blob Storage
    - Blob
    - Blob Storage
    - Blob
    - Blob Storage
    - Blob