Microsoft Azure Data Fundamentals (DP-900) Master Cheat Sheet

Here is the summary notes in accordance with the course content and Modules on Microsoft learn website.

***Explore Core Data Concepts:**

-Data can be categorized into 3 types.

-Structured data — ex : tabular data in rows and columns

-semi-structured data — ex : JSON, XML

-Unstructured data — ex: audio, video files, images etc.,

-Data is processed either in Batches or as in when the data arrives in real-time.

-Batch processing collects the data and then process it, on other hand streaming data processes it as in when data arrives.

-Batch processing data examples- CSV files, one months of sales date

-Stream processing data examples- online gaming data, data from sensors.

-Data processing solutions are of two broad categories: analytical systems, and transaction processing systems.

#Transactional system: (OLTP)

Records transactions. Transactions are a small, discrete unit of work in real time.

Ex: bank transactions. These systems are high-volume and handle millions of transactions in a single day.

#Analytical systems:(OLAP)

-concerned with capturing raw data, and using it to generate

insights.
-data ingestion- process of collecting raw data from different sources
-data transformation- convert the raw data into a meaningful data
-data querying- To analyze the data and run ad-hoc queries
-data visualization- generate reports, charts, plots for further data examination.

***Explore Relational DB in Azure:**

1.Azure SQL Database:

single Database- dedicated resources and managed by azure. Scale manually.
Elastic pool — similar to Single Database, except that by default multiple databases can share the same resources
Managed Instance — fully controllable instance of SQL Server in the cloud, can install multiple databases on the same instance.

***Explore Non -relational DB in Azure:**

Below services are created on Azure using an azure storage account.

1. Azure table storage:

-NoSQL key value model.
-Stores semi-structured data
-All rows in table must have a key and the columns vary from row to -row(remember it is not a relational DB)
-No concept of relationships, Stored Procedures, indexes or FK's (so data is de-normalized)
-for faster access, azure table storage splits its data into partitions.
-partitions helps group rows based on its keys and a table can have as many partitions as possible.

-So, a key in azure table storage has 2 parts to it. 1. partition key and 2. the row key that identifies the row uniquely.
-a table can have up to max of 252 columns
-max row size is 1MB
#advantages: simple to scale
-row insertion and retrieval is faster
-supports huge volumes of data. As you add rows to a table,
-Azure Table Storage automatically manages the partitions in a table and -allocates storage as necessary. You don't need to take any additional steps yourself.
-high availability
#disadvantages: There's no referential integrity
-Consistency needs to be given consideration as transactional updates across multiple entities aren't guaranteed
-Azure Table Storage is an excellent mechanism for:
-Examples include product catalogs for eCommerce applications, and customer information, where the data can be quickly identified and ordered by a composite key
-Capturing event logging and performance monitoring data
-if you need to analyze an ordered series of events and performance measures chronologically.

2. Azure Blob storage:
-allows to store unstructured data.
-for storing large binary object like images, audio files etc.
#Three types of Blob:
block blob: data is stored in blocks. The block is the smallest amount of data that can be read or written as an individual unit.
Block blobs are best used to store discrete, large, binary objects that

change infrequently.
page blob: A page blob is optimized to support random read and write operations; you can fetch and store data for a single page if necessary.
Append Blob: You can only add blocks to the end of an append blob; updating or deleting existing blocks isn't supported.
An append blob is a block blob optimized to support append operations.
-Blob storage has 3 access tiers:
hot — default one for frequently used data.
cool — infrequently used data
archive- historic data
-Azure Blob Storage is an excellent mechanism for:
-streaming video and audio,
-Storing data for backup and restore, disaster recovery, and archiving
-Storing data for analysis by an on-premises or Azure-hosted service
-Other features available with Azure Blob storage include:
#Versioning-You can maintain and restore earlier versions of a blob.
#Soft delete- This feature enables you to recover a blob that has been removed or overwritten, by accident or otherwise.
you can also Store, audit, and analyze changes to your objects, over any period of time
#Snapshots- A snapshot is a read-only version of a blob at a particular point in time.

3. Azure File Storage:
-allows you to share files with both apps running on cloud and also on-premise.

-you can upload files from portal or AzCopyUtility. To use Copy utility you need to generate a SAS (Shared access signature) token using portal.
-Azure File Sync service to synchronize locally cached copies of shared files with the data in Azure File Storage.
-offers two performance tiers — standard and premium
-premium offers more throughput at higher cost.
-All data is encrypted at rest, and you can enable encryption for data in-transit between Azure File Storage and your applications.
-allows 2000 concurrent reads to a shared file, but writes must be carefully managed using Lock file mechanism to maintain data consistency.

4. Azure cosmos DB:
-semi-structured data
-stores data as a partitioned set of documents. A document is a collection of fields, identified by a key.
-The fields in each document can vary, and a field can contain child documents.
-uses JSON to represent the document structure.
-Cosmos DB provides support for existing applications via API mongo DB API, Cassandra API, Gremlin API(Graph DB), SQL API and table API.
-Unlike Azure Table storage, documents in a Cosmos DB partition aren't sorted by ID. Instead, Cosmos DB maintains a separate index
-advantages: scalable, high availability, auto Index management, supports five well-defined consistency choices — strong, bounded staleness, session, consistent prefix, and eventual.
-Cosmos DB is highly suitable for the following scenarios:

-IoT and telematics, Gaming, Retail and marketing, web and mobile apps.

***Explore modern data warehouse analytics in Azure:**

1.Azure Data Factory :

-Its a data integration service. Does ETL / ELT
-Allows to get raw data in the form of batch/streaming data and helps clean and transform into meaningful data.
-It connects to different sources via Linked services.
-Its work is defined as a pipeline of operations (series of steps). pipelines can be triggered or run manually.
-Azure provides GUI for creating these pipelines.
-Azure Data Factory provides Poly Base support for loading data.
-Azure Data factory allows you to run your existing SSIS packages as part of a pipeline in the cloud via SSIS Feature Pack for Azure.
-SSIS Feature Pack for Azure support transfer to or from Azure storage, Azure Data Lake, and Azure HDInsight.

2.Azure Data Lake :

-A repo for large raw data that is not processed. Easy to load and read
it is a starting point for the ingested data to get stored.
-It organizes the data into directories and subdirectories
-Enables granular Role-Based Access Control (RBAC) on your data
-Is compatible with the Hadoop Distributed File System (HDFS)

3.Azure Data bricks:

-It is a Apache Spark environment running on Azure to provide big

data processing, streaming, and machine learning.
-It provides a graphical user interface where you can define and test your processing step by step.
-You can create Data bricks scripts and query data using languages such as R, Python, and Scala. You write your Spark code using notebooks.

4.Azure Synapse Analytics:

-Analytics engine.
-you can ingest data from multiple sources, process it and transform the data for analytical purposes.
-It uses poly base that enables synapse to retrieve data from relational and non-relational sources, such as delimited text files, Azure Blob Storage, and Azure Data Lake Storage.
-Azure Synapse Analytics leverages massively parallel processing (MPP) architecture which includes a control node(master) and pool of compute nodes(salves).
-The master sends the queries to be executed to the compute nodes, and the results are then sent back to the control node.
-Azure Synapse Analytics supports two computational models: SQL pools and Spark pools.
-You can only scale a SQL pool when it's not running a Transact-SQL query.
In a Spark pool, the nodes are replaced with a Spark cluster
you can run jobs on these spark nodes just like you run notebooks in data bricks.
SQL pool — can add more nodes manually
Spark pool — auto scaling of nodes is enabled
-Azure Synapse Analytics can consume a lot of resources.

-If you aren't planning on performing any processing for a while, you can pause the service.
-This action releases the resources in the pool to other users, and reduces your costs.

5. Azure Analysis:
-It does everything a synapse service can do but at a smaller scale and additional allows visualization of data.
-Smaller volumes of data (a few terabytes).
-Multiple sources that can be correlated.
-High read concurrency (thousands of users).
-Detailed analysis, and drilling into data, using functions in Power BI.
-Rapid dashboard development from tabular data.

6. Azure HD Insight:
-This is similar to that used by Synapse Analytics, except that the nodes are running the Spark processing engine rather than Azure SQL Database.

Module 1(Explore core data concepts)

- Data
  Structured: relational databases (stored in SQL Server or Azure SQL Database)
  Unstructured: audio and video files, and binary data files (BLOB)
- Semistructured: key-value stores and graph databases(key-value database stores data as a single collection without structure or relation) (Azure cosmodb)

- Access:
  - Read-only (management team of large org)
  - read/write (customers)
  - Ownership (db admin, Data analysts and data managers)
- *Data processing:*
  - Transactional (oltp): Splitting tables out into separate groups of columns like this is called normalization it can make querying more complex.
  - Analytical (oatp): big picture view of the information held in a database
- Relational db
- Non relational db
- A transaction is a sequence of operations that are atomic. This means that either all operations in the sequence must be completed successfully, or if something goes wrong, all operations run so far in the sequence must be undone.
- ACID (Atomicity, Consistency, Isolation, Durability)
- Many systems implement relational consistency and isolation by applying locks to data when it is updated. The lock prevents another process from reading the data until the lock is released. The lock is only released when the transaction commits or rolls back.
- Distributed db
- eventual consistency.
- Analytical workloads
- Batch processing and streaming
- Batch processing adv and disadv: adv(large vol, scheduled run) - disadv (time delay, minor issues can stop the process)
- Diff b/w streaming and batch: data scope, size, performance(latency), analysis

Module 2(Explore roles and responsibilities in the world of data)

Data job role

- *Database Administrators* (operational aspects of on-premises and cloud-based database) manage databases, assign permissions to users, implement policies, manage security, storing backup copies of data and restoring data in case of any failures.
- *Data Engineers* work with data, applying data cleaning routines and ingestion, identifying business rules, and turning data into useful information.
- *Data Analysts* explore and analyze data to create visualizations and charts to enable organizations to make informed decisions.

TASK AND RESPONSIBILITY

- *Db admin:*
  - Install, upgrade, db server & tools
  - Allocating, controlling, monitoring and modifying storage
  - Enrolling users and security
  - Backup, restore and archive
  - Generate reports
  - Tools: SQL Server Management Studio, pgadmin, mysql workbench, azure data studio
- *Data eng*:
  - Developing, constructing, testing, acquisition, examining, automating and maintaining db and data
  - Deploying sophisticated analytics programs, machine learning, and statistical methods.
  - improve data reliability, efficiency, and quality and research
  - Tools: Microsoft SQL Server , Azure SQL Database, Azure Databricks, and Azure HDInsight , cosmodb
- *Data Analyst*:
  - Creating charts and graphs, histograms, geographical maps

- Transforming, improving, and integrating data
- Finding hidden patterns using data and delivering info by creating rich graphical dashboards and reports
- Tools: excel, power bi

Module 3(Describe concepts of relational data)

Relational db:

- You design a relational db by creating a data model.
- The primary key indicates the column (or combination of columns) that uniquely identify each row
- Foreign Key are reference, or link to, the primary key of another table, and are used to maintain the relationships between tables
- All data is tabular. Entities are modeled as tables, each instance of an entity is a row in the table, and each property is defined as a column.
- All rows in the same table have the same set of columns. A table can contain any number of rows.
- Supports sql
- suited for OLTP applications
- Index
- contains a copy of this data in a sorted order, with pointers to the corresponding rows in the table and can create many indexes
- consume additional storage space and additional work can slow down operations and incur charges
- A clustered index physically reorganizes a table by the index key
- View
- virtual table based on the result set of a query **on-premises hosting vs cloud**

Iaas:

- installing and configuring the software, patching, taking backups, and restoring data
- virtual machine in the cloud
- best for migrations and applications requiring operating system-level access

PaaS

- Does not expose the underlying operating system and hardware to your applications
- Azure automatically creates the necessary virtual machines, networks, and other devices for your requirement
- Quickly handles this scaling for you,

Module 4 (Explore concepts of non-relational data)

Non relational db:

- enable you to store data in a very flexible manner
- don't impose a schema on data rather focus on the data itself
- store the information for entities in collections or containers
- Each entity should have a unique key value and are usually stored in key-value order
- advanced non-relational systems support indexing (Azure Cosmos DB)
- Semi structured:
- data that contains fields defined on a per-entity basis
- Json, Avro, ORC, and Parquet
- Avro is a row-based format. Each record contains a header(json) that describes the structure of the data(binary) in the record.
- ORC (Optimized Row Columnar format) organizes data into columns, Hive supports SQL-like queries over unstructured data
- Parquet is another columnar data format (row group)
- Unstructured data:

- store video and audio data as block blobs in an Azure Storage account
- A block blob only supports basic read and write operations.
- No sql
- key-value stores:
- key uniquely identifies the item, and the value(opaque) holds the data for the item.
- read and write data very quickly
- excellent choice for data ingestion
- Azure Table storage, cosmo db
- document databases:
- each document has a unique ID, but the fields in the documents are transparent to the dbms
- XML, YAML, JSON, BSON format or plain text
- enables you to query and filter data by using the values in these fields.
- Some create the document key automatically and support indexing to facilitate fast lookup
- Azure Cosmos DB implements a document database approach in its Core (SQL) API.
- column family databases:
- ORC and Parquet files
- denormalized approach to structuring sparse data
- column family database as holding tabular data comprising rows and columns, but you can divide the columns into groups known as column-families
- Apache Cassandra. Azure Cosmos DB supports the column-family approach through the Cassandra API.
- graph databases:
- store entities, but focuses on the relationships between entities
- nodes (entities), and edge (relationships between nodes)
- efficiently perform queries

- Azure Cosmos DB supports graph databases using the Gremlin API

Module 5: Explore concepts of data analytics :

- Data ingestion
- Data Processing
- ELT and ETL
- SQL Server Integration Services.and Azure Data Factory: Azure Data Factory is a cloud-based data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale.
- ETL processes that transform data visually with data flows, or by using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database.
- Reporting
- Business Intelligence (BI)
- Benchmarking: Comparison with other companies in the same industry.
- Data Visualization
- Most famous tool in Azure is Power BI for data visualization: you can connect to multiple different sources of data, and combine them into a data model
- Bar and column chart
- Line chart
- Matrix
- Key influencer
- Tree map
- Scatter
- Filed map
- Data analytics:
- Descriptive : what happened . By developing KPIs (Key Performance Indicators), these strategies can help track the success or failure of key objectives
- Diagnostic :why happened

- Predictive : what will happen in the future using neural networks, decision trees, and regression.
- Prescriptive: what action should we take to achieve a goal
- Cognitive: Cognitive analytics helps you to learn what might happen if circumstances change, and how you might handle these situations.It uses several NLP (Natural Language Processing) concepts to make sense of previously untapped data sources, such as call center conversation logs and product reviews.

Module 6: Explore relational data services in Azure:

- Data base
- Stored procedure: A stored procedure is a block of code that runs inside your database.
- A linked server is a connection from one database server to another. SQL Server can use linked servers to run queries on one server that can include data retrieved from other servers; these are known as distributed queries.
- IAAS: Infrastructure as a service e.g: azure virtual network
- PAAS: Platform-as-a-service e.g: Azure SQL Databases
- SAAS:Software-as-a-Service, e.g office 365
- Azure Data Services: Azure Data Services fall into the PaaS category. These services are a series of DBMSs managed by Microsoft in the cloud. Each data service takes care of the configuration, day-to-day management, software updates, and security of the databases that it hosts. All you do is create your databases under the control of the data service.
- Most famous database service is Azure SQL database
- Azure Database for sql server
- MariaDB server
- Postgre SQL server
- Microsoft also provide services for non relational dbms such as cosmos DB

- Azure Data Services ensure that your databases are available for at least 99.99% of the time.
- There are cost associated with the running database in azure data services
- Can;t shutdown the database and restart it later. This services are always on
- SQL Server on Azure Virtual Machines: SQL Server on Virtual Machines enables you to use full versions of SQL Server in the Cloud without having to manage any on-premises hardware.SQL Server running on an Azure virtual machine effectively replicates the database running on real on-premises hardware. Migrating from the system running on-premises to an Azure virtual machine
- lift-and-shift refers to the way in which you can move a database directly from an on-premises
- server to an Azure virtual machine without requiring that you make any changes to it.
- A hybrid deployment is a system where part of the operation runs on-premises, and part in the cloud
- Azure SQL Database :Azure SQL Database is a PaaS offering from Microsoft. You create a managed database server in the cloud, and then deploy your databases on this server.
- Single Database
- Elastic Pool database :This option is similar to Single Database, except that by default multiple databases can share the same resources, such as memory, data storage space, and processing power through multiple-tenancy.
- Managed instance
- Azure SQL Database is often used for:
- Modern cloud applications that need to use the latest stable SQL Server features.
- Applications that require high availability.
- Systems with a variable load, that need the database server to scale up and down quickly.

- SQL Database helps secure your data by providing encryption. For data in motion, it uses Transport Layer Security. For data at rest, it uses Transparent Data Encryption. For data in use, it uses Always Encrypted.
- Azure SQL Database Managed Instance:
- You have complete control over this instance, much as you would for an on-premises server. The Managed instance service automates backups, software patching, database monitoring, and other general tasks, but you have full control over security and resource allocation for your databases.Managed instances depend on other Azure services such as Azure Storage for backups, Azure Event Hubs for telemetry, Azure Active Directory for authentication, Azure Key Vault for Transparent Data Encryption (TDE) and a couple of Azure platform services that provide security and supportability features. The managed instances make connections to these services.
- MySQL:
- MariaDB: compatibility with oracle database, One notable feature of MariaDB is its built-in support for temporal data. A table can hold several versions of data, enabling an application to query the data as it appeared at some point in the past.

- PostgreSQL

- Azure Database for MySQL:
- High availability features built-in.
- Predictable performance.
- Easy scaling that responds quickly to demand.
- Secure data, both at rest and in motion.
- Automatic backups and point-in-time restore for the last 35 days.
- Enterprise-level security and compliance with legislation.
- Azure Database for MariaDB:
- Built-in high availability with no additional cost.

- Predictable performance, using inclusive pay-as-you-go pricing.
- Scaling as needed within seconds.
- Secured protection of sensitive data at rest and in motion.
- Automatic backups and point-in-time-restore for up to 35 days.
- Enterprise-grade security and compliance.
- Azure Database for PostgreSQL: same property as azure database for my sql.
- Azure Database for PostgreSQL single-server: Each tier supports different numbers of CPUs, memory, and storage sizes
- Azure Database for PostgreSQL Hyperscale (Citus): Data is split across nodes
- Use Azure database migration service to migrate on premise Mysql, mariaDB or postgresql to a database running the corresponding data services in Azure

Module 7: Explore provisioning and deployment in database service in azure:

- What is Provisioning? :
- Provisioning is the act of running a series of tasks that a service provider, such as Azure SQL Database, performs to create and configure a service.
- The act of increasing (or decreasing) the resources used by a service is called scaling.
- Tools to provision services:
- Azure portal: Display list of service specific pages before actual provisioning
- Azure command line interface(CLI): basic command prompt and powershell command we run on windows to automate service creation.
- Azure Powershell

- Azure Resource manager templates: describes the service (or services) that you want to deploy in a text file, in a format known as JSON
- Provisioning in Azure SQL database
- Provisioning PostgreSQL and MySQL:
- Hyperscale option for postgreSQL supports:
- Horizontal Scaling
- Query parallelization
- Excellent support for multi-tenant applications,
- real time operational analytics, and
- high throughput transactional workloads
- Configuring relational data services
- Configure connectivity and firewall
- Configure connectivity to virtual networks and on-premises computers
- Azure SQL Database communicates over port 1433
- A firewall rule of 0.0.0.0 enables all Azure services to pass through the server-level firewall rule and attempt to connect to a single or pooled database through the server.
- Configure connectivity from private endpoints.
- Configure Authentication
- Configure access control: who or what can access your resources
- Role assignment consist of three elements:
- Security principle
- Role definition: Collection of permission
- Owner
- Contributor
- Reader
- User access administrator
- Scope : lists set of resources that the access applies to
- You add role assignments to a resource in the Azure portal using the Access control (IAM) page
- Configure advance data security

- Configure Azure SQL Database
- An ACL(access control list) contains a list of resources, and the objects (users, computers, and applications) that are allowed to access those resources
- Connectivity from within Azure
- Connectivity from outside azure
- Configure DoS(Denial of Service) guard: DoSGuard actively tracks failed logins from IP addresses
- Configure Azure Database for PostgreSQL: Connections to your Azure Database for PostgreSQL server communicate over port 5432
- Configure read replicas : replicate data from an Azure Database for PostgreSQL server to a read-only server
- Configure Azure Database for MySQL
- Configure server parameter
- Configure read replicas: same as postgreSQL

Module 8: Query relational data in azure:

- SQL: structured query language
- Transact-SQL (T-SQL). This version of SQL is used by Microsoft SQL Server and Azure SQL Database.
- pgSQL. This is the dialect, with extensions implemented in PostgreSQL.
- PL/SQL. This is the dialect used by Oracle. PL/SQL stands for Procedural Language/SQL.
- Data Manipulation Language (DML): SELECT, INSERT, UPDATE, DELETE

- Data Definition Language (DDL): CREATE, ALTER, DROP, RENAME

- Query relational data in Azure SQL Database
- Retrieve connection information for azure sql DB
- Use the azure portal to query a DB

- Use SQLCMD to query a database
- USe Azure data studio: Azure Data Studio is a graphical utility for creating and running SQL queries from your desktop
- SQL server management studio
- Use SQL Server Data Tools in Visual Studio
- Query relational data in Azure Database for PostgreSQL
- Retrieve connection information for Azure Database for PostgreSQL
- Use psql to query a database
- psql commands include:
- \l to list databases.
- \dt to list the tables in the current database.

- Connect to PostgreSQL database using Azure Data Studio

- Query relational data in Azure Database for MySQL
- port : 3306

Module 9 (Explore non-relational data offerings in Azure)

- Azure Table Storage:
- items are referred to as rows(must have a key), and fields are known as columns
- store semi-structured data (schemaless)
- Denormalized data
- splits a table into partitions. Partitioning is a mechanism for grouping related rows, based on a common property or partition key
- Partitions are independent from each other
- you can include the partition key in the search criteria
- Items in the same partition are stored in row key order.

- quickly perform Point queries that identify a single row, and Range queries that fetch a contiguous block of rows in a partition
- The columns in a table can hold numeric, string, or binary data up to 64 KB in size. A table can have up to 252 columns, apart from the partition and row keys. The maximum row size is 1 MB
- Adv
- no need to map and maintain the complex relationships
- Fast operations like insertion deletion query
- Simple to scale
- Storing TBs of structured data capable of serving web scale applications
- Storing datasets that don't require complex joins, foreign keys, or stored procedures
- Capturing event logging and performance monitoring data
- Disadv
- Consistency isn’t guaranteed
- no referential integrity
- difficult to filter and sort on non-key data
- The data for each table is replicated three times within an Azure region. For increased availability
- transparently switch to a working replica while the failed replica is recovered
- configure security and role-based access control
- Azure Blob storage
- unstructured data, or blobs
- you create blobs inside containers
- Block blobs. handled as a set of blocks. used to store discrete, large, binary objects that change infrequently.
- Page blobs. organized as a collection of fixed size 512-byte pages. support random read and write operations, used to implement virtual disk storage for virtual machines

- Append blobs. a block blob optimized to support append operations updating or deleting existing blocks isn't supported
- Acces tiers
- The Hot tier is accessed frequently. data is stored on high-performance media.
- The Cool tier. has lower performance, accessed infrequently.
- The Archive tier. provides the lowest storage cost, but with increased latency for historical data that mustn't be lost, but is required only rarely, stored in an offline state
- A lifecycle management policy can automatically move a blob from Hot to Cool, and then to the Archive tier, as it ages and is used less frequently
- Serving images or documents directly to a browser
- Storing data for backup and restore, disaster recovery, and archiving, analysis and distributed access
- Versioning
- Soft delete
- Snapshots
- Change Feed
- Azure File storage
- Create and access files shares in the cloud
- exposes file shares using the Server Message Block 3.0 (SMB) protocol
- tools such as the AzCopy utility or azure portal
- Services: standard and premium (greater throughput)
- Migrate existing applications to the cloud (access data using file-based APIs)
- Share server data across on-premises and cloud(leverage the availability, durability, scalability, and geo redundancy built into the Azure storage platform)
- Integrate modern applications with Azure File Storage (rest api)
- Simplify hosting High Availability (HA) workload data

- Azure Cosmos DB
- multi-model NoSQL dbms
- manages data as a partitioned set of documents(collection of fields, identified by a key)
- use JSON
- provides APIs that enable you to access these documents
- SQL API. provides a SQL-like query language over documents
- Table API. use the Azure Table Storage API to store and retrieve documents
- MongoDB API enable a MongoDB application to run unchanged against a Cosmos DB database
- Cassandra API column family database management system provides a Cassandra-like programmatic interface for Cosmos DB
- Gremlin API. implements a graph database interface to Cosmos DB
- Unlike Azure Table storage, documents in a Cosmos DB partition aren't sorted by ID Instead, Cosmos DB maintains a separate index. This index contains not only the document IDs, but also tracks the value of every other field in each document
- all databases are replicated within a single region
- replication is transparent, and failover from a failed replica is automatic
- guarantees less than 10-ms latencies for both reads (indexed) and writes at the 99th percentile, all around the world
- all data in Cosmos DB is encrypted at rest and in motion
- Module 10 (Explore provisioning and deploying non-relational data services in Azure)
- provisioning non-relational data services
- Provisioning is the act of running a series of tasks that a service provider, it is opaque. Parameters initially set can be later modified.
- Tools
- The Azure portal

- The Azure command-line interface (CLI): run command using cloud shell
- Azure PowerShell
- Azure Resource Manager templates (deploy in json)
- Provision Azure Cosmos DB
- Azure CLI, Azure PowerShell, or an Azure Resource Manager template
- Databases and containers are the primary resource(storage space) consumers
- uses the concept of Request Units per second (RU/s) to manage the performance and cost of databases
- If you underprovision (by specifying too few RU/s), Cosmos DB will start throttling performance
- configuring non-relational data services
- You can connect to these services from an on-premises network, the internet, or from within an Azure virtual network
- Azure Private Endpoint is a network interface that connects you privately and securely to a service powered by Azure Private Link
- Private endpoint connections page for a service allows you to specify which private endpoints, if any, are permitted access to your service
- Azure Active Directory (Azure AD) provides superior security and ease of use over access key authorization
- Configure access control
- Azure role-based access control
- Allow one user to manage virtual machines in a subscription and another user to manage virtual networks.
- Allow a database administrator group to manage SQL databases in a subscription.
- Allow a user to manage all resources in a resource group, such as virtual machines, websites, and subnets.
- Allow an application to access all resources in a resource group.

- security principal is an object that represents a user, group, service, or managed identity that is requesting access to Azure resources.
- role definition, often abbreviated to role, is a collection of permissions
- Owner
- Contributor
- Reader
- User Access Administrator
- scope lists the set of resources that the access applies to
- Tools: Access control (IAM)
- Configure consistency in cosmo db
- Eventual. least consistent Changes won't be lost, they'll appear eventually changes could appear out of order. lowest latency and least consistency
- Consistent Prefix. ensures that changes will appear in order, although there may be a delay before they become visible
- Session. If an application makes a number of changes, they'll all be visible to that application, and in order
- Bounded Staleness. There's a lag between writing and then reading the updated data
- Strong: In this case, all writes are only visible to clients after the changes are confirmed as written successfully to all replicas
- Configure shared access signatures
- use shared access signatures (SAS) to grant limited rights to resources in an Azure storage account for a specified time period
- access resources such as blobs and files, without requiring that they're authenticated first
- Module 11(Manage non-relational data stores in Azure)
- data operations in Cosmos DB
- Data Explorer in the Azure portal to run ad-hoc queries

- Cosmos DB Data Migration tool to perform a bulk-load or transfer of data
- Azure Data Factory to import data from another source
- custom application that imports data using the Cosmos DB BulkExecutor library (make use of multiple concurrent threads to batch your data into chunks and load the chunks in parallel)
- Create your own application that uses the functions available through the Cosmos DB SQL API client library
- insufficient throughput capacity configured results in https error 429
- Query Azure Cosmos DB
- The Cosmos DB SQL API supports a dialect of SQL for querying documents using SELECT statements
- Manage Azure File storage

Module 12: Examine components of a Modern data warehouse :

- The process of combining all of the local data or gathering data from many different sources within an organization is known as data warehousing.
- The process of analyzing streaming data and data from the Internet is known as Big Data analytics.
- Azure Synapse Analytics combines data warehousing with Big Data analytics.
- online analytical processing (OLAP)
- Modern data Warehousing : It contains a mixture of relational and non relational data like media files, social media streams or internet of things.
- Tools we can use: Azure Data Factory, Azure Data Lake Storage, Azure Databricks, Azure Synapse Analytics, and Azure Analysis Services, power BI is used to analyze and visualize the data, generating reports and charts.
- gathers data from many different sources within an organization

- provide answers to complex queries, unlike a traditional relational database
- handle big data
- contain a mixture of relational and non-relational data, including files, social media streams, and Internet of Things (IoT) sensor data
- Combine batch and stream process: upto the second data and historical data are important for a company. Upto the second data is generated by steam process and historical data is generated by batch process.
- Azure Data factory: its a data integration service. Purpose of the azure data factory is to retrieve data from one or more data sources and to convert it into the format that you process.
- Work performed by azure data factory is defined as pipeline of operations
- Azure Data Lake storage:
- It's a repository for Large Quantities of Raw data.
- Data is fast load and processed because its raw & unprocessed.
- A data warehouse holds structured information but a data lake stores raw data.
- It's an extension of Azure Blob storage.It organizes files into directories and sub directories but blob storage only mimics a directory structure.
- It supports POSIX(Portable Opening system interface) and compatible with HDFS (Hadoop distributed file system)
- Azure DataBricks:
- Azure Databricks is an Apache Spark environment running on Azure to provide big data processing, streaming, and machine learning.

- Azure Databricks provides a graphical user interface where you can define and test your

processing step by step, before submitting it as a set of batch tasks.

- Azure Synapse Analytics:
- It's an analytical engine. Design to process large data very quickly.
- Ingest data from external sources and then transform and aggregate this data into a format suitable for analytics processing.its used to process data which we have read in and processed locally.
- Control node is the brain of architecture, like a front end which interacts with all applications. When we submit a process request it transforms it into smaller requests that run against distinct subsets of the data in parallel.
- We have to submit queries in Transact-SQL statements
- PolyBase enables you to retrieve data from relational and non-relational sources, such as delimited text files, Azure Blob Storage, and Azure Data Lake Storage.
- Azure Analysis Service
- Azure Analysis Services enables you to build tabular models to support online analytical processing (OLAP) queries. You can combine data from various sources (e.g lake , analytics , sql db, cosmos db).
- Difference between Analysis service and synapse analytics:
- Use Azure Synapse Analytics for:
- Very high volumes of data (multi-terabyte to petabyte sized datasets).
- Very complex queries and aggregations.
- Data mining, and data exploration.
- Complex ETL operations. ETL stands for Extract, Transform, and Load, and refers to the way in which you can retrieve raw data from multiple sources, convert this data into a standard format, and store it.
- Low to mid concurrency (128 users or fewer).

- Use Azure Analysis Services for:
- Smaller volumes of data (a few terabytes).
- Multiple sources that can be correlated.
- High read concurrency (thousands of users).
- Detailed analysis, and drilling into data, using functions in Power BI.
- Rapid dashboard development from tabular data.
- Azure HDInsight: Azure HDInsight is a big data processing service that provides the platform for technologies such as Spark in an Azure environment.
- Hadoop is an open source framework that breaks large data processing problems down into smaller chunks.
- Hive is a SQL-like query facility that you can use with an HDInsight cluster to examine data held in a variety of formats.

Module 13: Explore large scale data analytics :

- Ingest data using Azure Data Factory
- HTAP(Hybrid transactional analytical processing): analyze operational data into original location.
- Orchestration is the process of directing and controlling other services, and connecting them together, to allow data to flow between them.
- A linked service provides the information needed for Data Factory to connect to a source or destination
- A dataset in Azure Data Factory represents the data that you want to ingest (input) or store (output). If the data is structured, the data set specifies how data is structured.
- A pipeline is a logical grouping of activities that together perform a task. The activities in a pipeline define actions to perform on your data. It includes a lot of activity like looping data in for each loop , can use if condition to execute the activity conditionally, can map input into different formats as per output is required. A trigger enables you to schedule a pipeline to occur according to a planned schedule

- Ingest data using PolyBase
- Polybase is a feature of SQL server & azure synapse analytics that enables you to run Transact-SQL queries that read data from external data sources.It makes external data appear like tables in a sql db.
- Azure SQL Database does not support PolyBase.
- Ingest data using SQL Server Integration Services
- SQL Server Integration Services (SSIS) is a platform for building enterprise-level data integration and data transformations solutions.It is used to solve complex business problems.

- SSIS can extract and transform data from a wide variety of sources such as XML data files, flat files, and relational data sources, and then load the data into one or more destinations.

Module 14: Get started building with Power BI

- collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights.
- Power BI Desktop, an online SaaS (Software as a Service)
- Use Power BI
- Bring data into Power BI Desktop, and create a report.
- Publish to the Power BI service, where you can create new visualizations or build dashboards.
- Share dashboards with others, especially people who are on the go.
- View and interact with shared dashboards and reports in Power BI Mobile apps.
- Building blocks of Power BI
- Visualizations
- Datasets
- Reports

- Dashboard
- Tiles
- app is a collection of preset, ready-made visuals and reports that are shared with an entire organization.
- Power BI can just as easily connect to a whole assortment of software services (also called SaaS providers or cloud services): Salesforce, Facebook, Google Analytics, and more.