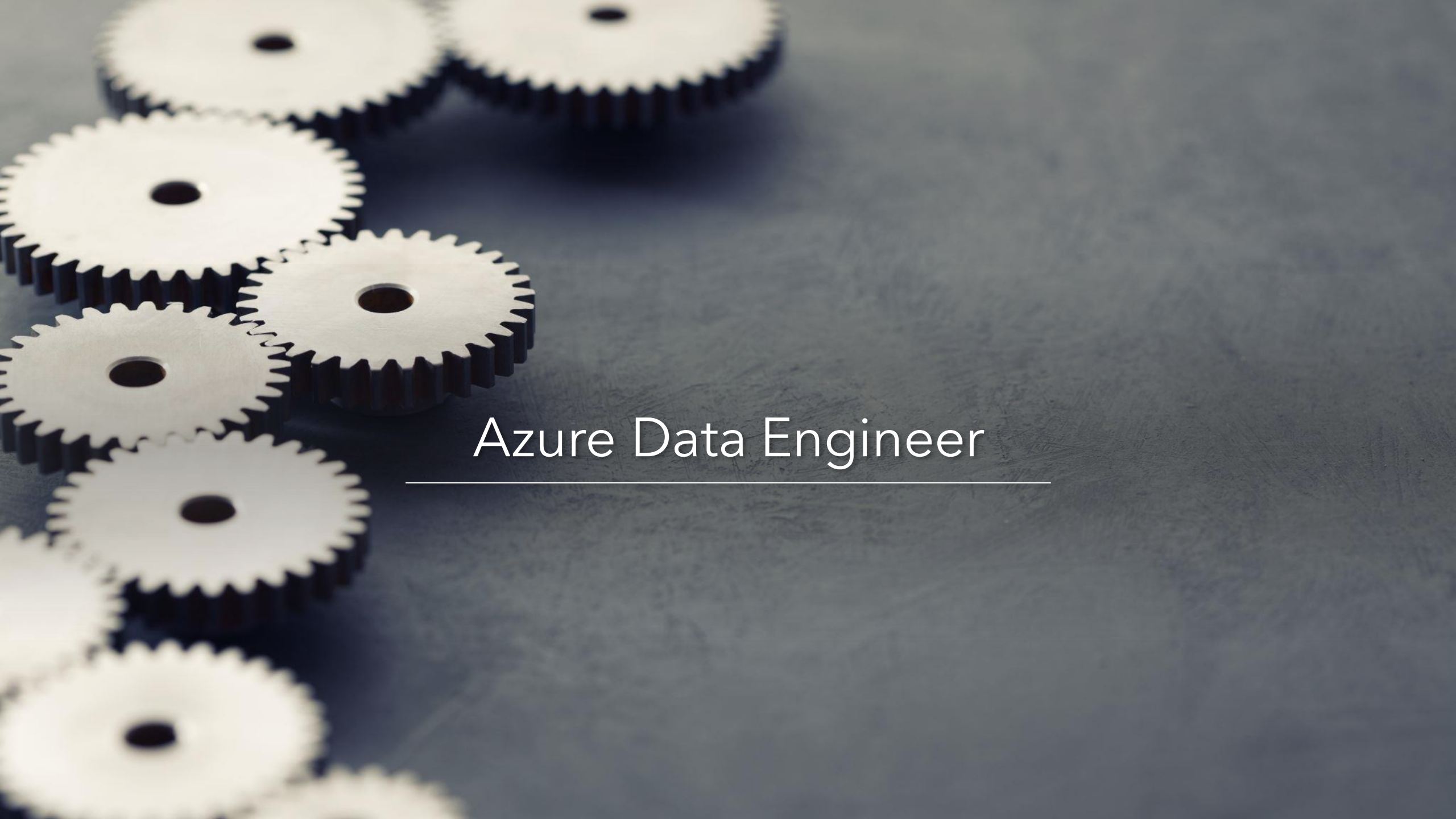


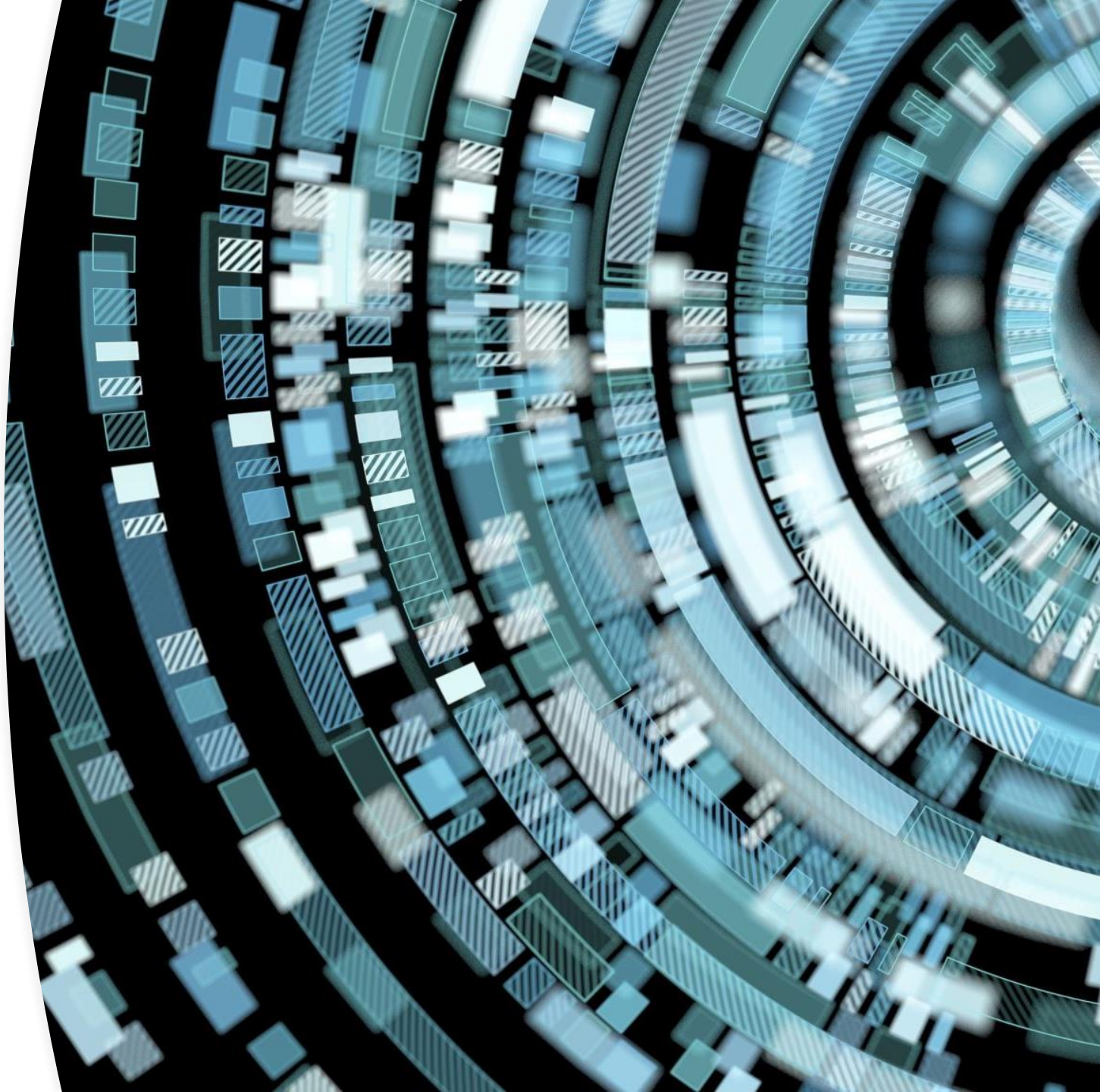
Azure Data Engineer

Pre-requisite

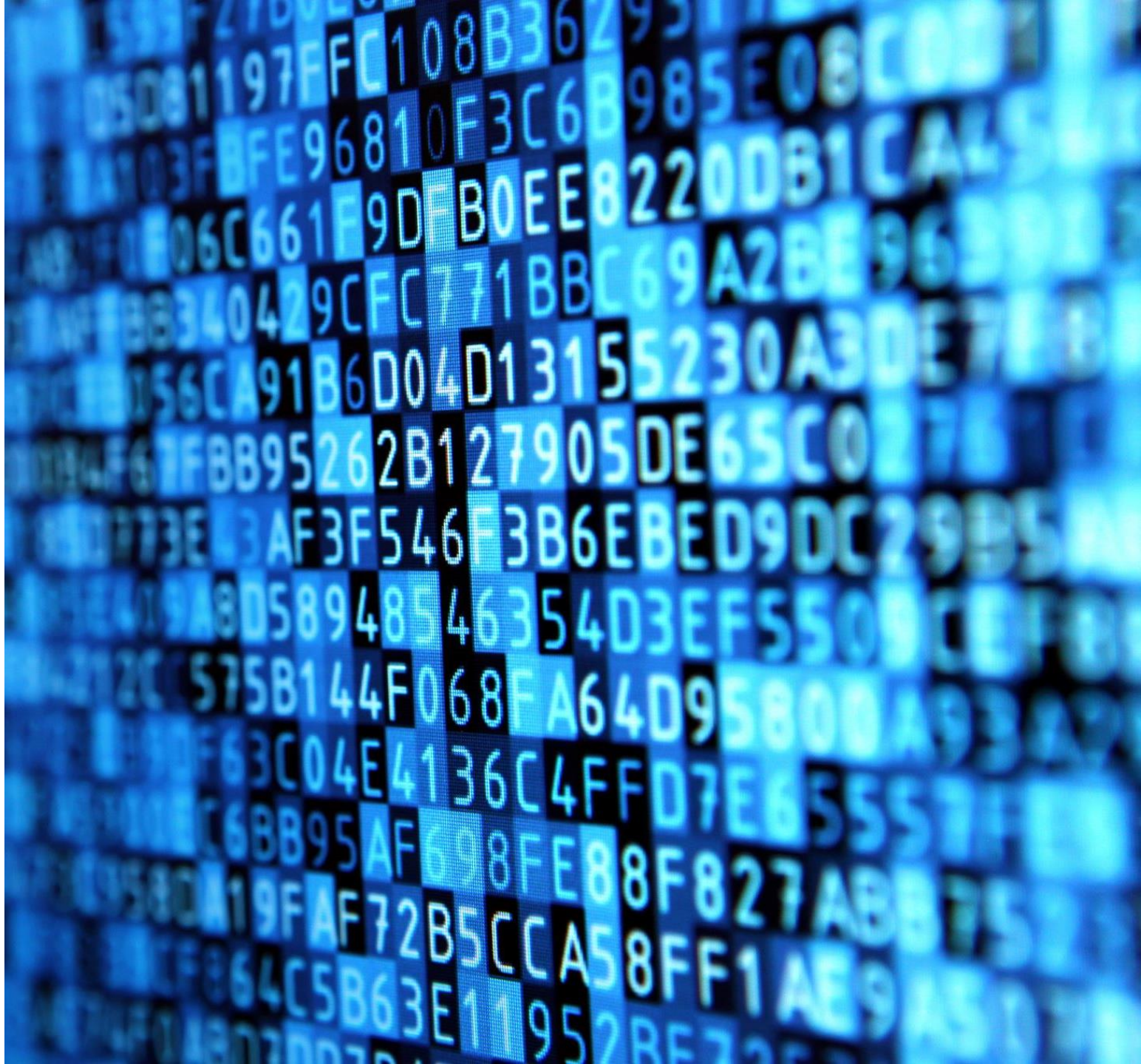
- Azure fundamentals
- Good overview of Azure EventHub, IOT Hub, IOT edge
- Implement and Use Azure key Vault
- Very good knowledge on Azure AD
- At multiple places, azure documentation lacks links to working code, in those cases write your own where possible.

Skill Measured

- Services in Scope
  - SQL Databases
  - Azure Synapse Analytics
  - Data Lake Gen2
  - Cosmos DB
  - Azure Databricks
  - Azure Datafactory
  - Stream Analytics
- Horizontals
  - Azure monitor
  - Diagnostics and Log Analytics
  - Optimization
  - Security
  - High Availability
  - Disaster Recovery



SQL Databases



Implementation Models

Deployment Models

- Managed Instance
  - High compatibility with SQL server
  - Size upto 8TB
  - Supports Private IP in VNET
  - Supports BYOL
- SQL Databases – Single and Elastic Pools
  - Low compatibility with SQL server
  - Size upto 100TB
  - Does not support Private IP
- SQL Virtual Machines
- Azure Doc – Feature comparison

Purchasing Models

- VCore
- DTU – Blended HW model
- Azure Doc – Purchasing model , Service Teirs

Elastic Pool

- Provides pool of resources shared by multiple single databases. Huge cost benefits can be derived if peak loads are scattered
- How Scaling Works:-
  - New compute instance is created
  - Switching of connections to new instance
  - <1min of disruption
- This page explains what are elastic pools, where to use them and exercise to creaet and use them - Azure Doc - Elastic Pool overview

Business Continuity

- Geo-Replication
  - Replicates data to same or other region
  - Supports read at secondary
  - Supports Multiple replicas
  - Requires connection string update
  - Supports only SQL databases
  - Azure Docs - overview
- Failover Groups
  - Failover multiple databases simultaneously, use with pool databases
  - Supports both SQL databases and managed instances
  - Does not support same region replication
  - No need to change connection string
  - Azure Docs – overview and failover group tutorials (5 tutorials at the time of writing)
- Backup and Recovery
  - Azure Docs - Overview and configure long term retention policy

Data Security

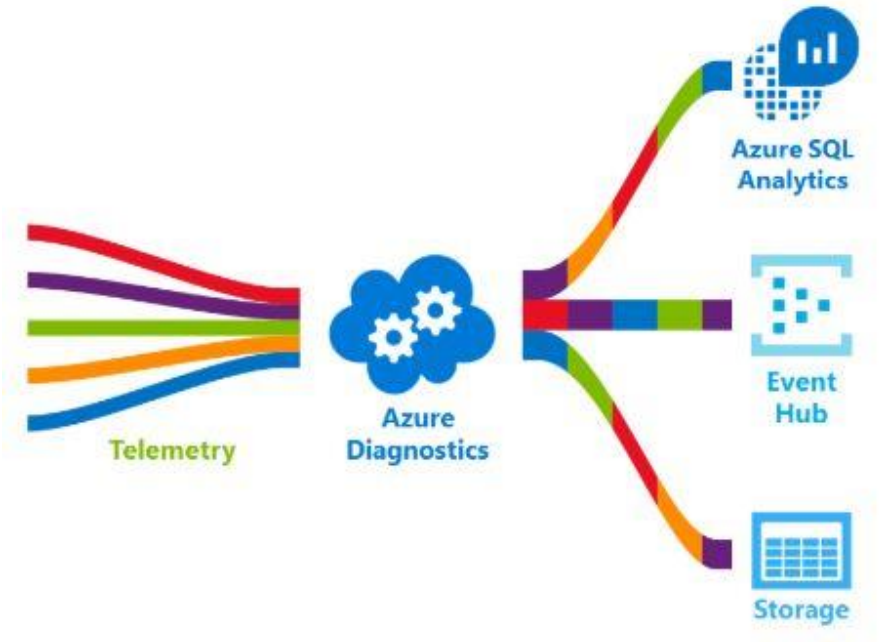
- Advanced data security
  - Provides discovery and classification
  - Provides Vulnerability protection and threat assessment
  - Azure Docs - Overview
- Audit
  - Audit policy can be defined at server level or database level. DB inherits server level policy.
  - Rule defined at both will cause duplicate event capture at DB
  - Three policies are default at server :-
    - BATCH_COMPLETION_GROUP
    - SUCCESSFUL_DATABASE_AUTHENTICATION_GROUP
    - FAILED_DATABASE_AUTHENTICATION_GROUP
  - Azure Doc - Overview
- Firewalls and Virtual Networks
  - Use firewall to restrict access to database from a single IP or IP ranges.
  - These are at server level not DB level
  - If using IP, make sure it is static IP.
  - Azure Docs – Overview
- Private end point
  - Create a private EP in VNET, then create network rule on server
  - Private Endpoint must exist in the same region as Serve
  - Azure Docs - Overview

Data Security

- Private Link
  - It provides private IP address to the configured azure service, in this case it is Azure SQL
  - Private End point requires network/firewall rule for SQL access, Private Link does not.
  - Azure Docs - Overview
- TDE
  - Customer managed key, which uses Keyvault integration
    - Azure Docs - Encryption with own key
  - Service managed key
  - Azure Docs - Overview
- Always encrypted
  - Provides encryption at rest when in database and also when it moves between client and server
  - Azure Docs – Always Encrypted
- Azure AD Authentication
  - Permissions can be managed using external / AD Groups
  - Link admin account to server
  - Create contained users same as AAD accounts
  - Azure Docs - Overview , Configure AAD

Data Masking

- Data masking functions
  - Credit Card - Shows only last 4 digits - xxxx-xxxx-xxxx-1234
  - Default – Fixed value 'X' for string, 0 for numbers, 01-01-1900 for date
  - Email – axx@xxxx.com
  - Random Number – Randomize based on From:To range of numbers.
  - Custom Text – [Exposed Prefix][mask][Exposed suffix]
    - To expose first 3 digits and last tow digits Ex. [3][XXXXX][2]
- Data Masking policy
  - SQL users excluded from masking I.e they see unmasked data
    - Admin users are always in exlusion list
  - Masking rule maps DB columns to masking functions
- Azure Doc - Overview
- This is applicable to Synapse as well



Optimize / Performance Tuning

- Azure Diagnostics: Type of telemetry available and how to export these. Important ones are Basic, Automated Tuning and SQLInsights.
  - Azure Docs - Overview
- Intelligent Insights – Azure Doc
- Automates Tuning
  - Three actions available Create Index, Drop Index and Force Last Good Plan.
  - Drop Index is disabled by default
  - Servers can inherit azure defaults and databases from server.
  - Azure Doc - Overview and how to implement

Monitor

- There is overlap between this and optimize
  - Azure Doc - Monitoring Overview
  - Azure Doc - Azure monitor logs for Azure SQL
  - Azure Doc - Monitor performance of pools

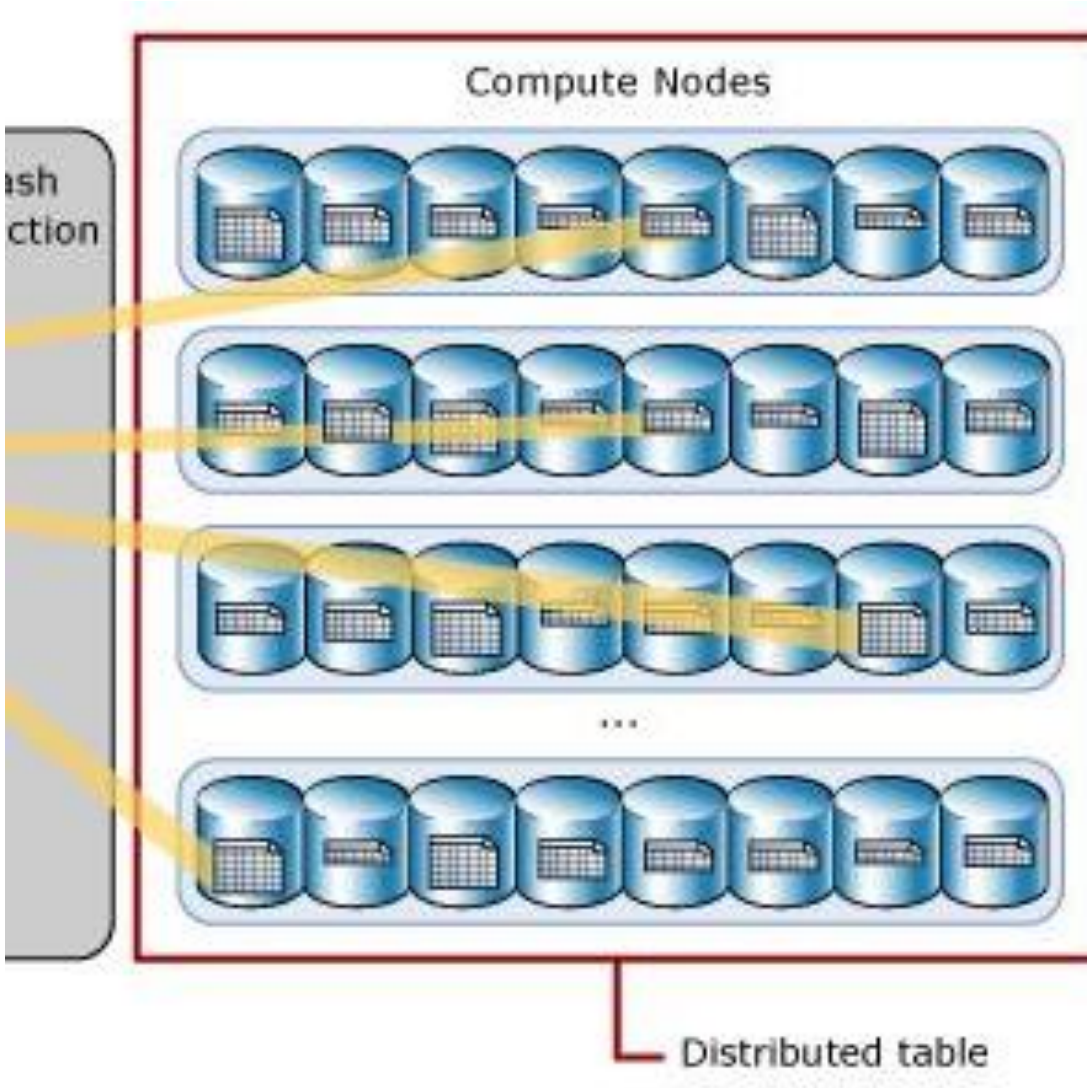
Azure Synapse

Get Started

- What is Azure Synapse?
  - Azure Doc - Overview
- Synapse Architecture
  - Azure doc - Overview
- Implement Azure Synapse
  - Azure Doc – Create and connect

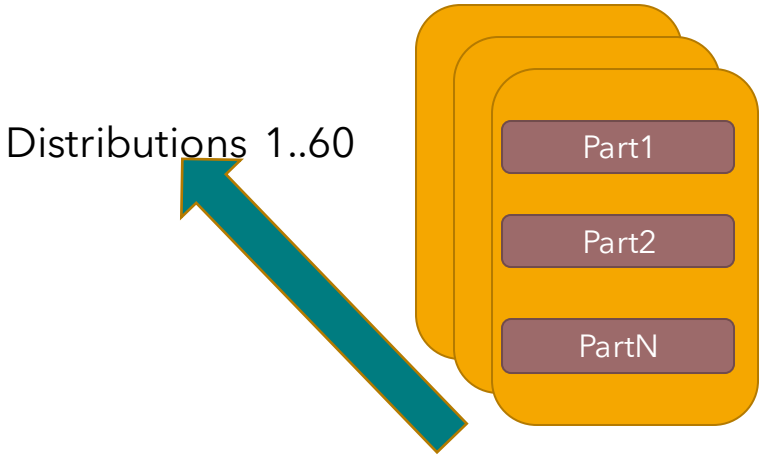
Data Distribution

- Data is stored across compute nodes in 60 distributions (or less) I.e if number of nodes is 60 , each nodes gets one distribution and is expensive and if you have single node all distributions are on that node (cheapest)
- Control node splits the query into 60 small queries to run on 60 distributions
- Choosing Distribution Column
  - This column cannot be updated
  - Must have many unique values, and distributes the data evenly across 60 partitions. Partition skew can lead to performance issues
  - Use a column from Group By, not from where clause
- To optimize JOIN performance, join columns must be hash distributed, use equal operator and must have same data type
- To change the distribution column, create new table as CTAS with new distribution column and then collect fresh stats on the table.
- Azure Doc – Table Distribution and replicated tables



Partitioning

- In Synapse, data is already distributed in 1-60 distributions. Partitioning further splits the data into partitions.
- Utilize partition switching to remove old data rather than using delete
- Partition column should be part of filter clause to improve query performance
  - This is different from distribution, where column is part of group by / aggregate clause
- Azure Doc - Table Partitioning



Data Loading

- Data Loading Best practices
  - Load to staging table, which is defined as heap or round-robin
  - Columnstore index requires large resources, so user should be member or medium or large resource class
  - Each rowgroup compresses 1M rows, anything less than 100K is sent to delta store and is inefficient.
  - Azure Doc – Data load guidance
- Example exercises
  - Using polybase external table to load data from Data lake to Synapse
  - Load with optimization
  - Using COPY command to laod data

Security

- Following concepts apply same as from SQL databases:-
  - AD Authentication
  - Network Security
  - Advanced Data Security & Auditing
  - Dynamic Data Masking and TDE
- Column level security – Azure Doc
- Row level security - Azure Doc

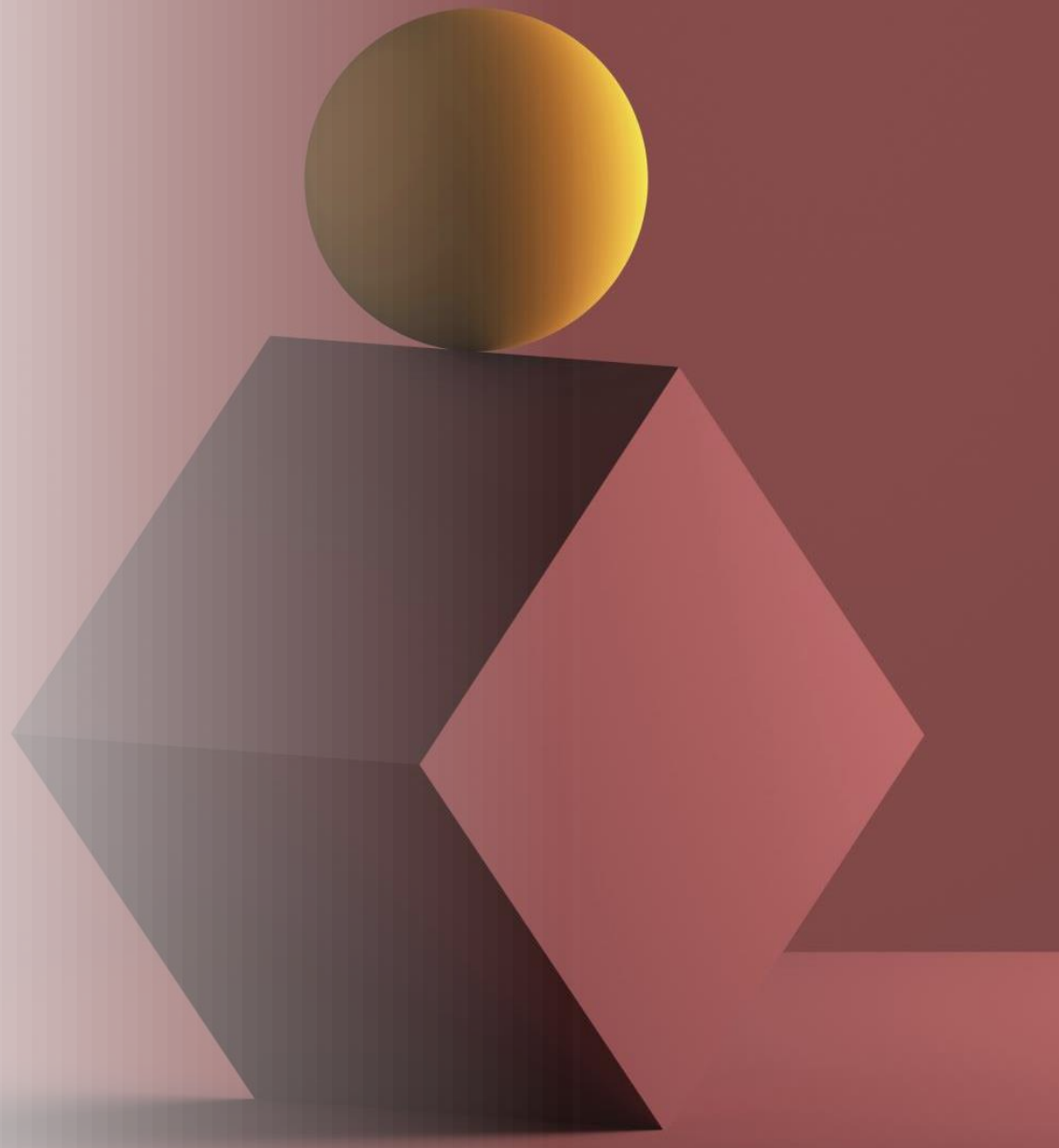
Optimize

- Hash distribute large tables
- Columnstore index is only suitable if each partition/distribution gets 1M rows I.e 100 partitions means 60(distributions)*100(partitions)*1M = 6Billion rows table !!!
- Best practices - Azure Doc

Monitor, Backup & DR

- Monitor and Logs – Azure Doc
  - Monitor using dynamic management views – Azure Doc
- Scaling Compute – Azure doc
  - Scale compute with azure functions – Azure Doc
- Backup and Recovery – Azure Doc
  - Restore from GEo backup – Azure Doc
- Tuning Recommendation – Azure doc

Azure CosmosDB



What is Cosmos DB?

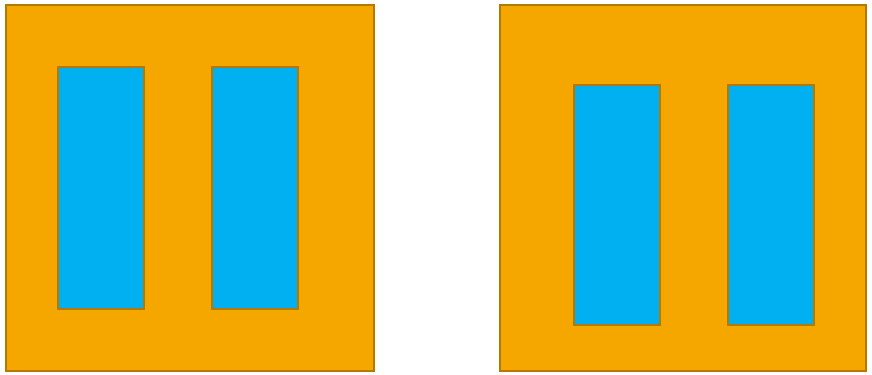
- Globally distributed multi-model database. Cosmos DB guarantees single-digit-millisecond latencies at the 99th percentile anywhere in the world, offers multiple well-defined consistency models to fine-tune performance, and guarantees high availability with multi-homing capabilities.
- Azure Cosmos DB is schema-agnostic. It automatically indexes all the data without requiring you to deal with schema and index management. It's also multi-model, natively supporting document, key-value, graph, and column-family data models.
- Azure Doc - Introduction, Implement

Data Distribution

- Database can be multi-regional
  - Azure Doc – Data distribution overview
- Consists of containers , which are partitioned by Key
- Replicaset is with data center / partition set is across multiple DC or regions and is composed of multiple RSs.
  - Azure Doc – Data Distribution in detail
- Change from single master to multiple master is without disruption
  - Azure Doc – Configure multiple regions, Using multi-master in application
- From the portal -> Replicate Data Globally -> enable multiple regions for read/write. Multi region write needs to be enabled separately
- Automatic failover can be set to multiple regions with priority set for each 1..N

Partitions

- Physical partitions hold one or more logical partitions
- Logical partitions are based on partition keys ex. Userid
- In addition to partition keys, each item has index ID. These two put together are item index
- Each physical partition provides 10000 rps throughput and 50GB
- Hot partition issue can happen, If load is not distributed evenly across partition key
- Partition key should have high cardinality
- For select heavy container, choose a key that appears in filters
- Azure Doc - Overview



Implement CosmosDB with Geo Distribution

- How to scale throughput globally – Azure Doc
- To configure multi-master / multi –region in application:-
  - Use PreferredLocations in connectionpolicy to specify list of regions in preferential order – Azure Doc
  - Use UseMultipleWriteLocations with setCurrentLocation to handle write operations to multiple regions dynamically – Azure Doc

Consistency Levels

Strong	The reads are guaranteed to return the most recent committed version of an item. A client never sees an uncommitted or partial write. Users are always guaranteed to read the latest committed write.
Bounded Staleness	The reads might lag behind writes by at most *"K"* versions (that is, "updates") of an item or by *"T"* time interval
	Provides strong consistency for single master , single region clients
Session	Within a single client session reads are guaranteed to honor the consistent-prefix, monotonic reads, monotonic writes, read-your-writes, and write-follows-reads guarantees.
	Clients outside the session perform either with consistent prefix or eventual
Consistent Prefix	Consistent prefix consistency level guarantees that read never see out-of-order writes.
Eventual	There's no ordering guarantee for reads. In the absence of any further writes, the replicas eventually converge.
	Eventual consistency is the weakest form of consistency because a client may read the values that are older than the ones it had read before.

Azure Doc - Overview, Choose Consistency Level

Consistency Levels and Latency

- Read and Write latency is always commited < 10ms
  - Multi-regions strong consistency is exception to this rule.

Selecting CosmosDB API

- Use CoreSQL for all cases except:-
  - Teams are using exisiting mongo, cassandra, table or graph APIs
  - There is a requirement to capture relationship among data, in this case us Graph
  - In addition, cassandra is best fit for fixed schema use cases and mongo for flexible schema
  - Azure Learn – Select cosmos db api

Monitor CosmosDB

- Azure monitor – Azure Doc
- Monitor Server side latency – Azure Doc
  - If high latency is seen for certain operations, then use diagnostic logs for checking the size of data returned
- Monitor Request Units – Azure Doc
- Diagnostic Logs – Azure Doc
- Using control plane logs – Azure Doc

Throughput

- Throughput is measured in RUs, and allocated in batch of 100 per sec.
- Read operation – 1KB data - is 1RU, write is 2RU
- If certain logical partition consumes more RUs than allocated to physical partition it is on, rate throttling will happen.
- Can be provisioned at both DB and container level
  - It is distributed evenly among objects ie. DB -> Container -> Physical partitions
- Azure Doc - Introduction, autoscale throughput, autoscale vs manual

Encryption using Key Vault

- Encryption can be either service managed or customer maanged
- For customer managed, first register the DocumentDB service, then create access policy in key vault for Cosmos DB to get/warp/unwrap key permissions.
  - Finally create cosmosdb account with key URI in encryption settings. This is at account level not DB level.
- Using CMK – Azure Doc

Data Access

- Access can be via AD IAM permissions or via Keys and Resource Tokens.
  - Account Management activities like master key rotation,global replication etc are available via AD only
  - Keys and resource token allow control of data operations which AD does not
  - Restrict user access to data operations - Azure Doc
- Master keys (primary & secondary) can be regenerated and rotated
  - Move secondary to primary and then generate new secondary key. Ensure all applications are using secondary key to connect
  - Resource tokens can be generated via mid-teir for end devices like mobile

Secure Data Access

- IAM
  - Cosmos DB operator cannot read data, but can admin account, db and containers. Neither can he access the keys
  - Cosmos DB admin changes can be locked down to prevent changes from key based access - "disableKeyBasedMetadataWriteAccess"
- IP address whitelisting
  - Access to cosmosdb can be limited to specific IP address or IP CIDR block
  - By using service endpoint access can be limited to certain subnet in VNET
  - This is similar to how this works for SQL Databases

Reference Architecture

- CosmosDB
- CosmosDB with IOT

Azure Data Lake

Architecture

- Based on blob storage, supports hadoop filesystem (built on Yarn)
- Hierarchical storage with file as unit of storage
- Both Blob and ADLS apis supported
- Supports access teirs and lifecycle policies – Hot, Cold(30 days),archive(180 days)
- Supports Diagnostics and events
- Is supported by data factory, databricks, eventhub, logic apps,ML, stream analytics HDInsight, Azure Data Explorer

Unsupported Features in Data Lake

- Custom domains not supported
- Logging to Azure monitor not supported
- Does not support snapshots

Data Access

- RBAC
- Shared Key and Shared Access Signature
- ACL on file and directories

- RBAC vs ACL
  - RBAC is resolved first and takes precedence. If access is approved based on RBAC then no ACL check is performed.
  - RBAC does not provide file / directry level access control
- Shared Key vs SAS
  - Shared key allows super user access, while SAS tokens have granular permissions along with duration permissions attached to it

Blob Storage Encryption and network access

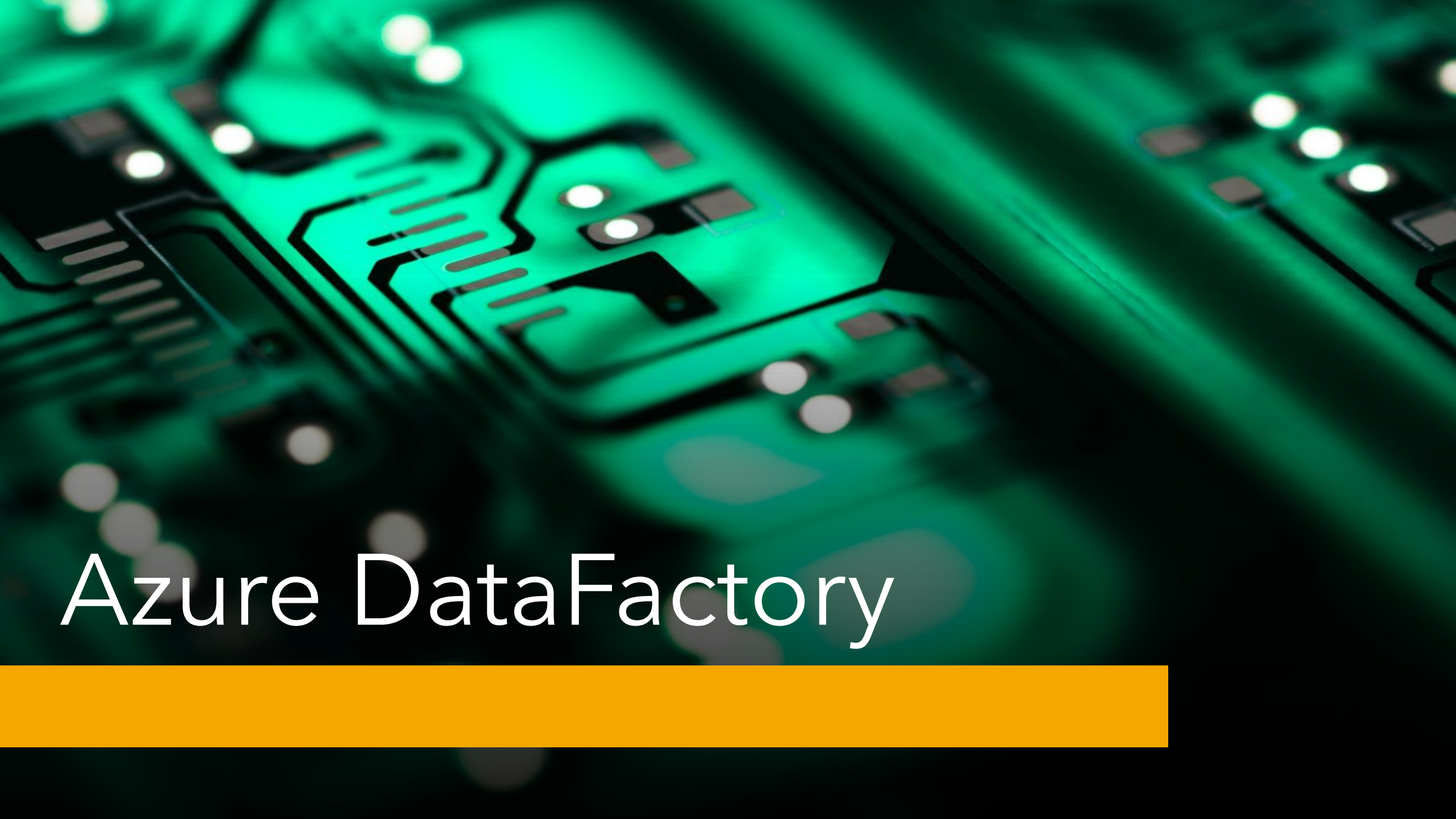
- Can use Microsoft managed or customer managed or customer provided
- Customer managed keys are in key vault in the same region as storage account
- Customer provided keys are passed along with requests and are managed by customers either on Az-Keyvault or any other vault
- Secure data transfer settings can be changed (default is enabled) from within configuration
- Use storage firewalls, network whitelists and private endpoints to secure your storage access

Data Redundancy Options

- LRS data is replicated 3 times in single DC/Zone, synchronously
- ZRS data is replicated 3 times in 3 zones with regions, synchronously
- GRS data is replicates to secondary regions in addition to LRS. Data is not available for read and is copied asynchronously
- GZRS data is replicated to secondary region in addition to ZRS. Data is not available for read and is copied asynchronously
- RA- GRS or RA-GZRS – use this option to make data available for reads in secondary regions
- Azure Doc – Data Redundancy Options

Blob Storage – Disaster Recovery and High Availability

- On regional failover, GRS is changed to LRS
- Use Last-sync-time to identify data lost
- You can change LRS again to GRS after failover
- Manual failover options is under geo-replication in storage account
  - This process updates the DNS entry
- Azure Doc – Disaster Recovery and Failover
- Azure Doc – Design Application for HA



Azure DataFactory

ADF Excercies

- Best way to go through ADF is to do hands-on, below are the links which cover required range of topics:-
- ADF Overview
- ADF Create / Implement
- ADF Using CMK
- ADF – COPY Data
- ADF – Mapping Data Flows
- ADF – Use Key Vault secrets
- ADF – e2e LAB

ADF – Integration Runtimes

- IR Overview and when to use which one
- IR – Create Azure IR
- IR – Create Self hosted IR

ADF – Triggers

- There are 3 types of triggers in ADF
  - Event based
    - Only `supports Data lake Gen2. This means we can use it to orchestrate batch but not stream processing.
  - Scheduled
  - Tumbling window
    - It is recurring event, but alows backfill runs and concurency controls – Read here
  - Azure Doc – Event based , Scheduled , Tumbling window


Azure Databricks

Azure Databricks on Microsoft Learn

Microsoft Learn – Databricks Overview

- Covers overview, create workspace, ceate notebook and attach to spark cluster

Microsoft Learn – Stream processing with Databricks

- Connect to Eventhub and process streaming data

Microsoft Learn – Security with KeyVault

Access Control

- Access Control in Databricks – Azure Doc
  - All 6 modules in this chapter cover access control for various parts of databricks.
  - Access is at two levels, first admin has to enable access controls, then these are used by relevant users to grant permissions. Enable Access Control

When to use what compute

- Azure Doc – Databricks Runtime
  - Read through all 4 sub-topics


Azure Stream Analytics

Overview and Implement

- Microsoft Learn - Overview and Implement azure stream analytics
- Azure Doc – when to use ?
  - For real-time alerts and dashboards, IOT Edge
- Input for stream analytics are:-
  - Event Hub
  - IOT hub
  - Blob storage

Stream Analytics Solutions

- Build with IOT Edge - Tutorial
  - Cloud part is responsible for job definition – input, output, query
  - IOT edge pushes the job to device
  - Stream analytics on Edge runs the job

- Process data from EventHub
  - Azure Doc – Process data from eventhub. For this exercise write your own code to send data to EventHub which SA can then process. Simple number streamer would do.

Solutions

Solution Exercises

- Azure Doc : Azure Datalake -> Databricks -> Synapse
- Azure Doc - Using ADF to transform CosmosDB Data
- Azure Doc – Processing events with Databricks