Structured Data. --> Relational Database
Semi Structured Data --> No Sql(key-value pairs, documents…..)
Unstructured Data --> Images, videos…..
Data Formats:
Relation database
NoSQL database
Analytical databases
Object/Block/File stores
Data Stores:


Structured Database:
Data stored in tables -- Rows and columns
Redefined schema 
Can define indexes for performance
OLTP (Online Transaction Processing)
OLAP (Online Analytics Processing)
Use cases:


Large no of users make large number of transactions
Use cases: Banking, ecommerce
Databases: Oracle, MySQL, SQL Server
Azure SQL Database:
Azure Database for MySQL:
Azure Database for PostgreSQL
Azure Managed Services:
Relational Databases for OLTP :

Applications allowing users to analyze petabytes of data
Use Cases: Reporting applications, Business Intelligence applications, Data Ware houses, Analytical systems
Provides a unified experience for developing end-to-end analytics solutions
Enabled massively parallel processing (MPP)
Run complex queries across petabytes of data
Azure Synapse Analytics: Petabyte-scale distributed data ware house 
Azure Managed Services:
Relational Database for OLAP:

OLTP and OLAP use similar data structures
Very different approach in how data is stored
Each table row is stored together
Efficient for processing small transactions
OLTP database use Row storage 
Each table column is stored together
High compression: Store petabytes of data efficiently
Distributed data: One table in multiple cluster nodes
Execute single query across multiple nodes -- complex queries can be executed efficiently
OLAP database use Columnar storage
Difference between OLTP and OLAP 



Semi Structured Data:
  
Semi Structured Data:
Data has some structured BUT not very strict
Semi Structured data is stored in NoSQL databases
Horizontally scale to petabytes of data with millions of TPS(Transactions per second)
Key-value pair like HashMap
Managed Services: Azure Cosmos DB Table API, Azure Table Storage 
Key-Value pairs: 
JSON documents
Managed Services: Azure Cosmos DB SQL API and Azure Cosmos DB MongoDB API
Document:
Stores data as a Graph in Graph database
Managed Service: Azure Cosmos DB Gremlin API
Graph
Data organized in to Rows and Columns
Can appear similar to relational database
Columns are divided into groups called column-family
Managed Service: Azure Cosmos DB Cassandra API
Column Family
Types of semi structured data

Azure Cosmos DB:
Managed Services:


Unstructured Data:
More like a hard disk attached to your system. Will be linked with 1 Virtual Machine
Azure Managed Service: Azure Disks
Block Storage:
Managed Service: Azure Files
File Storage: Ex: To share set of file in enterprise
Managed Service: Azure Blob Storage
Object Storage: Ex: Upload or Download objects using a REST API without mounting them onto VM




Relational Database:

Managed Services:


1
Azure SQL Database
2
Azure SQL Managed Instance
3
Azure Database for PostgreSQL
4
Azure Database for MySQL
5
Azure Database for MariaDB (retiring from Jan 2024)
Structured Query Language:

Data Definition Language: Create or Modify structure of the database objects
Create, Drop, Alter statements 
  
Create, Drop, Alter statements 

Data Manipulation language: Insert, update or delete data

Data Query Language: Perform queries on data

Data Control Language: Manage permissions and other controls 
Grant, Revoke 

Transaction Control Language: Control transactions with in a language
Commit, Rollback


Azure SQL Database:
Fully managed service for Microsoft SQL Server
99.99% availability
Built-in high availability, automatic updates and backups
Flexible and responsive serverless compute 
Hyperscale(up to 100 TB) storage
Transparent data encryption(TDE) - Data is automatically encrypted at rest




Authentication: SQL Server Authentication and Active Directory ( and MFA)

Cosmos DB API option:


1
Document DB
MongoDB API, SQL API
2
Key Value 
Table
3
Graph
Gremlin
4
Columnar
Cassandra
5



SQL API --> Cosmos DB --> Database --> Container--> item
MongoDB API --> Cosmos DB --> Database --> Collection --> Documents 
Table --> Cosmos DB -->Table --> Entities --> Key-Value
Gremlin --> Cosmos DB --> Database --> Graph --> Node or Edge
Cassandra --> Cosmos DB --> Key space--> Tables --> Rows



Azure Storage:


Azure Disks, Azure Files, Azure Blob Storage, Azure Table storage
It is managed storge service
Its highly available, durable, and massively scalable (up to few petabytes)  
Storage Types

  
Azure Disks: Block storage for Azure VMs
Azure Files: File shares for cloud and on-premises
Azure Blobs: Object store for text and binary data
Azure Queues: Decouple applications using messaging
Azure Tables: NoSQL store (very basic)
Storage Types
To create Azure Storage we should have Storage account as prerequisite

Durability : Can be achieved by have data in multiple location.

Option
Redundancy
Discussion
Locally redundant storge (LRS)
Three synchronous copies in same data center 
Least expensive and least availability
Zone redundant storage(ZRS)
Three synchronous copies in three AZs in the primary region
NA
Geo Redundant storage(GRS)
LRS + Asynchronous copy to secondary region (three more copies using LRS)
NA
Geo Zone Redundant storage(GZRS)
ZRS + Asynchronous copy to secondary region ( three more copies using LRS)
Most expensive and highest availability 



Block Storage: Hard disk attached to VM

Standard HDD: Recommended for Backups, non critical infrequent access
Standard SSD: Recommended for web servers, lightly used enterprise application and dev/test environments
Premium SSD: Recommended for production and performance sensitive workloads
Ultra disks (SSD): Recommended for IO intense workloads such as SAP HANA, top tier databases and other transaction heavy workloads 
Types:

Premium and Ultra provide very high availability

Managed vs Unmanaged Disks:
Azure handles storage
High fault tolerance and availability 
Managed disks are easy to use:
You need to manage storage account 
Disks stored in container (not docker container)
Unmanaged disks are old and trick ( if possible avoid )


Azure Files:
Media workflows need huge shared storage for things like video editing 
Enterprise users need quick way to share files in a secure and organized way
Use cases:
Managed File shares 
From cloud or on-premises
From different OS: Windows, Linux and MAC
Connect from multiple devices concurrently 
Support server message block and Network file system protocols

  
Support server message block and Network file system protocols
Shared files between multiple VMs (ex configuration files)

Azure Blob:

Object storage in Azure

Structure --> Storage Account > Container > Blob

Store massive volumes of unstructured data 

 Block Blobs: Store text or binary files ( Videos or archive)
Append Blobs: Store log files ( ideal for append operations) 
Page Blobs: Foundation for Azure Iaas Disks 
3 types of Blobs:

Designed for enterprise big data analytics (exabytes and hierarchical)
Low-cost, tiered storage with high availability / disaster recovery
Azure Data Lake Storage Gen2: Azure Blob storage enhanced


To pay cheaper price for objects stored in bloc storage has different access tiers

Access Tiers:
HOT: Store frequently accessed data
Cool: Infrequently accessed data stored for min 30 days
Low storage cost BUT highest access cost
Access latency: In hours
To access need to REHYDRATE(Change access type to HOT or Cool)
Archiver: Rarely accessed data stored for min 180 days

You can change access tier at any point of time 





---------------------------------------

Data Analytics:

Goal: Convert Raw Data in to intelligence 

Uncover trends and discover meaningful information
Find new opportunities and identify weakness
Increase efficiency and improve customer satisfaction
Make appropriate business decisions
Use Cases: 

Customer purchases, bank transactions, stock prices, weather data, monitoring devices, etc
Raw data can be from different sources 

Approach:

Ingestion ===> Data Processing ===> Store ( data warehouse or data lake) ===> Analyze
Ink Drawings
  
Ingestion ===> Data Processing ===> Store ( data warehouse or data lake) ===> Analyze

Ingestion: Capture Raw Data from various sources (Stream or Batch)

Data Processing: Raw data is not suitable to analytics so need to cleanse the data like clean ( remove duplicates), filter (remove anomalies) and aggregate data. Transform data to required format.

Data Storage: Store to data warehouse or data lake

Data Querying: Run queries to analyze data

Data visualization: Create visualization to make it easier to understand data and make better decisions.


Data Analysis Categories:

Based on historical / current data
Monitor status and generate alerts
Example: Generating reports
Descriptive Analytics: What's Happening ?

Take findings from descriptive analytics and dig deeper
Example: Why did sales increase last month?
Diagnostic analytics : Why happening?

Predict probability based on historical data
Mitigate risk and identify opportunities
Predictive Analytics: What will happen?

Use insights from predictive analytics and make data driven informed decisions
Still in early stages
Prescriptive Analytics: What action should we take?

Combine traditional analytics techniques with AI and ML features
Example: Speech to Text, Text to Speech….
Cognitive Analytics: Make analytics tools to think like humans



Big Data Terminology:

3V's of Big Data

Volume: Terabytes to Petabytes to Exabytes
Variety: Structured, Semi Structured, Unstructured
Velocity: Stream, Batch 

Data Warehouse Vs Data lake

Data stored in a format ready for specific analysis! ( Processed data)
Typically uses specialized hardware
Data Warehouse: We store processed data in to Data warehouses and it will be in Petabytes + Compute 

Typically object storage is used as data lake
Data Lake: Typically retains all Raw Data (Compressed)
  
Typically object storage is used as data lake
Flexibility while saving cost
Perform ad-hoc analysis on demand
Analytics and Intelligence services (even data warehouses ) can directly read from data lake


3 Azure Specific Services for Data Analytics
Data integration + Enterprise data warehouse + Data analytics 
Create SQL and Spark pools to analyze data
Azure Synapse Analytics: End to End analytic solutions 

Extract Transform Load 
Extract Load Transform
Data Integration

Azure Data Factory: Fully managed serverless service to build complex data pipelines
Unify data and create BI reports and Dashboards
Power BI: Create Visualization around data



Hadoop VS Spark VS Databricks

Hadoop Based Approach:
Runs on commodity servers with attached storage (Large clusters - thousands of nodes)
Hadoop Distributed File System (HDFS): Primary Data Storage
Enables massive parallelization
MapReduce: Apps to process data, can be written in Java, Python
HIVE: Query using SQL
Apache Hadoop: Create datasets with variety of data. Get intelligence.

Really fast: Can be up to 100 times faster than MapReduce(if you make sufficient memory available) 
Supports Java, Python, SQL, R, and Scala programming language
Run Data analytics, data processing, and machine learning workloads
Has become very popular and is offered as a separate service in most cloud platforms

Apache Spark: How about processing in-memory
Centralized platform for machine learning, streaming, and business intelligence workloads
Founded by creator of Apache spark
Automated Cluster management
Data Bricks: Web based platform for working with spark


Apache Parquet:

It’s a format of storing your data it uses Columnar Storage Format


Managed Services for Hadoop and Spark in Azure

Azure HDInsight: Managed Apache Hadoop Azure Service, process big data with Hadoop, spark

Azure Databricks: Managed Apache Spark Service, focused only on running Apache Spark workloads. Can consume data from Azure SQL Database, Event Hubs, Cosmos DB
  
Azure Databricks: Managed Apache Spark Service, focused only on running Apache Spark workloads. Can consume data from Azure SQL Database, Event Hubs, Cosmos DB

Other Apache Spark Integrations:
Azure Synapse Analytics: Can run spark jobs using "Apache spark for Azure Synapse"
Azure Data Factory: Run pipelines involving Azure services like Azure HDInsight, Azure Data bricks



Azure Synapse Analytics:
Helps to develop end-to-end analytics solutions
We can do Data Integration, acts as Data warehouse and perform analytics 
Allows to create pipelines, query using SQL technologies and run Spark Jobs
It has good integration with Power BI, also integrate CosmosDB and Azure ML
Also works with streaming data
Create Workspace 
Data Ingestion: Ingest data from 90+data sources
Stream data into SQL tables
Create Pipelines
Uses Storage account
Save data in different formats like parque, JSON, CSV
Data Storage:
Mix and match SQL and Spark 
2 consumption models dedicated and serverless
Recommended for complex reporting and data indexing using polybase
SQL pool can be paused to reduce cost
SQL Pool:
Create Spark data analytics notebooks
Create spark Jobs
Recommended for data preparation and ML
Spark Pool:
Data Processing:
Azure Synapse Analytics Workflow:



Azure Data Factory:

Fully managed serverless service to build complex data pipelines:
90 + build in connector
Big Data sources like Amazon  Redshift, Google Big Query
Enterprise data warehouses like Oracle, Exadata, Teradata
Ingest data from:
Integrate with compute service like Azure HDInsight, Azure Databricks, Azure Synapse
Analytics for data processing

Build data flows to transform data 
ETL and ELT and Data Integration





-------------------------------

  

SL No
Service
Description
1
SQL Server on Azure Virtual Machine
Provide full administrative  control over SQL server instance and underlying OS for migration to Azure
2
Azure SQL Database
Fully managed service for  SQL server , recommended for cloud born applications.
3 
Azure SQL Managed Instance
Full SQL server access and feature compatibility, recommended for migrating on-premises SQL server databases. Azure SQL managed instance ONLY features: Cross database queries, Database mail support, SQL server Agent
4
Azure Database for MySQL
Fully managed MySQL database
5
Azure Database for PostgreSQL
Fully managed PostgreSQL database
6
Azure Cosmos DB
NoSQL database. Globally distributed. Core(SQL), MongoDB, Table, Gremlin, and Cassandra APIs
7
Azure Cache for Redis
Managed service for Redis
8
Azure Database Migration Service
Migrate Databases to Cloud
9
Azure Disk Storage
Store disks attached to VM
10
Azure Blob storage
Store Unstructured data - video files, database archives etc
11
Azure File storage
Create file shares or file servers in the Cloud
12
Azure Queue storage
Decouple applications using Queue
13
Azure Table Storage
Store structure data using SQL approach. Key attribute store
14
Azure Data lake Storage
Data lake build on Azure Blob storage
15
Azure Databricks
Managed Apache Spark
16
Azure HDInsight
Managed Apache Hadoop
17
Azure Synapse Analytics
End-to-End analytics solutions
18
Azure Data Factory
Data integration, Fully managed serverless service to build complex data pipelines
19
Power Bi
Create Visualizations around data 
20
Event Hubs
Receive telemetry from millions of devices







