

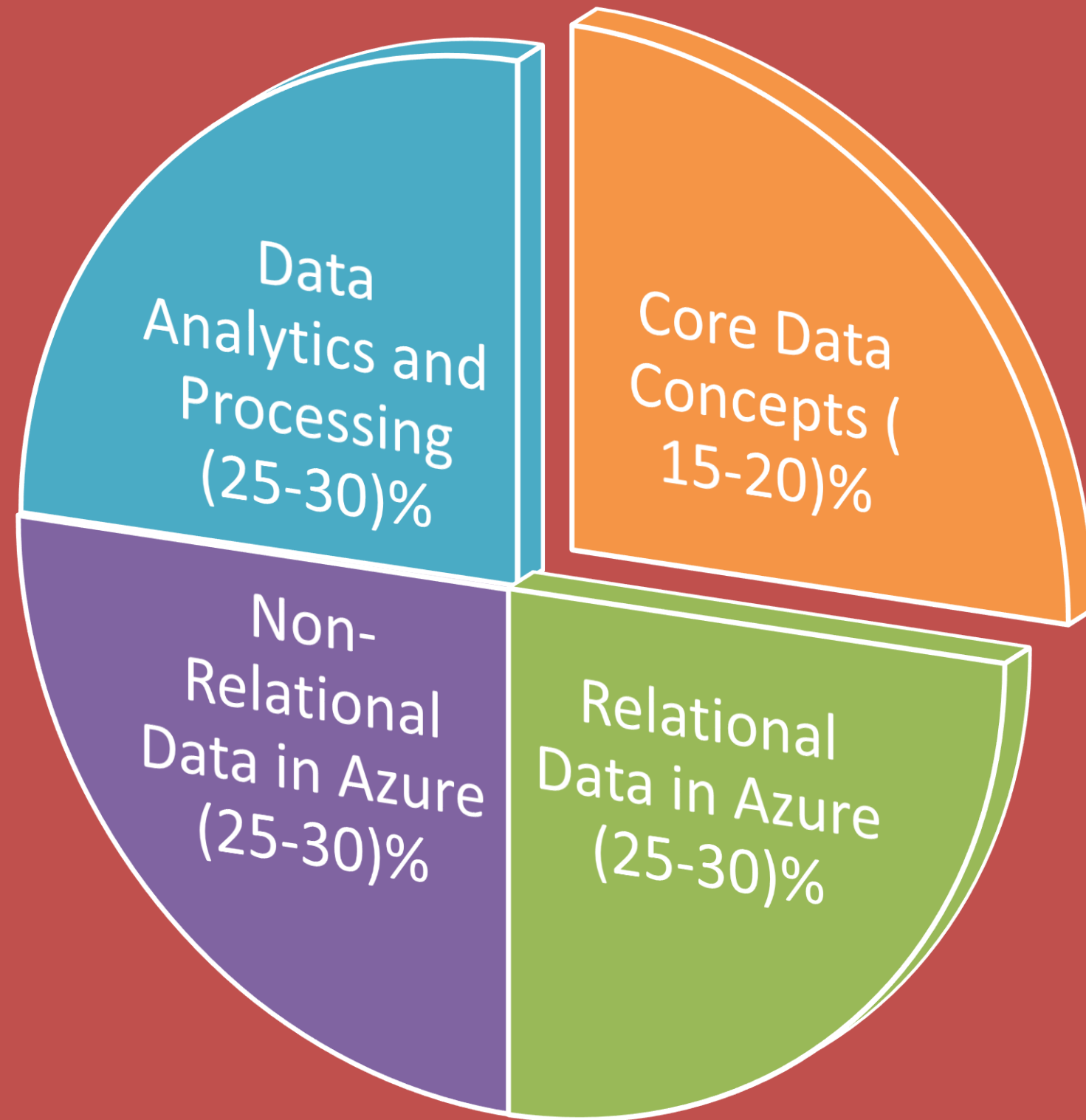
# DP-900 Azure Data Fundamentals



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Agenda

**Below topics will be covered**



- Core Data concepts
- Relational Data workload
- NOSQL Data Workload
- Data Analytics and Processing



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Core Data Concepts (15-20%)

What is Data?

Collection of facts such as numbers, descriptions, and observations used in decision making.

Structured

Semi-Structured

Unstructured

**Structured data** is typically tabular data that is represented by rows and columns in a database.

Databases that hold tables in this form are called relational databases

**Semi-structured data** is information that doesn't reside in a relational database but still has some structure to it. Examples include documents held in JavaScript Object Notation (JSON) format.

Not all data is structured or even semi-structured. For example, audio and video files, and binary data files might not have a specific structure. They're referred to as **unstructured data**.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Data processing

Data processing is simply the conversion of raw data to meaningful information through a process

Depending on how the data is ingested into your system, you could process each data item as it arrives, or buffer the raw data and process it in groups

Streaming

Batch Processing

Processing data as it arrives is called streaming

**Streaming Data:** When you play a video on Youtube, Netflix. The service streams the data through your browser In **real-time**.

Buffering and processing the data in groups is called batch processing.

**Batch processing:** Counting of votes in election where data is collected and counted in **batches**.

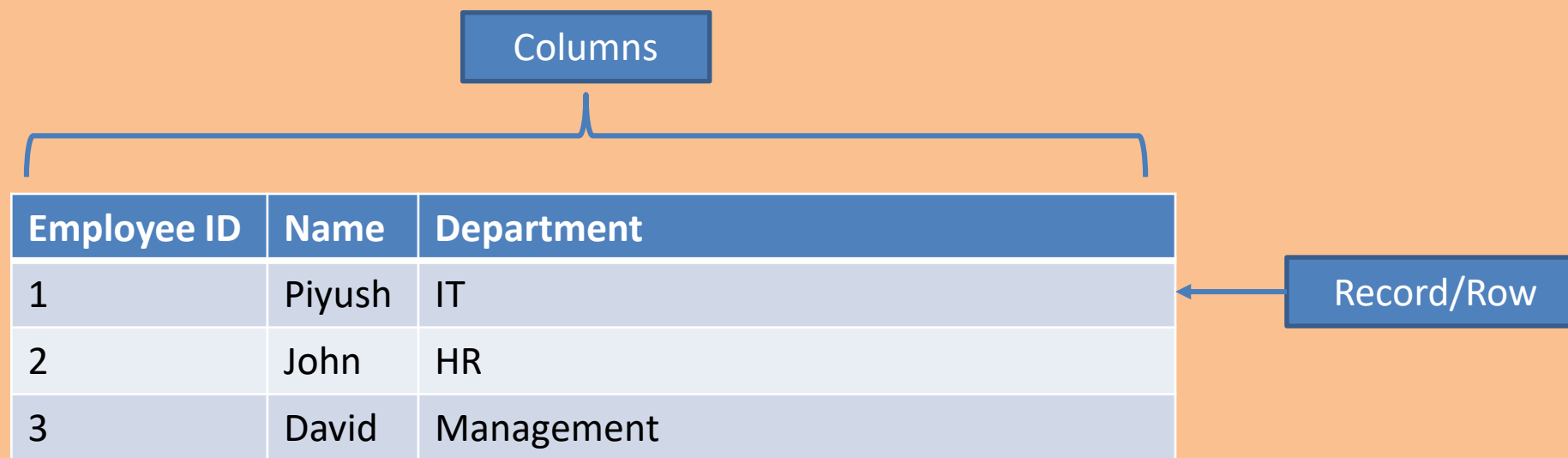


[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# RDBMS

Data represented in the form of rows and columns

Collection of related data entries are called Tables



The diagram illustrates a table structure. A blue box labeled "Columns" is positioned above the table, with a bracket indicating the header row. A blue box labeled "Record/Row" is positioned to the right of the table, with an arrow pointing to the first data row.

Employee ID	Name	Department
1	Piyush	IT
2	John	HR
3	David	Management

Collection of multiple tables and database objects : **Relational Database**



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Normalization



Store and organize relational data in most efficient manner



Improves data integrity



Create relationships between database tables



Enforces constraints and fixed schema

# SQL Commands

(DDL) Data Definition Language

Helps defining structure of database or schema

Defines how the data is stored in a database

Create	To create a database and its objects like (table, index, views, store procedure, function, and triggers)
Alter	Alters the structure of the existing database
Drop	Delete objects from Database( Tables , index, views)
Truncate	Removes all record from a Table
Comment	Add comments to a data dictionary
Rename	Rename a Database



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# DML (Data Manipulation Language)

Used to store, modify, retrieve, delete and update data in a database.

Select	Retrieve Data from a Database
Insert	Insert data into a table
Update	Update existing data within a table
Delete	Delete records from a database table



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



# Database Objects

Most of the major database engines offer the same set of major database object types:

## Table

Students

Student ID	Name	Age
121	Piyush	32
123	David	30
124	John	28

Grades

ID	Name	Grade	StudentID
101	Piyush	B	121
201	David	A	123
301	John	C	124

## Index

That helps improves the data retrieval speed

```
CREATE INDEX index_name ON table_name;
```

## View

The fields in a view are fields from one or more real tables in the database. ( Virtual Table)

```
CREATE VIEW view_name AS  
SELECT column1, column2, ...  
FROM table_name  
WHERE condition;
```

```
CREATE VIEW student_details AS  
SELECT s.Name, s.Age, g.Grades  
FROM students s, grade g  
WHERE s.studentID = g.studentID;
```

```
Select * from student_details
```

Name	Age	Grade
Piyush	32	B
David	30	A
John	28	C



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# SQL CONSTRAINTS

Rules enforced on data columns on a table.

These are used to limit the type of data that can go into a table.

These ensures the accuracy and reliability of the data in the database.

**NOT NULL Constraint** – Ensures that a column cannot have a NULL value.

```
CREATE TABLE table_name (  
    column1 datatype constraint,  
    column2 datatype constraint,  
    column3 datatype constraint,  
    ....  
);
```

```
CREATE TABLE students (  
    StudentID int NOT NULL,  
    Name varchar(255) NOT NULL,  
    FirstName varchar(255) NOT NULL,  
    LastName varchar(255)  
);
```



**DEFAULT Constraint** – Provides a default value for a column when none is specified.

```
CREATE TABLE students (  
    StudentID int NOT NULL,  
    Name varchar(255) NOT NULL,  
    Address varchar(255) DEFAULT 'India'  
);
```

**UNIQUE Constraint** – Ensures that all the values in a column are different.

```
CREATE TABLE students (  
    StudentID int NOT NULL UNIQUE,  
    Name varchar(255) NOT NULL,  
    FirstName varchar(255) NOT NULL,  
    LastName varchar(255)  
);
```



**PRIMARY Key** – Uniquely identifies each row/record in a database table.

```
CREATE TABLE students (  
  StudentID int PRIMARY KEY,  
  Name varchar(255) NOT NULL,  
  Address varchar(255) DEFAULT 'India'  
);
```

UNIQUE

+

NOT  
NULL

=

PRIMARY KEY

**FOREIGN Key** – Uniquely identifies a row/record in any another database table.

Primary Key

Foreign Key

Students

Student ID	Name	Age
121	Piyush	32
123	David	30
124	John	28

Grades

ID	Name	Grade	StudentID
101	Piyush	B	121
201	David	A	123
301	John	C	124



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

**CHECK Constraint** – ensures that all values in a column satisfy certain conditions.

```
CREATE TABLE students (  
    StudentID int NOT NULL,  
    Name varchar(255) NOT NULL,  
    FirstName varchar(255) NOT NULL,  
    Age int CHECK (Age>=18)  
);
```

**INDEX** – Used to create and retrieve data from the database very quickly.

```
CREATE INDEX index_name ON table_name;
```



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Data Integrity

- **Entity Integrity** – There are no duplicate rows in a table.
- **Domain Integrity** – Enforces valid entries for a given column by restricting the type, the format, or the range of values.
- **Referential integrity** – Rows cannot be deleted, which are used by other records.
- **User-Defined Integrity** – Enforces some specific business rules that do not fall into entity, domain or referential integrity.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# OLTP vs OLAP

Management of transactional data using computer systems

OLTP systems record business interactions as they occur in the day-to-day operation of the organization

Choose OLTP when you need to efficiently process and store business transactions and immediately make them available to client applications in a consistent way.

Business Transactions related to payments, orders, inventories etc.

Complex business analysis on large business databases.

It can be used to perform complex analytical queries without negatively affecting day to day business operations.

Choose OLAP, when you need to execute complex analytical and ad hoc queries without impacting your OLTP systems.

Reporting and forecasting, trend reports, market sentiments, recommendations and suggestions etc



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# IaaS

## Infrastructure as a Service

Gives full control over infra resources such as virtual machine /storage etc

You must take care of all the Admin tasks such as patching, upgrades, backups.

Pay-per-use

Azure VM, VNET,  
AWS EC2 servers

# PaaS

## Platform as a Service

Give runtime environment/platform To deploy application and Development tools.

Azure takes care of all the admin tasks including automated backups

Pay-per-service model

Azure DevOps, Azure Web App,  
OpenShift

# SaaS

## Software as a Service

Gives access to the end users

Azure takes care of all the admin tasks.

Pay-per-subscription model

DropBox, Office 365 , Teams



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



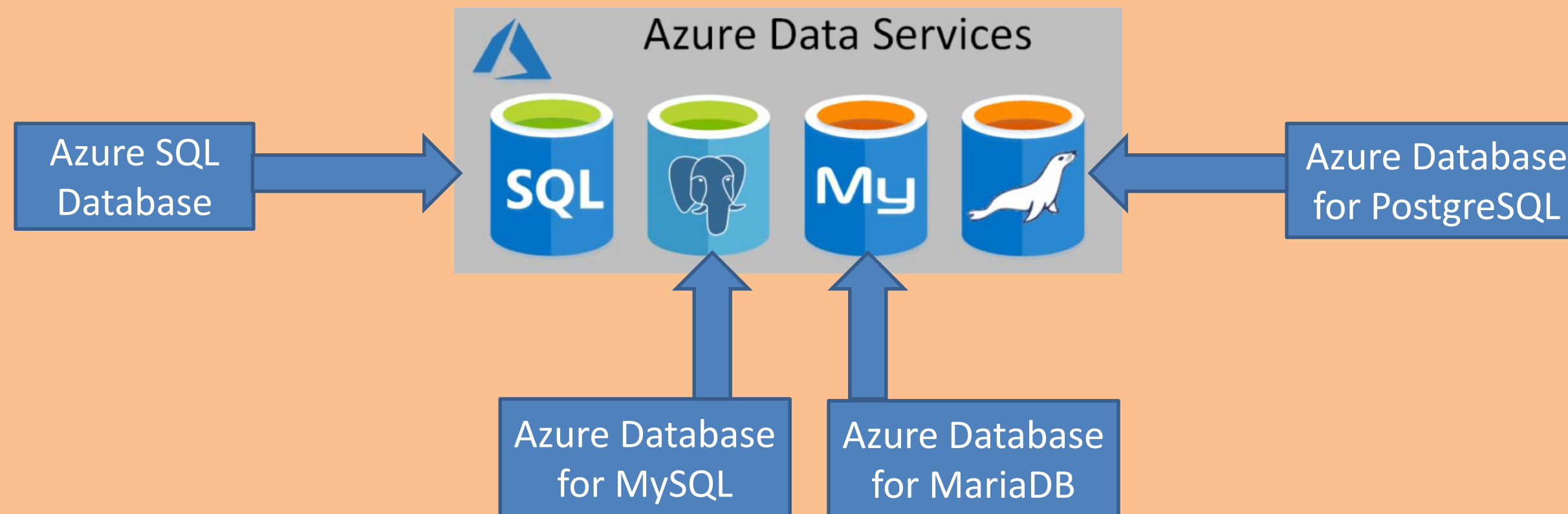


How to work with Relational Data on Azure (25-30%)

# Azure Data Services for RDBMS

Azure Data Services fall into the PaaS category.

These services are a series of DBMSs managed by Microsoft in the cloud.



Microsoft takes care of all your administrative tasks including server patching, backups and updates.

You have no direct control over the platform on which the services run.

By default, your DB is protected by a server level firewall



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure SQL Database ( PaaS)

Single Database

Elastic Pool

Managed Instance

This option enables you to quickly set up and run a single SQL Server database.(Cheapest)

By default, resources are pre-allocated, and you're charged per hour for the resources you've requested

You can also specify a serverless configuration. Your database automatically scales and resources are allocated or deallocated as required.

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.

The resources are referred to as a pool. You create the pool, and only your databases can use the pool.

You are charged per Pool.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure SQL Database ( PaaS)

## Managed Instance

Managed instance effectively runs a fully controllable instance of SQL Server in the cloud

You can install multiple databases on the same 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 instance has near 100% compatibility with SQL Server Enterprise Edition, running on-premises.**

**Consider Azure SQL Database managed instance if you want to lift-and-shift an on-premises SQL Server instance and all its databases to the cloud, without incurring the management overhead of running SQL Server on a virtual machine. (BYOL)**



**[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)**

# SQL Server in a Virtual Machine ( IaaS)

- 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
- You can easily move your on-premises SQL Database to Azure VM (Windows/Linux).
- This approach is suitable for migrations and applications requiring access to operating system features that might be unsupported at the PaaS level.
- SQL virtual machines are lift-and-shift ready for existing applications that require fast migration to the cloud with minimal changes.
- You get all the cloud benefits such as scalability, elasticity, high performance with no limitation of DBMS.
- You remain responsible for maintaining the SQL Server software and performing the various administrative tasks to keep the database running from day-to-day.



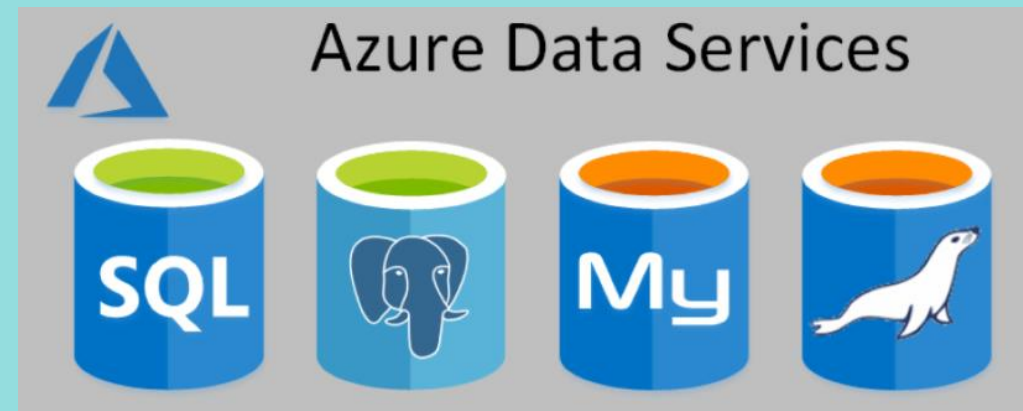
[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

IaaS

PaaS

SaaS

SQL Server in Virtual  
Machine



Azure SQL Database

Azure Database for MySQL

Azure Database for PostgreSQL

Azure Database for MariaDB

Single Database

Elastic Pool

Managed Instance



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



How to work with Non-Relational Data on Azure (25-30%)

# Non-Relational DB (NoSQL)

NoSQL database stands for “Not Only SQL” or “Not SQL.”

Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured data.

Doesn't follow fixed schema structure

Doesn't support features of a relational database

## Types of NoSQL Data Stores

Documents	High volume of JSON data
Graphs	Relationship between nodes and edges with graph
Key-Value	Multiple key-value pairs
Column based	Columns are divided into column families which hold related data
Object based	Unstructured/semi data storage for binary large object: images, videos, VM disk image





# Azure CosmosDB

- Azure Cosmos DB is a multi-model NoSQL database management system.
- Cosmos DB manages 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 partition keys for high performance/query optimization

- Example

```
## Document 1 ##
{
  "customerID": "101",
  "name":
  {
    "first": "Piyush",
    "last": "Sachdeva"
  }
}
```

```
## Document 2 ##
{
  "customerID": "102",
  "name":
  {
    "title" : "Mr"
    "firstname": "Piyush",
    "lastname": "Sachdeva"
  }
}
```



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# CosmosDB APIs

**SQL API** Enables you to run SQL queries over JSON data.

**Table API** This interface enables you to use the Azure Table Storage API to store and retrieve documents.

**MongoDB API** Many organizations run MongoDB(**document-based DB**) on-premises. You can use the MongoDB API for Cosmos DB to enable a MongoDB application to run unchanged against a Cosmos DB database or you can migrate MongoDB to CosmosDB in the cloud.

**Cassandra DB API** is a **column-based** DBMS ,the primary purpose of the Cassandra API is to enable you to quickly migrate Cassandra databases and applications to Cosmos DB.

**Gremlin API** The Gremlin API implements a **graph** database interface to Cosmos DB. A graph is a collection of data objects(Nodes) and directed relationships(Edges). Data is still held as a set of documents in Cosmos DB, but the Gremlin API enables you to perform graph queries over data.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure Table Storage

Azure Table Storage implements the NoSQL key-value model

In this model, the data for an item is stored as a set of fields, and the item is identified by a unique key.

Items are referred to as rows, and fields are known as columns.

Unlike RDBMS, it allows you to store unstructured data

Simple to scale and allows upto 5PB of data

Fast read/write as comparable to a relational DB, use partition key to increase performance.

Row insertion and data retrieval is fast.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure Blob Storage

Azure Blob storage is a service that enables you to store massive amounts of unstructured data, or blobs, in the cloud.

Many applications need to store large, binary data objects, such as images, video, virtual machine Images and so on. These are called Blobs.

Inside an Azure storage account, you create blobs inside containers(folders). You can group similar blobs together in a container.

## Block Blobs

Set of blocks

Each block vary in size,  
up to 100MB

Up to 100MB

## Page Blobs

Collection of fixed size pages

512-bytes each

Supports random read/write

## Append Blobs

Optimized to support append operations

You can only add blocks to the end of an  
append blob

Update/deleting existing blocks is not  
supported



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure Blob Storage: Access Tiers

## Hot Tier

The Hot tier is the default.

Used for Frequently access data.

Provide highest performance

Costliest among three

## Cool Tier

Used for infrequent data access

Cheaper than hot tier

Lower performance than hot tier

You can migrate your storage from Hot to cool tier to save storage cost.

## Archive Tier

Used for archival storage    Cheapest among all    Highest latency    Take hours for data retrieval

To retrieve a blob from the Archive tier, you must change the access tier to Hot or Cool.  
The blob will then be rehydrated.

You can read the blob only when the rehydration process is complete.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure File Storage

Azure File Storage enables you to create file shares in the cloud and access these file shares from anywhere with an internet connection.

Azure File Storage exposes file shares using the Server Message Block 3.0 (SMB) protocol.

Once you've created a storage account, you can upload files to Azure File Storage using the Azure portal, or tools such as the AzCopy utility.

Azure File Storage offers two performance tiers.

The **Standard tier** uses hard disk-based hardware in a datacenter

**Premium tier** uses solid-state disks. The Premium tier offers greater throughput, but is charged at a higher rate.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# NOSQL DB Suitable for?

Cosmos Cassandra API

Column based: When you need low latency, time-series, session details, telemetry data, analytics.

Object based: Store unstructured data or Blobs

Azure Blob Storage

Graph based: When you need to define relationship in form of graphs.

Cosmos Gremlin API

Key-Value: Data is accessed using a single key , used for caching, user profile mgt, session mgt.

Azure Table Storage

Cosmos Table API

Document: JSON documents for content/inventory mgt, product catalog

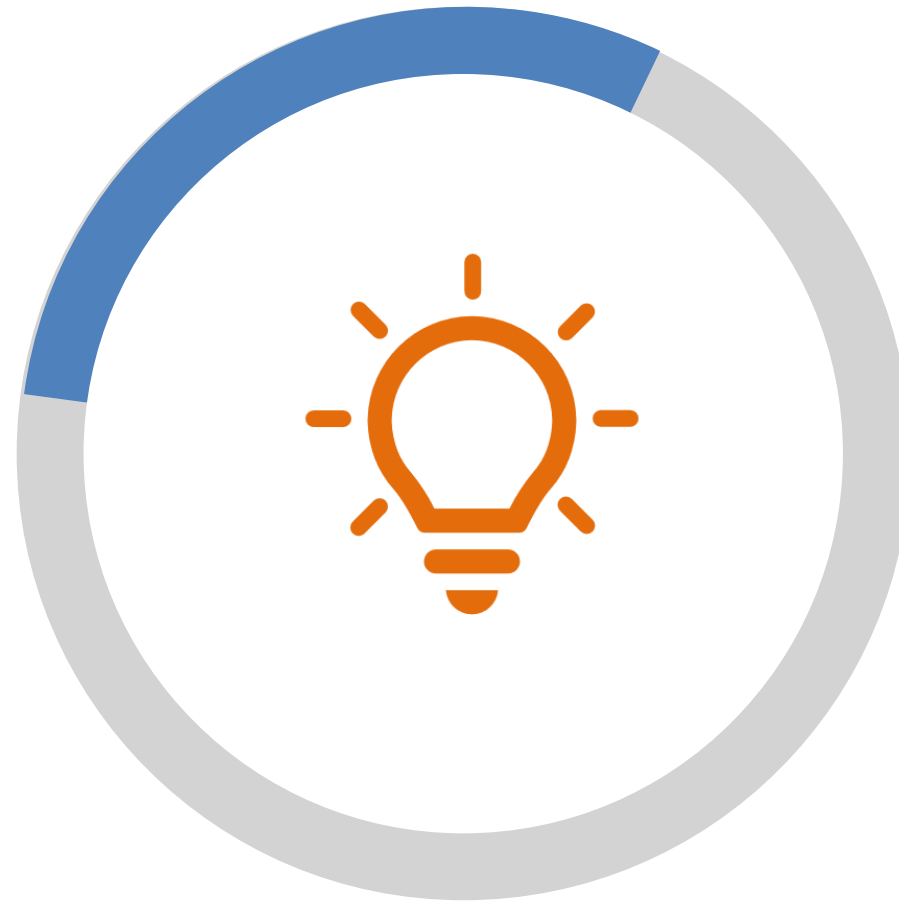
Cosmos SQL API

File share in the cloud , SMB 3.0 Protocol

Azure File Share



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



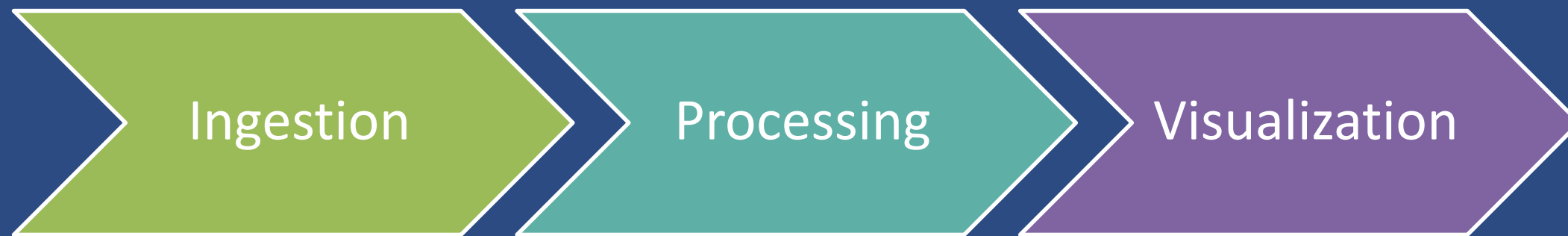
Analytics workload on Azure (25-30%)



# Data Analytics Core Concepts

Data analytics is concerned with examining, transforming, and arranging data so that you can study it and extract useful information

Data Analytics stages :



**Ingestion:** Taking the data from multiple sources into your processing system.

**Processing:** Transformation of data into more meaningful form

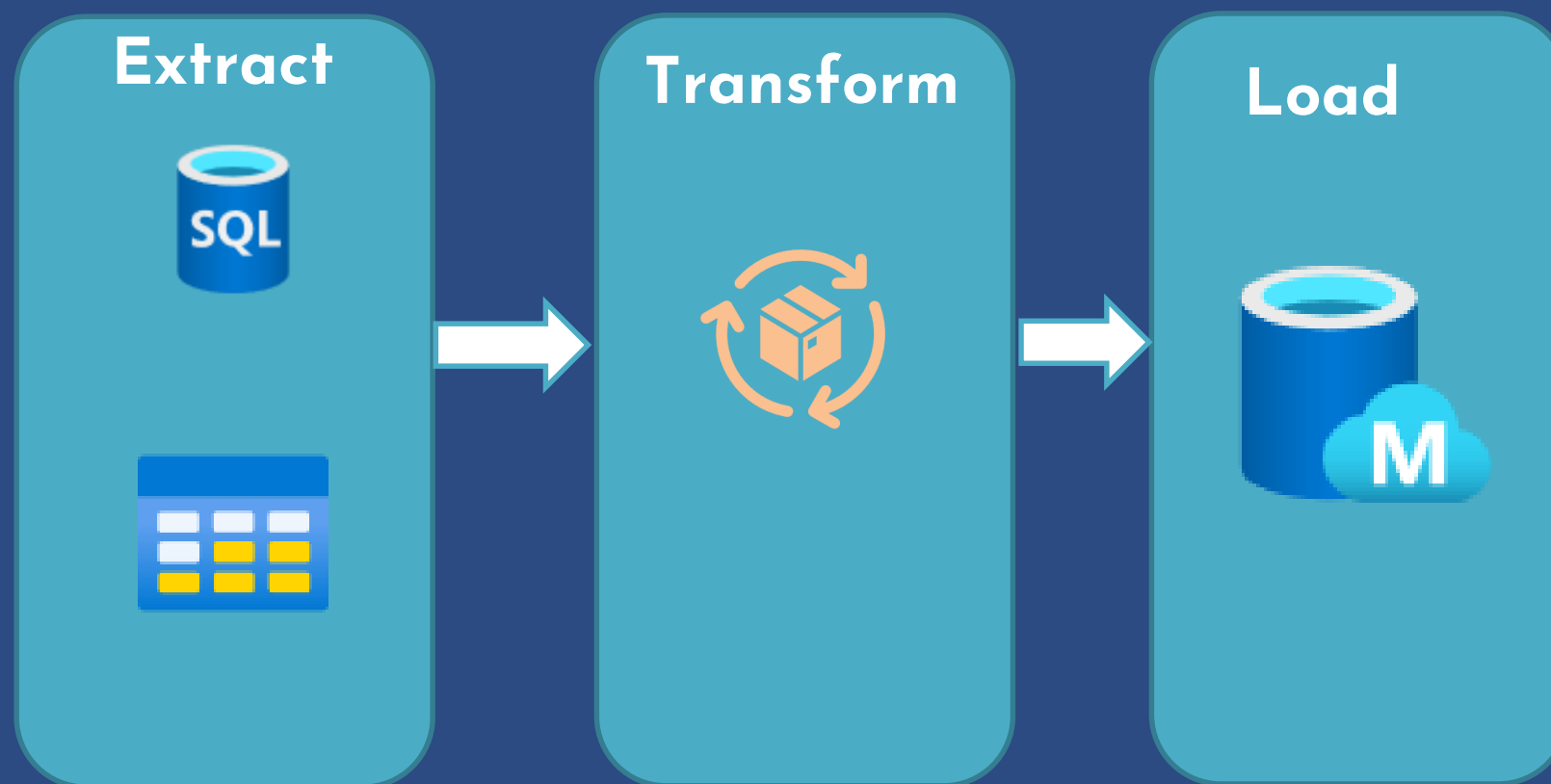
**Visualization:** Graphical representation of processed data in the form of graphs, diagrams, charts , Maps etc., for reporting and business intelligence purpose.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# ETL vs ELT

## ETL (Extract , Transform and Load)

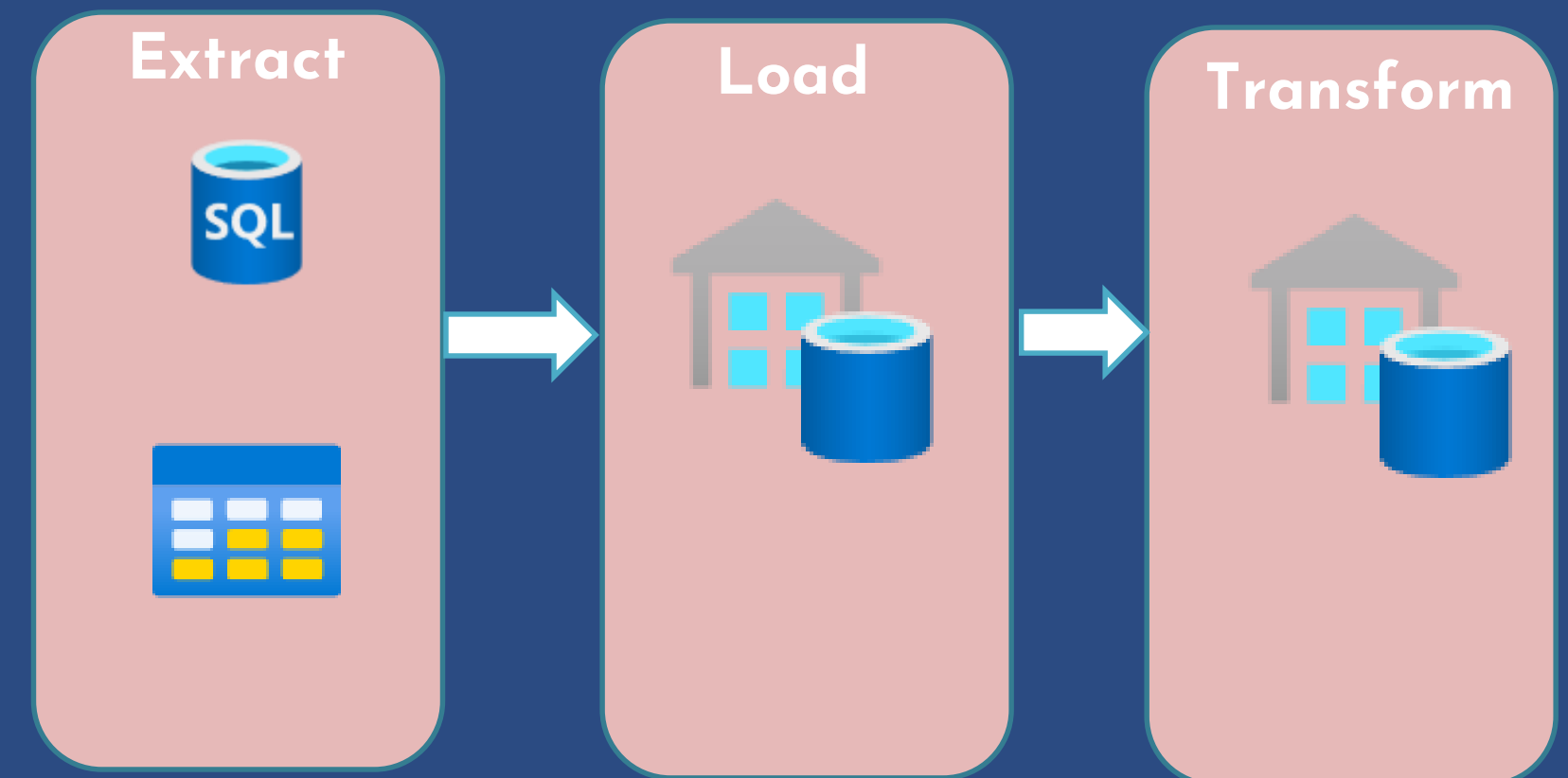


Data Ingestion  
Filtering  
Sorting

Aggregating  
Joining  
Cleaning  
De-dup

Validation

## ELT (Extract , Load and Transform)



Target data store is a data warehouse using either Hadoop Cluster or Azure Synapse Analytics.

Target datastore should be powerful enough to transform the data



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Data Analytics Techniques



## 1. DESCRIPTIVE

**What has happened, based on historical data**

Sales reports,  
profit and loss statements,  
quarterly earnings reports



## 2. DIAGNOSTICS

**why things happened.**

Comparison reports  
Drill-down reports



## 3. PRESCRIPTIVE

**What actions should we take to achieve a target**

Recommendation,  
Suggestions,  
Advise on best  
approach



## 4. PREDICTIVE

**What will happen in the future based on past trends**

Forecasting reports,



## 5. COGNITIVE

**What might happen if circumstances changes: AI/ML**

Self-driving cars,  
Video to audio conversion,  
Audio transcribing,

# Azure Tools for Data Analytics

Arm Template: To Automate Azure resource provisioning ( IaasC)

Azure CLI: Command line tool to interact with Azure resources

Azure Data Studio: Execute queries on SQL sever/big data cluster, restore a Db, execute admin tasks via sqlcmd/Powershell, Create and run SQL Notebooks

SSMS ( SQL Server Management Studio): complex admin task, platform configuration, security mgt, user mgt, vulnerability assessment, performance tuning, query Synapse Analytics

Sqlcmd: Command line SQL utility



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Data Warehousing

- Central Repository of data collected from one or more sources.
- Current and historical data used for reporting and analysis
- Can rename or reformat columns to make it easier for users to create reports
- Users can run reports without affecting the day-to-day business

## When to use data warehousing

When queries are long running and affect day to day operations

When data needs further processing (ETL or ELT)

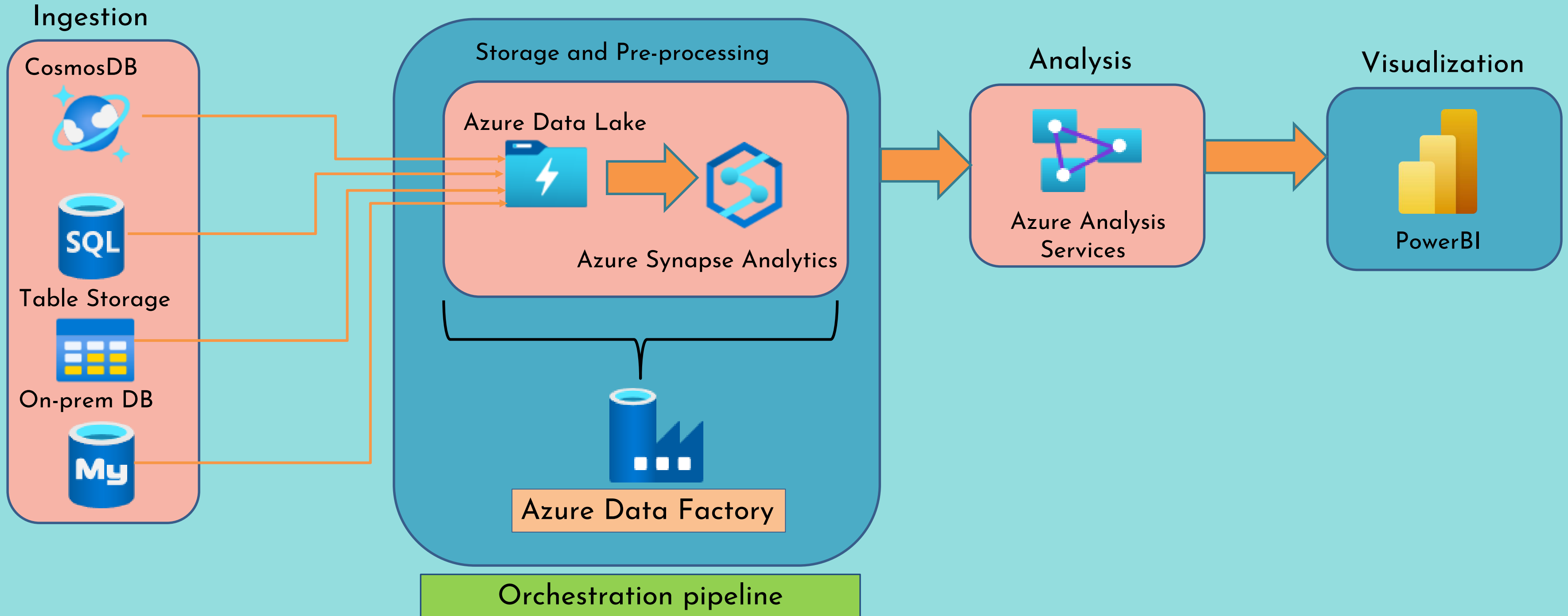
When you want to archive data (remove historical data from day-to-day system)

When you need to integrate data from multiple sources

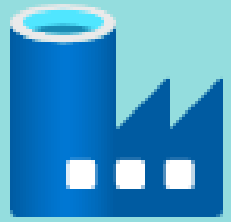


[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Data Warehousing Flow



# Azure Data Services for Data Warehousing



Azure Data Factory

Azure Data Factory is described as a data integration service. Responsible for **collection**, **transformation** and **storage** of data collected from multiple sources.

A logical grouping of activities to perform some task

A data factory can contain multiple pipelines

Sequential or parallel

## Pipeline Triggers



Scheduled trigger



Tumbling windows ( run as scheduled with the historical data)



Event-Based



Manual



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



# Azure Data Lake Storage

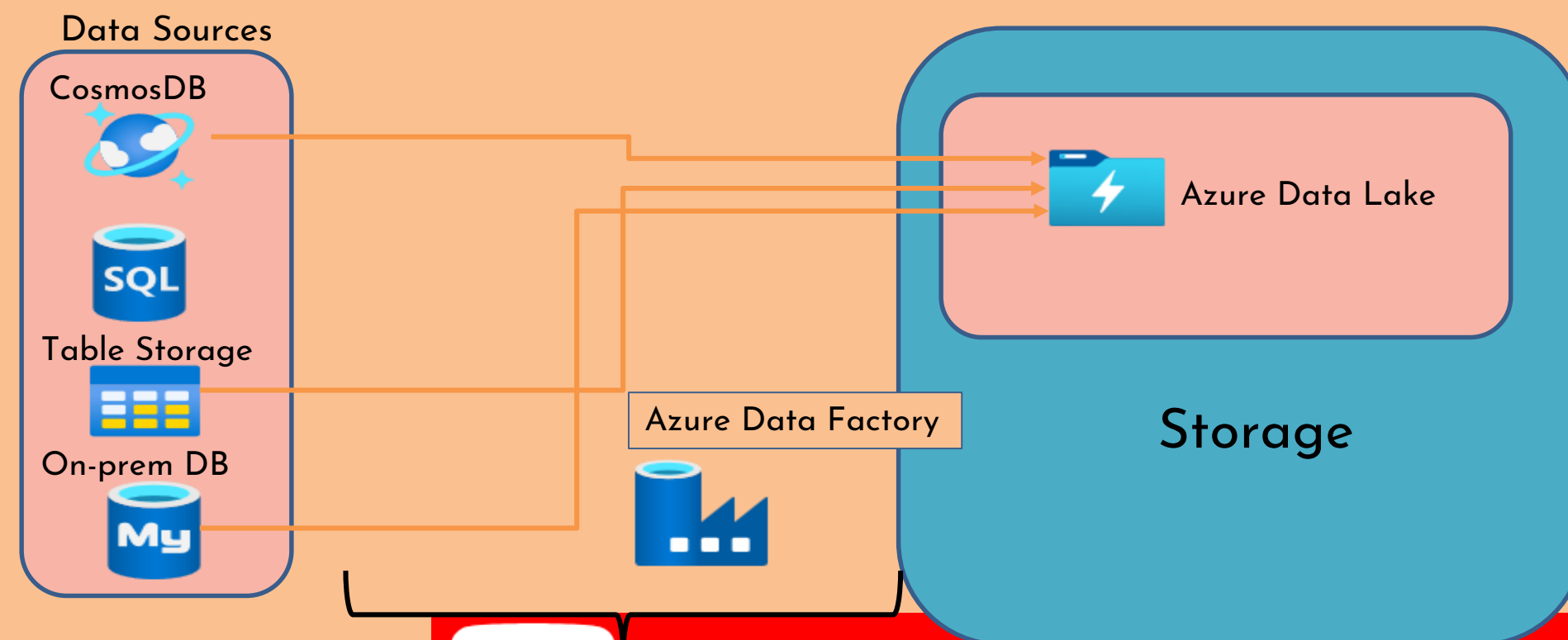
A data lake is a repository for large quantities of raw data

You can think of a data lake as a staging point for your ingested data, before it's transported and converted into a format suitable for performing analytics

Data Lake Storage organizes your files into directories and subdirectories for improved file organization. (Hierarchical Namespace)

Compatible with HDFS(Hadoop Distributed File System) used to examine huge datasets.

Role-Based Access Control (RBAC) on your data at file and directory level( POSIX access control list)



To implement azure Data Lake you need to have a storage account

It Stores data that is in parquet format

Data Ingestion

[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



# Azure Databricks

Azure Databricks is an Apache Spark environment running on Azure to provide big data processing, streaming, and machine learning.

Can consume and process large amounts of data very quickly.

Azure Databricks also supports structured stream processing

In this model, Databricks performs your computations incrementally, and continuously updates the result as streaming data arrives.

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.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



# Azure Synapse Analytics

You can ingest data from external sources, such as flat files, Azure Data Lake, or another database management systems, and then transform and aggregate this data into a format suitable for analytics processing

You can perform complex queries over this data and generate reports, graphs, and charts.

It stores and process the data locally for faster processing

This approach enables you to repeatedly query the same data without the overhead of fetching and converting it each time.

You can also use this data as input to further analytical processing, using Azure Analysis Services.

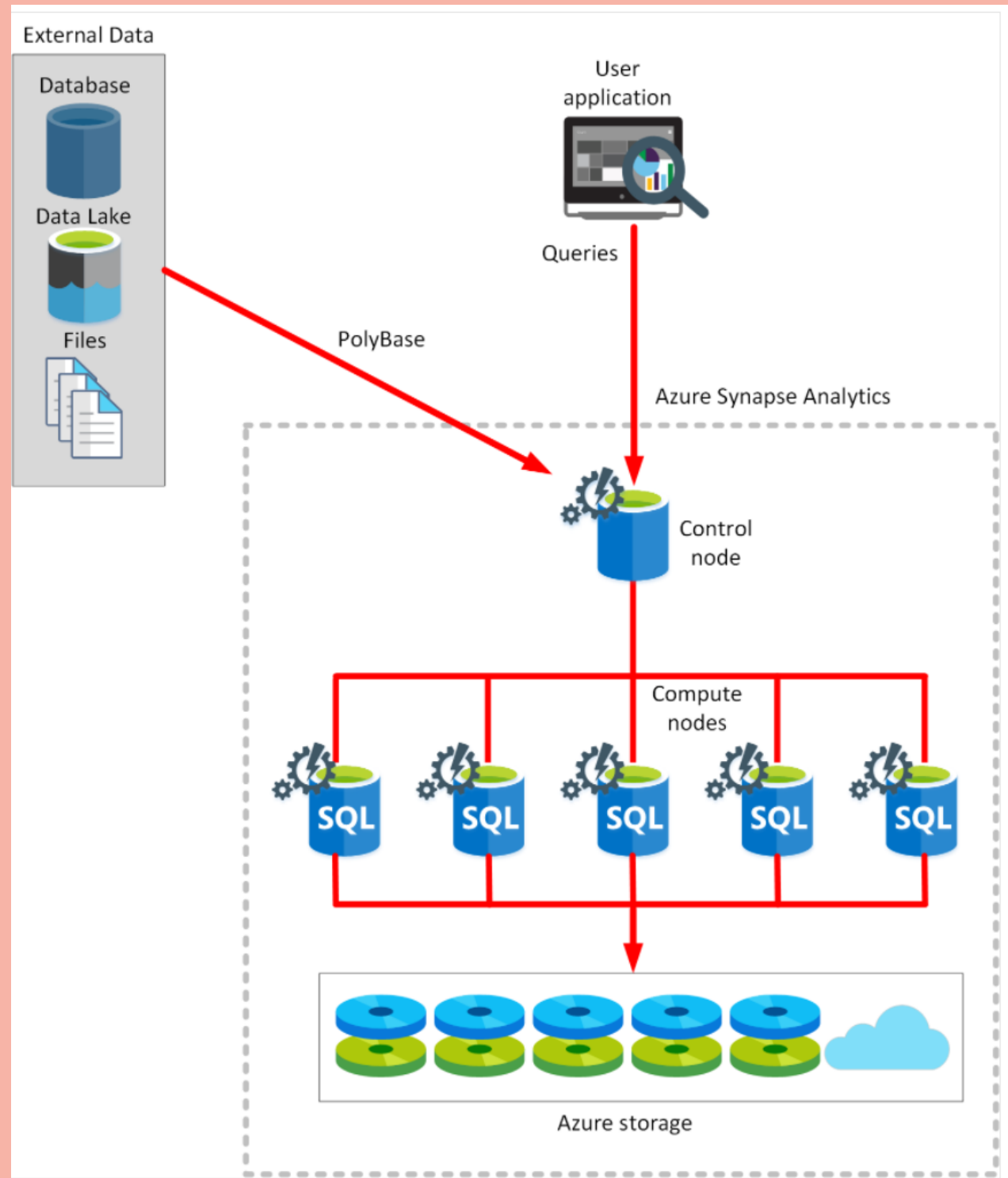
Azure Synapse Analytics leverages a massively parallel processing (MPP) architecture. This architecture includes a control node and a pool of compute nodes.

You can pause Azure Synapse Analytics to reduce cost.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure Synapse Analytics flow



It includes a control node and a pool of compute nodes

Control node receive the processing request from applications and distribute to compute nodes for parallel processing evenly.

Results from each node are then sent back to control node where it gets combined into overall result

It supports two computational models: SQL pools and Spark Pools

In a SQL pool, each compute node uses an Azure SQL Database and Azure Storage to handle a portion of the data.

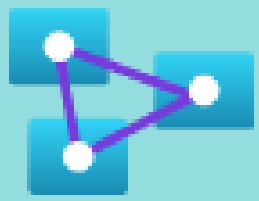
To receive data from multiple sources it uses a technology called PolyBase

It uses storage as it is a disk based processing engine and supports manual node scaling

Spark pools are optimized for in-memory processing and you can enable autoscaling of nodes.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



# Azure Analysis Service

**Azure Analysis Services** enables you to build tabular models to support OLAP queries.

You can combine data from multiple sources, including Azure SQL Database, Azure Synapse Analytics, Azure Data Lake store, Azure Cosmos DB, and many others.

You use these data sources to build models

A model is essentially a set of queries and expressions that retrieve data from the various data sources and generate results.

Analysis Services includes a graphical designer to help you connect data sources together and define queries that combine, filter, and aggregate data

## Recommended Usage

If you have large amounts of ingested data that require preprocessing, you can use Synapse Analytics to process the data and reduce into smaller datasets which can further be analyzed by Azure Analysis Service.



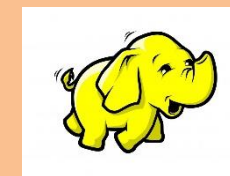
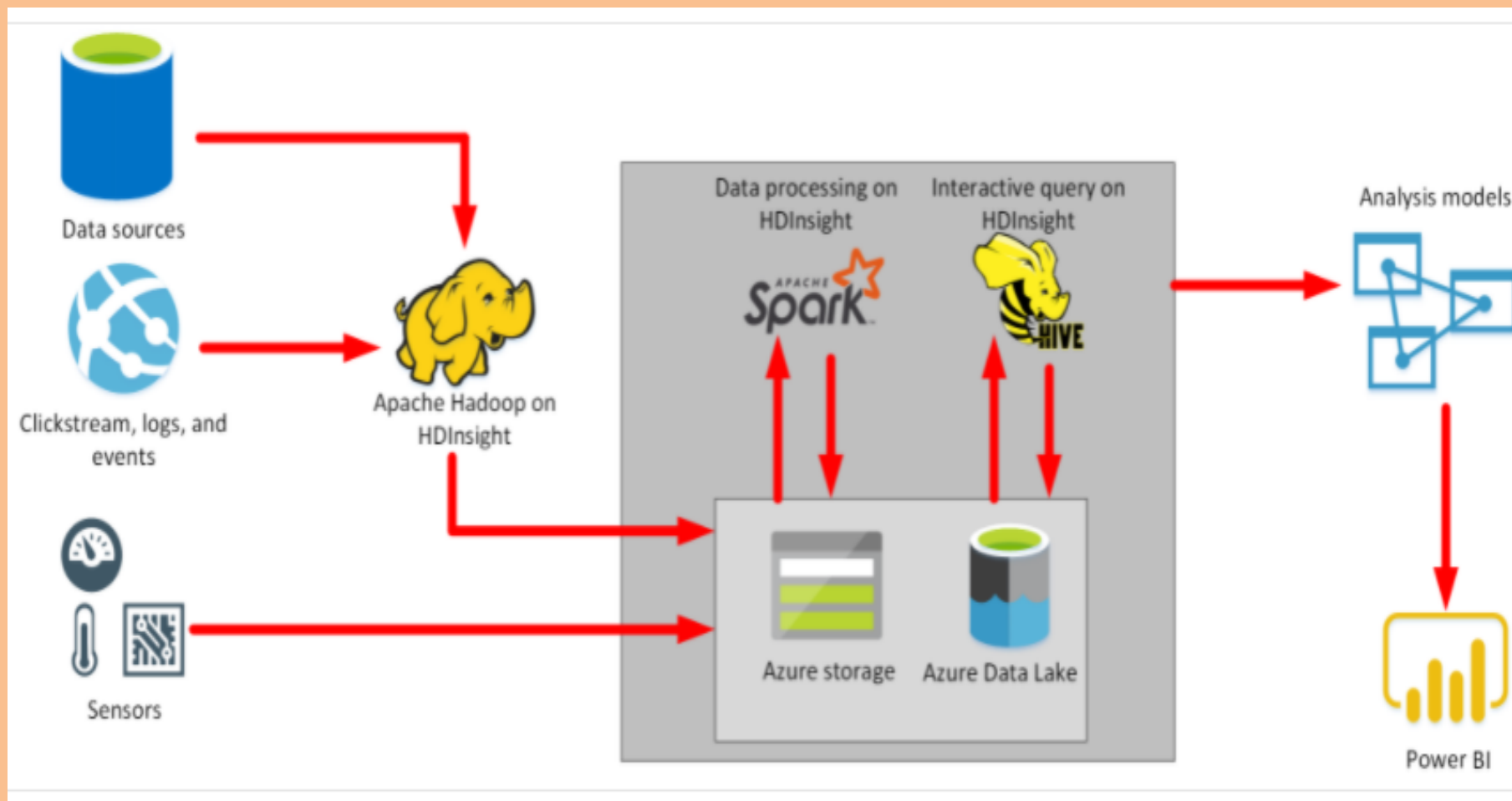
[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Azure HD Insight

Azure HDInsight is a big data processing service, that provides the platform for technologies such as Spark in an Azure environment

HDInsight implements a clustered model that distributes processing across a set of computers

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



Break down of data and distribute for processing



Data Processing



Create, load and query the data similar to PolyBase



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](http://WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH)

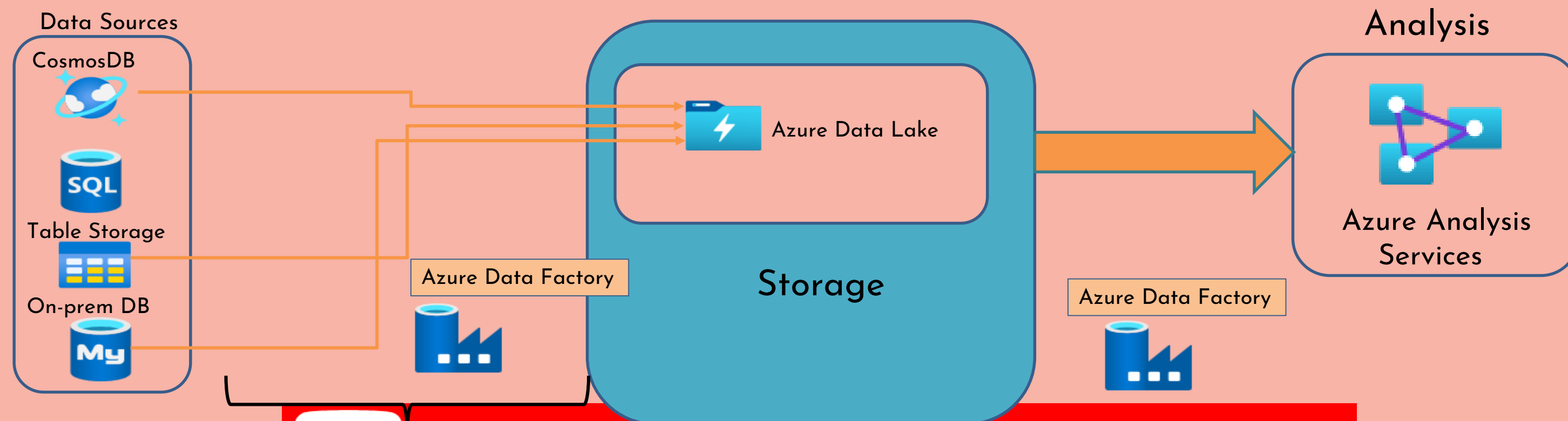
# Data Ingestion using Data factory

Azure Data Factory is a data ingestion and transformation service that allows you to load raw data from many different sources, both on-premises and in the cloud.

Data Factory can clean, transform, and restructure the data, before loading it into a repository such as a data warehouse.

Once the data is in the data warehouse, you can analyze it.

Azure Data Factory uses several different resources: linked services, datasets, and pipelines



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



# Linked Services

Data Factory moves data from a data source to a destination.

A linked service provides the information needed for Data Factory to connect to a source or destination

## Data Sets

A dataset in Azure Data Factory represents the data that you want to ingest (input) or store.

If your data has a structure, a dataset specifies how the data is structured.

For example, if you are using blob storage as input The dataset would specify which blob to ingest, and the format of the information in the blob (binary data, JSON, delimited text, and so on)



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

# Control Flow

To orchestrate a pipeline

# Integration Runtime

Compute environment for pipeline

# Trigger

That initiates the pipeline

# Mapping Data flow

Data flows allow data engineers to develop data transformation logic without writing code.



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)

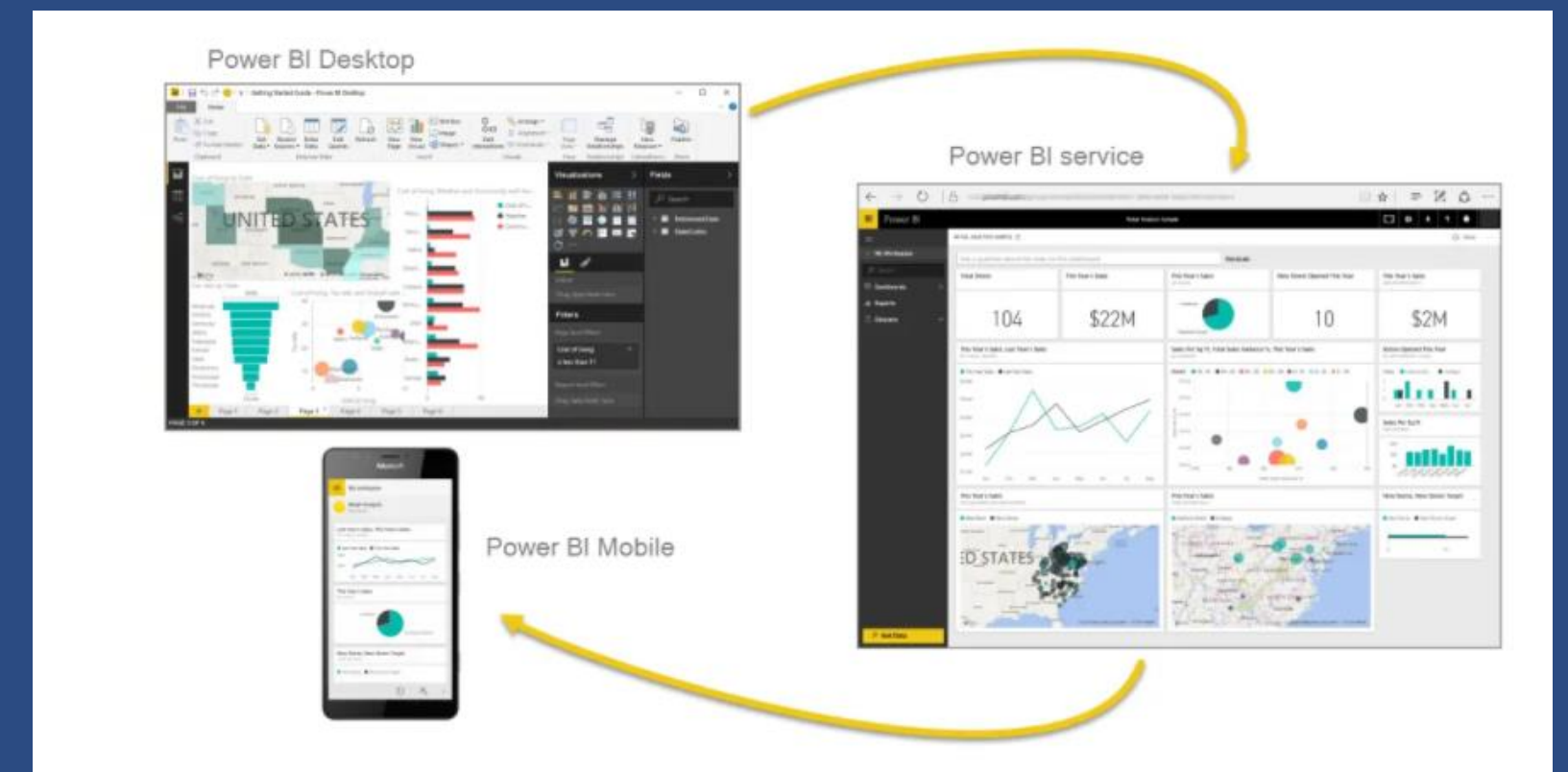


# Power BI

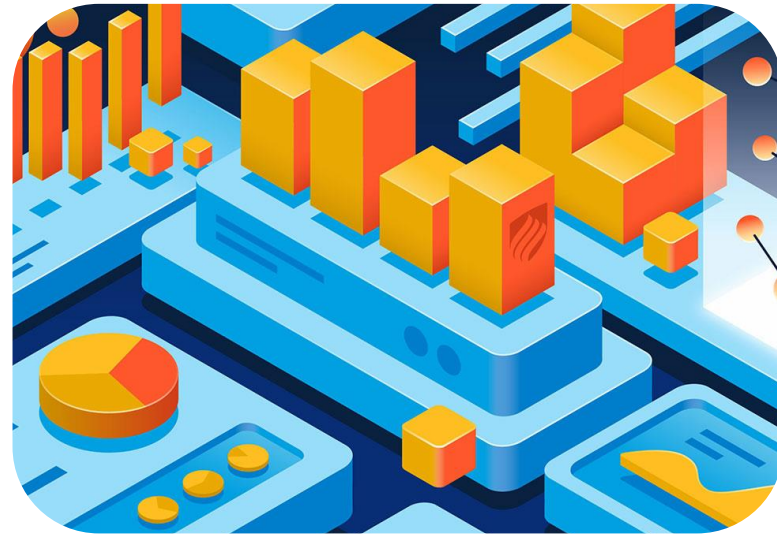
- Data visualization service which lets you generate dashboards, graphs and reports.
- Can consume data from various data sources to create interactive visualizations



## Parts of Power BI



[WWW.YOUTUBE.COM/TECHTUTORIALSWITHPIYUSH](https://www.youtube.com/techtutorialswithpiyush)



Visualizations



Datasets



Reports

# Building blocks of Power BI



Dashboards



Tiles

# Reports in PowerBI

## Paginated

---

Static Report

---

Printed and shared

---

Formatted

---

Contains data on multiple pages

---

Use Power BI report builder to create the paginated report

---

Share the report by Power BI service

## Interactive

---

Viewed on screen

---

Customized as per your requirements

---

More visuals

---

Make use of 'hover'

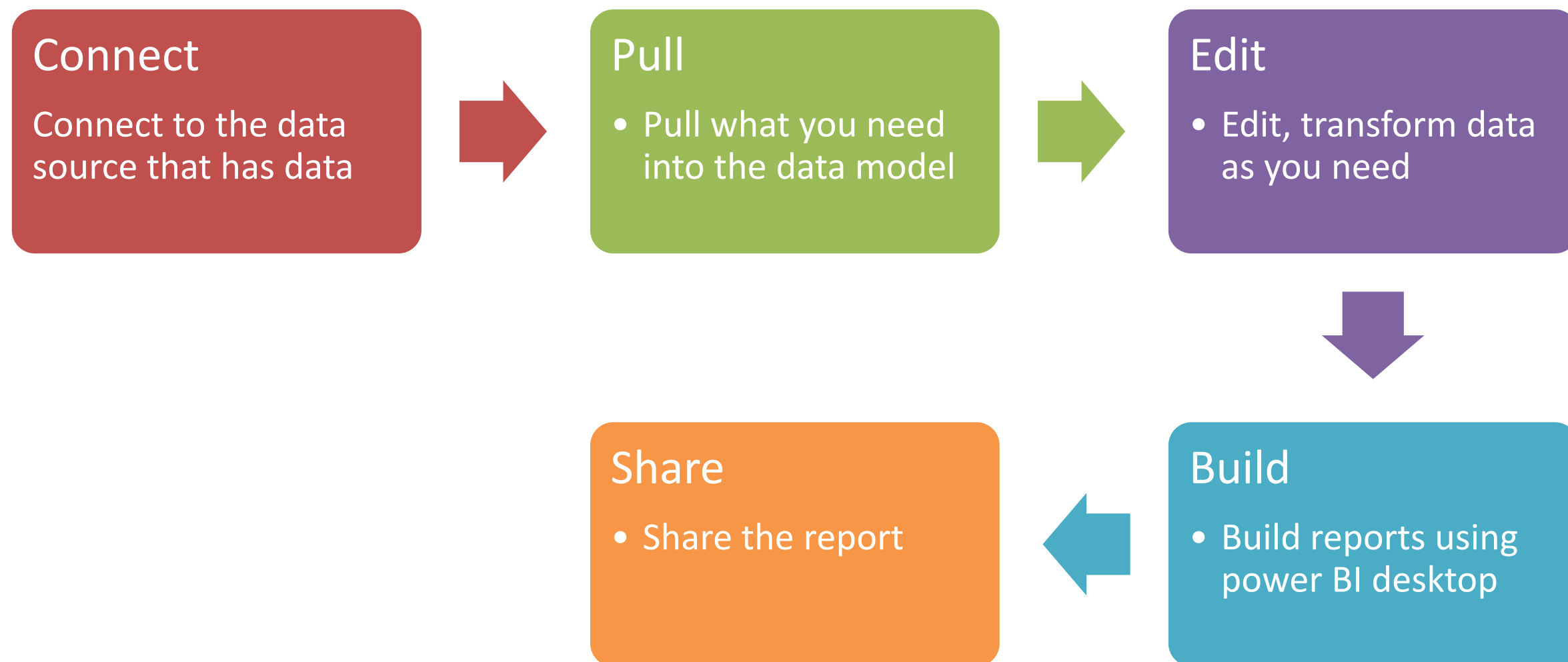
---

User can change layout of design

---

Use PowerBI server to serve the interactive reports. (Premium)

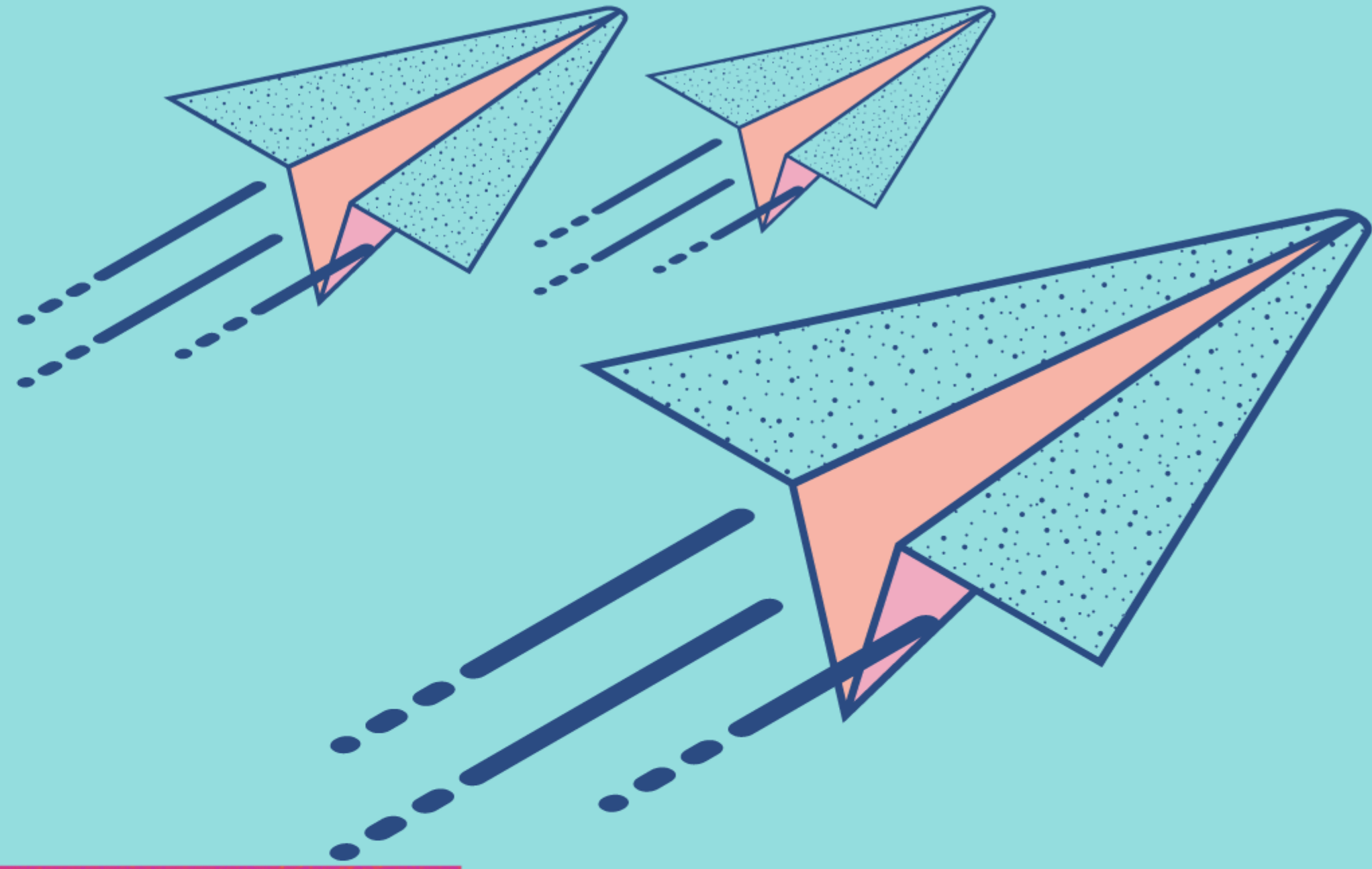
# Power BI content workflow





# Thank You For Your Patience

Watch Complete Video Below



 [www.linkedin.com/in/piyush-sachdeva/](https://www.linkedin.com/in/piyush-sachdeva/)

 [www.youtube.com/techtutorialswithpiyush](https://www.youtube.com/techtutorialswithpiyush)

