This documentation explains the role of the following Azure services in a typical data flow:

- Azure Data Factory – Data movement
- Azure Data Lake – Large-scale storage
- Azure Databricks – Data processing
- Azure Cosmos DB – Operational database
- Azure Synapse Analytics – Reporting and analytics

Each service has a specific responsibility, and together they enable efficient data engineering without overlap of concerns.

---

2. Azure Data Factory (ADF) – Data Ingestion & Movement

2.1 Overview

Azure Data Factory is a cloud-based ETL/ELT orchestration service whose primary role is to copy and move data continuously from one system to another.

Key principle:

Azure Data Factory does NOT process or transform data.

It only moves data.

2.2 Core Responsibilities

- Copy data from source systems to target systems
- Schedule and automate data movement
- Monitor data pipelines
- Support batch and near real-time ingestion

2.3 What Azure Data Factory Does NOT Do

- No complex transformations

- No aggregations
- No analytics or reporting
- No machine learning

All processing is delegated to services like Databricks or Synapse.

2.4 Supported Data Sources

ADF can connect to:

- On-premise databases
- Cloud databases
- APIs
- File systems
- SaaS platforms

2.5 Example Use Case

Copy sales data every 10 minutes from an on-premise SQL Server to Azure Data Lake for further processing.

---

3. Azure Data Lake – Centralized Data Storage

3.1 Overview

Azure Data Lake is a massive, low-cost storage system designed to store raw data at any scale.

Purpose:

Store everything, cheaply and reliably.

3.2 Key Characteristics

- Extremely scalable
- Low-cost storage
- Stores data in raw format
- Optimized for big data analytics

3.3 Types of Data Stored

Data Type	Examples
Structured	RDBMS tables
Semi-Structured	CSV, JSON, XML
Unstructured / NoSQL	Mongo-style documents, logs

3.4 Why Data Lake is Important

- Keeps raw, unmodified data
- Allows reprocessing if business rules change
- Acts as a single source of truth

3.5 Typical Zones in Data Lake

- Raw Zone – Original ingested data
- Processed Zone – Cleaned data
- Curated Zone – Analytics-ready data

---

4. Azure Databricks – Data Processing Engine

4.1 Overview

Azure Databricks is a data processing and analytics platform built on Apache Spark.

Think of Azure Databricks as:

Google Colab + PySpark + Enterprise ecosystem

4.2 Primary Role

- Process raw data
- Clean and transform data
- Perform aggregations
- Run PySpark jobs at scale

4.3 Technologies Used

- PySpark
- Apache Spark
- SQL
- Scala (optional)

4.4 Why Databricks?

- Handles huge datasets
- Distributed computing
- Very fast processing
- Tight integration with Azure services

4.5 Example Workflow

1. Read raw data from Azure Data Lake
2. Clean null values
3. Apply business logic
4. Write processed data back to Data Lake or Synapse

---

5. Azure Cosmos DB – Operational Database

5.1 Overview

Azure Cosmos DB is Azure's fully managed NoSQL database.

Used when low latency and high availability are required.

5.2 Key Features

- Globally distributed
- Very low latency
- Schema-less
- High availability

5.3 Supported Data Models

- Core (SQL) API
- MongoDB API
- Cassandra API
- Table API

5.4 Typical Use Cases

- User profiles
- IoT data
- Real-time applications
- High-traffic systems

---

6. Azure Synapse Analytics – Reporting & Analytics

6.1 Overview

Azure Synapse Analytics is used to analyze processed data and generate reports.

Final stage of the data pipeline

6.2 Core Responsibilities

- Analytical queries
- Data warehousing
- Business reporting
- Integration with Power BI

6.3 Data Sources for Synapse

- Processed data from Data Lake
- Databricks outputs
- External tables

6.4 Example Use Case

Generate monthly revenue reports for management dashboards.