AzureSynapse Analytics

Real World Projects

Interview Question

Reference c ont en t

Agenda

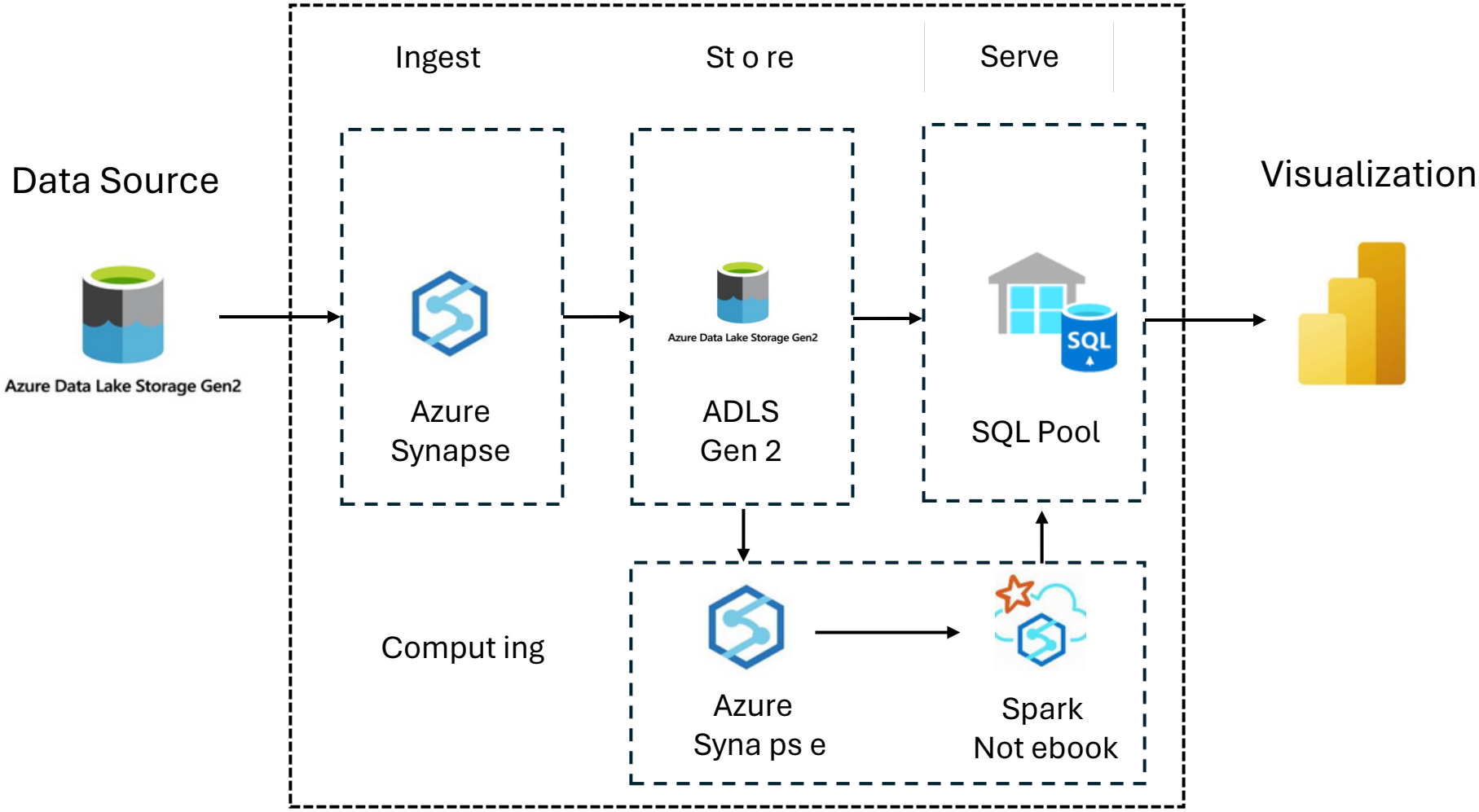
- Introduction to Azure Synapse Analytics
- Key Concepts and Architecture
- Poly base
- Distribution in SQL pool
- Synapse Spark & difference with data bricks
- Key difference in ADF and Azure Synapse Analytics
- Monitoring & Debugging
- Important Tips
- Real world projects
- Interview questions

Introduction to Azure Synapse Ana lytic s

- Unified analytics platform
- Combines big data and data warehousing
- Formerly Azure SQL Data Warehouse

Architecture

Azure Synapse Analytics



Azure Synapse Analytics
Focus: Unified Analytics Platform
Purpose: Combines data integration, big data, SQL, and data warehousing.

Feature	De scr ipt io n
Primary Use	End-to-end analytics: query, transform, visualize
Storage	Integrates with Data Lake + Dedicated SQL Pool
Querying	T-SQL (Dedicated & Serverless) + Apache Spark
Data Flows & Pipelines	Yes – similar to ADF
UI	Synapse Studio
Compute	SQL Pools, Spark Pools
Monitoring	Integrated with Synapse Studio

Key Features and Benefits

- Unified experience (SQL, Spark, Pipelines, Studio)
- On-demand or provisioned resources
- Integration with Power BI & Azure ML
- Security & compliance features
- Synapse Studio interface
- Dedicated SQL Pool vs Serverless SQL Pool
- Integration with Azure Data Lake Storage
- Pipelines, Spark Pools, Monitoring
- Ingestion using Synapse Pipelines
- Using COPY, PolyBase, Data Flows
- T-SQL support (dedicated/serverless)

What is PolyBase?

PolyBase is a technology in **Azure Synapse Analytics** (and SQL Server) that lets you **query external data** directly — **as if it were a local table** — without importing the data first.

It **connects SQL** to **external storage** (like Azure Data Lake, Blob Storage, or even Hadoop, Oracle, Teradata)

It **reads big files efficiently** using **parallel processing**.

Type	Description	When to use
Hash Distribution	Rows are distributed across nodes based on a hash function of a column value.	Best for large fact tables and joins on that column.
Round Robin Distr ibution	Rows are distributed evenly across nodes, without considering column values.	Best for staging tables, intermediate data, and simple loading.
Replicated Distr ibution	A full copy of the table is stored on **every** node.	Best for small dimension tables that are frequently jo in ed.

Fe at ure	De sc ripti on
Spark Pools	On-demand, scalable clusters managed by Synapse. No manual cluster setup needed.
Notebooks	Built-in notebooks for writing Spark code (PySpark, Scala, .NET Spark, SQL).
Language Support	Python (PySpark), Scala, C# (.NET Spark), SparkSQL.
Serverless-like Spark	You only pay for what you use. Clusters auto-pause when idle.
Data Access	Read/write from Azure Data Lake, Blob Storage, Cosmos DB, Synapse SQL tables.
ML and AI	Supports ML libraries (like MLlib) for Machine Learning workloads.
Security	Integrated with Azure Active Directory, RBAC, and encryption.
Monitoring	Track Spark job runs, resource usage, and logs within Synapse Studio.

🎯 Why Use Synapse Spark Instead of Azure Databricks?

Synapse Spark	Da ta bric ks
Integrated into Synapse Studio	Separate platform
Good for simple/medium complexity	Better for heavy ML, advanced big data
No deep MLFlow, Delta Lake features	Rich MLFlow, Delta Lake integration
Cost-effective for moderate use	More powerful, but usually costlier

Area	Azure Data Factory	Azure Synapse An aly t i cs
S cop e	ETL/ELT & orchestration	Full analytics lifecycle
Built-in Data Warehouse	❌	✅
Spark Support	❌	✅
Best For	Moving and transforming data	Analytics, querying, BI integration

Triggers and Scheduling

- Trigger types:
  1) Schedule – At defined time after fix interval
  2) Tumbling Window – On the completion of predecessor trigger
  3) Event-based – Ad hoc, based on event occurrence
- Pipeline chaining and dependencies
- Supports time zones and recurrence

SYNAPSE MONITORING PORTAL

VIEW PIPELINE RUNS AND ACTIVITY LOGS

CONFIGURE AL ER TS

INTEGRATE WITH AZURE LOG ANALYTICS

Security and Governance

- Use Managed Identity for secure au th enti cat ion
- Role-Based Access Control (RBAC)
- Monitor data lineage
- Audit logs for compliance

Best Practices

Design	Design modular pipelines
Use	Use parameterization for reusability
Op timize Monitor	Optimize data flow performance
	Monitor cost and activity
Try	Try to rerun pipeline from failed activity instead of all activity

Use Serverless SQL Pools for quick ad-hoc queries on massive raw files.

Use Dedicated SQL Pools for structured, high-performance analytics mo d el s.

Use Synapse Spark Pools for AI/ML and Big Data transformation.

Always use hash distribution for large fact tables and replicated distribution for small dimensional tables

Real-World Projects

- Build a pipeline for single source of truth for all enterprise data (sales, marketing, finance, operations).
- Combine customer data coming from different sources to create recommendation for future search.
- Real-time ingestion and analysis of digital devices data (example - smart watch) Ingest and Analyze patterns of transactional data to detect fraud or crime.
- Analyze supplier data, warehouse inventory, logistics to optimize supply chain routes and costs.
- Combine historic and real-time data to predict future requirement.
- Collect all data from social media platform to analyze any inflammatory/misguiding post.

Interview Questions

- Describe the process of creating a data pipeline What are the core components of Azure Synapse?
- Difference between Serverless SQL Pool and Dedicated SQL Pool?
- What is Poly Base and why is it used?
- Explain Synapse Notebooks.
- Explain how Spark Pools work inside Synapse.
- What is Distribution in Synapse SQL Pools? (Hash, Round Robin, R ep li ca ted)
- When should you use Serverless SQL Pool over Dedicated SQL Pool?
- How would you set up real-time analytics with Synapse?
- How would you migrate an on-premise data warehouse to Azure Syn a pse?
- Your serverless query is running very slow. How will you troubleshoot?