

# AZURE DATA ENGINEERING CRACK THE INTERVIEW

By Karthik J

<https://www.udemy.com/course/azure-data-engineering-interview-questions/?referralCode=41FC8A6544E97A2F4AB2>

itseasylearning@gmail.com  
BigDataAzure.com

The background is a dark blue gradient with faint, light blue circular patterns and a scale. The scale is a semi-circular arc with tick marks and numbers ranging from 40 to 260. The text 'BigDataAzure.com' is centered in the middle of the image.

# BigDataAzure.com

itseasylearning@gmail.com

# WHAT IS THIS COURSE ABOUT?

## This course will give you

- High confidence to face interview
- Guidance on how to articulate the architecture/design of your project
- Tips on how your answer to the point
- Quiz to test your knowledge
- Reference material links for self study

## This course will help you to

- Focus on relevant Azure services, not all you need to learn
- Self learning and practicing

## What you should **NOT** expect

- Questions and answers bank
- How to impress interviewer
- Communication and other soft skills
- Attentive, punctual etc qualities

## Before you buy this course

- Azure Data Services knowledge is required.
- Primary audience of this course is developers and architects.
- This course is available only in English language with no caption text



# STRUCTURE OF THE COURSE

Discussion point

Demo

Hints

Reference material to gain knowledge

Quiz per topic

# PRE-REQUISITES

Experience in working on Azure Data Services – Azure Data Factory, Storage services, Databricks, Azure SQL

If you do not have experience then I recommend below course on Udemy to gain project development experience

IT & Software > IT Certification > Big Data

## Fast track - Azure Data Engineering - Project Development

Azure Data Factory V2, Azure Data Lake, Azure Databricks, SQL Synapse Analytics, Azure Storage - Blob + ADLS, Azure SQL

4.2 ★★★★★ (51 ratings) 788 students



Grand opening

# TELL ME ABOUT YOUR EXPERIENCE ON AZURE



Candidate A



Candidate B



Candidate C



# TELL ME ABOUT YOUR EXPERIENCE ON AZURE



Candidate A

I have total xx number of years experience. I know Azure data factory, data lake, SQL server, Spark, Hadoop, Databricks. Also little bit of Scala.



Candidate B

I have total xx years experience. I worked on Azure data lake, data factory, Hive, Blob storage and adls. I worked on 2 projects for clients xx and xx. We copy data from multiple sources using ADF then store on storage, transform data in SQL ..... Long story



Candidate C

I have total xx years experience in Microsoft technologies, out of that I worked on Azure for xx years. I worked on Azure data engineering projects for big clients. I ingested data using batch transfer mode having size xx GB/TBs. We used ADF, Databricks, Storage services, Azure SQL. I pretty much liked it.

# TELL ME ABOUT YOUR EXPERIENCE ON AZURE

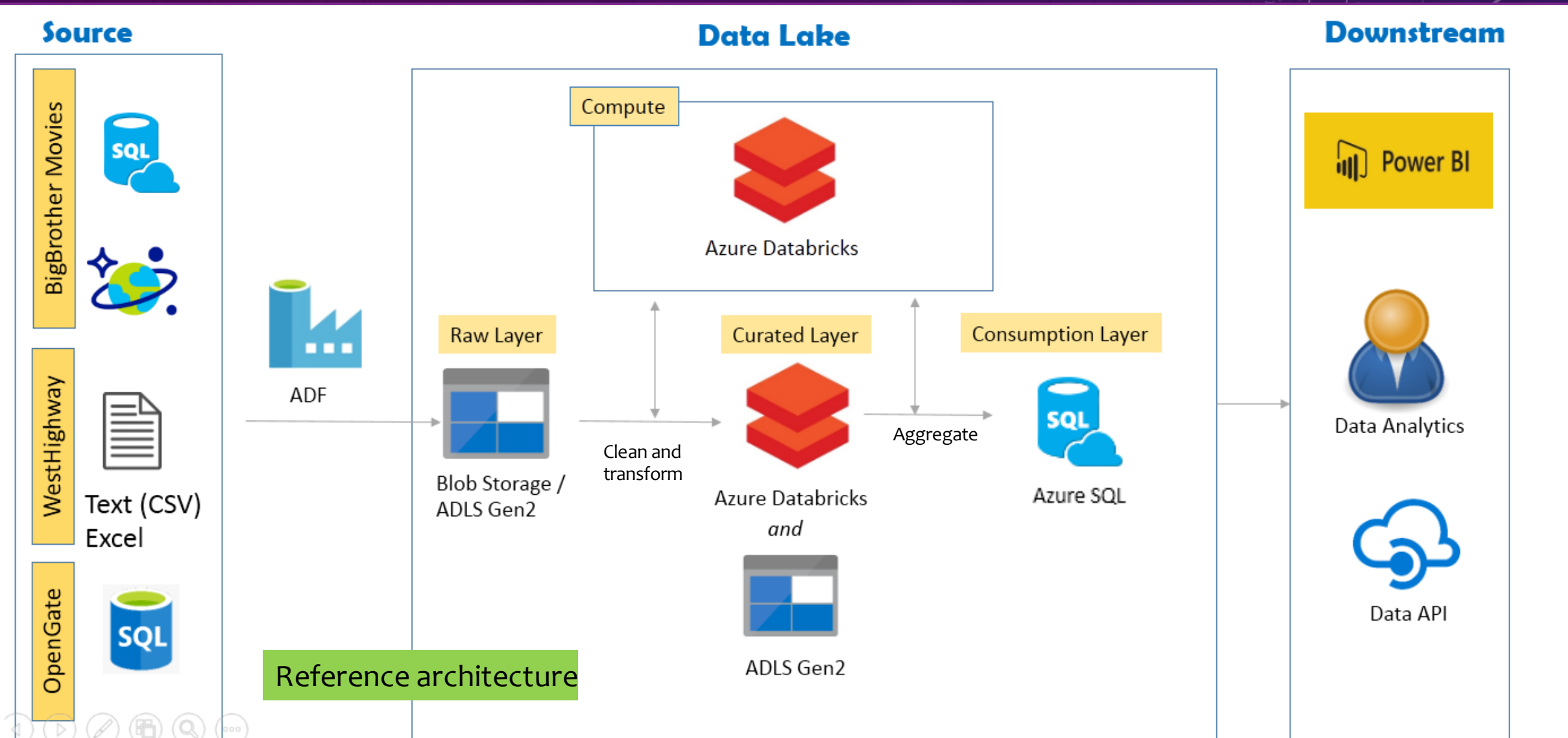
## Focus on:

- What was your role? – developer, senior developer, architect ....
- How many projects you have worked on
- Were you part of designing the solution?
- What was data size and volume?
- Batch mode and/or real time data ingestion?
- Structured / semi-structured / unstructured data
- Mention if you faced challenges or not. Don't describe those at this stage.
- List of Azure data services you have experience or confident on.

**Finish in max 2 mins**



# EXPLAIN YOUR PROJECT ARCHITECTURE



# EXPLAIN YOUR PROJECT ARCHITECTURE



Candidate A

Azure data factory copies the data from files on FTP server to blob as sink. Another copy activity reads data from blob and insert into SQL and Databricks. Then we call stored proc to transform the data.



Candidate B

There are multiple sources like database, files and web apis. ADF connects to these sources, pulls the data and stores to data lake. The data is stored inside date folders. HDInsight reads and transform data and stored in Hive tables. Jobs are scheduled on daily basis.



Candidate C

The project has layered architecture. Data is pulled from various source systems like SQL DB, files, Cosmos DB etc and kept in ingest layer in partitioned manner. Databricks is used to process the data and kept in curated layer then aggregated data is pushed to consumption purpose to SQL DB. Most of the data is structured. One time history load and daily incremental load.

# EXPLAIN YOUR PROJECT ARCHITECTURE

## Focus on:

- Your understanding of architecture
- How effectively Azure data services are used. What was the purpose? E.g. in ref architecture Databricks is used only for compute purpose
- What data is loaded? E.g. history data load, data load from heterogeneous sources etc
- Know consumers of your data – Visualization tools, data analysts, downstream applications

**Finish in max 2-3 mins**



# EXPLAIN YOUR PROJECT ARCHITECTURE

## Articulate benefits of architecture. Example:

- Decoupling of storage from compute and data processing
  - No compute power required to read data
  - Customized security roles can be defined and assigned
  - Any tool can analyze and process the data
- Centralized data architecture
  - No need to keep multiple copies for multiple purpose (viz. governance, EDW, auditing, EDM etc)
  - Cost effective
- Scalable
  - Data size no limit
  - Data from more source can be added easily
  - Real time data ingestion can plugged in
- Rest API can be leveraged to manage services and data

# OTHER STANDARD QUESTIONS

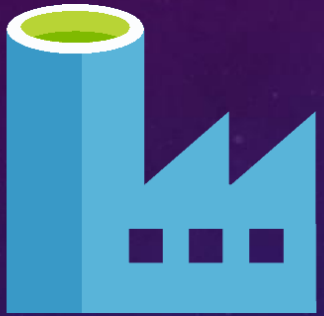
- What was team size?
- What was your role?
- What was project duration?
- Were there any difficulties in learning Azure? From where you learned Azure?
- Are you aware of current trends in the market?
- Difference between Data warehouse and data lake  
etc etc.....



# AZURE DATA FACTORY



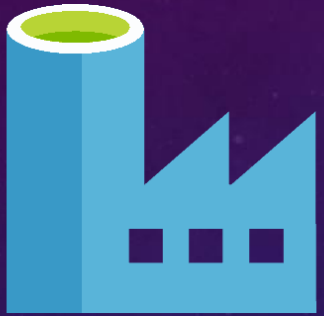
# AZURE DATA FACTORY



## BEST

## PRACTICES

# AZURE DATA FACTORY



## ERROR

## HANDLING

# AZURE DATA FACTORY



# INTERVIEW

# QUESTION



# AZURE DATA FACTORY



# MONITOR

# WINDOW

# AZURE DATA FACTORY



# ALL ABOUT COPY ACTIVITY

# AZURE DATA FACTORY

## What is Azure Data Factory?

- Do not start telling about pipelines, datasets and triggers.

Use below hints:

- It is orchestration tool. Calling ADF ETL or ELT tool is not appropriate.
  - Connects to various services like Databricks, Web, Function, Custom activity etc
- High number of connectors as source as well as sink. Not all source connectors can be used as sink.
- Capable to connect to on-prem data via Self hosted IR
- Building blocks: Pipeline, Datasets, Triggers, IR, Monitor, ControlFlow
- Inbuilt monitor with rich UI
- Integration Runtime : Cost is associated with Inbuilt IR.
- CI-CD supported thru Git
- Rest API and SDK support
- **Role based security support**



# AZURE DATA FACTORY

## What is Integration Runtime?

- Compute power to execute ADF code and to process the data
- Cost is associated (as per Data Integration Units used)
- Power (DIU) can be specified in ADF Copy Activity
- Own machines (nodes) can be attached to Self Hosted Integration Runtime (SHIR)
- SHIR can be shared between multiple ADFs
- Support to run legacy SSIS packages
- Advantages of SHIR:
  - Nodes are part of company network
  - Flexibility to upgrade drivers to connect to on-prem systems e.g. .NET driver for SAP OpenHub
  - Easily monitored and controlled

# AZURE DATA FACTORY

## What is difference between Parameters and Variables?

- Focus on advantages of parameters and variables instead of telling what are those
- Understand what to use where
- Parameters should mostly hold env dependent value (e.g. SQL server name) or dynamic values (data landing paths) which are supplied from outside of pipeline
- Variables should be used for internal functioning (e.g. building dynamic path to include date i.e. yyyy/mm/dd) or decision making. These should not be managed by external parties.

# AZURE DATA FACTORY

## Can you tell me about Copy Data activity?



Candidate A

It copies data from source to target. We have to specify from where data has to be copied e.g. SFTP and format (e.g. csv). On target also we have to specify type and format of data with additional settings like header = true..



Candidate B

It copies data from source to target using ADF connectors. Supports dynamic values as paths or table names etc. Configurable DIU and parallel connections can be specified for optimum performance.



# AZURE DATA FACTORY

## What is special in ADF Monitor?

- Supports monitoring pipeline executions, integration runtime health
- Annotations and Gantt chart
- Rest API and SDKs are available
- Set up alert notifications
- Integration with Log Analytics

# AZURE DATA FACTORY

## What are the best practices in ADF?

- Reusable (generic) dataset.
- Naming and objects arrangement
- Description field
- User properties and annotation
- Security via Key vault
- Notification on error (red path)
- Create template for pipelines

# AZURE DATA FACTORY

## Do you know how to ingest data using metadata?

See video from other course – Copy data from Azure SQL with Control Table. This applies to any source with enhancement of control table(s).

### Fast track - Learn Azure Data Engineering with Mini Project

Azure Data Factory V2, Azure Data Lake, Azure Databricks, SQL Synapse Analytics, Azure Storage - Blob + ADLS



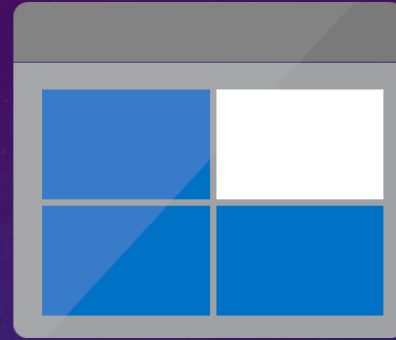
▶ [Hands-on][Project Work] Copy data from Azure SQL with Control Table



# AZURE DATA FACTORY

## How do you retrieve error details?

- Most of activities provide error as “output” or as “error”
- PreviousActivity.output.errors – collection of errors
- PreviousActivity.error – Direct error details
- Not all activities provide error in form of output or error e.g. Databricks Python activity
- The error message can be captured and used for logging and sending notification



# AZURE STORAGE SERVICES

# STORAGE SERVICES

## What is difference between Blob and ADLS?

	Blob Storage	ADLS Gen 2
Purpose	Store objects like media files, unstructured data, archival of data, log files, streaming data	Store large files (terabytes/petabytes), typically for analytics
Structure	Flat	Hierarchical
SAS URI	Yes at account, container and blob level	No (May add support in future)
Security	RBAC, Container/Blob level	RBAC, ACL, POSIX, Any level
Size Limit	Yes	No
SDK Support	Yes – Matured	Yes – Limited
REST API	Yes – Matured	Yes – Limited
Driver	wasb	abfs
Data stored as	Block Blob, Page Blob	Files
Append Data in Blob	Supported	Not supported
Tiers	Hot, Cool, Archive	Not supported
End points	blob.core.windows.net	dfs.core.windows.net

This

[questions/?referralCode=41FC8A6544E97A2F4AB2](https://www.bigtdataazure.com/questions/?referralCode=41FC8A6544E97A2F4AB2)



# STORAGE SERVICES

More detailed learning on storage with hands on

## Fast track - Learn Azure Data Engineering with Mini Project

Azure Data Factory V2, Azure Data Lake, Azure Databricks, SQL Synapse Analytics, Azure Storage - Blob + ADLS

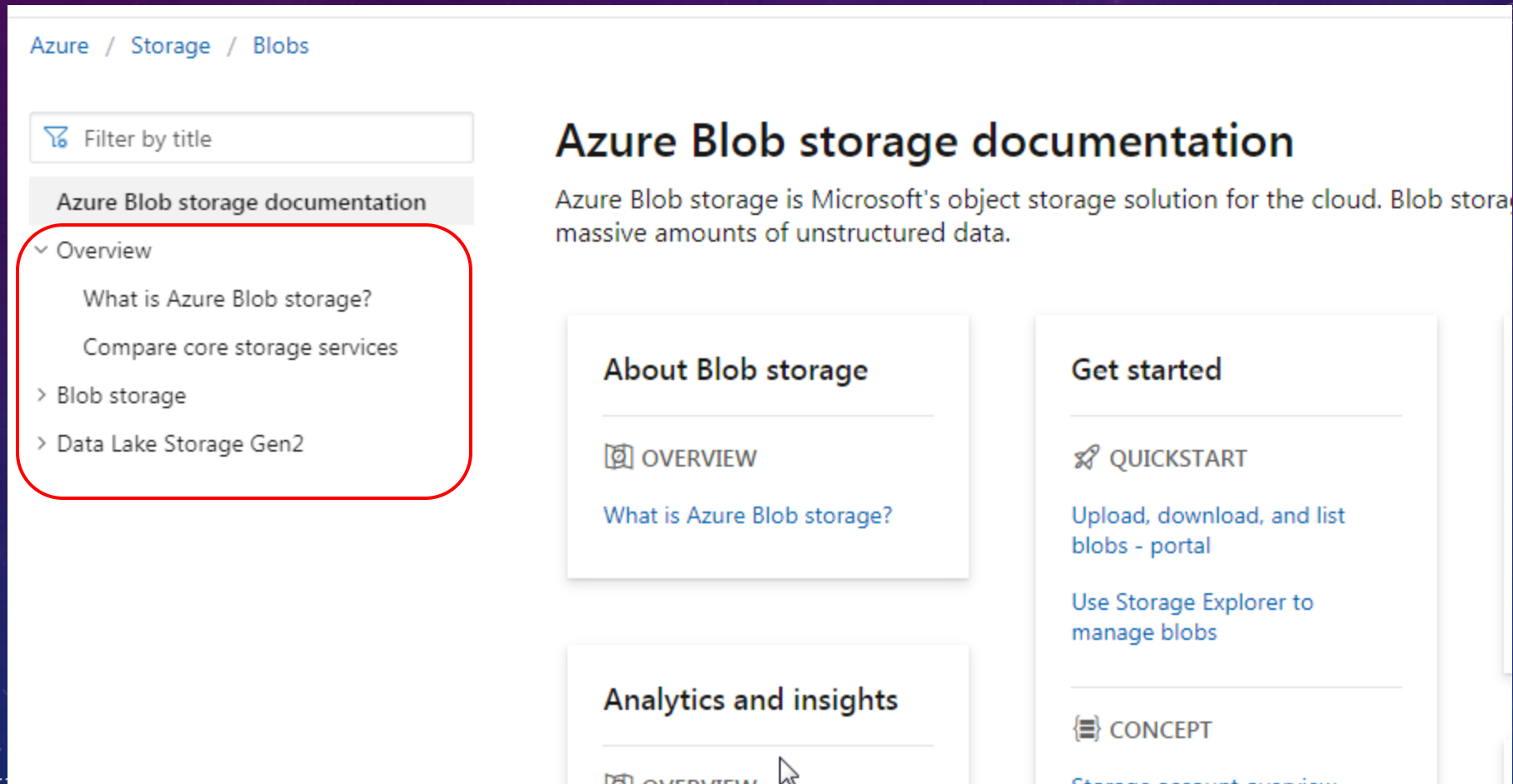


• [Hands-on] (PreSession) - Storage Services	10:05
• [Slides] - <a href="#">What and why storage comparison</a>	<a href="#">Preview</a> 08:06
• [Hands-on] Create Blob Storage and Learn More about it	15:07
• [Hands-on] Azure Storage Explorer - Account, Blobs and Security Management	16:46
• [Hands-on] Create and Manage Azure Data Lake Storage Gen 2	09:15
📁 [Assignment] [Readings] Storage Services	01:11

# STORAGE SERVICES

## What is difference between Blob and ADLS?

<https://docs.microsoft.com/en-us/azure/storage/blobs/>



Azure / Storage / Blobs

Filter by title

Azure Blob storage documentation

- Overview
  - What is Azure Blob storage?
  - Compare core storage services
- Blob storage
- Data Lake Storage Gen2

### Azure Blob storage documentation

Azure Blob storage is Microsoft's object storage solution for the cloud. Blob storage massive amounts of unstructured data.

#### About Blob storage

OVERVIEW

What is Azure Blob storage?

#### Get started

QUICKSTART

Upload, download, and list blobs - portal

Use Storage Explorer to manage blobs

#### Analytics and insights

OVERVIEW

# STORAGE SERVICES

## Tricky questions

- Can you not store media files on ADLS?
  - Yes, you can.
- Then why not?
  - Cost of ingress and egress is high on ADLS. Benefit of ADLS performance is on analytics.
- A third party team (outside of Azure env) wants to send data files to Azure. How?
  - Create container on Blob storage and share SAS URL. The team can POST data using HTTPS from any technology / programming language.
- Once I create BLOB storage, can I change it to ADLS later by changing properties?
  - No



# STORAGE SERVICES

## Why ADLS Gen2 faster than Blob storage?

Blob use WASB (Windows Azure Storage Blob) drivers which maps blob (files) each time. It adds load to code to maintain mapping of objects.

ABFS (Azure Blob File System) on other hand use Directories which reduce the vast number of operations required to retrieve data.

Tip: Instead of large number of small files, if you keep data in small number of large files, analytics performance improves by many folds.

# STORAGE SERVICES

## How do you archive the data?

- ADF Copy activity – compress the data, delete original
- Use of Lifecycle management in Azure Blob Storage
- Simply keep data as-is for longer duration – storage is very cheap



# AZURE DATABRICKS



# AZURE DATABRICKS

## Tell me about Azure Databricks

- Customized Spark engine with Apache Delta engine
- Enhanced for auto scaling, Machine learning, data management and streaming
- Offers rich UI to manage clusters, security and notebooks for code development
- Offers Delta tables – supports MERGE (ACID transactions), create views on top of tables, compressed data store
- Integrated with Azure storage, Key Vault services and Azure Active Directory
- Notebooks can be callable from ADF
- Notebook supports magic commands, version history, can be shared, returns value
- Good for data analysts - better visualization - graphs, charts in notebook
- Offer Rest API
- Job scheduler
- SQL Server like Security implementation
- Can be integrated with Git.

# AZURE DATABRICKS

## What did you not like about Azure Databricks?

- PaaS, not IaaS - customer who desire to manage infra, Databricks is not right choice
- Not every third party Spark library works on Databricks or does not work efficiently
- Can not migrate legacy Spark code as-is (security issues e.g. accessing inbuilt temp storage)
- Can not define keys and constraints on Delta lake table
- Does not support Kafka storage, Storm streaming.

# AZURE DATABRICKS



# BEST

# PRACTICES



# AZURE DATABRICKS

## Azure Databricks – Best practices (1/2)

- Development Guidelines
  - Organize notebooks in proper folder structure
  - Keep development and executable notebook copies in separate folders
  - Use widgets in development phase
  - Use only one language for all notebooks
  - Error handling using try/catch
  - Define standards to call spark code from ADF (notebooks, python, jar)
  - Mount storages instead of referring complete absolute storage path
  - Use of secret scope to access secured info (e.g DB server pwd/conn string)
  - Define strategy to interact with external systems (e.g. use Pyodbc to interact with SQL DB)
  - Create libraries for common functions.

# AZURE DATABRICKS

## Azure Databricks – Best practices (2/2)

- Performance Optimization (Delta Lake features)
  - Use Delta tables with proper partitions
  - Why - Delta tables supports ACID transactions, scalable metadata handling, compression
  - Compaction – to improve analytics performance by combining small files into large files (bin packing)
  - Optimization – Z-Ordering technique is kind of indexing but not indexing. Improves performance of select queries
  - Vacuum – Delete unreferenced files (generally these files are unreferenced after compaction process)
  - Analyzing Tables – Analyze the tables periodically to collect the statistics about tables which are used by query optimizer for better performance.



# AZURE HDINSIGHT



# AZURE HDINSIGHT

## Tell me about Azure HDInsight

- Fully IaaS – Full control on servers
- Offers Hadoop, Spark, Kafka, Storm, Interactive Query, Hbase clusters
- Easy to migrate on-prem code
- Offers auto sizing
- Rich UI (Ambari) to manage and monitor cluster and nodes
- RDBMS like Security support using Ranger and Azure Active Directory.

# AZURE HDINSIGHT

## What did you not like about Azure HDInsight?

- Costlier as it is IaaS. Billed even not used.
- Can not upgrade the runtime. Need to recreate cluster.
- No integration with Azure Key Vault
- More maintenance efforts in infra management
- Ranger is costly
- Support tickets are routed to Hortonworks.



# AZURE SQL DB

# AZURE SYNAPSE ANALYTICS



# AZURE SQL DB/DW (SYNAPSE ANALYTICS)

## Why do you want to use SQL DB or DW?

- Supports all features of on-prem SQL Server either thru Azure SQL DB or Managed instance
- Security at row and column level using AAD and SQL Login
- Easy to scale up and down
- Easy to manage using Azure portal
- Offers threat detection, encryption, data masking and vulnerability detection
- Azure SQL DB supports large data using Hyperscale
- Auto backup policy configuration
- Redundancy support across region
- SQL DW MPP (Massive Parallel Processing) architecture specially built for Big Data processing which separates storage and compute (distributed processing similar to Azure Databricks Spark)
- Excellent UI (Azure Portal) for monitoring utilization, load on server, fine tune queries, performance optimization suggestions, auto index creation etc.

# AZURE SQL DB

## What are the Azure SQL products you know?

### Azure SQL DB

- Virtual Server
- Only DB management
- Fully PaaS
- Default max size limit 4 TB
- Hyperscale upto 100 TB
- Supports Elastic Pools
- Auto backup

### SQL Managed Instance

- Can be added company network (VNET)
- Lift and shift database migration
- Fully PaaS
- Auto backup
- SSIS, SSAS and SSRS - No

### SQL on Azure VM

- SQL installation on VM
- Fully IaaS
- Easy lift and shift
- As good as on-prem
- Good for large databases, minimum time to migrate, third party softwares like SQL Sentry installation supported

# AZURE SQL DB

## What are the database migration tools you know?

### Azure Database Migration Service

- Cloud based service
- Internally uses Data Migration Assistant and SQL Server Migration Assistant
- Useful for large data migration (in terms of number of databases or size of databases)

### Data Migration Assistant

- SQL to Azure SQL migration
- Also analyses jobs, SSIS packages
- Desktop software

### Database Experimentation Assistant

- Used to upgrade SQL version

### SQL Server Migration Assistant

- Non SQL to SQL/Azure SQL migration
- Source – Access, DB2, MySQL, Oracle, SAP ASE



# AZURE SQL

## More on securing data

- Transparent Data Encryption – Data, logs and backups are encrypted, decrypted real time at rest. Bring your own key supported. This is default enabled for SQL DBs and manual switch on for Synapse Analytics
- Always encrypted – Encryption / decryption at client side, secure data in transit
- Data Protection –
  - Column level security
  - Row level security
  - Data masking
  - Manual Encryption.

# AZURE SQL DB/DW (SYNAPSE ANALYTICS)

## All about SQL DW (Synapse Analytics)?

- Sharding data via Hash distributed, Round robin and replicated tables
- Limitless storage
- Costly (you can pause it to save cost)
- Upcoming Synapse Studio is bundle of ADF, ADLS, Spark and DW
- Best for SQL developers who are already familiar with T-SQL language

## Any limitations or drawbacks?

- Very costly
- Various limitations like Polybase supports only 1 MB size per row, no support for serialized transactions, no support for UPDATE FROM etc
- No integration of Azure Key Vault and other services
- Can not link the other DB servers (instead use external data source)

Need to rely only on T-SQL

# SYNAPSE ANALYTICS



## BEST

## PRACTICES



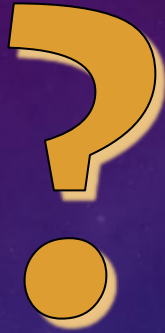
# AZURE SQL DB/DW (SYNAPSE ANALYTICS)

## Azure Synapse Analytics Best Practices

- Sharding data: Hash distributed - facts, Round robin – stage layer, replicated - dimensions
- Copy data from external source
- For fastest load use stage layer with round robin distribution
- Polybase limitation of max 1 MB per row. Prepare data with smaller size rows
- Scale up DW for large data load, scale down once data is loaded
- Create statistics periodically
- Minimize transaction size (avoid long running DML operations)
- Use smallest possible column size
- Source files: prefer parquet over flat files; multiple small files over few large files
- Use minimal logging to general small amount of logs and to increase I/O efficiency
- Rename tables instead of Delete / Update

# MORE ON AZURE DATA

- Azure Delta Lake – adoption of open source Delta storage layer (similar to Databricks Delta)
- Azure Data Explorer – To query and analyze real time data being ingested in event hub, IOT hub, Azure queue etc.
- Azure Data Studio – SQL Client + capability of running notebooks using Python and R on Spark cluster



# MISCELLANEOUS QUESTIONS



# MISCELLANEOUS QUESTIONS

1

## How do you improve performance of analytics?

How do you partition data?

### Storage (Blob/ADLS)

- Partition using folder structure
  - Subject area / org units
  - Region
  - Date (yyyy/mm/dd)
- Small number of large files
- Hadoop formats (e.g. parquet)

### Databricks / HDInsight

- Table partition
- Prefer static partitions over dynamic
- Bucketing (cluster by)
- Optimize
- Vacuum periodically
- Archive old data

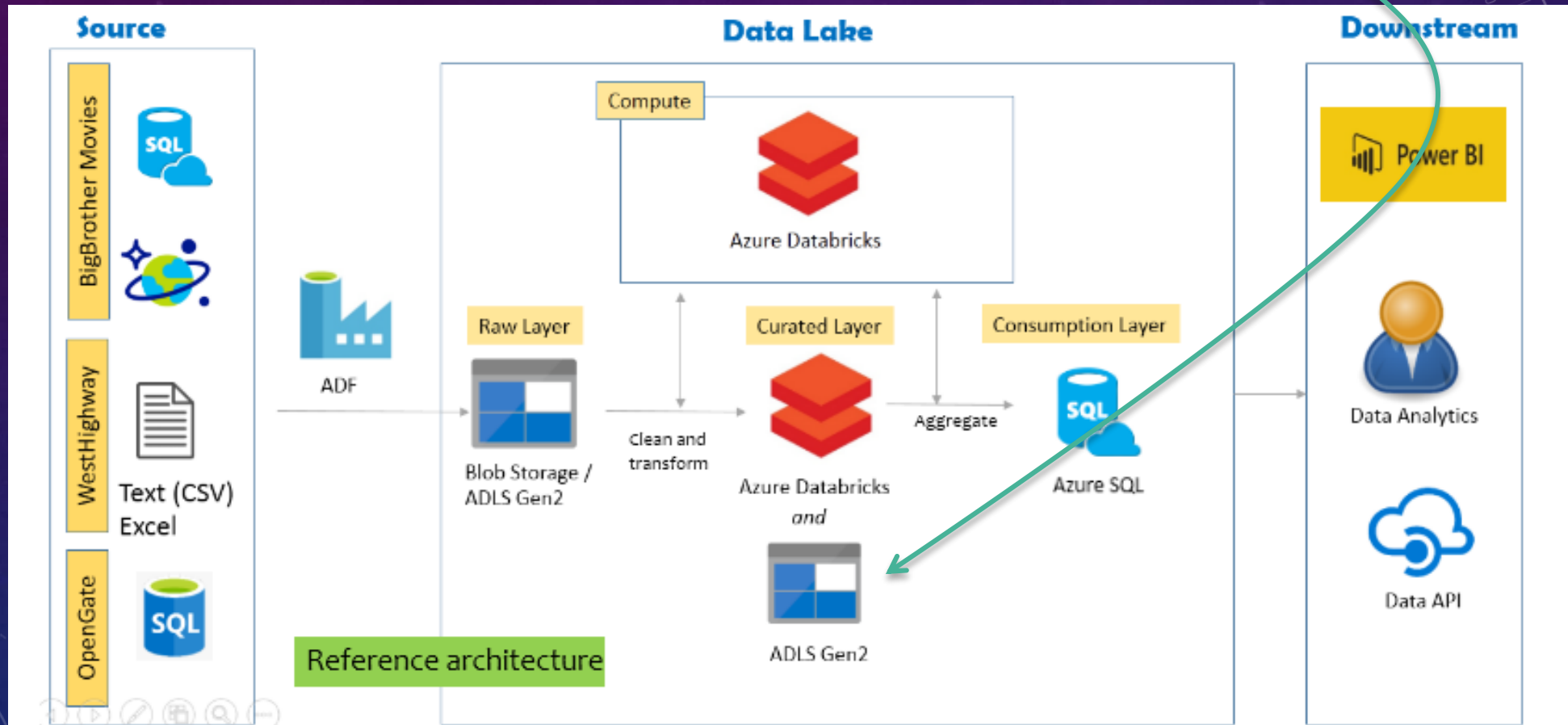
### SQL (DB & Synapse Analytics)

- Table partition
- Aggregate / summarize data
- Min use of external data source
- Indexing
- Sharding fact data using hash distribution
- Create statistics

# MISCELLANEOUS QUESTIONS

2

Why curated layer is ADLS Gen2? You could place SQL DB there....

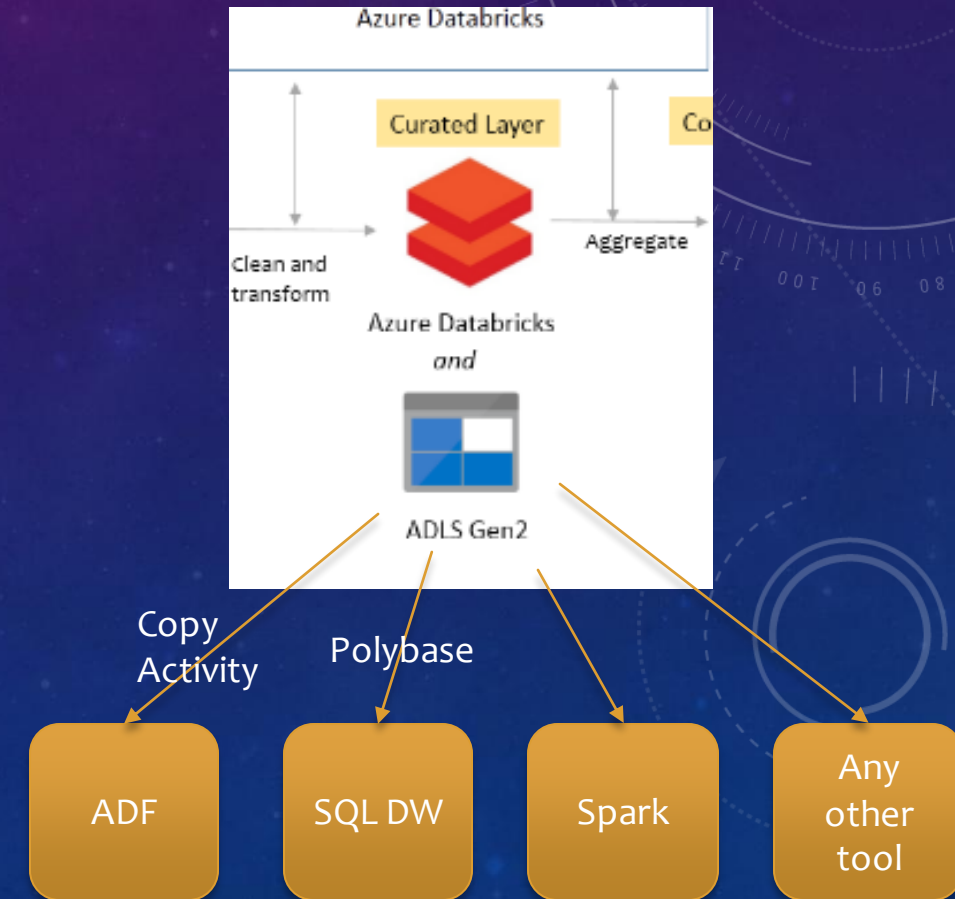


# MISCELLANEOUS QUESTIONS

2

Why curated layer is ADLS Gen2? You could place SQL DB there....

- ADLS Gen2 + Databricks combination is meant for Big Data processing
- Create delta tables pointing location to ADLS Gen2. Delta table file format is parquet which can be read by any external system (e.g. ADF, data analyst, reporting tools)
- To read data, no Databricks engine needed – saves compute cost
- Data is compressed requires less storage space
- ADLS storage is cheaper than SQL DB
- ADLS supports security at any level (POSIX like access control lists (ACL)).





# MISCELLANEOUS QUESTIONS

3

- Apart from ADF, how can you execute pipelines -- PowerShell, RestAPI, SDK, LogicApp
- What are various ways to copy data from Databricks to Azure SQL/DW? - From Spark pyodbc/jdbc, ADF Copy activity, SQL DW polybase if we have created external tables in Databricks, Azure SQL Library for Python
- Any challenges faced at any point of time? – Yes
  - Databricks excel file loading --> Third party lib was not performing well. So used panda. Converted Panda DF to Spark DF (spark.createDataFrame)
  - Data copy from Azure Databricks to SQL server was slow --> Scaled up DB, added batch size in jdbc, used SQL lib and pushed data using bulkcopy
  - Grouping of ADF activities to run parallel --> used foreach loop, supplied only 1 value to foreach
  - Library installation on HDInsight cluster --> No option, used "Action Scripts" from UI
  - ADF trigger was not considering default parameters --> Passed the params manually from trigger
  - ADF Databricks Activity was executing before lib installation on Databricks --> Added wait time in code as workaround. Also communicated the issue to platform team.
  - dbutils was not recognized inside lib --> accepted dbutils from user as input as parameter

# MISCELLANEOUS QUESTIONS

4

- Feed files are coming on monthly basis -- which trigger do you suggest - scheduled or event based  
--> If feed arrival schedule is defined then scheduled trigger. If the requirement is like near real time or no schedule of data arrival, use event based trigger
- Do you know cost saving techniques:
  - Auto shutdown/scale up/down DBX cluster
  - Pause SQL DW -- create data marts
  - Set DIU in ADF to Auto
  - Avoid unnecessary movement of large data files
  - Prefer Azure Function over logic apps. Logic apps is costlier than Azure functions.
- How do you read file from Blob storage and write back treating as File IO operation from Azure Databricks – Using Storage SDK for Python

**Keep note of each challenge faced during the execution of project**