Azure storage, ADF, Azure Databricks

Azure Storage
==========

Set 1:

what are the important azure storage services?

=> Azure Blob / ADLS
=> File
=> Queues
=> Azure tables

follow up question - which of these have you used?
blob / ADLS gen2

follow up question - what is the difference between Azure blob vs ADLS gen2

Azure blob storage
- object based storage solution for the cloud
- optimized to store massive amount of unstructured data
- stores data in a single hierarchy, also known as a flat namespace

Azure Datalake storage Gen2
- massively scalable, secure data lake functionality built on Azure blob storage
- is designed for big data analytics and offers a hierarchical file system.
- is more performant

follow up question - when will you prefer to use blob storage or will you always use ADLS gen2

Answer
- Serving images or documents directly to a browser
- streaming video and audio
- writing to log files
- storing data for backup and restore, disaster recovery and archiving

follow up question - how do you create an ADLS gen2 account

related question - lets say you have a website where user can upload some pdf documents/images. which storage solution you will choose for this?

Set 2:

what factors affect the cost of storage account in Azure?

Ideally there are a lot of factors which affect the cost

- Region (Geographic location)
- Redundancy (the more the redundancy the more the cost)
- account type (standard/premium)
- Data transfer cost - any data transferred out of an azure region
- access tiers - hot/cool/cold/archive
- transaction - number of read and write operations

follow up question - what is the difference between standard and premium account type?

premium storage accounts use SSD's (solid state drives) for low latency and high throughput (gives high performance)

follow up question - why the price vary based on region?
because the infra cost at different places would be different and because of that some regions might be costlier.

follow up question - okay so should we choose the region with lowest price?
if the region is far away then the latency would be high.

follow up question - Just now you mentioned about the access tiers, can you tell little bit more about them?

=> hot / cool/ cold/ archive

follow up question - how can you move the data between access tiers automatically? what is the business usecase?

=> using lifecycle management

follow up question - how to check how much storage account is costing?

=> cost management + billing => cost management => cost analysis

Set 3:

What are the different ways to provide the access for storage account.

Answer

=> account access key - entire account access (complete access)

=> SAS token (shared access signature)
to provide temporary and limited access to a specific resource, such as a blob or a container, without sharing the account keys.

time-limited access
granular permission (Read,write, delete etc)

=> Role based access control (IAM)

Azure active directory (microsoft Entra ID)
access control (IAM) - assign a role to the user/ service principal/ group

follow up question - why you have 2 access keys for your storage account?

microsoft recommends that you rotate your access keys periodically to help keep your storage account secured.

if possible, use azure key vault to manage the access keys.

if you are not using key vault, you will need to rotate your keys manually

2 keys are assigned so that you can rotate your keys.

this mechanism is for seamlessly changing access keys without service interruption.

follow up question - you talked about service principal, when to use it?

service principal are typically used when you need to authenticate and authorize an application or service to access azure resources.

follow up question - Lets say one app is generating some logs which need to be accessible to third party apps. it should be accessible for a specific number of days. how you would handle this requirement?

SAS token (shared access signature)

follow up question - Lets say you want to give read access to one of the team member & contributor access to another member. how you would achieve this?

follow up question - Let's say you have to give access to a lot of people for the files in your storage account. but this data is super critical. how would you ensure that you do not lose a file even if someone deletes it? Basically how do you ensure data protection?

=> soft delete

Follow up question - I got the point that you can set permissions for the users to access your files. But how do you ensure that your files are secured on the cloud. May be a person working in Azure can see it? right?

=> encryption of data at rest

Azure secures your data using various encryption methods, protocols and algorithms including double encryption. 256-bit AES encryption

microsoft managed keys
customer managed keys

if we are going with customer managed keys , then keys must be stored in azure key vault.

Azure Data Factory Interview questions
==========

Set 1
=======
what is ADF

=> Azure data factory is a managed cloud service thats built for complex ETL, ELT and data integration projects.

data migration from one data source to another
on premise to cloud data migration
Data Flows (transformation / nocode / minimal code)
Data orchestration service
Schedule the jobs

follow up question
==========
have you used ADF in any of your projects?

yes I have used ADF in one of my recent projects.

follow up question
==========
what was your usecase to use ADF?

you did data transfer from one data source to another.
created triggers so that this pipeline runs at every 24 hours.

follow up question
==========
lets say you have to transfer data from Azure SQL DB to ADLS gen2, how will you connect to the source and sink?

Linked Service (helps to connect to different data sources)
its like a connection string

follow up question
==========
okay, can you tell all the steps involved?

=> create 2 linked service
=> create datasets for your source and sink
=> create a ADF pipeline
=> create a copy activity
=> we can schedule this pipeline or trigger manually

follow up question
==========

you talked about copy activity, what it is?

helps you to copy the data from source to destination

follow up question
==========
what other activities you have used in your projects?

copy activity
data flow
databricks notebook
get metadata
lookup
filter
forEach
if condition
validation

follow up question
==========
you just mention about trigger, what types of triggers you can create?

=> helps to invoke the pipeline
=> determines when a pipeline execution needs to be kicked off
=> it can be used to schedule the pipeline on demand or based on any date time, frequency or based on some events in your storage account like file creation, deletion

schedule
tumbling window
storage events

follow up question
==========
what is the difference between scheduled triggers and tumbling window triggers?

Tumbling window triggers are a type of triggers that fires at a periodic time interval.

1 pm today

30 minutes

1 pm tomorrow

48 windows each of 30 minutes

1 pm - 1:30 pm (1:30 pm)
1:30 pm - 2:00 pm (2 pm)

windows of fixed size, non overlapping and contiguous time intervals

backfill scenario its quite handy

follow up question
==========
when you have triggers then whats the purpose of debug

debug is used to test the pipeline by manually executing it.
we can add debug points

Set 2
=======

you mentioned you used ADF for data transfer, how do you get the compute resources to perform this data transfer?

Integration Runtime (IR) is the compute infrastructure used by ADF to provide data integration capabilities.

copy the data
azure mapping data flows (ETL)

follow up question
==========
you just now talked about mapping dataflows what is that?

=> visually designed data transformations
=> no code/less code service
=> runs over the spark cluster (under the hood)

follow up question

==========
what are the different types of integration runtime

by default azure provide AutoResolveIntegrationRuntime

3 types
=======
=> Azure (when you are transferring data within cloud)
=> self hosted (onpremise to cloud)
=> Azure SSIS

follow up question
==========
Do you have to create that IR or its available?

there is an existing autoresolve integration runtime, however if we have a business usecase to create a different one then we can create

follow up question
==========
In which region default Azure IR is deployed?

For copy activity, a best effort is made to automatically detect your sink data store location, then use the IR in either the same region, if not then it will look for the closest one in the same geography. if the sink data stores region is not detectable then IR in the instance region is used instead.

follow up question
==========
why would someone create Azure IR as one is available by default?

If you have strict data compliance requirments and need to ensure that data do not leave a certain geography, you can explicitly create and Azure IR in a certain region and point the linked service to this IR

follow up question
==========
your company want to do cloud migration. you want to create a ADF pipeline, which Integration runtime will you use?

on premise -> cloud (self hosted)

follow up question
==========
what are the steps to create a self hosted Integration Runtime?

follow up question
==========
Lets say there are 3 different teams and each of them have a separte ADF instance. All of these teams have to do migration from on-prem to cloud. Do we have to create 3 different self hosted IR? Please explain

https://learn.microsoft.com/en-us/azure/data-factory/create-shared-self-hosted-integration-runtime-powershell

related question
==========
Assume that you are running a copy activity and it is working slow.
how will you improve the performance.

=> make sure if you have created an azure IR (in the same region)
=> in the copy activity setting we have DIU. by default it is auto. we can assign some high value since the start.