Azure Storage Solutions

1. What is Azure Storage?
   It’s a cloud storage solution for modern data storage scenarios.
   You can store the data of any format like images, audio, video, text files, csv, etc.
   You can access data stored in storage account through web url, programming, through rest api, PowerShell commands.

2. Benefits of Azure Storage?
   - Durable and highly available
   - Secure
   - Scalable
   - Managed
   - Accessible

3. What are different Azure Storage Data Services?
   Azure Blob Stores: Object store for text and binary data. Supports Big data analytics through ADLS.
   Azure Files: Managed file shares for cloud or on-premises deployments
   Azure Queues: A messaging store for reliable messaging between applications
   Azure Tables: A NOSQL store for schemeless storage of structured data
   Azure Disks: Block-level storage volumes for Azure VMs.

4. What is Azure Blob Storage?
   Used for storing unstructured data.
   - Serving images or documents directly to a browser.
   - Storing files for distributed access.
   - Streaming video and audio.
   - Writing log files.
   - Storing data for backup and restore, disaster recovery and archiving.

5. What is ADLS Gen2?
   Azure Data Lake Storage Gen2 is used for the big data analytics solution for the cloud.
   It has all the advantages of blob storage. It offers a hierarchical file system.

6. What are Azure Files?
   File Share Service in the cloud. It’s accessible via industry standard Server Message Block (SMB) protocol through windows, Linux and mac os clients or Network File System (NFS) protocol from Linux or Mac OS clients.
   Benefits- Shared Access, Fully Managed, Scripting and tooling, Resilient.

7. What is Azure Queue Storage?
   Service for storing large number of messages. Each queue message can be up to 64KB in size.
   Accessible from anywhere via authenticated calls using HTTP and HTTPS.

8. Does Azure Queue Storage Guarantee ordering?

No, Messages in storage queues typically FIFO but sometimes can be out of order.

9. What is Azure Table Storage?
   Service that stores NoSQL data. It's schemeless, fast and cost-effective. Can store up to TBs of structured data. Retrieval is faster using a clustered index.

10. Can storage account name be duplicated? Give Reason.
    No, Storage account name is part of the blob storage end-point URL. If the name gets duplicated, then two different storage account will have same endpoint. To avoid this problem the name must be unique.

11. What are the different types of blobs in azure storage?
    Block blobs- Store text and binary data. Max storage capacity is 190.7TiB.
    Append Blobs: Optimized for append operations. Used to store logging data from VMs. Updating and deleting existing blocks is not supported.
    Page Blobs: Store random access files up to 8TB, store Virtual Hard Drives files and serve as disks for azure VMs.

12. What are the factors affecting the cost of the Storage Accounts?
    Region: Geographical region in which your account is based.
    Account Type: Refers to types of storage account like standard, premium.
    Access Tier: Data storage pattern like Hot, Cool, Archive.
    Redundancy: How many copies of data are maintained at one time in multiple regions.
    Transactions: Refer to all read/write operations to azure storage.
    Data Egress: Data transfer across different regions.

13. What are different ways to authorize the data access of storage account?
    Account Access Key (Shared Key): It provides all access to the storage account with all services.
    SAS token: Shared Access Signature provides conditional access to type of service, time duration and various access methods.
    Azure AD: RBAC role back access control like reader, writer, contributor etc.

14. How azure ensures data security for storage account?
    By-default Azure uses Server-Side-Encryption (SSE) to automatically encrypt data when it is persisted in cloud. It uses 256-bit AES encryption. It can't be disabled.
    Encryption can be done using Microsoft managed keys or Customer managed keys.
    Keys must be stored inside Azure Key Vault or Azure Key Vault Managed Hardware Security Model (HSM).

15. How to ensure data protection in azure blob storage?
    Configure Azure Resource Manger lock on the storage account to prevent it from deletion.
    Enable container soft delete for the storage account to recover deleted container and its contents.
    For blob storage workloads, enable blob versioning to automatically save the state of your data each time it is overwritten.
    For ADLS workloads, take manual snapshots to save the state of your data at a particular time.

16. How does container soft delete works?
    When we enable container soft delete, we can specify the retention period for deleted containers i.e between 1 to 365days. By default, it's 7 days. We can restore deleted data by calling the RESTORE CONTAINER OPERATION.
17. What is AzCopy tool?
    It's a command line utility command to copy blobs or files from or to azure storage accounts.
    You can provide authorization credentials using AAD or SAS token.
18. Difference between ADLS and Azure blob storage?
    ADLS supports hierarchical namespace and we folder/directory structure. In blob storage it is flat file structure.
    ADLS used for Big Data workloads.
19. Why do you have two access keys (key1 & key2) in storage account?
    For security purposes if one of them is compromised, replace it with another key until first key is regenerated.

Azure Data Factory:

1. What is ADF and why do we need it?
   It's a pipeline orchestration tool provided by Microsoft azure to do data transfer/migration from one source to another. Used for ETL.

2. What is pipeline in the ADF?
   It's a set of activities to run in defined sequence to achieve any task.

3. What do you mean by data source in ADF?
   It's the source or destination system which contains the data to be used to transfer or send e.g ADLS, SQL server.

4. What is linked service in the ADF?
   It is the connection mechanism to connect with the source or the destination system provided connection string i.e authentication credentials.

5. What is Dataset in ADF?
   It's the reference to the data sources that is the structure and type of data to be used as an input or output.

6. What is integration runtime?
   It's the compute infrastructure service to connect with the source and the destination system for data transfer.

7. What is mapping data flows?
   Visually designed data transformation activities without writing any code, its executed over Apache spark clusters under the hood.

8. What are Triggers and different types?
   It's a scheduling mechanism to execute the ADF pipeline based on datetime, frequency and on some events like when a file gets inserted or deleted. They are:

   > **Tumbling Window Trigger:** Recurring trigger based on a time window.
   > **Schedule Trigger:** Recurring trigger based on a specified schedule.
   > **Event-based Trigger:** Triggered by an event, such as a file being created or deleted.

9. What is Copy Activity in ADF?
   It is a pipeline activity used for ETL purpose or lift and shift data from one source to another.

10. What is the difference between a trigger and a debug?
    Triggers help in automated pipeline execution whereas debug is used manually to do run test on the pipeline for entire pipeline or a particular point of activity to troubleshoot if any errors.
    Logs get generated separately for both during execution and triggers specifically used or Prod environments, debug for UAT testing and DEV purposes.

11. What are variables in pipeline?
    They are used to store the values for temporary purposes at pipeline level. They can be set during runtime for set variable activity.

12. What are Parameters in pipeline?
    The place holder of values to pass it to the pipelines during runtime for execution.

13. Difference between a pipeline variable and parameter?
    Variables values can be changed during the execution of the pipeline whereas parameters are used as an input to run the pipeline.

14. What is the global parameter?
    It's used when you have a property which is common across multiple pipelines, and they cannot be changed inside the pipeline level and can be accessed from the ADF pipeline with the provided account.

15. At how many levels of parameterization can be done in ADF?
    Linked Service level
    Dataset level
    Pipeline level
    ADF account level (Global Parameters)

16. What are ADF user properties?
    By adding user properties, you can view additional details about activities under activity runs.
    It is a key-value pair; key (user property name) and value (user property value).
    Per activity you can have 5 user properties. This can help us with monitoring and debugging purposes.

17. What are Annotations in ADF?
    They are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers.
    Easier search and filter for specific ADF resources.

18. What is the difference between ADF user-Properties and Annotations?
    Annotations are static values and tags that help in grouping and organizing objects such as pipelines, datasets, linked-services and triggers.
    Whereas user properties are defined within the activities and the values can change during execution.

19. What are different types of IRs(Integration Runtime)?
    Azure IR: For data transfer across azure resources
    Self-hosted IR: Between cloud and on-premises sources
    Azure SSIS IR – lift and shift operation between cloud and SSIS packages.

20. Is it mandatory to create IR in ADF?
    No, when you create an ADF resource AzureAutoresolveIR will be provisioned automatically for you.

21. How to call one pipeline from another pipeline?
    By using Execute Pipeline activity, add it inside parent pipeline and mention child pipeline to invoke in settings.

22. In which scenario we need to use the Linked-Self-hosted IR?
    Linked Self-hosted IR is the IR linked from external or different ADF account to current ADF account. If there's already a self-hosted IR created and being used by another ADF resource, instead of creating a new IR we can use that.

23. What are the scenarios where copy activity's mapping can be used?
    If the columns of the source and destination are not same i.e schema are not matching.

<u>Azure Databricks:</u>

1. What's the difference between transformation and action?
   In Transformation operation a new RDD is created on the top of existing RDD and its lazily evaluated until an action operation is called. Whereas action operation is used to get result or output from the existing RDD.

2. What do you mean by Lazy Evaluation?
   It means the execution will not start until an action is triggered.
   Transformations are lazy in nature i.e when we call some operation on RDD it doesn't get executed immediately. It adds them to a DAG of computation and only when driver requests some data, this DAG actually gets executed. Its the optimization technique of spark to reduce the no. of queries.

3. Difference between narrow and wide transformation?
   Compute data on single partition in narrow transformation so no data movement occurs across other partitions. Example filter, map, select.
   Wide transformation does computation with shuffling of data across different partitions. Specifically, during Joins and aggregation.

4. What is RDD and Dataframe?
   RDD is the distributed collection of data elements spread across many machines in the cluster. They are set of Java or Scala objects representing data.
   Dataframe is the distributed collection of data organized into named columns. Its conceptually same as table in a relational database.

5. What do you mean by partitions in spark and what are they?
   A partition is the logical division of data into small chunks stored on a node in the cluster.
   No. of partitions need to decided according to the cluster configuration and requirements of the application.
   They are two types: Hash Partitioning & Range Partitioning

6. What is shuffling and when it happens?
   It's the process of redistributing data across partitions that may lead to data movement across the executors. It occurs during joining dataframes or performing byKey operations like GroupBy, ReduceByKey.

7. Why do we need broadcast variables in spark?
   It is used to keep a read-only variable cached on each machine rather than shipping a copy of it with tasks to reduce communication costs.

8. What is the difference between delta lake and data lake?
   Data lake is the place where you can store all your data in append mode whereas in delta lake you can update/modify data with ACID and the latter supports time-travel.

9. What is Hive metastore?
   It’s the relational database. It stores metadata related to the tables/schemas you create to easily query big data using spark.
10. When to use Repartition in Spark?
    - When there is data skew, use repartition to increase the no. of partitions.
    - If you have enough memory and not properly utilized
    - When you have to decrease the no. of partitions due to less data size.