**Introduction**

Azure is a cloud platform launched on 1 February 2010. Azure is an online portal that enables you to manage and access various cloud services provided by Microsoft, which include storing and transforming data depending on our requirements. Azure is based on a pay-per-use model, and it is free to start with.



Azure provides over 200 cloud services and products, divided into 18 categories. These categories include networking, computing, storage, IoT, mobile, migration, analytics, artificial intelligence, containers, machine learning, management tools, integration, developer tools, databases, security, media identity, DevOps, and web services.

To perform well in interviews, one must know those skills and learn how to answer the company's interview questions.

Let's not make it boring and start discussing a collection of questions on one of the most popular cloud portals. Some might already know what it is, but we will discuss these in detail, considering everyone a newbie. Some of the most common Azure data engineer interview questions are given below.

Data Engineering involves designing, building, and maintaining data pipelines, databases and data infrastructure to have efficient data storage, processing, and retrieval thus making data analysis of big sets of data easier. It includes tasks like data collection, storage, transformation, etc. which is further used in various the decision-making purposes.

**Commonly Asked Azure Data Engineer Interview Questions With Answers**

**1. What is Data Engineering?**

The application of data collecting and analysis is the emphasis of data engineering. The information gathered from numerous sources is merely raw information. Data engineering helps in the transformation of unusable data into useful information. It is the process of transforming, cleansing, profiling, and aggregating huge data sets in a nutshell.

**2. What is Azure Synapse analytics?**

Azure Synapse is an enterprise service accelerating time to discernment across data storage and tectonic data networks. Azure Synapse combines the stylish of SQL(Structured Query Language) technologies used in enterprise data warehousing, Spark technologies used for big data, Pipelines for data integration and ETL/ ELT, and deep integration with distinct Azure services like Power BI, CosmosDB, and AzureML.

**3. Explain the data masking feature of Azure?**

Data masking helps in preventing unauthorized access to delicate data by enabling customers to assign how much of the delicate data to reveal with minimal impact on the application layer. Dynamic data masking limits acute data exposure by masking it to non-privileged users. It is a policy-based security feature that hides the delicate data in the result set of a query over designated [database](https://www.naukri.com/code360/library/what-is-database) fields. In contrast, the data in the database will not be changed.

A few data masking policies are:

* **SQL users excluded from masking** - A set of SQL users or Azure Active Directory identities that get unmasked data in the SQL query results. Users with administrator privileges are permanently banned from masking and seeing the original data without any mask.
* **Masking rules** - A set of rules defining the designated fields to be masked and the masking function used. The selected fields can be determined using a database schema, table, and column names.
* **Masking functions** - A set of methods that control data exposure for different scenarios.

**4. Difference between Azure Synapse Analytics and Azure Data Lake Storage?**

| **Azure Synapse Analytics** | **Azure Data Lake** |
| --- | --- |
| It is optimized for processing structured data in a well-defined schema. | It is optimized for storing and  processing structured and non-structured data. |
| Built on SQL(Structured Query Language) Server. | Built to work with Hadoop. |
| Built-in data pipelines and data streaming capabilities. | Handle data streaming using Azure Stream Analytics. |
| Compliant with regulatory standards. | No regulatory compliance |
| Used for Business Analytics. | Used for data analytics and exploration by data scientists and engineers |

**5. Describe various windowing functions of Azure Stream Analytics?**

A window in Azure Stream Analytics is a block of instant events that enables users to perform various operations on the event data. To analyze and partition a window in Azure Stream Analytics, There exist four windowing functions:

* **Hopping Window**: In these windows, the data segments can overlap. So, to define a hopping window, we need to specify two parameters:
  + Hop (duration of the overlap)
  + Window size (length of data segment)
* **Tumbling Window**: In this, the data stream is segmented into distinct time segments of fixed length in the tumbling window function.
* **Session Window**: This function groups events based on arrival time, so there is no fixed window size. Its purpose is to eliminate quiet periods in the data stream.
* **Sliding Window**: This windowing function does not necessarily produce aggregation after a fixed time interval, unlike the tumbling and hopping window functions. Aggregation occurs every time an existing event falls out of the time window, or a new event occurs.

**6. What are the different storage types in Azure?  
The following are the various advantages of the Java collection framework:**

| **Storage Types** | **Operations** |
| --- | --- |
| **Files** | Azure Files is an organized way of storing data on the cloud. The main advantage of using Azure Files over Azure Blobs is that Azure Files allows for organizing the data in a folder structure. Also, Azure Files is SMB (Server Message Block) protocol compliant, i.e., and can be used as a file share. |
| **Blobs** | Blob stands for a large binary object. This storage solution supports all kinds of files, including text files, videos, images, documents, binary data, etc. |
| **Queues** | Azure Queue is a cloud-based messaging store for establishing and brokering communication between various applications and components. |
| **Disks** | The Azure disk is used as a storage solution for Azure VMs (Virtual Machines) |
| **Tables** | Tables are NoSQL storage structures for storing structured data that does not meet the standard RDBMS (relational database schema). |

**7. What are the different security options available in the Azure SQL database?**

Security plays a vital role in databases. Some of the security options available in the Azure SQL database are:

* **Azure SQL Firewall Rules:** Azure provides two-level security. There are server-level firewall rules which are stored in the SQL Master database. Server-level firewall rules determine the access to the Azure database server. Users can also create database-level firewall rules that govern the individual databases’ keys.
* **Azure SQL TDE (Transparent Data Encryption):** TDE is the technology used to encrypt stored data. TDE is also available for Azure Synapse Analytics and Azure SQL Managed Instances. With TDE, the encryption and decryption of databases, backups, and transaction log files, happens in real-time.
* **Always Encrypted:** It is a feature designed to protect sensitive data stored in the Azure SQL database, such as credit card numbers. This feature encrypts data within the client applications using Always Encrypted-enabled driver. Encryption keys are not shared with SQL Database, which means database admins do not have access to sensitive data.
* **Database Auditing:**Azure provides comprehensive auditing capabilities along with the SQL Database. It is also possible to declare the audit policy at the individual database level, allowing users to choose based on the requirements.

**8. How data security is implemented in Azure Data Lake Storage(ADLS) Gen2?**

Data security is one of the primary concerns for most organizations for moving data to cloud storage. Azure data lake storage gen2 provides a multi-layered and robust security model. This model has 6 data security layers:

* **Authentication:** The first layer includes user account security. ADLS Gen2 provides three authentication modes, Azure Active Directory (AAD), Shared Access Token (SAS), and Shared Key.
* **Access Control:** The next layer for restricting access to individual containers or files. This can be managed using Roles and Access Control Lists (ACLs)
* **Network Isolation:** This layer enables administrators to manage access by disabling or allowing access to only particular Virtual Private Networks (VPNs) or IP Addresses.
* **Data Protection:**This is achieved by encrypting in-transit data using HTTPS(Hypertext Transfer Protocol Secure). Options to encrypt stored data are also available.
* **Advanced Threat Protection:**If enabled, ADLS Gen2 will monitor any unauthorized attempts to access or exploit the storage account.
* **Auditing:** This is the sixth and final layer of security. ADLS Gen2 provides comprehensive auditing features in which all account management activities are logged. These logs can be later reviewed to ensure the highest level of security.

**9. What are the various data flow partition schemes available in Azure?**

| **Partition Scheme** | **Explanation** | **Usage** |
| --- | --- | --- |
| **Round Robin** | It is the most straightforward partition scheme which spreads data evenly across partitions. | No good key candidates were available in the data. |
| **Hash** | Hash of columns creates uniform partitions such that rows with similar values fall in the same partition. | It is used to check for partition skew. |
| **Dynamic Range** | Spark dynamics range based on the provided columns or expression. | Select the column that will be used for partitioning. |
| **Fixed Range** | A fixed range of values based on the user-created expression for disturbing data across partitions. | A good understanding of data is required to avoid partition skew. |
| **Key** | Partition for each unique value in the selected column. | Good understanding of data cardinality is required. |

**10. Why is the Azure data factory needed?**

The amount of data generated these days is vast, coming from different sources. When we move this particular data to the cloud, a few things must be taken care of-

* Data can be in any form as it comes from different sources, and these various sources will transfer or channelize the data in different ways, and it can be in different formats. When we bring this data to the cloud or particular storage, we need to make sure that this data is well managed. i.e., you need to transform the data and delete unnecessary parts. As per moving the data is concerned, we need to make sure that data is picked from different sources and bring it to one common place, then stored, and if required, we should transform it into more meaningful.
* A traditional [data warehouse](https://www.naukri.com/Top%20Azure%20Data%20Engineer%20Interview%20Questions%20&%20Answers%20(2023)) can also do this, but certain disadvantages exist. Sometimes we are forced to go ahead and have custom applications that deal with all these processes individually, which is time-consuming, and integrating all these sources is a huge pain.
* A data factory helps to orchestrate this complete process into a more manageable or organizable manner.

**11. What do you mean by data modeling?**

Data Modeling is creating a visual representation of an entire information system or parts to express linkages between data points and structures. The purpose is to show the many types of data used and stored in the system, the relationships between them, how the data can be classified and arranged, and its formats and features. Data can be modeled according to the needs and requirements at various degrees of abstraction. The process begins with stakeholders and end-users providing information about business requirements. These business rules are then converted into data structures to create a concrete database design.

There are two design schemas available in data modeling:

* Star Schema
* Snowflake Schema

**12. What is the difference between Snowflake and Star Schema?**

Both are multidimensional models of the data warehouses. The main differences are:

| **Snowflake** | **Star Schema** |
| --- | --- |
| It contains 3D, sub-dimension, and fact tables. | It contains fact and dimension tables. |
| It is a type of bottom-up model. | It is a type of top-down model. |
| It uses both normalization and denormalization. | It does not use normalization. |
| In the snowflake schema, data redundancy is lower. | In the star schema, data redundancy is higher. |
| The design is very complex. | The design is straightforward. |
| Execution time for queries is high. | Execution time for queries is low. |
| It makes use of less space. | It makes use of more space. |

**13. What are the 2 levels of security in Azure data lake storage Gen2?**

The two levels of security available in Azure data lake storage Gen2 are also adequate for Azure data lake Gen1. Although this is not new, it is worth calling it two levels of security because it’s a fundamental piece for getting started with the Azure data lake. The two levels of security are defined as:

* **Role-Based Access Control (RBAC)**: RBAC includes built-in Azure roles such as reader, owner, contributor, or custom. Typically, RBAC is assigned due to two reasons. One is to permit the use of built-in data explorer tools that require reader permissions. Another is to specify who can manage the service (i.e., update properties and settings for the storage account).
* **Control Lists (ACLs)**: ACLs specify exactly which data objects a user may write, read, and execute (execution is required for browsing the directory structure). ACLs are POSIX (Portable Operating System Interface) - compliant, thus familiar to those with a Linux or Unix background.

**14. Explain a few important concepts of the Azure data factory?**

* **Pipeline**: It acts as a carrier in various processes occurring. An individual process is considered an activity.
* **Activities**: It represents the processing steps of a pipeline. A pipeline can have one or many activities. It can be a process like moving the dataset from one source to another or querying a data set.
* **Datasets:**It is the source of data or, we can say it is a data structure that holds our data.
* **Linked services**: It stores information that is very important when connecting to an external source.

**15. Differences between Azure data lake analytics and HDInsight?**

| **Azure Data Lake Analytics** | **HDInsight** |
| --- | --- |
| It is a software. | It is a platform. |
| Azure Data Lake Analytics creates essential computer nodes as on-demand instruction and processes the dataset. | HDInsight configures the cluster with predefined nodes and then uses a language like a hive or pig for data processing. |
| Azure data lake analytics does not give much flexibility in provisioning the cluster. | HDInsight provides more flexibility, as we can create and control the cluster according to our choice. |

**16. Explain the process of creating ETL(Extract, Transform, Load)?**

The process of creating ETL are:

* Build a Linked Service for source data store (SQL Server Database). Suppose that we have a cars dataset.
* Formulate a Linked Service for address data store which is Azure Data Lake Store.
* Build a dataset for Data Saving.
* Formulate the pipeline and attach copy activity.
* Program the pipeline by combining a trigger.

**17. What is Azure Synapse Runtime?**

[**Apache Spark**](https://www.naukri.com/code360/library/introduction-to-apache-spark) pools in Azure Synapse use runtimes to tie together essential component versions, Azure Synapse optimizations, packages, and connectors with a specific Apache Spark version. These runtimes will be upgraded periodically to include new improvements, features, and patches.

These runtimes have the following advantages:

* Faster session startup times.
* Tested compatibility with specific Apache Spark versions.
* Access to popular, compatible connectors and open-source packages.

**18. What is SerDe in the hive?**

Serializer/Deserializer is popularly known as SerDe. For IO(Input/Output), Hive employs the SerDe protocol. Serialization and deserialization are handled by the interface, which also interprets serialization results as separate fields for processing.

The Deserializer turns a record into a Hive-compatible Java object. The Serializer now turns this Java object into an HDFS (Hadoop Distributed File System) -compatible format. The storage role is then taken over by HDFS. Anyone can create their own SerDe for their own data format.

**19. What are the different types of integration runtime?**

* **Azure Integration Run Time:** It can copy data among cloud data repositories and it can express the exercise to a type of computing services like SQL server or Azure HDinsight where the transformation takes place
* **Self-Hosted Integration Run Time:**It is software with basically the equivalent code as Azure Integration Run Time. Except you install it on an on-premise instrument or a virtual machine in the virtual network. A Self Hosted IR can operate copy exercises between a data store in a private network and a public cloud data store.
* **Azure SSIS Integration Run Time:**With this, one can natively perform SSIS (SQL Server Integration Services) packages in a controlled environment. So when we elevate and shift the SSIS packages to the data factory, we work Azure SSIS Integration Run Time.

**20. Mention some common applications of Blob storage?**

Common works of Blob Storage consists of:

* Laboring images or documents straight to a browser.
* Saving files for shared access.
* Streaming audio and video.
* Collecting data for backup and recovery disaster restoration, and archiving.
* Saving data for analysis by an on-premises or Azure-hosted.

**21. What are the main characteristics of Hadoop?**

* It is an open-source structure that is ready for freeware.
* Hadoop is cooperative with the various types of hardware and simple to access distinct hardware within a particular node.
* It encourages faster-distributed data processing.
* It saves the data in the group, which is unconventional of the rest of the operations.
* Hadoop supports building replicas for every block with separate nodes.

**22. What is the Star scheme?**

Star Join Schema or Star Schema is the most manageable type of Data Warehouse schema. This is called a star schema because its construction is like a star. In this, the heart of the star may have one particular table and various connected dimension tables. This schema is practiced for questioning large data sets.

**23. How would you approve data to move from one database to another?**

The efficency of data and guaranteeing that no data is released should be of the highest priority for a data engineer. Hiring administrators examine this question to know your thought method on how validation of data would occur.

The candidate should be capable to talk about proper validation representations in different situations. For example, you could recommend that validation could be a simplistic comparison, or it can occur after the comprehensive data migration.

**24. Discriminate between structured and unstructured data?**

| **Parameter** | **Structured Data** | **Unstructured Data** |
| --- | --- | --- |
| **Storage** | DBMS (Database Management System) | Unmanaged file structure |
| **Standard** | ADO.net, ODBC, and SQL | STMP, XML, CSV, and SMS |
| **Scaling** | Schema scaling is hard. | Schema scaling is easy. |
| **Integration Tool** | ETL(Extract, Transform, Load) | Manual data entry or batch processing that incorporates codes |

**25. What do you mean by data pipeline?**

A data pipeline is a system for transporting data from one location (the source) to another (the destination), such as a data warehouse. Data is converted and optimized along the journey, and it eventually reaches a state that can be evaluated and used to produce business insights.  The procedures involved in aggregating, organizing, and transporting data are referred to as a data pipeline. Many of the manual tasks needed in processing and improving continuous data loads are automated by modern data pipelines.

**Frequently Asked Questions**

**How to prepare for the Azure data engineer interview?**

To prepare for an Azure data engineer interview, one should have a good understanding of Azure data management services such as Azure Data Factory, Azure Data Lake, Azure Stream Analytics, and Cosmos DB. Also, it's important to have a good understanding of big data concepts, data warehousing, and ETL process.

**What is the key role of a data engineer?**

The key role of data engineers is to design, build, and maintain data pipelines, databases and data infrastructure to have efficient data storage, processing, and retrieval thus making data analysis of big sets of data easier.

**What is the hiring process for a data engineer?**

The hiring process of a data engineer involves screening of resumes, conducting interviews to assess the technical knowledge and judge the problem solving and communication skills followed by an offer letter to the selected candidate.

**Conclusion**

Being well-prepared for an Azure Data Engineer interview is crucial for success in the competitive field of cloud computing. Mastering these 20+ commonly asked questions demonstrates a comprehensive understanding of Azure data services and their practical applications. You can also check out our other important articles:

**Recommended Readings:**

* [**Database Interview Questions**](https://www.naukri.com/code360/library/dbms-interview-questions)
* [**MySQL Interview Questions**](https://www.naukri.com/code360/library/mysql-interview-questions)
* [**SQL Query Interview Questions**](https://www.naukri.com/code360/library/sql-interview-questions)
* [**Excel Interview Questions**](https://www.naukri.com/code360/library/ms-excel-interview-questions)
* [**Managerial Round Interview Questions**](https://www.naukri.com/code360/library/managerial-round-interview-questions)
* [**System Design Interview Questions**](https://www.naukri.com/code360/library/system-design-interview-questions)

You can also consider our [Interview Preparation Course](https://www.naukri.com/code360/complete-interview-preparation-course) to give your career an edge over others.

Do upvote our blog to help others

Happy Learning!