DP-900 FAQ

Describe core data concepts (15–20%)

Practice questions based on these concepts

- Describe types of core data workloads
- Describe data analytics core concepts

1. **What is Data and why Data is a very important asset?**

```
Data is a collection of facts such as numbers, descriptions, and observations used
in decision making.In this competitive market, data is a valuable asset, and when
analyzed properly can turn into a wealth of useful information and inform critical
business decisions.
```

2. How many ways you can classify the data?

```
Structured
Semi-structured
Unstructured
```

3. What is Structured Data?

```
Structured data is typically tabular data that is represented by rows and columns
in a database. Databases that hold tables in this form are called relational
databases (the mathematical term relation refers to an organized set of data held
as a table). Each row in a table has the same set of columns.
```

4. What is Semi-structured Data?

```
Semi-structured data is information that doesn't reside in a relational database
but still has some structure to it. Examples include documents held in JavaScript
Object Notation (JSON) format.There are other types of semi-structured data as
well. Examples include key-value stores and graph databases.
A key-value store is similar to a relational table, except that each row can have
any number of columns.You can use a graph database to store and query information
about complex relationships.
A graph contains nodes (information about objects), and edges (information about
the relationships between objects).
```

5. What is the Unstructured Data?

```
Not all data is structured or even semi-structured. For example, audio and video files, and binary data files might not have a specific structure. They're referred to as unstructured data.
```

6. Azure provides different types of storage services based on the type of data. Is this true?

```
True

Depending on the type of data such as structured, semi-structured, or unstructured, data will be stored differently. Structured data is typically stored in a relational database such as SQL Server or Azure SQL Database.If you want to store unstructured data such as video or audio files, you can use Azure Blob storageIf you want to store semi-structured data such as documents, you can use a service such as Azure Cosmos DB.
```

7. What is called Provisioning?

```
The act of setting up the database server is called provisioning.
```

8. You can define several levels of access to your data in Azure. Is this true?

```
TrueRead-only access means the users can read data but can't modify any existing data or create new data.
Read/write access gives users the ability to view and modify existing data.
Owner privilege gives full access to the data including managing the security like adding new users and removing access to existing users.You can also define which users should be allowed to access the data in the first place.
```

9. What are the two kinds of Data processing solutions?

```
transactional system (OLTP)
analytical system (OLAP)
```

10. What is a transactional system?

A transactional system records transactions. A transaction could be financial, such as the movement of money between accounts in a banking system, or it might be part of a retail system, tracking payments for goods and services from customers. Think of a transaction as a small, discrete, unit of work.

11. What is an analytical system?

An analytical system is designed to support business users who need to query data and gain a big picture view of the information held in a database.Analytical systems are concerned with capturing raw data, and using it to generate insights. An organization can use these insights to make business decisions. For example, detailed insights for a manufacturing company might indicate trends enabling them to determine which product lines to focus on, for profitability.

12. What are the tasks that are involved in the analytical system?

Data Ingestion: Data ingestion is the process of capturing the raw data. This data could be taken from control devices measuring environmental information such as temperature and pressure, point-of-sale devices recording the items purchased by a customer in a supermarket, financial data recording the movement of money between bank accounts, and weather data from weather stations. Some of this data might come from a separate OLTP system. To process and analyze this data, you must first store the data in a repository of some sort. The repository could be a file store, a document database, or even a relational database.

Data Transformation/Data Processing: The raw data might not be in a format that is suitable for querying. The data might contain anomalies that should be filtered out, or it may require transforming in some way. For example, dates or addresses might need to be converted into a standard format. After data is ingested into a data repository, you may want to do some cleaning operations and remove any questionable or invalid data, or perform some aggregations such as calculating profit, margin, and other Key Performance Metrics (KPIs). KPIs are how businesses are measured for growth and performance.

Data Querying: After data is ingested and transformed, you can query the data to analyze it. You may be looking for trends, or attempting to determine the cause of problems in your systems. Many database management systems provide tools to enable you to perform ad-hoc queries against your data and generate regular reports.

Data Visualization: Data represented in tables such as rows and columns, or as documents, aren’t always intuitive. Visualizing the data can often be useful as a tool for examining data. You can generate charts such as bar charts, line charts, plot results on geographical maps, pie charts, or illustrate how data changes over time. Microsoft offers visualization tools like Power BI to provide rich graphical representation of your data.

13. What is called normalization?

```
The Process of spliting  into a large number of narrow, well-defined tables (a narrow table is a table with few columns), with references from one table to another, as shown in the image below. However, querying the data often requires reassembling information from multiple tables by joining the data back together at run-time.
```

14. You have a lot of customer data and you have decided to store this data in the relational database. What is the first thing you should do?

```
normalization
```

15. What are the drawbacks of normalization?

```
You split the information into tables. When you read this info you need to essemble this information at runtime by joins. These queries might be expensive sometimes.
```

16. Non-relational databases enable you to store data in a format that more closely matches the original structure. What is the disadvantage of this?

```
Some of the data is duplicated in the documentaed database. This duplication not only increases the storage required, but can also make maintenance more complex(you have to modify everywhere)
```

17. What are ACID principles?

```
Atomicity guarantees that each transaction is treated as a single unit, which
either succeeds completely, or fails completely. If any of the statements
constituting a transaction fails to complete, the entire transaction fails and the
database is left unchanged. An atomic system must guarantee atomicity in each and
every situation, including power failures, errors, and crashes.

Consistency ensures that a transaction can only take the data in the database from
one valid state to another. A consistent database should never lose or create data
in a manner that can’t be accounted for. In the bank transfer example described
earlier, if you add funds to an account, there must be a corresponding deduction of
funds somewhere, or a record that describes where the funds have come from if they
have been received externally. You can’t suddenly create (or lose) money.

Isolation ensures that concurrent execution of transactions leaves the database in
the same state that would have been obtained if the transactions were executed
sequentially. A concurrent process can’t see the data in an inconsistent state (for
example, the funds have been deducted from one account, but not yet credited to
another.)

Durability guarantees that once a transaction has been committed, it will remain
committed even if there’s a system failure such as a power outage or crash.
```

18. A transactional database must adhere to the ACID properties to ensure that the database remains consistent while processing transactions. Is this true?

```
True
```

19. What is eventual consistency and why do we need it?

```
Many systems implement relational consistency and isolation by applying locks to
data when it is updated. The lock prevents another process from reading the data
until the lock is released.A distributed database is a database in which data is
stored across different physical locations. It may be held in multiple computers
located in the same physical location (for example, a datacenter), or may be
dispersed over a network of interconnected computers.If you require transactional
consistency in this scenario, locks may be retained for a very long time,
especially if there's a network failure between databases at a critical point in
time. To counter this problem, many distributed database management systems relax
the strict isolation requirements of transactions and implement "eventual
consistency." In this form of consistency, as an application writes data, each
change is recorded by one server and then propagated to the other servers in the
distributed database system asynchronously. While this strategy helps to minimize
latency, it can lead to temporary inconsistencies in the data. Eventual consistency
is ideal where the application doesn't require any ordering guarantees.
```

20. What is Data processing and how many kinds?

```
Data processing is simply the conversion of raw data to meaningful information
through a process.Processing data as it arrives is called streaming.Buffering and
processing the data in groups is called batch processing.
```

21. What are the advantages and disadvantages of batch processing?

```
Advantages:
- Large volumes of data can be processed at a convenient time.
- It can be scheduled to run at a time when computers or systems might otherwise be
idle, such as overnight, or during off-peak hours.

Disadvantages:
- The time delay between ingesting the data and getting the results.
- All of a batch job's input data must be ready before a batch can be
processed.Even minor data errors, such as typographical errors in dates, can
prevent a batch job from running.
```

22. A real-estate website that tracks a subset of data from consumers' mobile devices, and makes real-time property recommendations of properties to visit based on their geo-location. How do you process this data?

```
streaming
```

23. What are the other differences between streaming and batch processing of data?

```
Data Scope: Batch data can process all the data in the dataset. Stream processing
typically only has access to the most recent data received, or within a rolling
time window (the last 30 seconds, for example).

Data Size: Batch data is suitable for handling large datasets efficiently. Stream
processing is intended for individual records or micro batches consisting of few
records.

Performance: The latency for batch processing is typically a few hours. Stream
processing typically occurs immediately, with latency in the order of seconds or
milliseconds. Latency is the time taken for the data to be received and processed.

Analysis: You typically use batch processing for performing complex analytics.
Stream processing is used for simple response functions, aggregates, or
calculations such as rolling averages.
```

24. How is data in a relational table organized?

```
Rows and Columns
```

25. What is an example of unstructured data?

```
Audio and Video files
```

26. What is an example of a streaming dataset?

```
Data from Twitter feeds
```

27. What are the roles in the world of data?

```
Database Administrator role:
A database administrator is responsible for the design, implementation,
maintenance, and operational aspects of on-premises and cloud-based database
solutions built on Azure data services and SQL Server. They are responsible for the
overall availability and consistent performance and optimizations of the database
solutions. They work with stakeholders to implement policies, tools, and processes
for backup and recovery plans to recover following a natural disaster or human-made
error.

Data Engineer role:
A data engineer collaborates with stakeholders to design and implement data-related
assets that include data ingestion pipelines, cleansing and transformation
activities, and data stores for analytical workloads. They use a wide range of data
platform technologies, including relational and nonrelational databases, file
stores, and data streams.

Data Analyst role:
A data analyst enables businesses to maximize the value of their data assets. They
are responsible for designing and building scalable models, cleaning and
transforming data, and enabling advanced analytics capabilities through reports and
visualizations.A data analyst processes raw data into relevant insights based on
identified business requirements to deliver relevant insights.
```

28. What are Database Administrator tasks and responsibilities?

```
https://docs.microsoft.com/en-us/learn/modules/explore-roles-responsibilities-
world-of-data/3-review-tasks-tools-for-database-administration
```

29. What is Azure Data Studio?

```
Azure Data Studio provides a graphical user interface for managing many different database systems. It currently provides connections to on-premises SQL Server databases, Azure SQL Database, PostgreSQL, Azure SQL Data Warehouse, and SQL Server Big Data Clusters, amongst others. It's an extensible tool, and you can download and install extensions from third-party developers that connect to other systems, or provide wizards that help to automate many administrative tasks.

https://docs.microsoft.com/en-us/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15
```

30. What is SQL Server Management Studio?

```
SQL Server Management Studio provides a graphical interface, enabling you to query data, perform general database administration tasks, and generate scripts for automating database maintenance and support operations.
```

31. What are Data Engineer tasks and responsibilities?

```
https://docs.microsoft.com/en-us/learn/modules/explore-roles-responsibilities-world-of-data/4-review-tasks-tools-for-data-engineering
```

32. What are some of the common tools that Data engineer uses?

```
sqlcmd utility, Azure Databricks, and Azure HDInsight, etc
```

33. What are Data Analyst tasks and responsibilities?

```
https://docs.microsoft.com/en-us/learn/modules/explore-roles-responsibilities-world-of-data/5-review-tasks-tools-for-data-visualization-reporting
```

34. What are some of the common tools that Data Analyst uses?

```
Power BI
```

35. Name one of the following tasks is the role of a database administrator?

```
restroing and backup
```

36. What are the characteristics of relational data?

```
All data is tabular. Entities are modeled as tables, each instance of an entity is
a row in the table, and each property is defined as a column.

All rows in the same table have the same set of columns.A table can contain any
number of rows.

A primary key uniquely identifies each row in a table. No two rows can share the
same primary key.

A foreign key references rows in another, related table. For each value in the
foreign key column, there should be a row with the same value in the corresponding
primary key column in the other table.
```

37. What is the primary key and foreign key?

```
The primary key indicates the column (or combination of columns) that uniquely
identify each row. Every table should have a primary key.
The columns marked FK are Foreign Key columns. They reference, or link to, the
primary key of another table, and are used to maintain the relationships between
tables. A foreign key also helps to identify and prevent anomalies, such as orders
for customers that don't exist in the Customers table.
```

38. How do you query the relational data?

```
Most relational databases support Structured Query Language (SQL). You use SQL to
create tables, insert, update, and delete rows in tables, and to query data.
```

39. Give an example of SQL?

```
SELECT CustomerID, CustomerName, CustomerAddress
FROM Customers
```

40. Why do use JOINS in SQL queries?

You can combine the data from multiple tables in a query using a join operation. A join operation spans the relationships between tables, enabling you to retrieve the data from more than one table at a time. The following query retrieves the name of every customer, together with the product name and quantity for every order they've placed. Notice that each column is qualified with the table it belongs to:

```
SELECT Customers.CustomerName, Orders.QuantityOrdered, Products.ProductName
FROM Customers
JOIN Orders
ON Customers.CustomerID = Orders.CustomerID
JOIN Products
ON Orders.ProductID = Products.ProductID
```

41. What are the most common use cases of relational databases?

Examples of OLTP applications that use relational databases are banking solutions, online retail applications, flight reservation systems, and many online purchasing applications.

42. What is an index?

When you create an index in a database, you specify a column from the table, and the index contains a copy of this data in a sorted order, with pointers to the corresponding rows in the table. When the user runs a query that specifies this column in the WHERE clause, the database management system can use this index to fetch the data more quickly than if it had to scan through the entire table row by row.

43. Why creating indexes make inserts or updates or deletes slow?

An index might consume additional storage space, and each time you insert, update, or delete data in a table, the indexes for that table must be maintained. This additional work can slow down insert, update, and delete operations, and incur additional processing charges.

44. You have a table that is read frequently and rarely updates or inserts. How do you increase the performance of the queries?

Creating a index

45. What is a view?

```
A view is a virtual table based on the result set of a query. In the simplest case, you can think of a view as a window on specified rows in an underlying table.
```

46. You can query the view and filter the data in much the same way as a table. Is this true?

```
True
```

47. What is IaaS and when should you use it?

```
IaaS is an acronym for Infrastructure-as-a-Service. Azure enables you to create a virtual infrastructure in the cloud that mirrors the way an on-premises data center might work.you're still responsible for many of the day-to-day operations, such as installing and configuring the software, patching, taking backups, and restoring data when neededThe IaaS approach is best for migrations and applications requiring operating system-level access. SQL virtual machines are lift-and-shift. That is, you can copy your on-premises solution directly to a virtual machine in the cloud. The system should work more or less exactly as before in its new location, except for some small configuration changes (changes in network addresses, for example) to take account of the change in environment.
```

48. What is Paas and when should you use it?

```
PaaS stands for Platform-as-a-service. Rather than creating a virtual infrastructure, and installing and managing the database software yourself, a PaaS solution does this for you. You specify the resources that you require (based on how large you think your databases will be, the number of users, and the performance you require), and Azure automatically creates the necessary virtual machines, networks, and other devices for you.
```

49. What is the benefit of using a PaaS service, instead of an on-premises system, to run your database management systems?

```
Increased scalabilityPaaS solutions enable you to scale up and out without having to procure your own hardware.
```

50. What are the key characteristics of non-relational data?

```
A key aspect of non-relational databases is that they enable you to store data in a very flexible manner. Non-relational databases don't impose a schema on data. Instead, they focus on the data itself rather than how to structure it. This approach means that you can store information in a natural format, that mirrors the way in which you would consume, query and use it.
```

51. Non-relational systems such as Azure Cosmos DB (a non-relational database management system available in Azure), support indexing even when the structure of the indexed data can vary from record to record. Is this true?

```
True
```

52. What are the use cases of the non-relational databases?

```
IoT and telematics: These systems typically ingest large amounts of data in frequent bursts of activity. Non-relational databases can store this information very quickly. The data can then be used by analytics services such as Azure Machine Learning, Azure HDInsight, and Microsoft Power BI. Additionally, you can process the data in real-time using Azure Functions that are triggered as data arrives in the database.

Retail and marketing: Microsoft uses CosmosDB for its own ecommerce platforms that run as part of Windows Store and XBox Live. It’s also used in the retail industry for storing catalog data and for event sourcing in order processing pipelines.

Gaming: The database tier is a crucial component of gaming applications. Modern games perform graphical processing on mobile/console clients, but rely on the cloud to deliver customized and personalized content like in-game stats, social media integration, and high-score leaderboards. Games often require single-millisecond latencies for reads and write to provide an engaging in-game experience. A game database needs to be fast and be able to handle massive spikes in request rates during new game launches and feature updates.

Web and mobile applications: A non-relational database such as Azure Cosmos DB is commonly used within web and mobile applications, and is well suited for modeling social interactions, integrating with third-party services, and for building rich personalized experiences. The Cosmos DB SDKs (software development kits) can be used build rich iOS and Android applications using the popular Xamarin framework.
```

53. What are the formats of semi-structured data?

JSON:

A JSON document is enclosed in curly brackets ({ and }). Each field has a name (a label), followed by a colon, and then the value of the field. Fields can contain simple values, or subdocuments (each starting and ending with curly brackets). Fields can also have multiple values, held as arrays and surrounded with square brackets ([and]). Literals in a field are enclosed in quotes, and fields are separated with commas.

Avro:

Avro is a row-based format. It was created by Apache. Each record contains a header that describes the structure of the data in the record. This header is stored as JSON. The data is stored as binary information. An application uses the information in the header to parse the binary data and extract the fields it contains. Avro is a very good format for compressing data and minimizing storage and network bandwidth requirements.

ORC

ORC (Optimized Row Columnar format) organizes data into columns rather than rows. It was developed by HortonWorks for optimizing read and write operations in Apache Hive. Hive is a data warehouse system that supports fast data summarization and querying over very large datasets. Hive supports SQL-like queries over unstructured data. An ORC file contains stripes of data. Each stripe holds the data for a column or set of columns. A stripe contains an index into the rows in the stripe, the data for each row, and a footer that holds statistical information (count, sum, max, min, and so on) for each column.

Parquet

Parquet is another columnar data format. It was created by Cloudera and Twitter. A Parquet file contains row groups. Data for each column is stored together in the same row group. Each row group contains one or more chunks of data. A Parquet file includes metadata that describes the set of rows found in each chunk. An application can use this metadata to quickly locate the correct chunk for a given set of rows, and retrieve the data in the specified columns for these rows. Parquet specializes in storing and processing nested data types efficiently. It supports very efficient compression and encoding schemes.

54. What are the NoSQL databases?

NoSQL (non-relational) databases generally fall into four categories: key-value stores, document databases, column family databases, and graph databases.

Key-Value store

```
A key-value store is the simplest (and often quickest) type of NoSQL database for inserting and querying data. Each data item in a key-value store has two elements, a key and a value. The key uniquely identifies the item, and the value holds the data for the item. The value is opaque to the database management system. Items are stored in key order.
```

Document database

```
A document database represents the opposite end of the NoSQL spectrum from a key-value store. In a document database, each document has a unique ID, but the fields in the documents are transparent to the database management system. Document databases typically store data in JSON format. they could be encoded using other formats such XML, YAML, JSON, BSON
```

Column family database

```
A column family database organizes data into rows and columns. Examples of this structure include ORC and Parquet files
In its simplest form, a column family database can appear very similar to a relational database, at least conceptually. The real power of a column family database lies in its denormalized approach to structuring sparse data.
```

Graph database

```
Graph databases enable you to store entities, but the main focus is on the relationships that these entities have with each other. A graph database stores two types of information: nodes that you can think of as instances of entities, and edges, which specify the relationships between nodes. Nodes and edges can both have properties that provide information about that node or edge (like columns in a table). Additionally, edges can have a direction indicating the nature of the relationship.https://docs.microsoft.com/en-us/learn/modules/explore-concepts-of-non-relational-data/4-describe-types-nosql-databases
```

55. What are the characteristics of the Key-value store?

```
* A query specifies the keys to identify the items to be retrieved.
* You can't search on values. An application that retrieves data from a key-value store is responsible for parsing the contents of the values returned.
* The value is opaque to the database management system.
* Write operations are restricted to inserts and deletes.
* If you need to update an item, you must retrieve the item, modify it in memory (in the application), and then write it back to the database, overwriting the original (effectively a delete and an insert).
```

56. What is the use case for the Key-value store?

```
The focus of a key-value store is the ability to read and write data very quickly. Search capabilities are secondary. A key-value store is an excellent choice for data ingestion, when a large volume of data arrives as a continual stream and must be stored immediately.
```

57. You are building a system that monitors the temperature throughout a set of office blocks and sets the air conditioning in each room in each block to maintain a pleasant ambient temperature. Your system has to manage the air conditioning in several thousand buildings spread across the country or region, and each building typically contains at least 100 air-conditioned rooms. What type of NoSQL datastore is most appropriate for capturing the temperature data to enable it to be processed quickly?

```
A key-value store
```

58. What Is Data Wrangling?

```
Wrangling is the process by which you transform and map raw data into a more useful format for analysis. It can involve writing code to capture, filter, clean, combine, and aggregate data from many sources.
```

59. What are the two important stages of data analytics?

data ingestion, and data processing.Data Ingestion
Data ingestion is the process of obtaining and importing data for immediate use or storage in a database. The data can arrive as a continuous stream, or it may come in batches, depending on the source. The purpose of the ingestion process is to capture this data and store it. This raw data can be held in a repository such as a database management system, a set of files, or some other type of fast, easily accessible storage.The ingestion process might also perform filtering and transformation at this stage.Data Processing
The data processing stage occurs after the data has been ingested and collected. Data processing takes the data in its raw form, cleans it, and converts it into a more meaningful format (tables, graphs, documents, and so on). The result is a database of data that you can use to perform queries and generate visualizations, giving it the form and context necessary to be interpreted by computers and used by employees throughout an organization.

60. What are ETL and ELT?

ETL stands for Extract, Transform, and Load. The raw data is retrieved and transformed before being saved. The extract, transform, and load steps can be performed as a continuous pipeline of operations. It is suitable for systems that only require simple models, with little dependency between items.ELT is an abbreviation of Extract, Load, and Transform. The process differs from ETL in that the data is stored before being transformed. The data processing engine can take an iterative approach, retrieving and processing the data from storage, before writing the transformed data and models back to storage. ELT is more suitable for constructing complex models that depend on multiple items in the database, often using periodic batch processing.

61. What is Reporting?

Reporting is the process of organizing data into informational summaries to monitor how different areas of an organization are performing. Reporting helps companies monitor their online business, and know when data falls outside of expected ranges. Good reporting should raise questions about the business from its end users. Reporting shows you what has happened, while analysis focuses on explaining why it happened and what you can do about it.

62. What is Business Intelligence?

The term Business Intelligence (BI) refers to technologies, applications, and practices for the collection, integration, analysis, and presentation of business information. The purpose of business intelligence is to support better decision making.

63. What is Data Visualization?

```
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to spot and understand trends, outliers, and patterns in data.
```

64. What are the most common forms of visualizations?

Bar and column charts:

```
Bar and column charts enable you to see how a set of variables changes across different categories.Line charts: Line charts emphasize the overall shape of an entire series of values, usually over time.
```

Matrix:

```
A matrix visual is a tabular structure that summarizes data. Often, report designers include matrixes in reports and dashboards to allow users to select one or more element (rows, columns, cells) in the matrix to cross-highlight other visuals on a report page.
```

Key influencers:

```
A key influencer chart displays the major contributors to a selected result or value. Key influencers are a great choice to help you understand the factors that influence a key metric.
```

Treemap:

```
Treemaps are charts of colored rectangles, with size representing the relative value of each item. They can be hierarchical, with rectangles nested within the main rectangles.
```

Scatter:

```
A scatter chart shows the relationship between two numerical values. A bubble chart is a scatter chart that replaces data points with bubbles, with the bubble size representing an additional third data dimension.
```

Filled map:

```
If you have geographical data, you can use a filled map to display how a value
differs in proportion across a geography or region.
```

65. What are the categories of data analytics?

Descriptive analytics

```
Descriptive analytics helps answer questions about what has happened, based on
historical data. Descriptive analytics techniques summarize large datasets to
describe outcomes to stakeholders.
```

Diagnostic analytics

```
Diagnostic analytics helps answer questions about why things happened. Diagnostic
analytics techniques supplement more basic descriptive analytics. They take the
findings from descriptive analytics and dig deeper to find the cause.
```

Predictive analytics

```
Predictive analytics helps answer questions about what will happen in the future.
Predictive analytics techniques use historical data to identify trends and
determine if they're likely to recur. Predictive analytical tools provide valuable
insight into what may happen in the future.
```

Prescriptive analytics

```
Prescriptive analytics helps answer questions about what actions should be taken to
achieve a goal or target. By using insights from predictive analytics, data-driven
decisions can be made. This technique allows businesses to make informed decisions
in the face of uncertainty.
```

Cognitive analytics

```
Cognitive analytics attempts to draw inferences from existing data and patterns,
derive conclusions based on existing knowledge bases, and then add these findings
back into the knowledge base for future inferences--a self-learning feedback loop.
Cognitive analytics helps you to learn what might happen if circumstances change,
and how you might handle these situations.
```

Describe how to work with relational data on Azure (25–30%)

Practice questions based on these concepts

- Describe relational data workloads
- Describe relational Azure data services
- Identify basic management tasks for relational data
- Describe query techniques for data using SQL language

66. We have IaaS, PaaS, SaaS. In which category that Azure data services fall into?

```
Azure Data Services fall into the PaaS category. These services are a series of DBMSs managed by Microsoft in the cloud. Each data service takes care of the configuration, day-to-day management, software updates, and security of the databases that it hosts.
```

67. What are currently available relational databases on Azure?

```
Azure SQL Database
Azure Database for MySQL servers
Azure Database for MariaDB servers
Azure Database for PostgreSQL servers
```

68. Using Azure Data Services reduces the amount of time that you need to invest to administer a DBMS. Is this true?

```
True
```

69. What is the availability of these Azure Data Services?

```
Azure Data Services ensure that your databases are available for at least 99.99% of the time.
```

70. How Azure data services cost you?

```
The base price of each service covers underlying infrastructure and licensing, together with the administration charges. Additionally, these services are designed to be always on. This means that you can't shut down a database and restart it later.
```

71. What is the limitation of Azure data services?

```
Not all features of a database management system are available in Azure Data Services. This is because Azure Data Services takes on the task of managing the system and keeping it running using hardware situated in an Azure datacenter. Exposing some administrative functions might make the underlying platform vulnerable to misuse, and even open up some security concerns. So, you have no direct control over the platform on which the services run.
```

72. How do you get more control than Azure Data Services allow?

```
you can install your database management system on a virtual machine that runs in Azure.
```

73. Migrating from the system running on-premises to an Azure virtual machine is no different than moving the databases from one on-premises server to another. Is this correct?

```
True
SQL Server running on an Azure virtual machine effectively replicates the database running on real on-premises hardware.
```

74. When you are running SQL Server on Virtual Machines who takes care of maintaining the SQL Server software and performing the various administrative tasks to keep the database running from day-to-day.

```
Customer
```

75. In Which situation you should run the Microsoft SQL Server on Virtual Machines?

```
This approach is optimized for migrating existing applications to Azure, or extending existing on-premises applications to the cloud in hybrid deployments.A hybrid deployment is a system where part of the operation runs on-premises, and part in the cloud. Your database might be part of a larger system that runs on-premises, although the database elements might be hosted in the cloud.
```

76. What are the capabilities of running SQL Server on Virtual machines?

Create rapid development and test scenarios when you do not want to buy on-premises non-production SQL Server hardware.Become lift-and-shift ready for existing applications that require fast migration to the cloud with minimal changes or no changes.
Scale up the platform on which SQL Server is running, by allocating more memory, CPU power, and disk space to the virtual machine. You can quickly resize an Azure virtual machine without the requirement that you reinstall the software that is running on it.

77. What are the options available when selecting the Azure SQL database?

Azure SQL Database is a PaaS offering from Microsoft. Azure SQL Database is available with several options: Single Database, Elastic Pool, and Managed Instance.

Single Database

This option enables you to quickly set up and run a single SQL Server database. You create and run a database server in the cloud, and you access your database through this server. Microsoft manages the server, so all you have to do is configure the database, create your tables, and populate them with your data. You can scale the database if you need additional storage space, memory, or processing power.

Elastic Pool

This option is similar to Single Database, except that by default multiple databases can share the same resources, such as memory, data storage space, and processing power. The resources are referred to as a pool. You create the pool, and only your databases can use the pool. This model is useful if you have databases with resource requirements that vary over time, and can help you to reduce costs.

Managed Instance

Managed instance effectively runs a fully controllable instance of SQL Server in the cloud. You can install multiple databases on the same instance. You have complete control over this instance, much as you would for an on-premises server.

78. What are the use cases of the Azure SQL Database?

```
Modern cloud applications that need to use the latest stable SQL Server features.
Applications that require high availability.
Systems with a variable load, that need the database server to scale up and down
quickly.
```

79. What are the features of the Azure SQL Database?

- Azure SQL Database automatically updates and patches the SQL Server software to ensure that you are always running the latest and most secure version of the service.
- The scalability features of Azure SQL Database ensure that you can increase the resources available to store and process data without having to perform a costly manual upgrade.
- The service provides high availability guarantees, to ensure that your databases are available at least 99.99% of the time.
- Azure SQL Database supports point-in-time restore, enabling you to recover a database to the state it was in at any point in the past.
- Databases can be replicated to different regions to provide additional assurance and disaster recovery
- Advanced threat protection provides advanced security capabilities, such as vulnerability assessments, to help detect and remediate potential security problems with your databases.
- It continuously monitors your database for suspicious activities, and provides immediate security alerts on potential vulnerabilities, SQL injection attacks, and anomalous database access patterns.
- SQL Database helps secure your data by providing encryption. For data in motion, it uses transport layer security. For data at rest, it uses transparent data encryption.

80. What is the use case for Azure SQL server managed instance?

```
Consider Azure SQL Database managed instance if you want to lift-and-shift an on-
premises SQL Server instance and all its databases to the cloud, without incurring
the management overhead of running SQL Server on a virtual machine.
```

81. What are MySQL, MariaDB, and PostgreSQL?

MySQL

```
MySQL started life as a simple-to-use open-source database management system. It's
available in several editions; Community, Standard, and Enterprise. The Community
edition is available free-of-charge, and has historically been popular as a
database management system for web applications, running under Linux. Versions are
also available for Windows.
```

MariaDB

```
MariaDB is a newer database management system, created by the original developers of MySQL. The database engine has since been rewritten and optimized to improve performance. MariaDB offers compatibility with Oracle Database (another popular commercial database management system). One notable feature of MariaDB is its built-in support for temporal data. A table can hold several versions of data, enabling an application to query the data as it appeared at some point in the past.
```

PostgreSQL

```
PostgreSQL is a hybrid relational-object database. You can store data in relational tables, but a PostgreSQL database also enables you to store custom data types, with their own non-relational properties. The database management system is extensible; you can add code modules to the database, which can be run by queries. Another key feature is the ability to store and manipulate geometric data, such as lines, circles, and polygons.
```

82. PostgreSQL has its own query language called pgsql. Is this true?

```
True
```

83. What are the deployment options for Azure Database for PostgreSQL?

Azure Database for PostgreSQL single-server

```
The single-server deployment option for PostgreSQL provides similar benefits as Azure Database for MySQL. You choose from three pricing tiers: Basic, General Purpose, and Memory Optimized. Each tier supports different numbers of CPUs, memory, and storage sizes—you select one based on the load you expect to support.
```

Azure Database for PostgreSQL Hyperscale (Citus)

```
Hyperscale (Citus) is a deployment option that scales queries across multiple server nodes to support large database loads. Your database is split across nodes. Data is split into chunks based on the value of a partition key or sharding key. Consider using this deployment option for the largest database PostgreSQL deployments in the Azure Cloud.
```

84. Scaling up or out will take effect without restarting the SQL database. Is this true?

```
True
```

85. What is the best way to transfer the data in a PostgreSQL database running on-premises into a database running Azure Database for PostgreSQL service?

```
Use the Azure Database Migration Services
```

86. When using an Azure SQL Database managed instance, what is the simplest way to implement backups?

```
Backups are automatically handled
```

87. You're responsible for all software installation and maintenance, and performing backups when SQL Server running on a virtual machine. Is this true?

```
True
```

88. Describe provisioning PostgreSQL and MySQL?

```
https://docs.microsoft.com/en-us/learn/modules/explore-provision-deploy-relational-database-offerings-azure/4-describe-provision-postgresql-mysql
```

89. Describe configuring Azure SQL Database, Azure Database for PostgreSQL, and Azure Database for MySQL?

```
https://docs.microsoft.com/en-us/learn/modules/explore-provision-deploy-relational-database-offerings-azure/6-configure-sql-database-mysql-postgresql
```

90. Consider the following SQL statement what is the table name and columns name?

```
SELECT *
FROM customers
WHERE username = 'contoso'
```

```
Table Name: customers
Column Name: username
```

91. Consider the following SQL statement what is the table name and columns name?

```
SELECT * FROM users ORDER BY email
```

```
Table Name: users
Column Name: email
```

92. What is the dialect used by Oracle?

```
PL/SQL
```

93. What is the dialect used by PostgreSQL?

```
pgSQL
```

94. What are the DML statements?

```
SELECT, INSERT, UPDATE, DELETE
```

95. What are DDL statements?

```
CREATE, ALTER, DROP, RENAME
```

96. When should we use DML statements?

```
You use DML statements to manipulate the rows in a relational table. These
statements enable you to retrieve (query) data, insert new rows, or edit existing
rows. You can also delete rows if you don't need them anymore.
```

97. By default, the SELECT, UPDATE and DELETE statements are applied to every row in a table. Is this true?

```
True
```

98. Which clause should you use with the SELECT, UPDATE, and DELETE statements to apply changes only for specific rows?

```
WHERE
```

99. Which clause should you use t sort the data in the select query?

```
ORDER BY
```

100. Which clause should you use to retrieve the related data from multiple tables?
JOIN

```
A join condition defines the way two tables are related in a query by:Specifying the column from each table to be used for the join. A typical join condition specifies a foreign key from one table and its associated primary key in the other table.Specifying a logical operator (for example, = or <>,) to be used in comparing values from the columns.
```

101. Consider the following SQL query which type of statement is this?

> INSERT INTO MyTable(MyColumn1, MyColumn2, MyColumn3)
> VALUES (99, ‘contoso’, ‘hello’)

```
DML
```

102. Consider the following SQL query which type of statement is this?

> CREATE TABLE MyTable (MyColumn1 INT NOT NULL PRIMARY KEY, MyColumn2 VARCHAR(50) NOT NULL, MyColumn3 VARCHAR(10) NULL);

```
DDL
```

Describe how to work with non-relational data on Azure (25–30%)

Practice questions based on these concepts

- Describe non-relational data workloads
- Describe non-relational data offerings on Azure
- Identify basic management tasks for non-relational data

103. What is the Azure service that implements the NoSQL key-value model?

```
Azure Table Storage
```

104. What is Azure Table Storage?

```
Azure Table Storage is a scalable key-value store held in the cloud. You create a
table using an Azure storage account.
In an Azure Table Storage table, items are referred to as rows, and fields are
known as columns.
```

105. You have semi-structured data and you want to store that data in the database as key-value pairs where the key is unique and columns can vary and each row holding the entire data for a logical entity. Which storage option should you select?

```
Azure Table Storage
```

106. Azure Table Storage tables have no concept of relationships, stored procedures, secondary indexes, or foreign keys. Is this true?

```
True
```

107. Why Azure Table Storage provides much faster access to the data you need?

```
Azure Table Storage provides much faster access to the data because the data is
available in a single row, without requiring that you perform joins across
relationships.To help ensure fast access, Azure Table Storage splits a table into
partitions.
```

108. What is a partition in Azure Table Storage?

```
Partitioning is a mechanism for grouping related rows, based on a common property
or partition key. Rows that share the same partition key will be stored together.
Partitioning not only helps to organize data, it can also improve scalability and
performance
```

109. Can these partitions grow or shrink as required?

```
YesPartitions are independent from each other, and can grow or shrink as rows are
added to, or removed from, a partition. A table can contain any number of
partitions.
```

110. How does partitioning help improving performance?

```
When you search for data, you can include the partition key in the search criteria.
This helps to narrow down the volume of data to be examined, and improves
performance by reducing the amount of I/O (reads and writes) needed to locate the
data.
```

111. What is Azure table storage key comprises of?

```
The key in an Azure Table Storage table comprises two elements:
- The partition key that identifies the partition containing the row
- The row key that is unique to each row in the same partition.
Items in the same partition are stored in row key order. If an application adds a
new row to a table, Azure ensures that the row is placed in the correct position in
the table.
```

112. What are point queries and range queries in Azure Table Storage?

```
In a point query, when an application retrieves a single row, the partition key
enables Azure to quickly hone in on the correct partition, and the row key lets
Azure identify the row in that partition. In a range query, the application
searches for a set of rows in a partition, specifying the start and end point of
the set as row keys. This type of query is also very quick, as long as you have
designed your row keys according to the requirements of the queries performed by
your application.
```

113. You need to define a schema for Azure Table Storage. Is this correct?

False

Azure Table Storage tables are schemaless. It's easy to adapt your data as the needs of your application evolve.

114. What are the advantages of using Azure Table Storage?

It’s simpler to scale. It takes the same time to insert data in an empty table, or a table with billions of entries. An Azure storage account can hold up to 500 TB of data.

A table can hold semi-structured data. There’s no need to map and maintain the complex relationships typically required by a normalized relational database.
Row insertion is fast. Data retrieval is fast, if you specify the partition and row keys as query criteria.

115. What are the disadvantages of using Azure Table Storage?

Consistency needs to be given consideration as transactional updates across multiple entities aren't guaranteed
There's no referential integrity; any relationships between rows need to be maintained externally to the tableIt's difficult to filter and sort on non-key data. Queries that search based on non-key fields could result in full table scans

116. What are the use cases of Azure Table Storage?

Storing TBs of structured data capable of serving web scale applications. Examples include product catalogs for eCommerce applications, and customer information, where the data can be quickly identified and ordered by a composite key. In the case of a product catalog, the partition key could be the product category (such as footwear), and the row key identifies the specific product in that category (such as climbing boots).
Storing datasets that don't require complex joins, foreign keys, or stored procedures, and that can be denormalized for fast access. In an IoT system, you might use Azure Table Storage to capture device sensor data. Each device could have its own partition, and the data could be ordered by the date and time each measurement was captured.
Capturing event logging and performance monitoring data. Event log and performance information typically contain data that is structured according to the type of event or performance measure being recorded. The data could be partitioned by event or performance measurement type, and ordered by the date and time it was recorded. Alternatively, you could partition data by date, if you need to analyze an ordered series of events and performance measures chronologically. If you want to analyze data by type and date/time, then consider storing the data twice, partitioned by type, and again by date. Writing data is fast, and the data is static once it has been recorded.

117. Azure Table Storage is intended to support very large volumes of data, up to several hundred TBs in size. Is this correct?

```
True
```

118. You need to create a storage account before creating an Azure Table Storage. Is this correct?

```
True
```

119. What is Azure Blob Storage?

```
Azure Blob storage is a service that enables you to store massive amounts of
unstructured data, or blobs, in the cloud.
```

120. You need to create a storage account before creating an Azure Blob Storage. Is this correct?

```
True
```

121. What are the different types of the blob that Azure Blob Service Supports?
Block blobs:

```
A block blob is handled as a set of blocks. Each block can vary in size, up to 100
MB. A block blob can contain up to 50,000 blocks, giving a maximum size of over 4.7
TB. The block is the smallest amount of data that can be read or written as an
individual unit. Block blobs are best used to store discrete, large, binary objects
that change infrequently.
```

Page blobs:

```
A page blob is organized as a collection of fixed size 512-byte pages. A page blob
is optimized to support random read and write operations; you can fetch and store
data for a single page if necessary. A page blob can hold up to 8 TB of data. Azure
uses page blobs to implement virtual disk storage for virtual machines.
```

Append blobs:

```
An append blob is a block blob optimized to support append operations. You can only add blocks to the end of an append blob; updating or deleting existing blocks isn't supported. Each block can vary in size, up to 4 MB. The maximum size of an append blob is just over 195 GB.
```

122. You control who can read and write blobs inside a container at the container level. Is this true?

```
True
```

123. How many access tiers that Blob Storage provides?

The Hot Tier:

```
The Hot tier is the default. You use this tier for blobs that are accessed frequently. The blob data is stored on high-performance media.
```

The Cool tier:

```
This tier has lower performance and incurs reduced storage charges compared to the Hot tier. Use the Cool tier for data that is accessed infrequently. It's common for newly created blobs to be accessed frequently initially, but less so as time passes. In these situations, you can create the blob in the Hot tier, but migrate it to the Cool tier later. You can migrate a blob form the Cool tier back to the Hot tier.
```

The Archive tier:

```
This tier provides the lowest storage cost, but with increased latency. The Archive tier is intended for historical data that mustn't be lost, but is required only rarely. Blobs in the Archive tier are effectively stored in an offline state. Typical reading latency for the Hot and Cool tiers is a few milliseconds, but for the Archive tier, it can take hours for the data to become available. To retrieve a blob from the Archive tier, you must change the access tier to Hot or Cool. The blob will then be rehydrated. You can read the blob only when the rehydration process is complete.
```

124. You are saving objects into Blob Storage with Hot tier and you want to move these files after 6 months to the Archive tier as we no longer need to access the files. How do you handle this scenario?

```
You can create lifecycle management policies for blobs in a storage account. A
lifecycle management policy can automatically move a blob from Hot to Cool, and
then to the Archive tier, as it ages and is used less frequently (policy is based
on the number of days since modification).
```

125. What are the use cases of Azure Blob Storage?

```
Serving images or documents directly to a browser, in the form of a static website.
Storing files for distributed access
Streaming video and audio
Storing data for backup and restore, disaster recovery, and archiving
Storing data for analysis by an on-premises or Azure-hosted service
```

126. What should you do if you want to maintain and restore earlier versions of a blob?

```
Versioning
```

127. Which feature should you enable if you want to recover a blob that has been removed or overwritten by accident?

```
Soft Delete
```

128. What are snapshots?

```
A snapshot is a read-only version of a blob at a particular point in time.
```

129. What is the Change Feed option in Blob Storage Service?

```
The change feed for a blob provides an ordered, read-only, record of the updates
made to a blob.
```

130. What is Azure File Storage?

```
Azure File Storage enables you to create files shares in the cloud, and access
these file shares from anywhere with an internet connection.
```

131. What are the features of Azure File Storage?

```
Azure File Storage enables you to share up to 100 TB of data in a single storage
account.
The maximum size of a single file is 1 TiBAzure File Storage supports up to 2000
concurrent connections per shared file.
```

132. How do you upload files into Azure File Storage?

```
You can upload files to Azure File Storage using the Azure portal, or tools such as
the AzCopy utility. You can also use the Azure File Sync service to synchronize
locally cached copies of shared files with the data in Azure File Storage.
```

133. How many tiers that Azure File Storage Offers?

```
Azure File Storage offers two performance tiers. The Standard tier uses hard disk-
based hardware in a datacenter, and the Premium tier uses solid-state disks. The
Premium tier offers greater throughput, but is charged at a higher rate.
```

134. What are some of the use cases of Azure File Storage?

```
Migrate existing applications to the cloud.Share server data across on-premises and
cloud.Integrate modern applications with Azure File Storage.Simplify hosting High
Availability (HA) workload data.
```

135. You shouldn't use Azure File Storage for files that can be written by multiple concurrent processes simultaneously. Why?

```
Don't use Azure File Storage for files that can be written by multiple concurrent
processes simultaneously. Multiple writers require careful synchronization,
otherwise the changes made by one process can be overwritten by another. The
alternative solution is to lock the file as it is written, and then release the
lock when the write operation is complete. However, this approach can severaly
impact concurrency and limit performance.
```

136. What is Azure Cosmos DB?

```
Azure Cosmos DB is a multi-model NoSQL database management system. Cosmos DB
manages data as a partitioned set of documents. A document is a collection of
fields, identified by a key. The fields in each document can vary, and a field can
contain child documents. Many document databases use JSON (JavaScript Object
Notation) to represent the document structure. In this format, the fields in a
document are enclosed between braces, { and }, and each field is prefixed with its
name.Cosmos DB is a highly scalable database management system. Cosmos DB
automatically allocates space in a container for your partitions, and each
partition can grow up to 10 GB in size. Indexes are created and maintained
automatically.
```

137. What are the different APIs that Cosmos DB supports?

SQL API:

```
This interface provides a SQL-like query language over documents, enable to
identify and retrieve documents using SELECT statements.Table API. This interface
enables you to use the Azure Table Storage API to store and retrieve documents. The
purpose of this interface is to enable you to switch from Table Storage to Cosmos
DB without requiring that you modify your existing applications.
```

MongoDB API:

```
MongoDB is another well-known document database, with its own programmatic
interface. Many organizations run MongoDB on-premises. You can use the MongoDB API
for Cosmos DB to enable a MongoDB application to run unchanged against a Cosmos DB
database.
```

Cassandra API:

```
Cassandra is a column family database management system. This is another database
management system that many organizations run on-premises. The Cassandra API for
Cosmos DB provides a Cassandra-like programmatic interface for Cosmos DB. Cassandra
API requests are mapped to Cosmos DB document requests.
```

Gremlin API:

```
The Gremlin API implements a graph database interface to Cosmos DB. A graph is a
collection of data objects and directed relationships. Data is still held as a set
of documents in Cosmos DB, but the Gremlin API enables you to perform graph queries
over data. Using the Gremlin API you can walk through the objects and relationships
in the graph to discover all manner of complex relationships
```

The primary purpose of the Table, MongoDB, Cassandra, and Gremlin APIs is to support existing applications. If you are building a new application and database, you should use the SQL API.

138. Cosmos DB guarantees less than 10-ms latencies for both reads (indexed) and writes at the 99th percentile, all around the world. Is this true?

```
True
```

139. What are some of the scenarios where CosmosDB is suitable?

IoT and telematics:

```
These systems typically ingest large amounts of data in frequent bursts of
activity. Cosmos DB can accept and store this information very quickly. The data
can then be used by analytics services, such as Azure Machine Learning, Azure
HDInsight, and Power BI. Additionally, you can process the data in real-time using
Azure Functions that are triggered as data arrives in the database.
```

Retail and marketing:

```
Microsoft uses CosmosDB for its own e-commerce platforms that run as part of
Windows Store and Xbox Live. It's also used in the retail industry for storing
catalog data and for event sourcing in order processing pipelines.
```

Gaming:

```
The database tier is a crucial component of gaming applications. Modern games
perform graphical processing on mobile/console clients, but rely on the cloud to
deliver customized and personalized content like in-game stats, social media
integration, and high-score leaderboards. Games often require single-millisecond
latencies for reads and write to provide an engaging in-game experience. A game
database needs to be fast and be able to handle massive spikes in request rates
during new game launches and feature updates.
```

Web and mobile applications:

```
Azure Cosmos DB is commonly used within web and mobile applications, and is well
suited for modeling social interactions, integrating with third-party services, and
for building rich personalized experiences. The Cosmos DB SDKs can be used to build
rich iOS and Android applications using the popular Xamarin framework.
```

140. What are the elements of an Azure Table storage key?

```
Partition key and row key
```

141. When should you use a block blob, and when should you use a page blob?

```
Use a page block for blobs that require random read and write access. Use a block
blob for discrete objects that change infrequently.
```

142. Why might you use Azure File storage?

```
To enable users at different sites to share files.
```

143. You are building a system that monitors the temperature throughout a set of office blocks and sets the air conditioning in each room in each block to maintain a pleasant ambient temperature. Your system has to manage the air conditioning in several thousand buildings spread across the country/region, and each building typically contains at least 100 air-conditioned rooms. What type of NoSQL datastore is most appropriate for capturing the temperature data to enable it to be processed quickly?

```
Send the data to an Azure Cosmos DB database and use Azure Functions to process the
data.
```

144. What are the several tools that you can use to provision services?

```
The Azure portal.
The Azure command-line interface (CLI)
Azure PowerShell
Azure Resource Manager templates
```

145. What is replication and how many options we have when provisioning a Storage account?

```
Data in an Azure Storage account is always replicated three times in the region you specify as the primary location for the account. Azure Storage offers multiple options for how your data is replicated in the primary region and secondary region:

Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the region. LRS is the least expensive replication option, but isn't recommended for applications requiring high availability.

Geo-redundant storage (GRS) copies your data synchronously three times within a single physical location in the primary region using LRS. It then copies your data asynchronously to a single physical location in the secondary region. This form of replication protects you against regional outages.

Read-access geo-redundant storage (RA-GRS) replication is an extension of GRS that provides direct read-only access to the data in the secondary location. In contrast, the GRS option doesn't expose the data in the secondary location, and it's only used to recover from a failure in the primary location. RA-GRS replication enables you to store a read-only copy of the data close to users that are located in a geographically distant location, helping to reduce read latency times.
```

146. The default connectivity for Azure Cosmos DB and Azure Storage is to enable access to the world at large. You can connect to these services from an on-premises network, the internet, or from within an Azure virtual network. Is this correct?

```
True
```

147. What are the options for protecting Azure resources such as storage account, Azure cosmos DB, etc?

```
Azure Private Endpoint
Firewalls and virtual networks
Configure authentication
Configure access control
Configure advanced security
```

148. When you configure CosmosDB for replication what is the default behavior?

```
By default, only the region in which you created the account supports write operations; the replicas are all read-only.
```

149. Replication is asynchronous for Cosmos DB. Is this true?

```
True

Replication is asynchronous, so there's likely to be a lag between a change made in one region, and that change becoming visible in other regions.
```

150. What is eventual consistency?

```
This option is the least consistent. It's based on the situation just described. Changes won't be lost, they'll appear eventually, but they might not appear immediately. Additionally, if an application makes several changes, some of those changes might be immediately visible, but others might be delayed; changes could appear out of order.
```

151. What is Consistent Prefix Option?

```
This option ensures that changes will appear in order, although there may be a delay before they become visible. In this period, applications may see old data.
```

152. What is the Session option?

```
If an application makes a number of changes, they'll all be visible to that application, and in order. Other applications may see old data, although any changes will appear in order, as they did for the Consistent Prefix option. This form of consistency is sometimes known as read your own writes.
```

153. What is the bounded Staleness option?

```
There's a lag between writing and then reading the updated data. You specify this staleness either as a period of time, or number of previous versions the data will be inconsistent for.
```

154. What is a strong consistency option?

```
All writes are only visible to clients after the changes are confirmed as written successfully to all replicas. This option is unavailable if you need to distribute your data across multiple global regions.
```

155. Which consistency option of CosmosDB provides the lowest latency and least consistency?

```
Eventual Consistency
```

156. What is the shared access signature?

```
You can use shared access signatures (SAS) to grant limited rights to resources in an Azure storage account for a specified time period. This feature enables applications to access resources such as blobs and files, without requiring that they're authenticated first.
```

157. What is a security principal?

```
An object that represents a user, group, service, or managed identity that is requesting access to Azure resources.
```

158. What is the advantage of using multi-region replication with Cosmos DB?

```
Availability is increased.
```

159. What is the operator that you use as part of the SELECT clause to eliminate duplicates in the result data?

```
DISTINCT
```

160. Name some of the aggregate functions?

```
COUNT(p): This function returns a count of the number of instances of field p in the result set.

SUM(p): This function returns the sum of all the instances of field p in the result set. The values of p must be numeric.

AVG(p): This function returns the mathematical mean of all the instances of field p in the result set. The values of p must be numeric.

MAX(p). This function returns the maximum value of field p in the result set.

MIN(p). This function returns the minimum value of field p in the result set.
```

161. A container provides a convenient way of grouping related blobs together, and you can organize blobs in a hierarchy of folders inside a container, similar to files in a file system on the disk. Is this true?

```
True
```

162. What is a command-line utility optimized for transferring large files (and blobs) between your local computer and Azure File storage?

```
AzCopy
```

Describe an analytics workload on Azure (25–30%)

Practice questions based on these concepts

- Describe analytics workloads
- Describe the components of a modern data warehouse
- Describe data ingestion and processing on Azure
- Describe data visualization in Microsoft Power BI

163. Describe modern data warehousing?

```
A data warehouse gathers data from many different sources within an organization. This data is then used as the source for analysis, reporting, and online analytical processing (OLAP). The focus of a data warehouse is to provide answers to complex queries, unlike a traditional relational database, which is focused on transactional performance.

Data warehouses have to handle big data. Big data is the term used for large quantities of data collected in escalating volumes, at higher velocities, and in a greater variety of formats than ever before. It can be historical (meaning stored) or real time (meaning streamed from the source). Businesses typically depend on their big data to help make critical business decisions.
```

164. What is Azure Data Factory?

```
Azure Data Factory is described as a data integration service. The purpose of Azure Data Factory is to retrieve data from one or more data sources, and convert it into a format that you process.
```

165. Your data might contain dates and times formatted in different ways in different data sources. You can use ___to transform these items into a single uniform structure.

```
Azure Data Factory
```

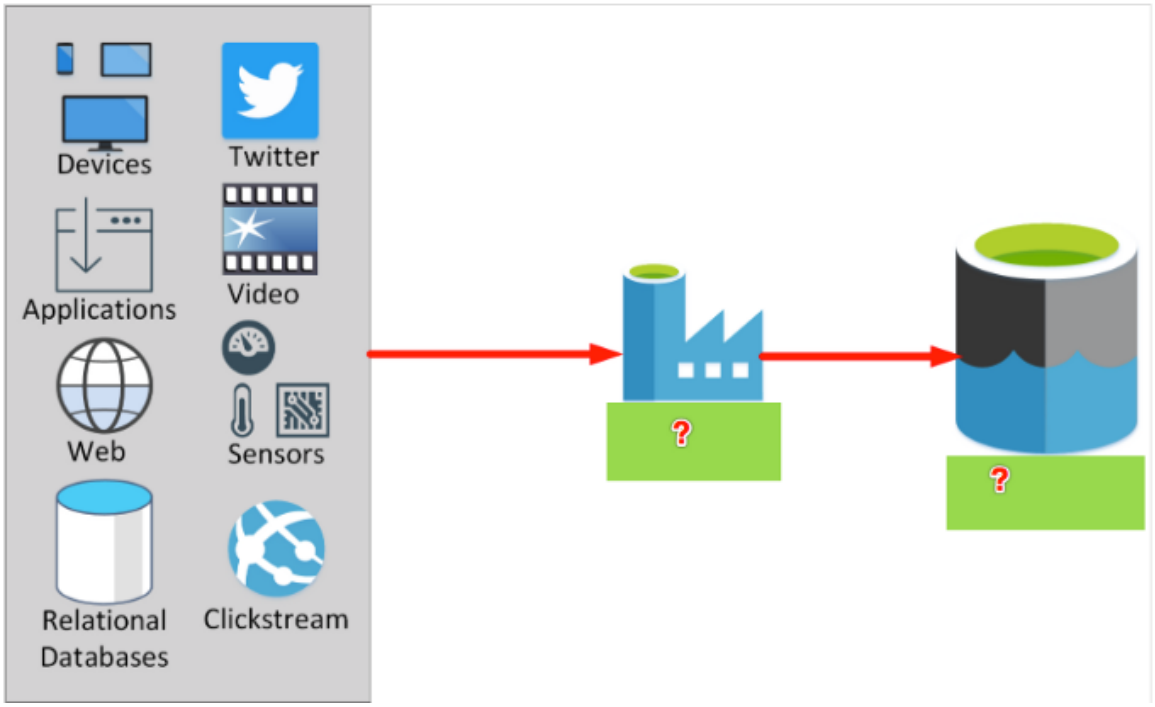
166. A pipeline can run continuously, as data is received from the various data sources. You can create pipelines using the graphical user interface provided by Microsoft, or by writing your own code. Which service is this?

```
Azure Data Factory
```

167. What is Azure Data Lake Storage?

```
A data lake is a repository for large quantities of raw data. Because the data is
raw and unprocessed, it's very fast to load and update, but the data hasn't been
put into a structure suitable for efficient analysis. You can think of a data lake
as a staging point for your ingested data, before it's massaged and converted into
a format suitable for performing analytics.
```

168. You are taking the data from different sources for data warehousing and processing. What are the services you should use in the missing parts of the following data warehouse solution?




```
Azure Data Factory
Azure Data Lake Storage
```

169. Azure Data Lake Storage is essentially an extension of Azure Blob storage, organized as a near-infinite file system. Is this true?

```
True
```

170. _____organizes your files into directories and subdirectories for improved file organization. Blob storage can only mimic a directory structure.

```
Data Lake Storage
```

171. What is Azure Databricks?

```
Azure Databricks is an Apache Spark environment running on Azure to provide big
data processing, streaming, and machine learning.
```

172. What is Azure Synapse Analytics?

```
Azure Synapse Analytics is an analytics engine. It's designed to process large
amounts of data very quickly.
Using Synapse Analytics, you can ingest data from external sources, such as flat
files, Azure Data Lake, or other database management systems, and then transform
and aggregate this data into a format suitable for analytics processing. You can
perform complex queries over this data and generate reports, graphs, and charts.
```

173. ____leverages a massively parallel processing (MPP) architecture.

```
Azure Synapse Analytics
```

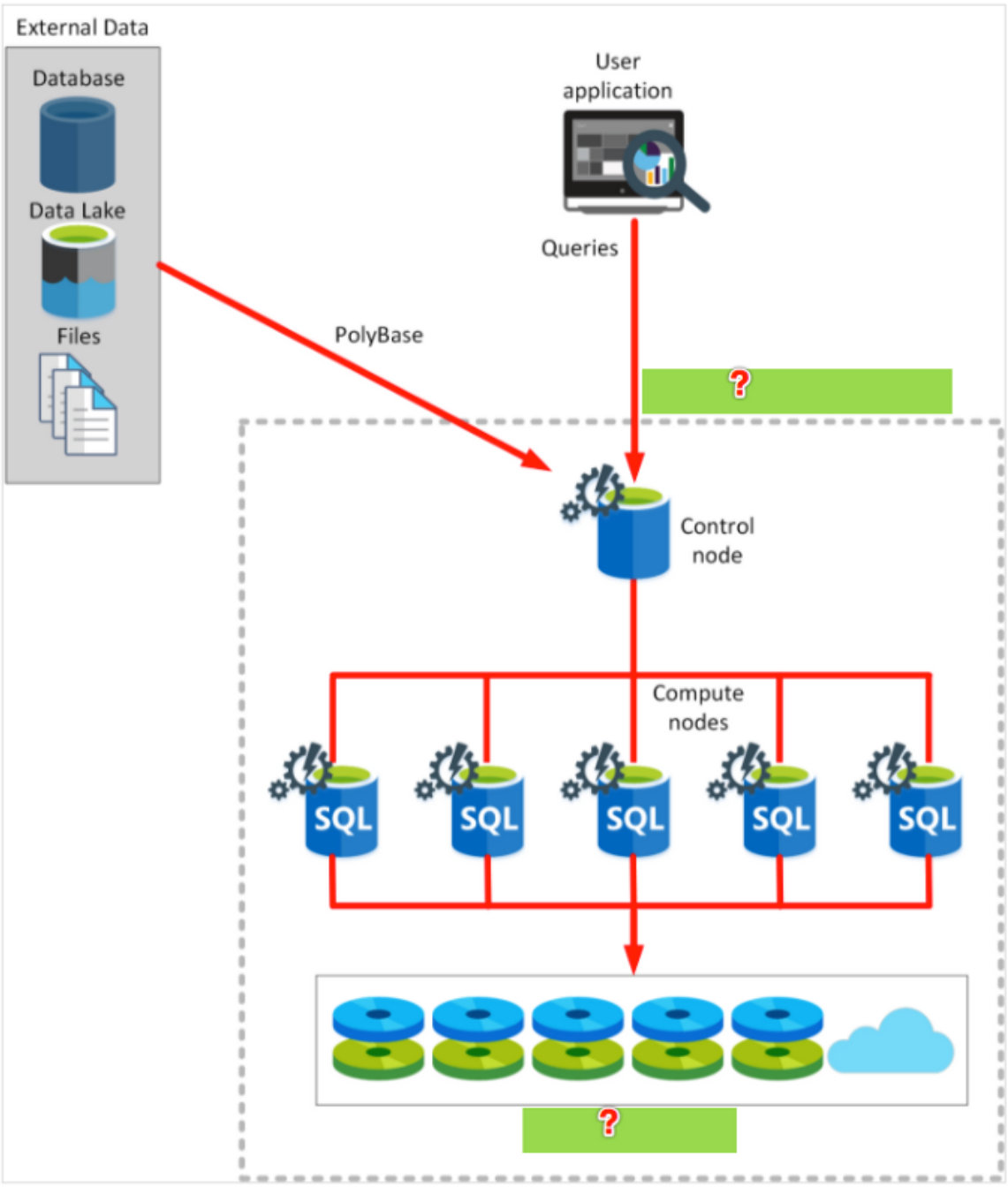
174. Azure Synapse Analytics includes a control node and a pool of compute nodes. Explain the architecture?

```
The Control node is the brain of the architecture. It's the front end that
interacts with all applications. The MPP engine runs on the Control node to
optimize and coordinate parallel queries. When you submit a processing request, the
Control node transforms it into smaller requests that run against distinct subsets
of the data in parallel.The Compute nodes provide the computational power. The data
to be processed is distributed evenly across the nodes. Users and applications send
processing requests to the control node. The control node sends the queries to
compute nodes, which run the queries over the portion of the data that they each
hold. When each node has finished its processing, the results are sent back to the
control node where they're combined into an overall result.
```

175. Azure Synapse Analytics supports two computational models. What are those?

```
SQL pools
Spark pools.
```

176. Azure Synapse Analytics supports two computational models. We are using the SQL pool in the following design. What is missing in the following design?




```
Azure Synapse Analytics
Azure Storage
```

177. What is polybase in the above design?

```
Azure Synapse Analytics uses a technology named PolyBase. PolyBase enables you to retrieve data from relational and non-relational sources, such as delimited text files, Azure Blob Storage, and Azure Data Lake Storage. You can save the data read in as SQL tables within the Synapse Analytics service.
```

178. In a Spark pool, the nodes are replaced with a __?

```
Spark cluster
```

179. What is Azure Analysis Services?

```
Azure Analysis Services enables you to build tabular models to support online
analytical processing (OLAP) queries. You can combine data from multiple sources,
including Azure SQL Database, Azure Synapse Analytics, Azure Data Lake store, Azure
Cosmos DB, and many others.You use these data sources to build models that
incorporate your business knowledge. A model is essentially a set of queries and
expressions that retrieve data from the various data sources and generate results.
The results can be cached in-memory for later use, or they can be calculated
dynamically, directly from the underlying data sources.
```

180. What is the difference between Analysis Services and Synapse Analytics?

```
Azure Analysis Services has significant functional overlap with Azure Synapse
Analytics, but it's more suited for processing on a smaller scale.
```

181. Use ___ for Very high volumes of data (multi-terabyte to petabyte sized datasets) and Very complex queries and aggregations.

```
Azure Synapse Analytics
```

182. Use ____ for Smaller volumes of data (a few terabytes) and Multiple sources that can be correlated.

```
Azure Analysis Services
```

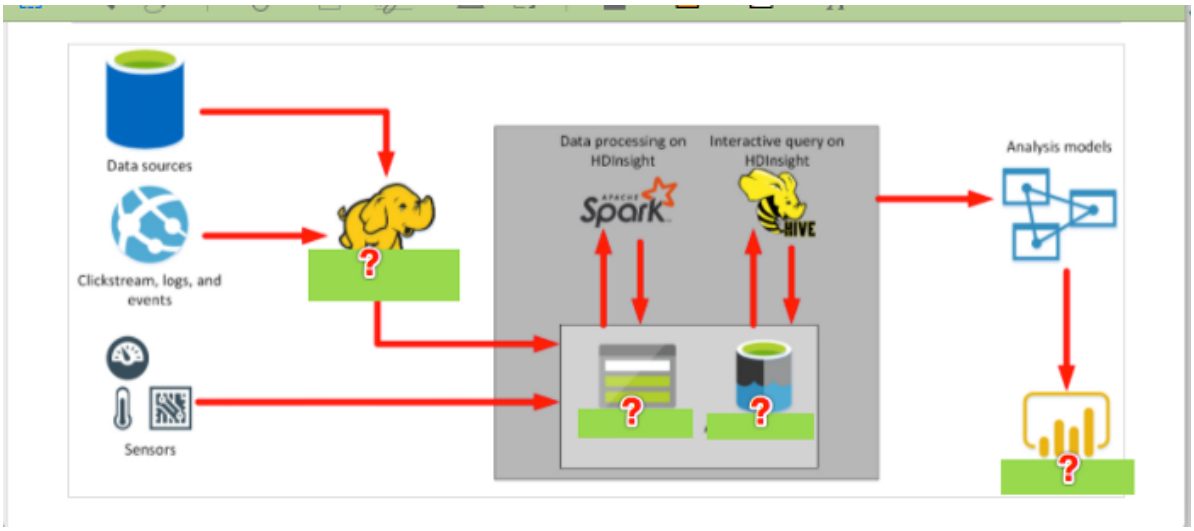
183. Can you combine Analysis Services with Synapse Analytics?

```
Yes
Many scenarios can benefit from using Synapse Analytics and Analysis Services
together. If you have large amounts of ingested data that require preprocessing,
you can use Synapse Analytics to read this data and manipulate it into a model that
contains business information rather than a large amount of raw data. The
scalability of Synapse Analytics gives it the ability to process and reduce many
terabytes of data down into a smaller, succinct dataset that summarizes and
aggregates much of this data. You can then use Analysis Services to perform
detailed interrogation of this information, and visualize the results of these
inquiries with Power BI.
```

184. What is Azure HDInsight?

```
Azure HDInsight is a big data processing service, that provides the platform for
technologies such as Spark in an Azure environment. HDInsight implements a
clustered model that distributes processing across a set of computers. This model
is similar to that used by Synapse Analytics, except that the nodes are running the
Spark processing engine rather than Azure SQL Database.
```

185. What are the missing in the following design?




```
Apache Hadoop on HDInsight
Azure Storage
Azure Data Lake
Power BI
```

186. When should you use Azure Synapse Analytics?

```
To perform very complex queries and aggregations
```

187. What is the purpose of data ingestion?

```
To capture data flowing into a data warehouse system as quickly as possible
```

188. What is the primary difference between a data lake and a data warehouse?

```
A data lake holds raw data, but a data warehouse holds structured information
```

189. Which component of an Azure Data Factory can be triggered to run data ingestion tasks?

```
Pipeline
```

190. When might you use PolyBase?

```
To query data from external data sources from Azure Synapse Analytics
```

191. Which of these services can be used to ingest data into Azure Synapse Analytics?

```
Azure Data Factory
```

192. You have a large amount of data held in files in Azure Data Lake storage. You want to retrieve the data in these files and use it to populate tables held in Azure Synapse Analytics. Which processing option is most appropriate?

```
Synapse SQL pool
```

193. Which of the components of Azure Synapse Analytics allows you to train AI models using AzureML?

```
Synapse Spark
```

194. In Azure Databricks how do you change the language a cell uses?

```
The first line in the cell is %language. For example, %scala.
```

195. What is the common flow of activity in Power BI?

```
Bring data into Power BI Desktop and create a report, share it to the Power BI
service, view and interact with reports and dashboards in the service and Power BI
mobile.
```

196. Which of the following are building blocks of Power BI?

```
Visualizations, datasets, reports, dashboards, tiles.
```

197. A collection of ready-made visuals, pre-arranged in dashboards and reports is called what in Power BI?

```
An app.
```

198. Power BI consists of three main elements. What are those elements?

```
Power BI Desktop
Power BI service
Power BI Mobile
```

199. What are the basic building blocks of Power BI?

```
Visualizations – A visual representation of data, sometimes just called visuals
Datasets – A collection of data that Power BI uses to create visualizations
Reports – A collection of visuals from a dataset, spanning one or more pages
Dashboards – A single-page collection of visuals built from a report
Tiles – A single visualization on a report or dashboard
```

200. An ___ is a collection of preset, ready-made visuals and reports that are shared with an entire organization.

```
App
```

201. What is the common flow when using Power BI?

```
1. Bring data into Power BI Desktop, and create a report.
2. Publish to the Power BI service, where you can create new visualizations or
build dashboards.
3. Share dashboards with others, especially people who are on the go.
4. View and interact with shared dashboards and reports in Power BI Mobile apps.
```