CLOUDYML

WWW.CLOUDYML.COM

100+ DATA ANALYTICS INTERVIEW QNAs

Akash Raj
Data Scientist

1. What is Data Integrity?

Data Integrity is the assurance of accuracy and consistency of data over its entire life-cycle and is a critical aspect of the design, implementation, and usage of any system which stores, processes, or retrieves data. It also defines integrity constraints to enforce business rules on the data when it is entered into an application or a database.

2. What is the Difference Between Joining and Blending in Tableau?

Combining the data from two or more different sources is data blending, such as Oracle, Excel, and SQL Server. In data blending, each data source contains its own set of dimensions and measures. Combining the data between two or more tables or sheets within the same data source is data joining. All the combined tables or sheets contain a common set of dimensions and measures.

3. What is slicing in Python?

When data is ingested into Power BI, it is basically stored in Fact and Dimension tables.
Fact tables: The central table in a star schema of a data warehouse, a fact table stores quantitative information for analysis and is not normalized in most cases.
Dimension tables: It is just another table in the star schema that is used to store attributes and dimensions that describe objects stored in a fact table.

4. What is the difference between NOW() and CURRENT_DATE() in SQL?

NOW() returns a constant time that indicates the time at which the statement began to execute. (Within a stored function or trigger, NOW() returns the time at which the function or triggering statement began to execute.
The simple difference between NOW() and CURRENT_DATE() is that NOW() will fetch the current date and time both in format 'YYYY-MM_DD HH:MM:SS' while CURRENT_DATE() will fetch the date of the current day 'YYYY-MM_DD'.

5. What's a Fourier transform?

A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. The Fourier transform finds the set of cycle speeds, amplitudes, and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain—it's a very common way to extract features from audio signals or other time series such as sensor data.

6. What is x-velocity in Power Pivot?

X-Velocity is the in-memory analytics engine behind Power Pivot that loads and handles huge data in Power BI. It stores data in columnar storage that results in faster processing.

7. Gantt chart in Tableau ?

A Tableau Gantt chart illustrates the duration of events as well as the progression of value across the period. Along with the time axis, it has bars. The Gantt chart is primarily used as a project management tool, with each bar representing a project job.

8. What in Excel is a macro?

An Excel macro is an algorithm or a group of steps that helps automate an operation by capturing and replaying the steps needed to finish it. Once the steps have been saved, you may construct a Macro that the user can alter and replay as often as they like.

9. What is the lambda function in Python?

Python Lambda Functions are anonymous function means that the function is without a name. As we already know that the def keyword is used to define a normal function in Python. Similarly, the lambda keyword is used to define an anonymous function in Python.
Eg. lambda_cube = lambda y: y*y*y

10. What is the difference between SQL and MySQL?

Autoencoders are artificial neural networks that learn without any supervision. Here, these networks have the ability to automatically learn by mapping the inputs to the corresponding outputs.
Autoencoders, as the name suggests, consist of two entities:
Encoder: Used to fit the input into an internal computation state
Decoder: Used to convert the computational state back into the output

11. What are Filters in Power BI?

The term "Filter" is self-explanatory. Filters are mathematical and logical conditions applied to data to filter out essential information in rows and columns. The following are the variety of filters available in Power BI:

- Manual filters
- Auto filters
- Include/Exclude filters
- Drill-down filters
- Cross Drill filters

12. What is concurrency control in DBMS?

This is a process of managing simultaneous operations in a database so that database integrity is not compromised. The following are the two approaches involved in concurrency control:
Optimistic approach – Involves versioning
Pessimistic approach – Involves locking

13. What is a checkpoint in DBMS and when does it occur?

A checkpoint is a mechanism where all the previous logs are removed from the system and are permanently stored on the storage disk. So, basically, checkpoints are those points from where the transaction log record can be used to recover all the committed data up to the point of crash.

14. What are groups in Tableau?

A group is a combination of dimension members that make higher level categories. For example, if you are working with a view that shows average test scores by major, you may want to group certain majors together to create major categories.

15. How are nested IF statements used in Excel?

The function IF() can be nested when we have multiple conditions to meet. The FALSE value in the first IF function is replaced by another IF function to make a further test.

16. What are the ways to detect outliers?

Box Plot Method: According to this method, the value is considered an outlier if it exceeds or falls below 1.5*IQR (interquartile range), that is, if it lies above the top quartile (Q3) or below the bottom quartile (Q1).

Standard Deviation Method: According to this method, an outlier is defined as a value that is greater or lower than the mean ± (3*standard deviation).

17. What is a Recursive Stored Procedure?

A stored procedure that calls itself until a boundary condition is reached, is called a recursive stored procedure. This recursive function helps the programmers to deploy the same set of code several times as and when required.

18. What is the shortcut to add a filter to a table in EXCEL?

The filter mechanism is used when you want to display only specific data from the entire dataset. By doing so, there is no change being made to the data. The shortcut to add a filter to a table is Ctrl+Shift+L.

19. What is DAX in Power BI?

DAX stands for Data Analysis Expressions. It's a collection of functions, operators, and constants used in formulas to calculate and return values. In other words, it helps you create new info from data you already have.

20. What is the Difference Between a Shallow Copy and Deep Copy in python?

Deepcopy creates a different object and populates it with the child objects of the original object. Therefore, changes in the original object are not reflected in the copy. copy.deepcopy() creates a Deep Copy. Shallow copy creates a different object and populates it with the references of the child objects within the original object. Therefore, changes in the original object are reflected in the copy. copy.copy creates a Shallow Copy.

21. How can you remove duplicate values in a range of cells?

To delete duplicate values in a column, select the highlighted cells, and press the delete button. After deleting the values, go to the 'Conditional Formatting' option present in the Home tab. Choose 'Clear Rules' to remove the rules from the sheet. 2. You can also delete duplicate values by selecting the 'Remove Duplicates' option under Data Tools present in the Data tab

22. Define shelves and sets in Tableau?

Shelves: Every worksheet in Tableau will have shelves such as columns, rows, marks, filters, pages, and more. By placing filters on shelves we can build our own visualization structure. We can control the marks by including or excluding data.

Sets: The sets are used to compute a condition on which the dataset will be prepared. Data will be grouped together based on a condition. Fields which is responsible for grouping are known assets. For example – students having grades of more than 70%.

23. What data sources can Power BI connect to?

The list of data sources for Power BI is extensive, but it can be grouped into the following:
Files: Data can be imported from Excel (.xlsx, xlxm), Power BI Desktop files (.pbix) and Comma Separated Value (.csv).
Content Packs: It is a collection of related documents or files that are stored as a group. In Power BI, there are two types of content packs, firstly those from services providers like Google Analytics, Marketo, or Salesforce, and secondly those created and shared by other users in your organization.
Connectors to databases and other datasets such as Azure SQL, Database and SQL, Server Analysis Services tabular data, etc.

24. What are the different integrity rules present in the DBMS?

The different integrity rules present in DBMS are as follows:
Entity Integrity: This rule states that the value of the primary key can never be NULL. So, all the tuples in the column identified as the primary key should have a value.
Referential Integrity: This rule states that either the value of the foreign key is NULL or it should be the primary key of any other relation.

25. What are some common clauses used with SELECT query in SQL?

Some common SQL clauses used in conjuction with a SELECT query are as follows:
WHERE clause in SQL is used to filter records that are necessary, based on specific conditions.
ORDER BY clause in SQL is used to sort the records based on some field(s) in ascending (ASC) or descending order (DESC).
GROUP BY clause in SQL is used to group records with identical data and can be used in conjunction with some aggregation functions to produce summarized results from the database.
HAVING clause in SQL is used to filter records in combination with the GROUP BY clause. It is different from WHERE, since the WHERE clause cannot filter aggregated records.

26. What is the difference between count, counta, and countblank in Excel?

The count function is very often used in Excel. Here, let's look at the difference between count, and it's variants - counta and countblank.

1. COUNT
It counts the number of cells that contain numeric values only. Cells that have string values, special characters, and blank cells will not be counted.

2. COUNTA
It counts the number of cells that contain any form of content. Cells that have string values, special characters, and numeric values will be counted. However, a blank cell will not be counted.

3. COUNTBLANK
As the name suggests, it counts the number of blank cells only. Cells that have content will not be taken into consideration.

27. What is Density-based Clustering?

Density-Based Clustering is an unsupervised machine learning method that identifies different groups or clusters in the data space. These clustering techniques are based on the concept that a cluster in the data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.
Partition-based(K-means) and Hierarchical clustering techniques are highly efficient with normal-shaped clusters while density-based techniques are efficient in arbitrary-shaped clusters or detecting outliers.

28. How to create empty tables with the same structure as another table?

To create empty tables:
Using the INTO operator to fetch the records of one table into a new table while setting a WHERE clause to false for all entries, it is possible to create empty tables with the same structure. As a result, SQL creates a new table with a duplicate structure to accept the fetched entries, but nothing is stored into the new table since the WHERE clause is active.

29. What is a Parameter in Tableau? Give an Example

A parameter is a dynamic value that a customer could select, and you can use it to replace constant values in calculations, filters, and reference lines.
For example, when creating a filter to show the top 10 products based on total profit instead of the fixed value, you can update the filter to show the top 10, 20, or 30 products using a parameter.

30. How will you write the formula for the following in Excel?

Multiply the value in cell A1 by 10, add the result by 5, and divide it by 2.
To write a formula for the above-stated question, we have to follow the PEDMAS Precedence.
The **correct answer** is ((A1*10)+5)/2.
Answers such as =A1*10+5/2 and =(A1*10)+5/2 are not correct. We must put parentheses brackets after a particular operation.

31. Define the term 'Data Wrangling

Data Wrangling is the process wherein raw data is cleaned, structured, and enriched into a desired usable format for better decision making. It involves discovering, structuring, cleaning, enriching, validating, and analyzing data. This process can turn and map out large amounts of data extracted from various sources into a more useful format.

32. What are the best methods for data cleaning?

Create a data cleaning plan by understanding where the common errors take place and keep all the communications open. Before working with the data, identify and remove the duplicates. This will lead to an easy and effective data analysis process.Focus on the accuracy of the data. Set cross-field validation, maintain the value types of data, and provide mandatory constraints.Normalize the data at the entry point so that it is less chaotic. You will be able to ensure that all information is standardized, leading to fewer errors on entry

33. Explain 4 steps to use CTE in sql.

All CTE starts with "with" clause.

After with you need to define CTE name and the field names. For instance in the below code snippet I have 3 fields Count,Column and Id. The name of CTE is "MyTemp".

Once you have defined CTE we need to specify the SQL which will give the result for the CTE.

Finally you can use the CTE in your SQL query

34. What are the various types of refresh options provided in Power BI?

Package refresh - This synchronizes your Power BI Desktop or Excel file between the Power BI service and OneDrive, or SharePoint Online.
Model or data refresh - This refreshes the dataset within the Power BI service with data from the original data source.

- **Tile refresh** - This updates the cache for tile visuals every 15 minutes on the dashboard once data changes.
- **Visual container refresh** - This refreshes the visible container and updates the cached report visuals within a report once the data changes.

35. What are Ensemble Methods?

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Random Forest is a type of ensemble method. The number of component classifier in an ensemble has a great impact on the accuracy of the prediction, although there is a law of diminishing results in ensemble construction.

36. What are constraints in SQL?

Constraints are the rules that we can apply on the type of data in a table. That is, we can specify the limit on the type of data that can be stored in a particular column in a table using constraints. NOT NULL, UNIQUE, DEFAULT, PRIMARY KEY, FOREIGN KEY, CHECK are the different constraints in SQL.

37. How do you apply a single format to all the sheets present in a workbook?

To apply the same format to all the sheets of a workbook, follow the given steps:
Right-click on any sheet present in that workbook
Then, click on the Select All Sheets option
Format any of the sheets and you will see that the format has been applied to all the other sheets as well

38. Define the term 'Data Wrangling.

Data Wrangling is the process wherein raw data is cleaned, structured, and enriched into a desired usable format for better decision making. It involves discovering, structuring, cleaning, enriching, validating, and analyzing data. This process can turn and map out large amounts of data extracted from various sources into a more useful format.

39. What are the best methods for data cleaning?

Create a data cleaning plan by understanding where the common errors take place and keep all the communications open. Before working with the data, identify and remove the duplicates. This will lead to an easy and effective data analysis process.Focus on the accuracy of the data. Set cross-field validation, maintain the value types of data, and provide mandatory constraints.Normalize the data at the entry point so that it is less chaotic. You will be able to ensure that all information is standardized, leading to fewer errors on entry

40. What is the difference between HAVING and WHERE clauses?

WHERE Clause is used to filter the records from the table based on the specified condition. HAVING Clause is used to filter records from the groups based on the specified condition. WHERE Clause can be used without GROUP BY Clause HAVING Clause cannot be used without GROUP BY Clause. WHERE Clause implements in row operations. HAVING Clause implements in column operation. WHERE Clause cannot contain aggregate function. HAVING Clause can contain aggregate function

41. Explain how relationships are defined in Power BI Desktop?

Relationships between tables are defined in two ways:
Manually - Relationships between tables are manually defined using primary and foreign keys.
Automatic - When enabled, this automated feature of Power BI detects relationships between tables and creates them automatically.

42. Mention the order of operations used in Excel while evaluating formulas.

The order of operations in Excel is referred to as PEMDAS. Shown below is the order of precedence while performing an Excel operation.

- Parentheses
- Exponentiation
- Division/Multiplication
- Addition
- Subtraction

43. What is map function in Python?

map function executes the function given as the first argument on all the elements of the iterable given as the second argument. If the function given takes in more than 1 arguments, then many iterables are given

44. How many report formats are available in Excel?

There are three report formats available in Excel; they are:
1. Compact Form
2. Outline Form
3. Tabular Form

45. What are sets in Tableau?

Sets are custom fields that define a subset of data based on some conditions. A set can be based on a computed condition, for example, a set may contain customers with sales over a certain threshold. Computed sets update as your data changes. Alternatively, a set can be based on specific data point in your view

46. What is the difference between DROP and TRUNCATE commands?

DROP command removes a table and it cannot be rolled back from the database whereas TRUNCATE command removes all the rows from the table.

47. Define the story in Tableau?

The story can be defined as a sheet which is a collection of series of worksheets and dashboards used to convey the insights of data. A story can be used to show the connection between facts and outcomes that impacts the decision-making process. A story can be published on the web or can be presented to the audience.

48. How to fetch unique records from a table in SQL?

SQL DISTINCT clause is used to remove the duplicated columns from the result set.
The distinct keyword is used with the select keyword in conjunction. It is helpful when we avoid duplicate values present in the specific columns/tables. The unique values are fetched when we use the distinct keyword.

49. How are Pivot tables used to filter data in Excel?

You can filter data according to your requirements with Excel Pivot tables. Place the field on which you want the data to be filtered. Then open the drop-down list of the field you put in the Filter area from the pivot table and choose your line.

50. What is R2? What are some other metrics that could be better than R2 and why??

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared does not measure goodness of fit. R-squared does not measure predictive error. R-squared does not allow you to compare models using transformed responses. R-squared does not measure how one variable explains another. Some better metrics that could be better than R2 are:

Mean Squared Error (MSE).
Root Mean Squared Error (RMSE).
Mean Absolute Error (MAE)

51. What is the curse of dimensionality??

The curse of dimensionality basically means that the error increases with the increase in the number of features. It refers to the fact that algorithms are harder to design in high dimensions and often have a running time exponential in the dimensions..

52. What are advantages of plotting your data before performing analysis ?

It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables

53. How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean? ?

A 95% confidence interval, for example, implies that were the estimation process repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

54. How do you deal with some of your predictors being missing?

Simple approaches include taking the average of the column and use that value, or if there is a heavy skew the median or mode might be better. A better approach, you can perform regression or nearest neighbor imputation on the column to predict the missing values. Then continue on with your analysis/model.

55. What is a sensitivity analysis in the decision making process?

Sensitivity analysis is a method for predicting the outcome of a decision if a situation turns out to be different compared to the key predictions. It helps in assessing the riskiness of a strategy. Helps in identifying how dependent the output is on a particular input value.

56. How do you interpret the data using statistical techniques

Most Important Methods For Statistical Data Analysis Mean.
1. Standard Deviation.
2. Regression.
3. Sample Size.
4. Determination.
5. Hypothesis Testing.

57. Explain the KNN imputation method.

A KNN (K-nearest neighbor) model is usually considered one of the most common techniques for imputation. It allows a point in multidimensional space to be matched with its closest k neighbors. By using the distance function, two attribute values are compared. Using this approach, the closest attribute values to the missing values are used to impute these missing values.

58. What is Map Reduce?

MapReduce facilitates concurrent processing by splitting petabytes of data into smaller chunks, and processing them in parallel on Hadoop commodity servers. In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application.

59. What is a Pivot Table?

A pivot table is a table of grouped values that aggregates the individual items of a more extensive table within one or more discrete categories.

60. Difference between 1-Sample T-test, and 2-Sample T-test?

The 2-sample t-test takes your sample data from two groups and boils it down to the t-value. The process is very similar to the 1-sample t-test, and you can still use the analogy of the signal-to-noise ratio. Unlike the paired t-test, the 2-sample t-test requires independent groups for each sample.

61. variance and covariance difference?

Variance and covariance are mathematical terms frequently used in statistics and probability theory. Variance refers to the spread of a data set around its mean value, while a covariance refers to the measure of the directional relationship between two random variable.