What kind of application you build?
Can you talk about that application?
What was your role and responsibility for that?
Roles and responsibilities related to Django app developer?
Which language you used?
Since how long have you been you working in Python?
Why you are leaving this company?
What kind of work you did in data engineering?
What kind of ETL work you did?
What was the technical stack?
How good you are in data engineering?
How you ensure you are writing quality code?
Have you written test cases?
Have you participated in code review? What did you check?
What is UAT?
What you did in code review?
Have you worked in Azure?
How to build pipeline in Azure?
Do you have any idea about [unclear]?
What kind of databases you have worked on?
Diff between SQL and NoSQL?
How you decide to use SQL or NoSQL?
Give me example when you use and you don't?
Features of MongoDB?
1. Machine Learning and Models
2. Overfitting
3. Training Set and Testing Set in Data
4. How to Handle Missing and Corrupted Data
5. How to Handle Missing Values
6. Confusion Matrix
7. Deep learning
8. Data types in Python
9. Decorators
10. Pickling and unpickling
11. OOPs concept
12. Second maximum in list of tension numbers
13. Second largest number in a list of 10 numbers, write simple code using functions
14. CI/CD pipeline
1. ETL Process
2. Getting data real-time with emp_age values like 25A and 25B - how to fetch age only?
3. Fetching age only in PySpark
4. Join large and small DataFrames in PySpark
5. Difference between two DataFrames with 1000 records each
6. Subtract records in PySpark
7. How to create a flat file
8. Types of Immigration
9. SQL query for the 2nd highest salary
10. Azure Data Factory (ADF)

1. What is dax query?
2. What is db utils in pyspark?
3. Read a CSV in pyspark.
4. Merge schema.
5. When we use (unclear or incomplete query)?
6. To write merge schema in pyspark.
7. 2nd highest salary in pyspark.
8. Same using rank.
9. Every day we are getting source file 10 col and today received extra 1 col and I don't want to that col and process automatically.
10. What is db utils in pyspark?
11. Incremental refresh in powerbi.
12. Which parameter is used to incremental refresh?
1. AWS Glue-Related Questions:
2. What is AWS Glue, and how does it work?
3. How do you handle schema changes in AWS Glue?
4. How does AWS Glue differ from AWS Data Pipeline?
5. How do you optimize AWS Glue jobs for performance?
6. What are Glue Crawlers, and how do they work?
7. What is the difference between Glue DynamicFrames and DataFrames?
8. Explain how AWS Glue integrates with Athena, Redshift, and S3.
9. What are Glue Workflows, and when do you use them?
10. How do you handle large datasets in AWS Glue?
11. How do you troubleshoot AWS Glue job failures?
12. Slow Changing Dimensions (SCD) and Data Warehousing Questions:
13. What are Slow Changing Dimensions (SCD), and why are they important in data warehousing?
14. Can you explain the different types of Slow Changing Dimensions (SCD)?
15. How would you implement Type 2 SCD in an ETL pipeline using AWS Glue?
16. How would you implement Type 3 SCD in a Glue ETL pipeline?
17. What is the difference between Type 1, Type 2, and Type 3 SCD?
18. What is a Hybrid (Type 6) SCD, and when would you use it?
19. How do you manage the size and performance of tables when dealing with Type 2 SC
In which scenarios would you use a separate history table (Type 4) in data warehousing?
General Data Integration and ETL Process Questions:
How do you ensure data quality during ETL processes using AWS Glue?
Can you describe an ETL pipeline where you needed to combine multiple data sources? How did you use AWS Glue for this?
What is the role of the Glue Data Catalog, and how does it improve the ETL process?
How do you handle error handling and retries in AWS Glue?
How would you handle versioning and testing in AWS Glue jobs?
How do you manage dependencies and scheduling of multiple Glue jobs?
Performance Optimization and Cost Management:
What are some common performance bottlenecks in AWS Glue, and how do you address them?
How do you monitor the cost and performance of AWS Glue jobs?
How do you handle large-scale joins and aggregations in Glue ETL jobs to optimize for performance?
1. Machine Learning and Models
2. Overfitting
3. Training Set and Testing Set in Data
4. How to Handle Missing and Corrupted Data
5. How to Handle Missing Values
6. Confusion Matrix

7. Deep learning
8. Data types in Python
9. Decorators
10. Pickling and unpickling
11. OOPs concept
12. Second maximum in list of tension numbers
13. Second largest number in a list of 10 numbers, write simple code using functions
14. CI/CD pipeline
1. Technical challenges faced in the recent project
2. Teradata
3. Scd
4. How we do current indicator
5. I want a convert starschema to snowflake schema
6. Use of snowflake schema
7. Factless fact table
8. Non clustered index
9. Intersept and inner join
10. View and materialized view
11. Window functions in sql
12. Window function
13. How LIKE operator is case sensitive or not in SQL
14. Give me all questions asked by me

1. *What kind of analysis you have done in your project?*
2. *Syntax in pandas to check null value.*
3. **What is the output of isnull()?**
4. *How to get the total null values in a whole CSV file?*
5. *What are the problems you faced while performing the EDA part?*
6. *What is your approach for handling missing values?*
7. *What is encapsulation?*
8. *What is polymorphism?*
9. *What is inheritance?*
10. *How do you have the option in Python for both single and multiple inheritance?*
11. *What is a crawler in AWS?*
12. *Can I edit the schema once created in AWS Glue?*
13. *What is the workflow of how data goes from you to S3?*
14. *What is a data warehouse?*
15. *What is a stored procedure?*
16. *What is a view?*
17. *I have millions of records, and I want to extract those records using MySQL or MSSQL. The query is taking more time. What approach should I use?*
18. *I am not using any structured database.*
19. *What is a data warehouse?*
20. *What is a stored procedure?*
21. *What is a view?*
22. *I have millions of records, and I want to extract them. The query is taking more time. What approach should I use?*
23. *I am not using any structured database.*
24. *What is encapsulation?*
25. *What is polymorphism?*

26. *What is inheritance?*
27. *How do you have the option in Python for both single and multiple inheritance?*
28. *What is a crawler in AWS?*
29. *Can I edit the schema once created in AWS Glue?*
30. *What is the workflow of how data goes from you to S3?*
31. *What is a data warehouse?*
32. *What is a stored procedure?*
33. *What is a view?*
34. *I have millions of records, and I want to extract those records using MySQL or MSSQL. The query is taking more time. What approach should I use?*
35. *I am not using any structured database.*
36. *How do you handle missing values in your data?*

1. During the ETL process, if the schema of the input changes, how will you handle it?

2. Single table called abc with columns orderid, customerid, and order_amount. I want the top 10 customers based on the largest order in terms of amount.

3. How to optimize SQL query?

4. If I want o migrate data from on-premise to AWS, how will you do that?

5. Lambda function code in Python to trigger an S3 bucket, read, process, and store the output in another bucket.

Star schema.
Data architect
Is it necessary to have numerical data in fact table.
Example of fact table what would be information in that table .
Which tool you used for etl processing
Steps to taken data engineer to ML
That make me to switch your career in ML
Can you tell me supervised and unsupervised learning and example of it ..
What is application of unsupervised
model where it can be applied
What is clustering and real time example
What is linear regression
What is logistics regression
What is hypothesis testing
What is p value
What is probality value
Have heard of z-test and t-test
What is class imbalance data
How will you evaluate a data ..
What is decision tree , Advantage and disadvantage
How to handle overfitting.
Handling null values in pyspark and How to checks in pyspar
Indexing in SQL
Data masking
Differnce between tuple and list in Python
What is a generator in Python?
How to convert a DataFrame to CSV?

What exactly is ETL?
In a table, I have id as a column with values 1, 2, 3, 4. I want the output as 1abc, 2abc, 3abc, 4abc in SQL.
I have duplicate records and I want to check duplicate records and count the duplicates.
What is the use of CTE?
I want to see a CTE for the student table with ID and name.
What is data partitioning?
Write a query for the rank function. We have id and marks in 2 columns, and we want to use the rank function.
I have the name "Aditi" and I want the output as "adi".
I have a list in Python with duplicates, and I want only unique values.
What is shell scripting?
What are double underscore methods in Python, and what is the use of dunder methods?
Difference between is and == in Python.
Can we rename a column in Pandas?
In a list, I have values like 364, and I want the cube of each output.
What are SQL procedures and functions?
Suppose I want to create a table and copy the schema without copying the data.
1. Intro
2. Experience with data warehouse ,aws ,tableau ,plsql
3. What is materialized schema
4. With clause
5. What is merge query
6. What window function have u worked on
7. What is rank and dense rank
8. Partioning and indexing
9. What is difference between star schema and snowflake schema
10. What is scd
11. Scd and its type type 0 type 1 type 2 type 3
12. What is a natural key?
13. What is candidate key
14. What is surrogate key
15. When you run pipeline how will handle etl
16. Bridge table
17. Which will be loaded first dimension or fact table
18. What is like in sql
19. What services u used in aws
20. Use of s3
21. What are different storage classes in s3
22. Advantages of redshift
23. How will u search a string in a file
24. How to find a string from file which is present in multiple directories
25. What is seed(unix command)
26. If you want to search a keyword in
27. How much exp in control enviroment
28. Libraries in python
29. What is decorator
30. What is generator
31. Define a list of integers 11,5,2,9,32.find min and max values from the list without using min/max functions.

32. write sql query
table 1 :columns id name,salary
table 2: columns emp_id mgr_id
find the name of manager of the employees with the highest salary
1. Lambda in Python
2. Synchronous and Asynchronous
3. Lambda in AWS
4. Limitations of AWS Lambda
5. TeraData to S3 - AWS service used
6. Steps for transferring data from TeraData to S3
7. How would you connect to TeraData in AWS? Which library would you use?
8. What is Dynamic Glue?
9. What is Lazy Evaluation in PySpark?
10. What happens when we submit a job to a cluster in Spark and how does it get divided?
11. Narrow and Wide transformations in Spark
12. Joins in SQL
13. Joins in Spark
14. Joins specific in Spark
15. Broadcast join
16. Query to get the top 3 employees from each department and age group based on performance rating using PySpark
17. How to read data from a file in PySpark?
18. PySpark steps for getting top 3 employees based on performance rating and age group
19. SQL query for getting top 3 employees based on performance rating, department, and age group from a CSV file in PySpark
20. How to find the factorial of a number in Python?
1. *Output Parquet:* How to transform and write a DataFrame to 10 Parquet files?
2. *What is adaptive query execution (AQE)?*
3. *Advanced features of Spark?*
4. *Small file issue in Spark?*
5. *What is data skewness?*
6. *How does AQE help in resolving data skewness?*
7. *What is shuffle data?*
8. *Default mode of shuffle partitioning?*
9. *Partitioning a 30GB CSV file into 5GB chunks—how many partitions are needed?*
10. *Default partition size in Spark?*
11. *City with the 3rd highest number of 4-wheelers and 3rd highest number of 2-wheelers?*
12. *SQL query to find the above results?*
13. *Why use ORDER BY for 4-wheelers?*
14. *What is PARTITION BY in SQL?*
15. *City with the 3rd highest number of 4-wheelers?*
16. *What is indexing in SQL?*
17. *Difference between UNION and UNION ALL?*
18. *How is GROUP BY different from a window function?*
19. *Number of columns returned by GROUP BY vs. window function?*
20. *What is _init_ in Python?*
21. *What are decorators in Python?*
22. **Difference between ls and ls -l commands?**
23. *What is multiple inheritance?*
24. *Does Python support multiple inheritance?*
25. **Transforming "aabbeeeeddddaaaeee" to "abedae"—code for the transformation?**

26. *What AWS services have you used?*
27. *What is EMR in AWS Glue?*
28. *Different storage types in S3?*
29. *What is a cold start in AWS Lambda?*
30. *Difference between Spot and On-Demand Instances in AWS?*
Iterate 2 list together at the same time"
"Does it affect the data (ReIndexing in pandas)"
"Append in pandas"
"Deep copy shallow copy"
"Static and class methods"
"Is python Call by value and call by function"
"Overriding in python"
"Generators"
"Iterators"
"ReIndexing in pandas"
"Does it affect the data (ReIndexing in pandas)"
"Append in pandas"
"Deep copy shallow copy"
"Static and class methods"
1. How good you are handling missing data?
2. There were 500 records coming from the source and there are 24 records that should be in your record but they are not there. What will you do?
3. There were 500 records coming from the source and there are 24 records that should be in your record but they are not there. What will you do, how will you apply mean, median, and mode?
4. How much you rate in SQL?
T1 | T2 — What is the count of INNER JOIN, LEFT JOIN, RIGHT JOIN, FULL JOIN, and CROSS JOIN with given data?
5. Query to find duplicates in SQL.
6. By using window functions, how to find employees who have the 2nd highest salary?
7. When do we use Autoloader?
8. What is enforcement and evolution?
9. How does Databricks manage fault tolerance?
10. I am loading 1 billion records from source to target, and I want it to be processed, but one record has a constraint violation. What will happen to my half records that are processed?
11. What will happen to my half records that are processed if there is a constraint violation?
12. In S3 data is stored in various partition data get process vaery soon and one data is getting to process very long what will be the reason ?
13. What is data skewness?
14. How do you handle data skewness?
15. I am loading 1 billion records from the source to the target. I want it to be processed, but 1 record has a constraint violation. What will happen to my half records that are processed?
16. What will happen to my half records that are processed if there is a constraint violation?
17. How do you handle missing data in a dataset with 500 records, and 24 records are missing?
18. What will happen to my data when there's a constraint violation during a large batch load?
19. How do you handle a dataset with skewed data or uneven distribution of records?
20. How do you rate yourself in SQL?
21. What is the impact of data skewness, and how do you handle it in a distributed environment?

Dimensional model
Multiple currency and exchange rate and where to put currency in fact table or dimension table
What is SCD (Slowly Changing Dimension)?
Difference between Snowflake schema and Star schema
What is Pandas?
Working on large datasets in Pandas
How to optimize performance in large datasets?
Abstraction
Difference between class and instance variables
Class and
Overloading and Overriding
object
Map in Python
Purpose of split() in Python
What is NumPy?
Users table and User Cards table SQL questions
List the users who have not transacted in the last 7 days, ordered by user ID.
Calculate the running total of users based on the enroll date.
State-code-wise count of users, debit cards, and credit cards.
Why do we use table aliases (u and uc)?
Give moderate temperature from a list (e.g., [75, 76, 78])
Find the second largest number in a list (e.g., [2, 3, 1, 2, 5])
What is the difference between dimensions and lookup tables?
What is a factless fact table?
Degenerate dimension: (This is a topic you asked for explanation about.)
Difference between Star Schema and Snowflake Schema?
How do you write an optimized query for a large dataset with billions of records?
Full Outer Join: (You asked for an explanation about full outer join.)
When do we use a self join?
How to find duplicates in a table using SQL?
What is the difference between a dimension table and a lookup table?
1. Any complex transformation you done in your banking project?
2. What kind of data sources you had?
3. Any performance issue you faced?
4. What transformation you did in AWS Glue?
5. How you do if data quality is bad?
6. How you used SCD?
7. Late arriving dimension in SCD?
8. Do you load late arriving data into fact table or not?
9. How does testing happen when the project is finished?
10. Table emp has 2 columns emp_no and emp_name emp no as 1, 2, 3 and emp name as A, B, C. I want output which has 4 columns in which 2 columns are added next_emp_no and next_emp_name and values will?
11. I want same output using PySpark.
1. *How to perform validation before dumping data into two buckets, one for selected and one for rejected records?*
2. *How to print details of employees who got at least two awards in the last year, with tables: employee and award?*
3. *How do you run a Spark application?*
4. *What is the spark-submit command with its options?*
5. *How to determine the number of stages in a Spark application?*

6. *How to increase the number of parallel tasks running in a Spark application?*
7. *How does Spark determine the number of stages in an application?*
8. *How can we run a Spark application faster and check the code?*
9. *How does Spark determine the number of stages in an application?*
10. *How to increase the number of parallel tasks in a Spark application?*
11. *What file formats have you used in your project?*
12. *How is Parquet different from CSV?*
1. *Parquet file*
2. *How does partitioning make your query run more efficiently?*
3. *What parameters you used in Airflow?*
4. *Branch operator in Airflow*
5. *What is an object in OOPs Python?*
6. *How does OOP give us more flexibility?*
1. *You are working on a local system and you want to take it on production and local is not on AWS and we want to take this on AWS as a stand-alone product, how would you take it? What would be your pipeline?*
2. *What is a good option, EC2 or any other?*
3. *Which one will be good, AWS or EC2, and why?*
4. *Why EC2 over serverless?*
5. *What is the role of ECS?*
6. *In loan management, where can we use ECS?*
7. *How will you build a CI/CD pipeline, i.e., we want to automate the pipeline?*
8. *Python program which will create a list of numbers between 1 to 100, i.e., any 5 random numbers between 1 to 100.*
9. *Write it in a function.*
10. *Mutable and immutable data types in Python.*
1. Tools similar to Databricks
2. What is partition in PySpark job execution?
3. How many partitions to divide into in PySpark?
4. What is data skewness in PySpark job execution?
5. Example of data imbalance (data skewness) in PySpark
6. Why will data go into a single partition?
7. What is persist() in PySpark and how to use it?
8. What is Unity Catalog?
AWS Glue Component
Data Catalog
Lifecycle Policy of S3
Purpose of AWS Glue
How to Extract Data from Oracle
Services Used for Data Injection
Joins in SQL
Seat Table Example and Transformation
Odd Values to Be Replaced by Even Values
Values to Be Replaced by Even Values and Vice Versa (SQL)
Replace Student Name and Not IDs (SQL)
If Last ID Is Odd, Keep the Value as Same Using Window Function
Wide and Narrow Transformation
Architecture of Spark
DataFrame
RDD
When *to Use RDD and When to Use DataFrame*

1. How many buckets can be created in s3?
2. What are zone and regions
3. Minimum and max size in s3 bucket?
4. Auto scaling?
5. How will you reduce the slowness?
6. EIP?
7. Comparison with elastic IP?
8. How to secure s3?
9. Poliicies of s3?
10. Main elements of bucket policies?
11. Snowball?
12. Redshift?
13. Access control vs security controls?
14. RDD?
15. SparkSession?
16. UDF?
17. We have a catalog and in that 4 schemas I want to get the schema wise table how many table in each schema using spark?
18. Schema table count: A 12, B 20
19. Query to remove duplicate in sql?
20. Nth highest salary?
21. Simple SQL query for nth highest salary?
22. SCD?
23. Narrow and wide transformation?
24. Rank dense rank diff?
25. SCD 1,2 ,3 difference

1. How to handle long queries with joins that take 1-2 hours to execute?
2. How to get runtime logs in MySQL?
3. How to get logs for query execution in MySQL?
4. How to create a custom layer in AWS Lambda?
5. How to handle slow-running queries in Pyspark?
6. How to pull data from S3 to Athena?
7. How to give input in AWS Lambda?
8. What happens when you insert a new key-value pair in a Python dictionary with the same key?
9. How to handle large SQL queries and joins in an optimized way?

1. *What are the transformations you have done in the ETL process?*
2. *Explain Spark architecture.*
3. *What is parallel processing?*
4. *What type of cluster have you used in your database notebook?*
5. *What is lazy evaluation in Spark?*
6. *How many jobs get created for an action in Spark?*
7. *Can you explain DAG (Directed Acyclic Graph)?*
8. *What is the difference between DataFrame and Dataset in Spark?*
9. *Why did you use DataFrame and not Dataset in your project?*
10. *Explain the different types of transformations in Spark.*
11. *What are the different optimization methods in Spark?*
12. *What are the different types of partitioning in Spark?*
13. *I have a string 'aaabbcccdddd'. I want the number of occurrences for each alphabet.*
14. *There is 'aaabbcdd'. I want to print the 1st non-repeating character.*
15. *Write a function to filter out numbers from a string and return only the alphabets.*

What is diff lambda and glue, and why are you using Glue?
How much data are you using?
PySpark architecture explanation.
Particular job is taking a lot of time. How to resolve the issue?
How does cache memory work in PySpark? If it fails in between, what happens?
What is broadcast join, and how does it work?
Best practices for creating a database and multiple tables in Redshift.
What should be the name of an S3 bucket?
Given tables A and B, divide column C as A/B in SQL. How would you write the query?
How can I display a float value in decimal places?
SQL query to find the student name who got the second-highest marks in the last four years.
How to find the longest common prefix among a list of strings (e.g., ["flower", "float", "flow"])?
How to optimize Python code?
SQL query to calculate the average marks for each class.
How to write the result of average marks for each class into a CSV file in PySpark?
1. Technical features of Spark?
2. Architecture of Spark?
3. Memory optimization of Spark?
4. I have 50 million rows in my business. When the business queries it, I have records of many years. I want to access all data but want to reduce query time. What should I do?
5. Difference between partitioning and bucketing?
6. Can I bucket the partitioning data?
7. Write a simple Python code for this pattern:

```
*

**

*

**

***

**
```

8. Write an SQL code for rank?
9. Steps of execution of Spark Job in the background?
10. How are jobs, stages, and tasks executed in Spark?
11. What are transformations and actions in Spark?
12. What does wide transformation do in Spark?
13. When scheduling happens, what gets created in Spark?
14. If I have 2 workers, each with 5 executors and every executor has 10 cores, how many tasks will be running in parallel?
15. I want to perform a task with 2 TB of load. Will I be able to do this with 2 workers, 5 executors, and 10 cores?
16. How can Spark achieve the task with 2 TB of data?
17. years of experience as a Data Architect working with multiple data engineers
1. Can you describe a use case or project you've delivered that was based on data architecture? What were the requirements and the business use case? What valuable insights did you gain from the requirements provided to you?

2. How do you manage data transformations from silver to gold in your architecture? What tools or technologies are you using in each layer (silver vs gold)?
3. Can you discuss any specific visualization tools you've worked with in your projects?
4. Have you had exposure to AWS services? If so, how have you utilized them in your architecture?
5. For data ingestion, how have you leveraged AWS Glue, and how does it fit with the AWS layer and Snowflake layer in your architecture?
6. Can you describe any experiences you've had working with RSS services or similar real-time streaming data sources?
7. How do you ensure data governance is maintained throughout the ETL process?
8. We often face challenges while ingesting data into Snowflake. How do you optimize performance for data loading in Snowflake?
9. Can you explain how you handle parallel processing in Snowflake to improve performance?
10. Can you elaborate on the benefits and challenges of using a multi-cluster architecture in Snowflake?
11. How does virtualization work in Snowflake, particularly with regard to storage and compute resources?
12. What are some challenges you've encountered when working with Snowflake architecture, and how did you resolve them?
13. When facing challenges with data ingestion, particularly when dealing with data from multiple sources and in different formats, how do you handle the complexity?
14. Take a scenario where there is duplicate data for a product or user. How would you handle this situation in your data pipeline?
15. What is the best algorithm or approach to handle data duplication, and why do you prefer using surrogate keys in such cases?
16. How do you ensure data integrity across your pipelines and systems?
17. How do you optimize SQL queries in your architecture? Do you avoid complex functions, and if so, why? How do window functions fit into your optimization strategy?
18. Can you explain your approach to micro-partitioning in Snowflake? How does it help with query optimization and storage efficiency?
19. Can you explain the difference between a data lake and a data warehouse? When would you use each, and why?
20. Have you faced any challenges while working with data science (DS) or machine learning (ML) models? Can you share some of your experiences?
21. How do you tackle scenarios where you encounter null values or dirty/clunky data in your pipelines?
22. Have you worked with image data in your data engineering projects? If so, how did you process and handle it?
23. In your architecture, you've used catalog tables. Could you explain if RDS or DynamoDB could be used in place of a data catalog? When would you prefer one over the other?
24. If you were to design a data pipeline in AWS, how would you approach this task? What services would you use, and why?
25. How do you ensure fault tolerance in your data pipelines, especially when handling large-scale data?
26. Given two migration scenarios — one using EC2 for data migration and the other using SageMaker and Lambda — which would you prefer and why?
27. How do you design data engineering use cases or pipelines, particularly in terms of scalability and performance?
28. When using Step Functions, you have the option of choosing synchronous or asynchronous execution. How do you decide which one to use in your workflow?

29. What are the limitations of AWS Lambda, and how do you work around them in your architecture?
30. If you were to use spot instances for daily jobs, would that increase your cost? How would you manage cost optimization in this case?
31. For large-scale data migration, such as transferring 1TB of data, how would you design the architecture? Which AWS services would you choose for this process?
32. In terms of data pipeline design, would you prefer using Databricks or AWS Glue? What are the key differences between the two in your opinion?
33. Can you explain the main differences between Redshift and Snowflake? How do you decide which to use for a given use case?
34. Have you worked with generative AI (GenAI) in machine learning models? If so, how did you integrate it into your workflow

==========

ADF and Linked Services
How do you build connected pipelines in Azure Data Factory (ADF)?
How many types of linked servicesare available in ADF?
Source and Destination Systems
What were your source and destination systems in a recent project? Were they file systems, databases, or a mix of both?
Handling Large Files
If you have a huge file to process, how would you handle it efficiently?
Platforms and Tools
Which platforms (on-premise or cloud) and languages have you used for data engineering?
SQL Joins and Relationships
Given two tables with ID columns containing data like (1, 2, NULL, NULL) and (1, 2, 3, NULL), how many records would you get with:
An inner join?
A left join?
A full outer join?
Slowly Changing Dimensions (SCD)
What is SCD?
Can you explain SCD Type 6?
Why is SCD Type 5 not commonly used?
Data Integration
What is VCNF (if applicable to your context)?
Have you worked with Snowflake? How would you integrate data from multiple cloud providers in Snowflake?
Can Snowflake fetch data directly from Amazon S3?
Scheduling and Credentials
Do you use scheduled pipelines or trigger them manually?
Have you used a secrets management tool like Azure Key Vault or AWS Secrets Manager for Databricks credentials?
Data Modeling
How have you designed many-to-many relationships in data models?
What is the necessity of normalization in database design?
If all data can be stored in a single table, why do we split it across multiple tables?
Parent-Child Relationships

What is a cascading delete in database design?
How have you handled issues in parent-child relationships caused by cascading deletes?
Views in Databases
Why would you use a normalized view instead of a materialized view, or vice versa?
How do you ensure consistency between normalized and materialized views?
Data Migration
When migrating data from a source to a sink, and receiving incremental daily updates, how do you handle the increasing data load?
Bulk Queries
Have you worked with bulk queries? How did you implement them in your projects?
Architectures and Concepts
What is a medallion architecture, and how does it apply to data engineering?
What are facts and dimensions in a data warehouse?
Data Governance and Security
Why do you think data governance is important?
Can you explain the concept of data masking and its use cases?
Scalability and Optimization
How do you ensure scalability, robustness, and optimization in your data engineering pipelines?
Data Lake vs. Data Warehouse
How is a data lake different from a data warehouse?
OLAP vs. OLTP
Can you explain the differences between OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing)?
Pyspark
Getting data real-time with emp_age values like 25A and 25B - how to fetch age only? Fetching age only in PySpark
Join large and small DataFrames in PySpark
Difference between two DataFrames with 1000 records each
Subtract records in PySpark
What is db utils in pyspark?
Read a CSV in pyspark.
Merge schema.
When we use (unclear or incomplete query)?
To write merge schema in pyspark.
2nd highest salary in pyspark.
What is Lazy Evaluation in PySpark?
What happens when we submit a job to a cluster in Spark and how does it get divided?
Narrow and Wide transformations in Spark
Joins in Spark
Joins specific in Spark
Broadcast join
Query to get the top 3 employees from each department and age group based on performance rating using PySpark.
How to read data from a file in PySpark?
SQL query for getting top 3 employees based on performance rating, department, and age group from a CSV file in PySpark.
Anti join in spark
Broadcast join
Difference between dataframe, dataset, RDD in spark
What factors determine parallelism in spark

If we have 4 executors and 80 partitions, how many jobs can run in parallel
If narrow transformation apply to a single node, how can we utilize parallel processing in spark
Why map and filter considered as narrow transformation?
Difference between context and session into spark?
Entry point of spark?
Do data frame process in parallel or not?
Internal architecture in spark?
Left anti join?
Semi join?
Diff between caching and persist?
Incremental loads in pyspark? How to handle incremental load in pyspark?
collect() in pyspark
show and display difference
How to identify and troubleshoot slow-running data in PySpark
Optimization techniques in pyspark
When to use broadcast joins
What are shuffles and partitions and bucket partition?
What happens when we submit a job to a cluster in Spark and how does it get divided?
What is adaptive query execution (AQE)?
Advanced features of Spark?
Small file issue in Spark?
What is data skewness?
How does AQE help in resolving data skewness?
What is shuffle data?
Default mode of shuffle partitioning?
Partitioning a 30GB CSV file into 5GB chunks—how many partitions are needed?
Default partition size in Spark?
How do you run a Spark application?
What is the spark-submit command with its options?
How to determine the number of stages in a Spark application?
How to increase the number of parallel tasks running in a Spark application?
How does Spark determine the number of stages in an application?
How can we run a Spark application faster and check the code?
How does Spark determine the number of stages in an application?
How to increase the number of parallel tasks in a Spark application?
How does partitioning make your query run more efficiently?
Diff between SQL and NoSQL?
How you decide to use SQL or NoSQL?
What kind of ETL work you did?
ETL Process
Slow Changing Dimensions (SCD) and Data Warehousing Questions:
What are Slow Changing Dimensions (SCD), and why are they important in data warehousing?
Slow Changing Dimensions (SCD) and Data Warehousing Questions:
What are Slow Changing Dimensions (SCD), and why are they important in data warehousing
Can you explain the different types of Slow Changing Dimensions (SCD)?
How would you implement Type 2 SCD in an ETL pipeline using AWS Glue?
How would you implement Type 3 SCD in a Glue ETL pipeline?
What is the difference between Type 1, Type 2, and Type 3 SCD?
What is a Hybrid (Type 6) SCD, and when would you use it?
How do you manage the size and performance of tables when dealing with Type 2 SCD?
In which scenarios would you use a separate history table (Type 4) in data warehousing?

General Data Integration and ETL Process Questions:
How do you ensure data quality during ETL processes using AWS Glue?
Can you describe an ETL pipeline where you needed to combine multiple data sources? How did you use AWS Glue for this?
What is the role of the Glue Data Catalog, and how does it improve the ETL process?
How do you handle large-scale joins and aggregations in Glue ETL jobs to optimize for performance?
I want a convert starschema to snowflake schema
Use of snowflake schema
Factless fact table

During the ETL process, if the schema of the input changes, how will you handle it?

Is it necessary to have numerical data in fact table.
Example of fact table what would be information in that table .
Which tool you used for etl processing
What exactly is ETL?
What is materialized schema
What is difference between star schema and snowflake schema
What is scd
Scd and its type type 0 type 1 type 2 type 3
When you run pipeline how will handle etl
Which will be loaded first dimension or fact table