Azure Data Engineer Interview Questions

1) Explain about yourself with details about roles and responsibilities
2) Difference between Control Plane and Data Plane. What services/resources will come under Control Plane and Data Plane. For ex: data, DBFS etc.
3) If you create cluster with 4 VM, in which Plane they will be allocated? Control Plane or Data Plane
4) What are different types of cluster?
5) Different cluster access modes and explain
6) Tell me all the ADF Activities you know/used
7) Difference between RDD and Data Frame
8) What are Data Frames and Datasets in Databricks/Spark
9) Tell me about Unity Catalog
10) Tell me about DLT (Delta Live Tables)
11) There is a container with many files and in different formats such as JSON, CSV, Excel. How do you create single pipeline to copy all these different files to destination folder in ADF.
12) Python code: Given a list something [-2, 0, 1, 1, -1] write a program that gives unique number such as a+b+c=0. where a, b, and c are numbers from the above list
13) What is the difference between data lake and delta lake
14) In which scenario you consider using Data Flow or Databricks for transformations/processing
15) What all languages are supported in Databricks
16) What is the difference between Spark Context and Spark Session. When do you use each?

1) What is the integration runtime in ADF?
2) How to Delete files more than 30 days old?
3) How will you update new records without changing the old records?
4) How to archive files after load?
5) How to use Wildcard Characters to get metadata Activity?
6) What is meant by Metadata?
7) How does Master Slave architecture work?
8) Write a python code to show the highest salary of an employee.
9) Write a python program to find the users who have been active for the last three days?
10) Can you please elaborate on Skewed data?
11) Any problems or issues you faced during your project. (Be prepared for this question- Answer: Missing values, null columns, Skewed data , dependencies, etc you can tell as per your experiences)

1) How to pass parameters to Databricks Notebooks in ADF?

https://lnkd.in/gyDpFGCj

More info on dbutils.widgets:
https://lnkd.in/g_bysN2a

2) What is the difference between Job Cluster and All-Purpose Cluster and how to create these in Azure data bricks?

Ans : All-Purpose compute: Used to analyse data collaboratively using an interactive notebook. You can create, terminate, and restart this compute using the UI, CLI, or REST API. This cluster best suited for DEV activities.
Job compute : Used to run fast and robust automated jobs. The Databricks job scheduler creates a job compute when you run a job on a new compute. The compute terminates when the job is complete. You cannot restart a job compute. This cluster is best suited for prod environment.

More resources below:
https://lnkd.in/gp8nHhsA

https://lnkd.in/ge7reKU5

3) How to handle Copy activity failures in ADF?

Ans: We can enable Fault Tolerance property in copy data activity settings to skip incompatible rows.

https://lnkd.in/gSYXjuDa

4) What is the use of Shared Self Hosted integration runtime?

Ans:
We can reuse existing Self Hosted Integration run time in multiple data factories by granting the permission to the ADFs.

https://lnkd.in/gJK_vmBm

https://lnkd.in/gz2TgxFU

5) How to load the files having size greater than 10mb using ADF?

https://lnkd.in/gj2qUFR3