SAMPLE

Interview questions for Data Science

EURON

1 End-to-End Pipeline Overview

General Questions:

1. Can you explain the end-to-end machine learning pipeline?
2. What are the key stages of an ML pipeline?
3. How do you ensure data quality in a machine learning pipeline?
4. How do you handle missing data in an ML pipeline?
5. What are common data ingestion techniques?

Technical Questions:

6. What are the differences between data ingestion and data storage?
7. What are the pros and cons of batch vs. stream processing for data ingestion?
8. How do you implement feature engineering as part of an ML pipeline?
9. What role does data validation play in a production pipeline?
10. How do you ensure reproducibility in an ML workflow?

Scenario-Based Questions:

🔹 **Scenario 1:**

💡 *Your company is building a fraud detection system for online transactions. What would be the key components of your ML pipeline, and how would you implement real-time fraud detection?*

🔹 **Scenario 2:**

💡 *You receive highly unstructured data from multiple sources (APIs, logs, CSV files, databases). How would you design an ML pipeline to preprocess and use this data for training models?*

---

2 Batch vs. Real-Time Data Pipelines

General Questions:

11. What is the difference between batch and real-time data processing?
12. When would you prefer batch processing over real-time processing?
13. What are the challenges in designing real-time data pipelines?
14. What are some common frameworks used for batch and real-time data processing?
15. What are event-driven architectures? How do they help in real-time pipelines?

Technical Questions:

16. How would you implement a real-time data pipeline using Kafka and Spark Streaming?
17. What are the limitations of Apache Spark for real-time processing?
18. How do you ensure data consistency in a real-time processing system?
19. How do watermarking and windowing help in stream processing?
20. What is the role of CDC (Change Data Capture) in real-time pipelines?

Scenario-Based Questions:

🔹 **Scenario 3:**

💡 *Your company wants to monitor customer engagement in real-time to recommend personalized content. How would you design the data pipeline? What technologies would you use?*

🔹 **Scenario 4:**

💡 *You have a daily sales report that needs to be processed and aggregated for insights. Would you use a batch or real-time pipeline? Justify your choice.*

3️⃣ Microservices & Containerization (Docker, Kubernetes)

General Questions:

21. What are microservices, and why are they useful in ML deployment?
22. How does containerization help in deploying ML models?
23. What are the key benefits of using Docker in ML workflows?
24. What is Kubernetes, and how does it help in deploying ML models?
25. How does Kubernetes ensure fault tolerance in ML deployments?

Technical Questions:

26. How would you deploy a trained ML model using Docker?
27. What is the role of Kubernetes pods and services in ML model deployment?
28. How do you handle model versioning with microservices?
29. What are Kubernetes Operators, and how do they help in MLOps?
30. How do you implement rolling updates in Kubernetes for ML models?

Scenario-Based Questions:

🔻 **Scenario 5:**

💡 *You have trained multiple models for an e-commerce recommendation engine. How would you deploy them as microservices and scale based on demand?*

🔻 **Scenario 6:**

💡 *You deployed an ML model using Kubernetes, but users report slow response times. How would you troubleshoot and optimize it?*

4️⃣ Modern Data Stack

General Questions:

31. What are Data Lakes, Data Warehouses, and Lakehouses? How are they different?
32. What is Delta Lake, and why is it important in modern data architectures?
33. What is the difference between ETL and ELT? Which one is preferable for ML pipelines?
34. What are the key differences between Redshift, Snowflake, and BigQuery?
35. What is dbt, and how does it help in modern data engineering?

Technical Questions:

36. How do you design an efficient ETL pipeline for an ML model?
37. What are the advantages of using a Lakehouse architecture over a traditional Data Warehouse?
38. How does Apache Airflow help in orchestrating ML workflows?
39. How would you handle schema evolution in a Data Lake?
40. How do you manage data governance in a Data Lakehouse?

Scenario-Based Questions:

🔻 **Scenario 7:**

💡 *Your team is migrating from an on-premises data warehouse to a cloud-based Lakehouse. How would you approach this transition?*

🔻 **Scenario 8:**

💡 *You need to implement an ETL pipeline to ingest and transform customer data daily. Would you choose batch processing or an ELT-based approach? Why?*

5️⃣ MLOps & CI/CD Pipelines

General Questions:

41. What is MLOps, and how does it improve ML model lifecycle management?
42. What are the key differences between CI/CD and CI/CD/CT in ML pipelines?
43. How do you track ML model versions?
44. What are the challenges of deploying ML models in production?
45. What is drift detection, and how do you monitor for data drift?

Technical Questions:

46. How do you implement model monitoring using Prometheus or Grafana?
47. What is the role of feature stores in MLOps?
48. How do you retrain a model automatically when new data arrives?
49. What is a shadow deployment, and why is it used in ML models?
50. How do you ensure reproducibility in ML experiments using MLflow?

Scenario-Based Questions:

🔹 **Scenario 9:**

💡 *You deployed a model that performs well initially but degrades over time. How would you monitor and retrain it?*

🔹 **Scenario 10:**

💡 *Your ML team wants to enable continuous training (CT) for a fraud detection model. How would you design the CI/CD/CT pipeline?*