prominent Academy

CGI

DATA ENGINEER

PYSPARK

SCENARIO-BASED INTERVIEW QUESTIONS

www.prominentacademy.in

- Scenario: A new version of a PySpark job must be deployed without interrupting ongoing processes. Describe your approach to achieving a seamless rollout.
- Scenario: A dataset is distributed across hundreds of files in a cloud storage bucket. How would you optimize the reading process to avoid small-file overhead?
- Scenario: Write PySpark code to implement a custom partitioner for unevenly distributed keys in a dataset.
- Scenario: Write PySpark code to implement a distributed PageRank algorithm for a graph dataset.
- Scenario: Describe how to process a dataset with overlapping time windows and aggregate metrics for each window.
- Scenario: A dataset must comply with regulations requiring the masking of sensitive information before processing. How would you enforce this in PySpark?
- Scenario: A PySpark job must enforce unique constraints across multiple datasets. How would you design the pipeline to validate and deduplicate the data?
- Scenario: A PySpark application requires generating synthetic data for testing purposes. Write a pipeline to create synthetic datasets with configurable parameters.

- Scenario: A PySpark application requires comparing two large datasets for near-duplicates (e.g., fuzzy matching). How would you approach this?

- Scenario: A pipeline involves multiple actions, but only the last action triggers execution. Describe how Spark's lazy evaluation impacts performance.

- Scenario: Write PySpark code to handle a job that must scale down gracefully when cluster resources are reduced dynamically.

- Scenario: Explain how to join a high-throughput streaming dataset with a slowly updating reference table in PySpark.

- Scenario: A pipeline requires calculating percentile values for large datasets. Explain how to implement this efficiently in PySpark.

- Scenario: A pipeline must read and write data to a REST API in batches. How would you implement this in PySpark?

- Scenario: A pipeline trains and evaluates multiple machine learning models in parallel. How would you scale this process?

- Scenario: Explain how to debug a PySpark application using logs and visualizations from Spark History Server.

- Scenario: A PySpark pipeline writes output to Parquet files, but querying the files is slow. How would you optimize the Parquet file storage (e.g., partitioning, compression)?
- Scenario: A large dataset needs to be joined with a small dataset frequently. Write PySpark code to broadcast the small dataset.
- Scenario: Write PySpark code to calculate the clustering coefficient of nodes in a distributed graph dataset.
- Scenario: A real-time event stream must detect anomalies using PySpark Structured Streaming. How would you design this?
- Scenario: Explain how to handle out-of-order events in a PySpark Structured Streaming application.
- Scenario: A PySpark application requires retry logic for writing data to an unreliable sink. How would you implement this?
- Scenario: A pipeline needs to log sensitive transformations but redact sensitive information from the logs. How would you implement this?
- Scenario: Explain how to handle timezones effectively in PySpark when processing timestamped datasets.
- Scenario: A hybrid architecture requires combining Apache Spark with Apache Beam. How would you integrate them?

- Scenario: A PySpark job requires appending and overwriting different partitions in Delta Lake. How would you manage this?
- Scenario: A PySpark pipeline processes multiple sources, and one source intermittently becomes unavailable. How would you handle this?
- Scenario: Explain how to implement a sliding window join between two datasets with time-based keys.
- Scenario: Write PySpark code to train a decision tree model and evaluate it using a custom metric.
- Scenario: A PySpark application must write to a Postgres database in parallel. Write the code for this.
- Scenario: A streaming pipeline must emit alerts when certain thresholds are breached. Write PySpark code to implement this.
- Scenario: A pipeline must use PySpark to process data across multi-cloud environments. How would you set this up?
- Scenario: A job processes a Delta Lake table with millions of small files. How would you optimize this?
- Scenario: A streaming application processes data from a Kafka topic with high throughput. How would you tune the PySpark job?

#AzureSynapse #DataEngineering #InterviewPreparation #JobReady #MockInterviews #Deloitte #CareerSuccess #ProminentAcademy

prominent Academy

❌Think your skills are enough?

Think again—these Azure Data engineer scenario-based questions could cost you your data engineering job.

In a recent interview at many big MNC's, one of our students faced scenario-based questions related to data engineering, and many candidates struggled to answer them correctly. These questions are designed to test your real-world knowledge and ability to solve complex data engineering problems.

Unfortunately, many students failed to answer these questions confidently. The truth is, preparation is key, and that's where Prominent Academy comes in!
We specialize in preparing you for spark and data engineering interviews by:
✅ Offering scenario-based mock interviews
✅ Providing hands-on training with data engineering features
✅ Optimizing your resume & LinkedIn profile
✅ Giving personalized interview coaching to ensure you're job-ready
Don't leave your future to chance!

📞Call us at **+91 98604 38743** and get the interview prep you need to succeed