Data Architect Skills Test

Section 1: Data Warehouse Design & Performance

Use Case: You are designing a data warehouse hosted in SQL Server on AWS RDS. Data arrives daily from multiple hospitals in flat files on S3, containing patient encounters, billing, and diagnoses. Executives require Tableau dashboards with fast load times and drill-down capabilities.

Tasks:

1. Design a star schema including:
   a. Fact table: FactPatientEncounters
   b. Dimension tables: DimDate, DimPatient, DimProvider, DimDiagnosis, DimHospital
2. Identify two performance strategies (e.g., indexing, partitioning, materialized views)
3. Describe how you would use AWS Glue to load S3 data into SQL Server, covering:
   a. Schema enforcement
   b. Handling of late-arriving data
4. Recommend either SSAS Tabular or Redshift as the analytics engine. Justify your recommendation and provide a scenario where the other option might be preferable.

Section 2: Production Issue Troubleshooting

Use Case: An overnight AWS Glue job failed, and no data was loaded into Redshift. Morning dashboards show blank data for the “Billing Exceptions” report.

Tasks:

1. List the first three actions you would take to troubleshoot the issue.
2. Explain your approach to recover the missing data and avoid duplication.
3. Draft an email to internal stakeholders summarizing the issue, impact, and resolution timeline.

Section 3: Business & Team Collaboration

Use Case: A business user requests "weekly summaries of denied insurance claims by provider group and region." The request is vague, and the team is mid-sprint.

Tasks:

1. Write three clarifying questions to ask the business user.
2. Draft a Jira story with title, description, and two acceptance criteria.
3. Describe how you would coordinate with the project manager and QA engineer to deliver the request.

Section 4: CI/CD & Test-Driven Development

Use Case: You are developing a Python ETL job to extract data from S3 and load it into a SQL Server staging table. It must be deployed with Jenkins or AWS CodePipeline and follow TDD principles.

Tasks:

1. Describe a CI/CD pipeline that:
   a. Executes unit tests
   b. Deploys code to Lambda or EC2
   c. Uses secure credential handling
   d. Sends alerts on failure
2. Write a pytest unit test for a function validate_row() that checks:
   a. Non-null patient_id
   b. Valid encounter_date in YYYY-MM-DD format
   c. Include one valid and one invalid test case

Section 5: T-SQL Proficiency

Use Case: You need to write logic for aggregating and troubleshooting patient appointment data.

Tasks:

1. Write a T-SQL query to return the top five providers by number of missed appointments in the past 30 days.
2. Suggest an optimization strategy if performance degrades with high row volume.
3. Add a TRY...CATCH block to log runtime errors to an ErrorLog table.

Section 6: AI-Augmented Development

Use Case: Your organization supports the use of approved AI tools (Amazon Bedrock, OpenAI APIs) to accelerate development.

Tasks:

1. Describe two use cases where you would leverage AI in your workflow.
2. Write an AI prompt to help explain a complex SQL join to a junior analyst.
3. List two precautions you would take when using AI tools in regulated environments.

Section 7: Documentation

Task: Document the ETL pipeline described in Section 4, including:

- Purpose
- Data sources and targets
- Core transformations
- Performance considerations
- Error handling approach
- Key assumptions

Use bullet points or markdown formatting.

Bonus Section: Python + SQL Integration Challenge (Optional)

Use Case: You want to process a batch of records from S3 in Python, validate each one, and insert valid rows into SQL Server.

Task: Write a Python snippet that:

- Connects to SQL Server using pyodbc or SQLAlchemy
- Inserts records in batches
- Skips invalid rows and logs them to a file