AWS + Python + PySpark Real-time Scenarios for Data Engineers

1. Simple S3 → Transform → S3 Pipeline

Skill Focus: Basic PySpark transformations, AWS S3 integration, Python automation.
Scenario Steps:
- Company stores raw CSV sales data in s3://company-raw/sales/YYYY/MM/DD/
- Read raw data from S3 using PySpark.
- Perform basic cleaning (remove nulls, fix date formats).
- Write cleaned data back to s3://company-processed/sales/ in Parquet format.

Services & Tools: AWS: S3 | Processing: PySpark (local or EMR/Glue) | Automation: Python script or Glue job

2. Log Processing from S3 to Redshift

Skill Focus: Data modeling, incremental loads, Python AWS SDK (boto3).
Scenario Steps:
- Application logs uploaded daily to S3.
- Parse logs in PySpark to extract timestamp, user_id, action.
- Load transformed data into Amazon Redshift fact tables.
- Automate job daily with AWS Lambda or Glue Workflow.

Services & Tools: AWS: S3, Redshift, Glue, Lambda | Processing: PySpark | Automation: Python boto3

3. Real-time Stream Processing from Kinesis

Skill Focus: Streaming data processing, window functions in PySpark.
Scenario Steps:
- E-commerce site streams user click events to Amazon Kinesis Data Streams.
- Consume stream in Spark Structured Streaming.
- Aggregate clicks by user_id in a 5-minute sliding window.
- Store aggregated results in S3.

Services & Tools: AWS: Kinesis Data Streams, S3, Glue Catalog | Processing: PySpark Structured Streaming

4. Data Lake with Partitioning & Glue Catalog

Skill Focus: Partitioning, schema evolution, query optimization.
Scenario Steps:
- Sensor readings in JSON format in S3.
- Read and clean data in PySpark.
- Save in partitioned Parquet format by year/month/day.
- Create/update Glue Data Catalog table for Athena.

Services & Tools: AWS: S3, Glue Data Catalog, Athena | Processing: PySpark

5. Data Validation & Quality Checks

Skill Focus: Data quality frameworks (Great Expectations), Python validation scripts.
Scenario Steps:
- Validate marketing campaign data before loading to Redshift.

- No nulls in campaign_id, valid dates, spend > 0.
- Write pass/fail reports to S3.
- Send email via AWS SES if validation fails.

Services & Tools: AWS: S3, SES, Lambda | Processing: PySpark, Great Expectations (optional)

6. Incremental ETL from RDS to S3

Skill Focus: Change Data Capture (CDC), scheduling, incremental loads.

Scenario Steps:

- PostgreSQL RDS stores order transactions.
- Pull only new orders using last_updated timestamp.
- Append to S3 in Parquet format.
- Run daily via Glue Job or Airflow.

Services & Tools: AWS: RDS, S3, Glue | Processing: Python (pandas + boto3) or PySpark

7. Machine Learning Data Prep

Skill Focus: Feature engineering in PySpark for ML models.

Scenario Steps:

- Join purchase history with user demographics.
- Create aggregate features: total spend, last purchase date, categories purchased.
- Save as feature store in S3 for ML training.

Services & Tools: AWS: S3, SageMaker (optional) | Processing: PySpark

8. IoT Data Pipeline

Skill Focus: Handling high-volume data, compression, time-series analysis.

Scenario Steps:

- IoT sensors send temperature data to Kinesis Firehose → S3.
- Transform raw JSON to time-series friendly schema.
- Store compressed Parquet for cost savings.
- Generate daily summaries for anomalies.

Services & Tools: AWS: Kinesis Firehose, S3, Glue Catalog, Athena | Processing: PySpark Structured Streaming