DAY 6 SUMMARY — SQL ■ Python ■ Azure ■ ADF Debugging & Best Practices ■

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

1. SQL PRACTICE ✔

• Filtering with WHERE

• JOINs (INNER JOIN)

• Aggregation (SUM + GROUP BY)

• Window functions: DENSE_RANK(), ROW_NUMBER()

• Subqueries for filtering top-N per group

• Debugging SQL errors: window function rules & alias issues

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

2. PYTHON & PANDNDAS PRACTICE ✔

Functions:

• get_positive_numbers()

• remove_nulls(df, column)

• customer_summary(df): total_spent, avg_quantity, product_count

Concepts:

• no loops needed in Pandas

• groupby + named aggregations

• vectorized operations

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

3. AZURE FUNDAMENTALS FLASH DRILLS ✔

• Linked Service = connection definition

• Managed Identity = best practice authentication

• Integration Runtime = compute engine

• Schedule Trigger = run every X hours

• Retry logic = configured at activity level

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

4. ADF DEBUGGING SKILLS ✔

Key Debugging Clues:

• filesRead = 0 → wildcard or path issue

• AuthorizationPermissionMismatch → RBAC missing for managed identity

• OperationTimeoutError → too many files OR weak IR

• Null parameter errors → dataset parameter name mismatch

• Lookup → ForEach: @item().column

You learned to read:

• Activity output JSON

• Error messages

• Dependency failures

• IR-level failures

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

5. ADF BEST PRACTICES (FULL PROFESSIONAL VERSION) ✔

PIPELINE DESIGN

• Keep pipelines modular: raw → processed → curated

• One pipeline = one purpose

• Use parameters, avoid hardcoding

• Reuse datasets

DATASET & LINKED SERVICE

• Linked Services hold credentials

• Datasets should be generic + parameterized

• Use parameters: container, folder, file_pattern, file_name

DYNAMIC EXPRESSIONS

• Use concat(), pipeline().parameters, item(), dataset().parameters

• Always format dates explicitly

LOGGING & MONITORING

• filesRead for wildcard debugging

• Log rowsCopied, dataRead, dataWritten

• Use Azure Monitor for alerts

RETRY, DEPENDENCY & ERROR HANDLING

• Configure retry on activities, not pipelines

• Use Succeeded / Failed dependencies

• Validate → IfSuccess → Notebook

→ IfFail → AlertEmail

METADATA-DRIVEN PIPELINES

• Lookup → ForEach → Copy

• Use @item().folder and @item().file_pattern

• Dynamic curated paths with concat()

PERFORMANCE OPTIMIZATION

• Prefer Parquet over CSV

• Avoid too many small files

• Allocate IR properly

• Narrow wildcard patterns

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

6. REVIEWED EXERCISES ✔

Q1–Q9: All correct (one improved)

You demonstrated:

• Clear understanding of ADF architecture

• Strong debugging reasoning

• Correct SQL & Pandas practices

• Enterprise-level pipeline thinking

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

END OF DAY 6 ■