

Customer I/O Formats Pack

JSON • CSV • Parquet • XML — practical patterns for customer data

01_json_customers_explicit_schema.py

JSON (flat) with explicit schema, permissive mode, corrupt-record capture, and basic DQ checks.

02_json_nested_multiline_explode.py

Multiline nested JSON with arrays, explode, and Top-K per-customer windowing.

03_csv_customers_options.py

CSV with header, schema, timestamp parsing, normalization and partitioned write.

04_parquet_partition_pruning.py

Parquet partition discovery and pruning for time/country filters.

05_parquet_schema_evolution.py

Parquet schema evolution using mergeSchema for additive columns.

06_upsert_foreachBatch.py

Delta-style upsert via foreachBatch (conceptual MERGE snippet included).

07_xml_customers_basic.py

XML basics with `rowTag=customer`; requires `spark-xml` package.

08_xml_customers_nested.py

XML nested example: attributes, arrays, and schema exploration.

09_csv_vs_parquet_perf.py

CSV vs Parquet size/perf sketch (illustrative).

10_jsonlines_customers.py

JSON Lines (one record per line) fast path.