PSV → Parquet Processing Pipeline

Python Script (master_psv_to_pq_v6.py)

This script converts PSV files into Parquet files with strict schema enforcement, batching, and merging. It ensures parallel-safe job execution and logs progress and failures. Key Features: - Schema enforcement: converts columns to consistent types (string, float, Int64, datetime). - Batch processing: processes files in groups with append-safe Parquet writing. - Parallel-safe: each job writes to its own isolated folder. - Merge step: combines job outputs into clean final Parquet files. - Logging: job-level logs, failed rows, and merge issues are tracked.

Bash Script (Driver)

The bash script orchestrates parallel execution, splitting workload across jobs, launching the Python processor, and merging final outputs. Logs and summaries are also created. Workflow Steps: 1. Sets parameters (input/output dirs, station list, freq, batch size, jobs). 2. Activates Python environment and creates dirs. 3. Splits station list into job chunks. 4. Launches jobs in parallel with isolated job dirs. 5. Waits for completion and merges final Parquet files. 6. Collects logs: processed rows, failures, merges, summary.

End Result

- Final Data: Clean, merged Parquet files in FINAL_OUTPUT_DIR. - Job Outputs: Temporary PQ chunks in TMP_OUTPUT_DIR (can be deleted later). - Logs: • Job logs (job_XY_pq_processed_log.txt) • Merge log (merge_log.txt) • Failed rows (failed_rows.pq) • Failed merges (failed_merge_pq.txt) • Summary (summary.txt)

Pipeline Overview Diagram