Quick Ingest & Exploratory Data Analysis

The Foundation for Machine Learning on ATTEST Logs

Objective:

- Collect and standardize archived ATTEST test execution logs into structured formats.
- Clean data by removing duplicates, handling missing values, and normalizing fields.
- Perform Exploratory Data Analysis to understand test execution volumes, failure distributions, and trends.
- Identify anomalies and data quality gaps to ensure machine learning readiness.
- Deliver a clean dataset and comprehensive EDA report with actionable insights

Data Ingestion(Log Archive SetupData)

- Gather historical test execution logs from the ATTEST system archive
- Organize log files into a consistent folder hierarchy by date and test suite
- Transform raw log data into standardized, machine-readable formats such as CSV
- Extract key information fields for analysis:
  - Accurate timestamps of each test run
  - Detailed Device Under Test (DUT) attributes
  - Unique test case identifiers
  - Clear test result statuses (Pass, Fail, Abort)

Log Data Preprocessing:

=> Loaded dataset using Pandas.

=> Standardized status values → `PASS`, `FAIL`, `ABORT`.

=> Cleaned invalid or empty rows.

=> Extracted timestamps and derived run_date.

=> Parsed test_case_id from logs.

=> Extracted DUT, OS version, and config details.

=> Filled or inferred missing environment info.

Data Cleaning:

- **Data Refinement:** Filled missing values (forward/backward fill) and added “No Error” for passed tests.
- **Filtering & Optimization:** Removed duplicates, unknown DUTs, and unnecessary columns.
- **Finalization:** Validated data quality and saved the cleaned, ML-ready dataset (`logs_preprocessed.csv`).
- **Removed:**

  Unknown or generic DUT entries

  Duplicate Rows

  Unnecessary columns(error_category, dut_version)

Exploratory Data Analysis:

- **Volume Analysis:** Measured execution counts by suite, DUT, and time period.
- **Failure Distribution:** Identified top failure types and their occurrence frequency.
- **Execution Trends:** Visualized pass/fail rates over time to track performance stability.
- **Anomaly Detection:** Spotted unusual spikes or irregular run patterns in test results.

Data Profiling and Report:

1. Data Quality Summary:

- **Coverage per column:** High completeness across dataset
- **Missing values:** Minimal and manageable
- **Duplicates:** Checked and removed

2. Critical Fields for ML Readiness:

- Test Case ID | DUT | Config | Error Logs | Run Date

3. Data Validation Insights:

- DUT version logs complete
- Error messages consistentl captured
- Dataset ready for ML analysis

Outcomes:

Well-Organized Dataset: Reliable features with minimal missing or corrupt data

Insights: Failure modes and testing trends over time

Anomaly Detection: Identification of outliers for deeper investigation

ML Readiness: Clear understanding of critical fields and data gaps

Actionable Recommendations: Provides guidance for improving test coverage and data quality

Thanks