**ETL Development and Testing Plan**

**1. Overview**

This document outlines the **ETL (Extract, Transform, Load) development and testing plan** for enhancing an existing GCCA dataset in SAS by adding new columns and ensuring data integrity throughout its lifecycle.

**2. ETL Development Plan**

**2.1 Scope**

The ETL process aims to:

* Enhance the dataset by adding **six new columns**.
* Implement **formula-based transformations** for four of these columns.
* Document **data lineage** to track data flow.
* Ensure robust testing to maintain data integrity.

**2.2 Dataset Lifecycle Considerations**

* **Initial Load:** New columns added, transformation rules applied.
* **Incremental Loads:** Ensuring schema compatibility with future updates.
* **Archival & Retention:** No historic backfill required, ensuring smooth data progression.
* **Error Handling & Logging:** Implement monitoring and alerts for failures.

**2.3 Functional Requirements**

* Develop a **data lineage document** tracking data transformations from source to final dataset.
* **Add six new columns** to the existing dataset.
  + **Four columns** will contain calculated values based on predefined formulas.
  + **Two columns** will store raw or reference data.
* Ensure new schema integration aligns with downstream consumers.

**2.4 Non-Functional Requirements**

* **Unit testing:** Verify correctness of **all new columns** using primary key-based validation.
* **Integration testing:** Ensure schema compatibility and transformation accuracy across datasets.
* **Performance optimization:** Maintain processing efficiency without excessive resource utilization.
* **No historic backfill:** The ETL process will apply changes prospectively only.

**3. Testing Plan**

**3.1 Unit Testing**

**Objective:** Validate correctness at the column level, ensuring formula-based transformations work as expected.

| **Test Case** | **Test Scenario** | **Expected Outcome** |
| --- | --- | --- |
| UT-01 | Verify presence of all 6 new columns | Columns exist in dataset |
| UT-02 | Validate formula calculations | Output matches expected formula logic |
| UT-03 | Ensure NULL handling for new columns | NULL values handled as per specifications |
| UT-04 | Validate primary key uniqueness | No duplicates introduced |
| UT-05 | Check datatype consistency | Data types match schema definitions |

**3.2 Integration Testing**

**Objective:** Ensure smooth integration with downstream processes and schema consistency.

| **Test Case** | **Test Scenario** | **Expected Outcome** |
| --- | --- | --- |
| IT-01 | Verify schema compatibility | No schema conflicts detected |
| IT-02 | Validate dataset joins with existing records | Joins work as expected |
| IT-03 | Ensure no breaking changes in reports | Reports process data correctly |
| IT-04 | Confirm incremental processing works | Only new/modified records are processed |

**3.3 Performance & Regression Testing**

**Objective:** Ensure ETL efficiency and data consistency across runs.

| **Test Case** | **Test Scenario** | **Expected Outcome** |
| --- | --- | --- |
| PT-01 | Validate processing time within SLA | Meets performance benchmarks |
| PT-02 | Ensure minimal memory consumption | Resource utilization is within limits |
| RT-01 | Compare data before and after ETL changes | No unintended changes in old records |

**3.4 Error Handling & Logging**

* Implement logging mechanisms in SAS to capture transformation errors.
* Ensure alerts for critical failures (e.g., formula computation issues, schema mismatches).

**4. Deployment & Monitoring**

* **Pre-Deployment Validation:** Final testing before production rollout.
* **Production Deployment:** Ensure ETL jobs are scheduled and monitored.
* **Post-Deployment Monitoring:** Validate outputs for a defined period.
* **Incident Management:** Establish rollback plans for major failures.

**5. Documentation & Handoff**

* Data lineage document completed and reviewed.
* Test cases and validation reports archived for audit.
* Handoff to downstream users or BI teams.

**6. Conclusion**

This ETL development and testing plan ensures the successful implementation of **new columns, formula-based transformations, and schema integration** while maintaining data integrity, performance efficiency, and auditability.