**GSCM 521**

**Abdi Alamir**

**ETL Framework and Process Documentation**

**1. Introduction**

The ETL (Extract, Transform, Load) process is critical for integrating and analyzing data from diverse sources. This document delineates the ETL framework for linking three data files: Banner Grade Data, Canvas Gradebook Data, and Canvas Interaction Data.

**2. Data Understanding and Exploration**

**Banner Grade Data:**

- Description: Contains student grades, course details, instructor information, and student attributes.

- Cleaning and Transformation Needed:

- Check for missing values in critical fields like Student ID, Course ID, and Grade.

- Standardize the format for fields like Term and Instructor.

- Ensure consistency in data types across fields.

- Key Fields: Random Student, Course, Term.

**Canvas Gradebook Data:**

- Description: Provides student identifiers and login information.

- Cleaning and Transformation Needed:

- Handle missing or inconsistent data in fields like Random Student and SIS User ID.

- Ensure uniqueness of identifiers.

- Key Fields: Random Student.

**Canvas Interaction Data:**

- Description: Records student interactions with course content.

- Cleaning and Transformation Needed:

- Validate data integrity in fields like Student ID and Course ID.

- Standardize format for dates and times.

- Key Fields: Random Student, Course ID, Term.

**3. Data Linkage**

**Banner Grade Data and Canvas Interaction Data:**

Link based on the concatenated key of Random Student, Course, and Term. This linkage ensures that student grades are associated with the corresponding course interactions.

**Canvas Gradebook Data and Canvas Interaction Data:**

Link based on the Random Student identifier. This allows for the association of student login information with their interactions within courses.

**Banner Grade Data and Canvas Interaction Data Linkage Details:**

Concatenate Random Student, Course, and Term to form a unique key for both datasets.

Implement fuzzy matching algorithms to handle slight variations in course names or terms.

Use cross-referencing techniques to verify the accuracy of the linked data.

Handle cases where students may have incomplete or missing records in either dataset.

**Canvas Gradebook Data and Canvas Interaction Data Linkage Details:**

Match Random Student identifiers between the two datasets.

Implement validation checks to ensure consistency between student IDs.

Handle cases where students may have multiple entries or different identifiers across datasets.

**4. Conclusion**

The refined ETL process considers the variability of Student ID across courses and terms, ensuring a robust linking strategy. Stakeholders can effectively analyze the integrated data for meaningful insights.

**Data Dictionary**

**1. Banner Grade Data:**

- Subject: The subject of the course.

- Course: The course identifier.

- Title: The title of the course.

- Term: The academic term when the course was taken.

- Random Student: A unique identifier for the student.

- Random ID: An identification code associated with the student.

- Grade: The grade obtained by the student in the course.

- Instructor: The instructor of the course.

- CRN: Course Reference Number.

- Reg\_Status: Registration status of the student.

- Unduplicated\_Race: Race of the student.

- Citizenship: Citizenship status of the student.

- Overall\_UG\_Credits: Overall undergraduate credits.

- Overall\_GR\_Credits: Overall graduate credits.

- Credits\_in\_Class\_Term: Credits earned by the student in the specific term.

- 1st\_Term\_at\_PSU: The first term the student attended at the institution.

- Latest\_Term\_with\_Reg\_Activity: The latest term with registration activity.

- Level: Level of the student (e.g., freshman, sophomore).

- Latest\_Class: The most recent class taken by the student.

- Major\_1\_in\_Class\_Term: Student's major during the class term.

- Latest\_Major\_1: Student's latest major.

- Random Email: Email address associated with the student.

**2. Canvas Gradebook Data:**

- Random Student: A unique identifier for the student.

- Random SIS User ID: Student Information System User ID.

- Random SIS Login ID: Student Information System Login ID.

**3. Canvas Interaction Data:**

- Student Id: Student identifier.

- Random Student Name: Student's name.

- Random Sortable Name: Sortable name of the student.

- Random Student SIS ID: Student Information System ID.

- Section Id: Identifier for the course section.

- Section Name: Name of the course section.

- Course Id: Identifier for the course.

- Course Name: Name of the course.

- Content Type: Type of content (e.g., assignment, quiz).

- Content Name: Name of the content.

- Times Viewed: Number of times the content was viewed.

- Times Participated: Number of times the student participated.

- Start Date: Date when the interaction started.

- First Viewed: Date of the first view.

- Last Viewed: Date of the last view.

=>This data dictionary provides a clear definition of each field within the datasets, facilitating understanding and analysis.