EQAO Grade 9 Student Performance Assessment in Ontario School Districts Report

November 16, 2019

Expert Team
Shaheda Choudhury
Solomon Ejigu
Nha Nguyen
Rodi Tountcheva
Gemma Verlezza-Ortega

TABLE OF CONTENTS

1. INTRODUCTION

The purpose of the Extraction, Transformation, and Load (ETL) Technical Report is to capture details that pertain specifically to ETL portion of the data pipeline that is to be used in a data science project. This however does keep in mind the final target objective while performing the ETL.

1.1 Summary

The objective of this project is to use a data set of the results of the Grade 9 Education Quality and Accountability Office (EQAO) administers standardized tests in math in order to explore, transform and load so that it is in an easy to understand format for the audience of the data.

The business problem is to clearly present and identify connections with results of the data, from the results of the EQAO Exams of 9th grade students collected from the Toronto, Waterloo and Peel school districts boards during the 2013 - 2014 school year.

1.2 Scope

The raw data came in the form of CSV files from Kaggle which was titled “Student Performance” which includes 2 data sources with 12 columns and 86 rows.

The first csv named “school_data” contains fields with descriptions which are:

1. School - The unique name of the school.
2. School ID - Unique numeric identifier for each school
3. Board - The name of the board.
4. Board ID - A unique numeric identifier for each board.
5. Num Student - The number of students in the school.
6. Level 1% - Level 4% - The percent of students achieving each level.
    a. Level 3 = Provincial standards
    b. Level 4 = Exceeds standards
7. Num F - The number of female students.
8. Num M - The number of male students.
9. Num Responses - The number of students responding to the survey.

The second csv named “response_data” contains fields with descriptions which are:

1. School ID: A unique numeric identifier for each school
2. Q1(%) - Q11(%): The percentage of students answering "agree" or "strongly agree" to each statement Q1-Q11. The questions are as stated below:
    - Q1: I like mathematics

- Q2: I am good at mathematics
- Q3: I am able to answer difficult mathematics questions
- Q4: Mathematics is one of my favourite subjects
- Q5: I understand most of the mathematics I am taught
- Q6: Mathematics is an easy subject
- Q7: I do my best in mathematics class
- Q8: The mathematics I learn now is useful for everyday life
- Q9: The mathematics I learn now helps me do work in other subjects
- Q10: I need to do well in mathematics to study what I want later
- Q11: I need to keep taking mathematics for the kind of job I want after I leave school

Source: https://www.kaggle.com/hdawkins/student-performance#school_data.csv

1.3 Technologies and resource contributions

Team Members Contributions:

- Shaheda Choudhury - Transforming the data by creating pandas dataframes.
- Solomon Ejigu - Transforming the data by running SQL Queries.
- Nha Nguyen - Transforming the data by running SQL Queries.
- Rodi Tountcheva - Extracting data from (kaggle) raw form and cleaning it for uploading into PG Admin.
- Gemma Verlezza-Ortega - Transforming the data by running SQL Queries.

Resources used:

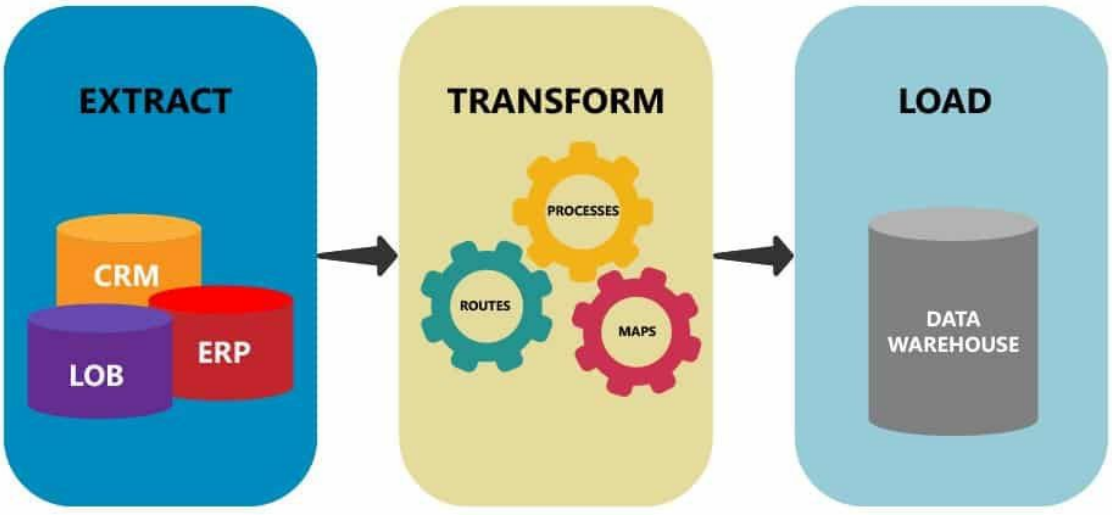
- Kaggle
- Jupyter Notebook
- Python
- Postgres
- PGADMIN4
- SQL
- Google
- Youtube
- Microsoft Excel
- quickdatabasediagrams.com

1.4 Definitions, Acronyms and Abbreviations

- Extract: The process of reading the data from various data sources. We extracted the data from the Kaggle web site. We searched the web to learn more about the organization that is collecting the data. The data is collected because often it is found in different places. In this case it was 2 csv files and the data information web

page.

- Transform: The process of converting the extracted data from the data sources to data that can be imported into a database. Below are the steps we took to transform the data:
    - Downloaded the data, school_data.csv and response_data.csv
    - Used quickdatabasediagrams.com to create the entity relationship diagram (ERD) for the data, with keys (both primary and foreign), data types, and relationships.
    - Created another csv file to store the survey questions, we filled the csv file with data from the general information page
- Load: The process of loading the data into a database management system. The steps we took are:
    - Generated the SQL statements from the quickdatabasediagrams.com website from our ERD.
    - Used pgAdmin to create the database in PostGres (datanwarehouse), school_evaluation.
    - Ran the SQL statements from the Transform stage to create the tables.
    - Imported the three csv files into each of the tables
    - Use SQL Query and Python (Jupyter Notebook) to retrieve, update, and run queries against the data.



ETL - Extract, Transform, Load

2. ETL DETAILS

This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.

2.1 Data Import/Extract Sources and Method

For this project the extract is how we research the data set to work on. First, we found the “Student performance” data set from Kaggle which has 2 CSV files which contains survey data from 9th grade students in Toronto, Waterloo and Peel district school boards from 2013 – 2014.

2.2 Data Acquisition

The dataset is obtained from Kaggle. It is a static dataset. Since the student evaluation is completed on an annual basis, the dataset needs to be updated on an annual basis. To obtain the data, we download it from Kaggle (https://www.kaggle.com/hdawkins/student-performance). The dataset is composed of two csv files, response_data.csv and school_data.csv.

INSERT DATA DESCRIPTION HERE

2.3 Data Transform

After downloading the CSV file, we used quickdatabasediagrams.com website to create the entity relationship diagram. Next, we used the site to generate SQL statements which we then used to create tables in a Postgres database. Finally, after creating the tables, we inserted the csv files into the tables using the import function in the pgAdmin4 admin app.

We created a new file, questions_data.csv, to store the list of survey questions. We then import the csv to the question_data table in our database.

2.4 Data Integrity

We reviewed the data for invalid or missing data. The data source is updated on an annual basis. A data set refresh annually is necessary to have an up to date data.

2.5 Data Refresh Frequency

The local dataset will need to be reset once a year since students take the evaluation survey once a year.

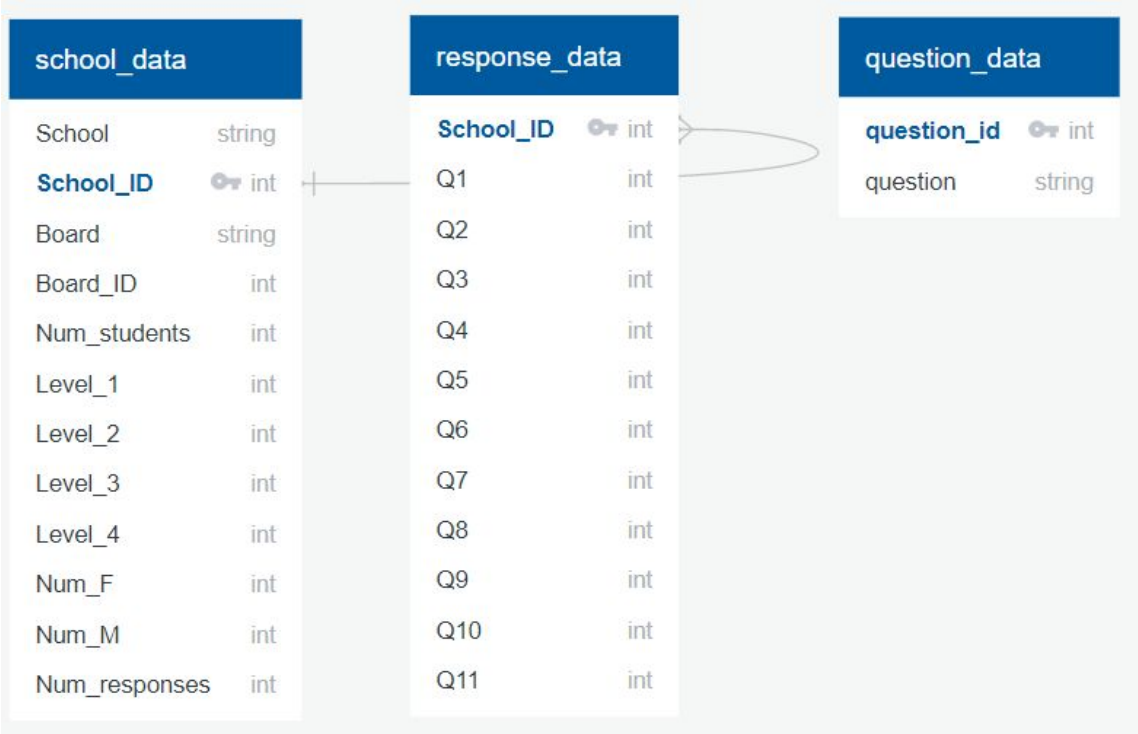
2.6 Data Security

We stored the data in a Postgres database on a laptop computer. Access to the database is secured by a password. Since the data is public data, and does not contain any personal identifiable information (PII) we are not required to follow any other regulatory compliance authorities.

Changes database is backed up automatically on a nightly basis, and full backups are completed on a weekly basis and stored on a members GITHub account.

2.7 Data Loading and Availability

The data schema, shown as an entity relationship diagram is below. A user can use the diagram to see the relationships, fields, data types, and keys of the tables in the database.



The user may use pgAdmin or Jupyter Notebook/Lab to read/access the data. If using the pgAdmin, connect to the school evaluation database using a given password. If accessing

via Jupyter Notebook, use SQLAlchemy to connect and run the queries against the data.

3. DATA QUALITY

The results from the assessments enhance the learning process and improve the instructional practice. They provide an opportunity for teachers and students alike to identify areas of understanding and misunderstanding. With this knowledge, students and teachers can build on the understanding and seek to transform misunderstanding into significant learning. Time spent on assessment will then contribute to the goal of improving the mathematics learning of all students.

Mathematics assessments can help both students and teachers improve the work the students are doing in mathematics. Students need to learn to monitor and evaluate their progress. When students are encouraged to assess their own learning, they become more aware of what they know, how they learn, and what resources they are using when they do mathematics. Almost 71% of students of all schools achieved level 3 which means they met the standards while another 12% of students achieved level 4 which means that they exceeded the standards.