ETL Group Project: Data Integration and Transformation

Assigned: Sunday, October 5, 2025
Due: Sunday, October 19, 2025 (11:59 PM)
Presentations: Monday, October 20 & Wednesday, October 22
Groups: 2–3 students per team — no solo submissions
Total Points: 100

Overview

In this project, your team will design and implement a Python ETL (Extract, Transform, Load) pipeline that combines data from three different sources:

1. A CSV file
2. A SQL database table
3. A Web API returning JSON data

Your final goal is to clean, merge, and analyze these datasets in a pandas DataFrame, document your process, and present your results.

Learning Objectives

- Apply the ETL process using Python
- Integrate CSV, SQL, and API (JSON) data
- Perform data cleaning and transformation
- Create a combined DataFrame for analysis and visualization
- Work collaboratively in small teams using professional data practices

---

Deliverables

1. ETL Jupyter/Colab Notebook – a working, documented pipeline
2. 300–500 word Team Summary – explaining your process and findings
3. Team Contributions Log – outlining each member's role
4. 5–7 Minute Group Presentation – demonstrate your data, ETL process, and insights (Oct 20 or 22)

Class Schedule & Checkpoints

Date	Focus	Notes
Mon Oct 6	Project kickoff & brainstorming	In-class overview of ETL and project setup
Wed Oct 8	Checkpoint 1: Group Formation (10 pts)	Form teams, choose topic, submit group info
Fri Oct 10	Checkpoint 2: Data Description (15 pts)	Submit one-page data description (sources, size, attributes)
Mon Oct 13	Reading Day – No Class	Independent group work
Wed Oct 15	Checkpoint 3: Zoom Check-In (15 pts)	Individual group meetings with instructor
Week of Oct 13	Presentation Sign-ups	Choose presentation slot for Oct 20 or 22
Sun Oct 19	Final ETL Submission (35 pts)	Submit completed notebook and reflection
Mon Oct 20	Presentations – Part 1 (15 pts)	In-class group presentations
Wed Oct 22	Presentations – Part 2 (15 pts)	Remaining groups present

---

Checkpoint Details

Checkpoint 1 – Group Formation (10 pts)

Due: Wednesday, Oct 8 (in class)

- Team name
- Member names and assigned roles (API, SQL, visualization, etc.)
- One-sentence topic idea or domain
  All students must be in a group to receive credit.

Checkpoint 2 – Data Description (15 pts)

Due: Friday, Oct 10 (submit via Canvas or notebook)
Submit a 1-page description including:

- Dataset names and sources (with URLs or file paths)
- Approximate size (# rows, file size)
- Key attributes or columns

- How the datasets relate and can be joined

Checkpoint 3 – Project Goals & Integration Plan (15 pts)

Due: Wednesday, Oct 15

- Write a Project Goals section in your notebook describing:
    - The main research question or analytical purpose
    - How CSV, SQL, and API data will be combined
    - Planned transformation steps
- Attend a 10–15 minute Zoom check-in with the instructor

Checkpoint 4 – Work Session Progress (10 pts)

When: Week of Oct 13
Show visible code progress in your notebook, including partial extraction and transformation work. Schedule a Group Zoom Session with Prof W…no class on the 15th so you can continue to do group work.

Final Submission (35 pts)

Due: Sunday, Oct 19 (11:59 PM)
Submit in Canvas:

- Completed ETL notebook
- 300–500 word team summary
- Team contributions log

Presentations (15 pts)

Dates: Monday, Oct 20 & Wednesday, Oct 22
Each team will give a 5–7 minute presentation demonstrating:

- The three data sources
- The ETL process and transformations
- One or more key insights or visualizations
- Reflection on collaboration and challenges
  Presentation sign-ups open during the week of Oct 13. All members must participate.

Grading Breakdown

Component	Points
Checkpoint 1 – Group Formation	10
Checkpoint 2 – Data Description	15
Checkpoint 3 – Zoom Goal Check-in	15
Checkpoint 4 – Work Session Progress	10
Final ETL Submission	35
Presentation	15
Total	**100 pts**

Late Policy

Late work loses 10% per day unless prior arrangements are made.

Resources & Inspiration

Your group will need three complementary data sources — one each from CSV, SQL, and an API (JSON). Below are sample topics and trusted data sources to help you get started.

1. Possible Project Topics

- Health & Environment – air quality vs. weather, COVID-19 trends, pollution impacts
- Economy & Business – housing prices vs. rates, cryptocurrency sentiment, stock data
- Sports & Performance – stats vs. salaries, weather and game outcomes, Olympic data
- Education & Demographics – graduation rates vs. income, school funding, census data
- Pop Culture & Media – box office and reviews, Spotify/YouTube trends, book analytics

2. Recommended Data Sources

CSV Data

- Kaggle Datasets – https://www.kaggle.com/datasets
- data.gov – https://data.gov
- Google Dataset Search – https://datasetsearch.research.google.com/
- Our World in Data – https://ourworldindata.org/

- UCI Machine Learning Repository – https://archive.ics.uci.edu/ml/index.php

SQL Databases

- Chinook Database – https://github.com/lerocha/chinook-database
- Sakila Database – https://dev.mysql.com/doc/sakila/en/
- Northwind Database – https://github.com/jpwhite3/northwind-SQLite3
- Google BigQuery Public Datasets – https://cloud.google.com/bigquery/public-data
- SQLite Online (browser-based) – https://sqliteonline.com/

APIs (JSON)

- WeatherAPI – https://www.weatherapi.com/
- Open-Meteo – https://open-meteo.com/en/docs
- NOAA Climate Data – https://www.ncdc.noaa.gov/cdo-web/webservices/v2
- NewsAPI – https://newsapi.org/
- TheSportsDB – https://www.thesportsdb.com/api.php
- OMDb API – https://www.omdbapi.com/
- Alpha Vantage – https://www.alphavantage.co/
- CoinGecko – https://www.coingecko.com/en/api
- SpaceX API – https://api.spacexdata.com/v4/launches
- PokéAPI – https://pokeapi.co/
- Open Library API – https://openlibrary.org/developers/api

3. Data Pairing Ideas

Theme	CSV Source	SQL Source	API Source
Weather & Retail	Daily sales CSV	Northwind orders table	WeatherAPI forecasts
Movies & Reviews	Box office CSV	Sakila films	OMDb API
Sports Analytics	Player stats CSV	SQL team data	TheSportsDB API
Public Health	CDC CSV data	Hospital/region database	COVID-19 API

Cryptocurrency Trends	Historical crypto CSV	SQLite transactions	CoinGecko API

4. Hints for Choosing Data

- Pick data with common keys (date, location, ID) for easier joins.
- Choose manageable sizes (under ~1 GB total).
- Verify the API supports JSON output and works with Python's requests library.
- Test your SQL query early to ensure the schema fits your idea.
- Keep a clear story — you'll present your findings concisely.