Conestoga College

School of Applied Computer Science & Information

Technology

SENG8081 - Case Studies

Canadian Job Market Trends Analysis

Anuroopa Balachandran

Bin Hu

Ce Chen

Xiaoman Yang

June 22, 2025

**Abstract**

In today’s rapidly evolving economic landscape, data-driven insights into job market trends are essential for policymakers, businesses, and job seekers to make informed decisions. This project presents a comprehensive analysis of the Canadian job market by integrating real-time job postings with historical labour force statistics from government APIs and open datasets. Leveraging a robust Big Data pipeline, the system processes and correlates key metrics—including employment rates, sectoral growth, regional disparities, and emerging skill demands—to uncover actionable trends.

By combining structured labour market data (Statistics Canada) with unstructured job posting descriptions (scraped from industry platforms), the project identifies high-growth industries, regional employment hotspots, and shifting skill requirements. Advanced analytics and interactive dashboards (Tableau) visualize these trends, enabling stakeholders to detect economic signals, anticipate labor shortages, and assess recovery patterns post-recession.

The solution emphasizes scalability, with automated data collection (Python, APIs), structured storage (SQL Server), and machine learning-ready preprocessing (extension). Ultimately, this project serves as a foundation for evidence-based workforce planning, helping bridge the gap between labor supply and demand in Canada’s dynamic economy.

**Table of contents**

[**Introduction** 4](#_Toc201526673)

[**Big Data System Diagram** 4](#_Toc201526674)

[**Data Research and Integration** 5](#_Toc201526675)

[**Data Collection** 5](#_Toc201526676)

[**Data Storage and Maintenance** 7](#_Toc201526677)

[**Project Timeline** 10](#_Toc201526678)

[**References** 10](#_Toc201526679)

# **Introduction**

This project analyzes trends in Canada’s job market by integrating real-time job postings with historical labour force statistics. Our goal is to identify high-growth sectors, regional employment disparities, and emerging skill demands to support policymakers, businesses, and job seekers.

A graph with a red arrow going down

AI-generated content may be incorrect.

# **Big Data System Diagram**

Below is an overview of our data pipeline:

[Data Sets – historic and real time]

↓

[Python Backend: Data Ingestion & Cleaning]

↓

[SQL Server: Structured Storage]

↓

[Tableau: Visualization & Dashboarding]

Key components:

* **Data Sources**: Job postings (web-scraped/Kaggle), labour force stats, industry jobs.
* **Processing**: Python (Pandas, SQLAlchemy) for cleaning/merging.
* **Storage**: SQL Server for structured time-series data.
* **Output**: Interactive dashboards for trend analysis.

# **Data Research and Integration**

**Data Sources**

Among the collected datasets, below are the final datasets filtered after careful observation to use for our project.

* **Job Postings:**
  + Real-time job listings with salary information
  + Skill requirements and location data
  + Source: Web-scraped/Kaggle.
* **Labour Force Stats**:
  + Monthly employment, unemployment, and participation rates.
  + Source: Statistics Canada.
* **NAICS Industry Jobs:**
  + Annual job counts by industry classification
  + Geographic distribution of employment
  + Source: NAICS Dataset

**Integration Plan**

* Merge datasets on **location** (province) and **time** (month/year).
* Use **sector** (e.g., IT, Healthcare) to align job demand with labour supply trends.

# **Data Collection**

**Data Sources & Tools**

1. **Job Postings Dataset (**CLEANED\_job\_postings.csv**)**
   * **Source**: Aggregated from multiple platforms:
     + **Kaggle**: Pre-cleaned job postings dataset for Canadian market.
     + **Web Scraping**: Python (BeautifulSoup, Scrapy) to extract real-time postings from google search results.
   * **Fields Collected**:
     + Job title, company, location, salary range, sector (e.g., IT, Healthcare), work type (remote/hybrid), and detailed descriptions.
   * Metadata: Posted date, applications count, experience level.
2. **Labour Force Statistics (**CLEANED\_labour\_force\_stats.csv**)**
   * **Source**: Statistics Canada’s Labour Force
   * **Fields Collected**:
     + Monthly employment rates, population estimates (ages 15+), and breakdowns by province, gender, and industry.
     + Seasonally adjusted metrics for trend consistency.
3. **NAICS Industry Jobs** (CLEANED\_industry\_jobs.zip)
   * **Source**: Government Canada – Job Bank Employment by Industry
   * **Fields Collected**:
     + Annual counts of paid worker jobs across Canada.
     + Breakdown by North American Industry Classification System (NAICS) codes.
     + Includes high-level sectors (e.g., *All industries*, *Manufacturing*, *Health care and social assistance*) and also supports categories (e.g., *IT services*)

**Cleaning Process**

* Cleaning datasets by fixing missing, null or error values.
* Standardize raw data formats for loading into the job\_market\_trends database.

A screen shot of a computer

AI-generated content may be incorrect.

* **Automated Pipelines**:
  + **Job Postings**: Scheduled Python scripts run weekly to scrape and append new postings to the dataset.
  + **Labour Stats**: API calls via requests library to fetch monthly updates; data validated against historical records.
* **Challenges & Solutions**:
  + **Inconsistent Formats**: Job titles/salaries varied.
    - *Solution*: Standardized salary ranges to hourly/yearly rates using regex; imputed missing values with sector medians.
  + **API Limits**: Statistics Canada throttled frequent requests.
    - *Solution*: Implemented caching with SQLite to store interim results and avoid redundant calls.

# **Data Storage and Maintenance**

**Storage Architecture**

1. **Database**: Microsoft SQL Server
   * **Schema Design**:
     + **Database:** job\_market\_trends
     + **Tables**:
       1. labour\_force\_stats : Labour force characteristics by region, gender, age group
       2. industry\_jobs: NAICS industry job counts across years and sectors
       3. job\_postings: Detailed job postings scraped from LinkedIn or Google Jobs.

A screenshot of a computer program

AI-generated content may be incorrect. A screen shot of a computer program

AI-generated content may be incorrect.

1. **Maintenance Practices**

**Ingestion Strategy**

* Use load\_data.py to insert or append CSV data to MySQL
* Column normalization and null checks handled before insert
* Timestamp columns converted appropriately

A screen shot of a computer

AI-generated content may be incorrect.

**Update Schedule**

* Monthly: Labour Force & Industry Jobs (from StatsCan, Open Canada)
* Daily/Weekly:Job postings via automated scraper

**Data Hygiene**

* De-duplication using primary keys
* Missing values filled with defaults or flagged for review
* Normalization ensures analytical consistency

**Backups**

* Weekly dump using MySQL dump to backups
* Retain minimum 3 historical snapshots

**📤 Integration Points**

* Data is used by:
  + *load\_data.py* for ingest
  + Dashboards for trends and forecasting
  + APIs for real-time insights

**Scalability Considerations**

* **Current Storage**: ~10GB (job postings + labour stats for 2020–2023).
* **Future Scaling**:
  + **Horizontal Partitioning**: Split job\_postings by province if data grows beyond 50GB.
  + **Cost Optimization**: Move cold data (pre-2022) to Azure Data Lake for cheaper long-term storage.

# **Project Timeline**

|  |  |  |
| --- | --- | --- |
| **Date** | **Deliverable** | **Responsible** |
| May 20 | Data Researched and Collected | Team |
| June 4 | Data cleaning | Anuroopa, Chen, Mila |
| June 17 | Data storage and maintenance | Bin, Mila, Chen |
| Jun 18 | Midterm report | Bin, Anuroopa |
| July 10 | Further cleaning of data |  |
| July 20 | Testing and solved errors |  |
| July 20 | System Diagram |  |
| July 25 | Analysis & visualization |  |
| August 10 | Final Report |  |

Git hub link*: https://github.com/SENG8081/SENG8081-S25-Team1*

A screenshot of a computer

AI-generated content may be incorrect.

# **References**

1. **Statistics Canada.** (n.d.). *Table 14-10-0287-01 Labour force characteristics by sex and detailed age group, monthly, seasonally adjusted*. Government of Canada.  
   <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410028701>
2. **Open Government, Government of Canada.** (n.d.). *Job vacancies by industry, Canada, unadjusted for seasonality, annual*.  
   <https://open.canada.ca/data/dataset/a70fcb3b-9a57-4e10-8372-9016935fc5d9>
3. **Arshkon.** (2023). *LinkedIn Job Postings [Data set]*. Kaggle.  
   <https://www.kaggle.com/datasets/arshkon/linkedin-job-postings>
4. **Google Careers / Google Jobs.** (n.d.). *Google Search Job Snippets (Scraped via Automation)*. Retrieved via programmatic browsing.  
   https://www.google.com/