Using Extract, Transform, and Load Framework and Data Visualization Tools to Enhance Career Services for Analytics Master's Program Student

Putranegara Riauwindu(✉) and Vladimir Zlatev

Boston University, Boston, MA 02215, USA
{putrangr,abamet}@bu.edu

Abstract. Tailored industry and occupation information for analytics graduates is vital to make a well-informed career decision, especially for Boston University Metropolitan College (BU MET) Applied Business Analytics students and graduates. This paper proposes an Extract Transform and Load (ETL) framework and Data Visualization method to provide students with easy-to-use and intuitive occupation information.

Multiple analytics-related industry and occupation data were extracted and aggregated from third-party sources, primarily from Lightcast and US Government Official Data. The resulting data underwent manipulation using Microsoft Power Query and Microsoft Excel and were stored in Microsoft SharePoint, with structured data in a flat table and unstructured data in a standalone file with a URL generated for linking the data. A relational database schema was then created to connect the ETL data output for visualization and analysis.

Interactive and user-friendly visualizations were created in Microsoft Power BI, resulting in two dashboards providing students with current information on the job market landscape: (i) Analytics Career Prospect, which offers data on top occupations, salary and wage information, job posting trends, required skills information, hiring industries and companies' information, education information, and job location; and (ii) Job Market Consultation, which provides a more in-depth analysis of required skills, industry performance and description, and specific job information reports such as Industry Insight, Industry Snapshot, Industry Supply Chain, Industry staffing pattern, and job posting analytics.

The resulting two dashboards provide "one-stop" search places for career research and shorten the cycle time of tedious searches.

Keywords: ETL · Microsoft Power BI · Relational Database · Structured & Unstructured Data · Data Manipulation · Career Development · Job Posting Analytics · Industry Insights

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
T. Zlateva and G. Tuparov (Eds.): CSECS 2023, LNICST 514, pp. 347–359, 2023.
https://doi.org/10.1007/978-3-031-44668-9_27

1 Introduction

1.1 Analytics: A Trending Career Prospects in a Data-Driven World

The analytics field is becoming increasingly important in our data-driven world, making it a trending career prospect for individuals seeking to impact their chosen industry significantly. With the rise of big data, according to Davenport and Patil (2012), companies are looking for experts in analytics to help them understand and make sense of the vast amounts of information they are collecting [1]. As a result, careers in analytics are in high demand, with a wide range of job opportunities available in various industries. From finance and healthcare to marketing and sports, the applications of analytics are virtually limitless.

According to a study by McKinsey Global Institute (Manyika et al., 2011), by 2025, the demand for professionals with analytics skills could exceed the supply by as much as 50 to 60 percent [2]. This trend emphasizes the need for individuals to pursue a career in analytics to meet the growing demand for this skill set. By doing so, they can take advantage of the numerous job opportunities available in this field.

1.2 Research Scope and Objectives

Given the narration above, conducting a job search in analytics can be difficult and complex. While there is a growing demand for analytics professionals, job postings in the field are also increasing rapidly. However, the job postings from one employer to another can differ significantly, making it challenging for analytics graduates to conduct initial research on where to apply, what kind of skills are required for various jobs, what possible occupations are available, which companies and industries are hiring, and where the employment locations. Hudson (2021) suggests that this complexity can make it challenging for graduates to navigate the job market and find the right fit for their skills and career goals [3].

This paper aims to propose a framework to help analytics students and graduates, especially in Boston University Metropolitan College Applied Business Analytics Master Program (BU MET ABA), in navigating the complex job market by providing provide easy to use and intuitive occupation and job reviews. The graduates can use the information to make a well-informed career decision after graduation, specifically to meet the following objectives:

1. The information must provide students with a list of potential occupations for ABA graduates in various locations in the US industry landscape.
2. The information must provide information about job trends for ABA graduates in different US industries.
3. The information must provide statistics on the potential occupations for ABA graduates within the US industry landscape in various locations, including but not limited to salary, wage, education, and experience.

The selected research framework is demonstrated with data for analytics-related industries and occupations in the United States (US) as dictated by the North America Industry Classification Standard (NAICS) and presented for the Finance, Insurance, and Pharmaceutical industries.

2 Literature Review

2.1 Career Analytics

Providing analytics graduates with insights and information about the job market and available occupations to provide them a competitive advantage in the rapidly evolving field is not a novel effort in an academic setting. This effort reflects the growing demand for analytics professionals and the complexity of navigating the job market, as employers' requirements and job postings can vary significantly.

For example, a study conducted by Wilbur and Angela Stanton (2020) suggests that the authors have performed "Career Analytics" by analyzing the skills required for an entry-level analytics position, focusing on data science, data analytics, and business analytics. The authors also identify the most in-demand job titles and functions and compare the required hard skills, soft skills, software skills, and credentials for each area. The authors wrap up the research by providing educators with recommendations on preparing students for careers in analytics [4].

Wilkins (2021) also researched the data analytics job market [5]. The author identified the most in-demand skill sets in analytics, providing insights for students, organizations, and universities. The authors argued that the findings would guide students in mastering the most relevant skills for the job market. At the same time, for the company, the information will aid them in understanding the most sought-after skills and competing for hiring. Lastly, the authors also suggested that for universities offering data analyst-related courses, the output could align their curriculum to meet the job market's needs.

The similarity between the two examples is that they both discuss the industry requirements for analytics graduates so that the students can prepare for the knowledge/skills "gap" and acquire a competitive advantage. Still, both types of research only capture the information on a specific period, resulting in a static research report. This paper will provide students with interactive and up-to-date information about the analytics industry landscape.

2.2 Extract, Transform, and Load Framework

Voluminous analytics-related industry and occupation data residing in multiple sources, a standardized framework to aggregate and collect all different data from different sources would need to be adopted to ensure scalability. The Extract, Transform, and Load (ETL) framework would suit this purpose well.

According to P. Vassiliadis' conceptual model for ETL, ETL tools are specialized tools designed to address data warehouse homogeneity, cleaning, and loading issues (Kimball & Caserta, 2004) [6]. Multiple commercial tools provide ETL functionality as a one-stop solution for all the ETL framework processes [7]. One example of a commercial ETL tool is Informatica, which offers holistic functionality to automate the data pipeline from multiple sources to the data warehouse or data lakes [8]. While most of the commercial ETL tools provide a broad range of advanced ETL functionality and cover the operations from upstream to downstream, Vassiliadis (2009) argues that ETL processes involve several key steps, including the extraction of relevant data from sources, the transportation of this data to a specific area of the data warehouse, and

the transformation of the data to comply with the structure of the target relation. ETL processes also entail the isolation and cleansing of problematic tuples to ensure adherence to business rules and database constraints, and ultimately, the loading of the cleansed, transformed data to the proper relationship in the warehouse, along with the refreshing of any indexes and materialized views [9].

This paper aims to provide a solution by leveraging the existing BU MET infrastructure while minimizing the use of external commercial ETL tools and maximizing the ETL process framework discussed by Vassiliadis (2009), which will be discussed further in the Methodology section.

2.3 Database Schema and Design

As previously highlighted in Sect. 2.2, the diverse nature of data types and sources necessitates a database schema and design to connect and comprehend the extracted and transformed data effectively.

Data can be either structured or unstructured and may come from various sources. Structured data is well-organized and can be stored in a traditional relational database management system (RDBMS) for easy querying using SQL. On the other hand, unstructured data has no pre-defined format and cannot be easily stored in relational tables. Unstructured data is the fastest-growing type of data and includes data from various sources such as images, sensor data, web chats, social media messages, videos, documents, log files, and email data [10].

A collection of structured or unstructured data, organized with a central focus, is known as a database. When dealing with computer databases, the tool used to input and modify data is referred to as either a database program or a database management system (DBMS), as opposed to a manual paper-based system [11]. In comparison, a schema is a structure that connects different information within a database to create a logical connection between the data.

The proposed in this paper framework uses both structured and unstructured data with the unstructured data projected into a meta-structured table containing the unstructured data and its associated link to the hosting place.

2.4 United States Labor Landscape

Let's introduce several essential characteristics of the labor landscape in the US:

Industry Classification. Various industries in the United States are classified into several categories based on their characteristics and nature, and the classification for these industries adheres to the North American Industry Classification System. The North American Industry Classification System (NAICS) is a standardized system used by Federal statistical agencies in the US to classify business establishments, gather statistical data, and analyze the US business economy. NAICS divides the economy into 20 sectors. Industries within these sectors are grouped according to the production criterion [12].

Occupation Classification. The Standard Occupational Classification (or SOC) system is a federal statistical standard employed by federal agencies for classifying workers

into occupational categories, aiming to collect, compute, or publish data. Workers are assigned to one of 867 distinct occupations based on their description. Detailed occupations are combined to create 459 broad occupations, 98 minor groups, and 23 major groups, making classification more efficient. Occupations with similar job responsibilities, and sometimes comparable skills, education, and/or training, are grouped in the SOC [13].

Registered Business. According to the latest data from the US Census Bureau, there were 6.1 million employer firms in the United States in 2019. Of those businesses, 99.7% had fewer than 500 employees, 98.1% had fewer than 100 employees, 89.0% had fewer than 20 employees, and 78.5% had fewer than ten employees. When nonemployer businesses are included, the share of firms with fewer than 20 workers increases to 98.0%, and those with fewer than ten employees represent 96.0%. Out of approximately 32.6 million businesses in the US, only 20,868 had 500 or more employees, indicating that America's economy is primarily comprised of small companies [14].

Skills Classification. There are no official skill classifications from the US government but Lightcast, one of the leading labor market aggregators and analytics, has suggested skill classification and definition. The Open Skills Library by Lightcast defines skills as abilities related to particular tasks or knowledge of specific subjects and tools obtained through education or experience. The library categorizes each skill as specialized, common, or certifications [15].

Specialized Skills, also known as technical skills or hard skills, are competencies that are mostly necessary within a specific occupation or enable an individual to carry out a particular task. Examples of specialized skills include "NumPy" or "Hotel Management."

Common Skills refer to the skills widely used across various occupations and industries, encompassing both learned skills and personal attributes. These skills may include "Communication" or "Microsoft Excel" and are also known as competencies, soft skills, or human skills.

Certifications refer to qualifications that are recognized by industry or educational bodies, such as a "Cosmetology License" or a "Certified Cytotechnologist" designation. These certifications indicate that the individual has achieved specific knowledge or expertise in a particular field or skill.

3 Methodology and Process

The proposed methodology and framework to achieve the research objective involves combining a simple Extract, Transform, and Load (ETL) framework and data visualization. This approach is distinctive because it is expected to generate dynamic output that evolves by adding new information. The resulting interactive output is designed to enable end-users to access relevant information directly instead of static, report-style output that remains unchanged as new data points emerge.

This paper's proposed methodology and framework could be summarized as Extracting and Collecting Data, Transforming Data, Loading Data to Cloud Storage, Connecting the Loaded Data, and Presenting the Data. The whole framework is illustrated and summarized in Fig. 1 below.

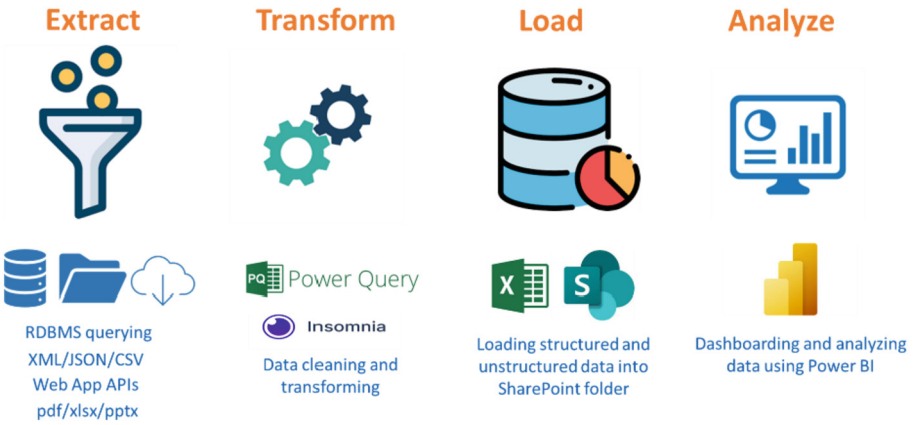


Fig. 1. Summary of Proposed Framework and Methodology to Enhance Career Services for Analytics Master's Program Student

3.1 Database Schema and Cloud Storage Creation

Before executing the ETL framework, a layout or Database Schema plan must be developed to meet the stated objectives first. The corresponding specific folders in the Cloud storage to host the data that went through the ETL framework must also be created first.

The database schema was created by connecting five main entities identified based on the business objective (Industry, Company, Skills, Occupations, and Location) to each other using a Relational Database model. Each entity had child entities, and the schema was made by grouping entities based on their relationships. The groups, or clusters, were then connected internally and externally, with different levels of connection (cardinality and optionality). A "bridge" entity was created to avoid complex relationships between groups. It connects two or more groups with many-to-many relationships in either one-to-many or many-to-one relationships.

The entities in the database schema were divided into two categories: "fact entities" and "dimension entities." Fact entities store transaction data from sources like Lightcast.io and the Bureau of Labor and Statistics, such as posted salary and job count within a specific industry. Dimension entities help gather information about the measures being taken, such as the industry and company. Fact entities are labeled with an "f" and dimension entities with a "d". The database schema that was developed according to the above explanation is available at the Appendix 1 Database Schema Career Services for Analytics Master's Program Student.

3.2 Collecting and Extracting Data

After the database schema was created, the first step of the ETL framework involved collecting and extracting data from various sources, primarily from Lightcast, the Bureau of Labor Statistics, the United States Census Bureau, EBSCO Information Services, and other internet sources. Data were acquired by querying third-party sources'

RDBMS, connecting to Application Programming Interfaces (API), and directly downloading in various formats, including Extensible Markup Language (XML), JavaScript Object Notation (JSON), Comma-Separated Values (CSV), Excel File (XLSX), Portable Document Format (PDF), and Power Point File (PPTX).

The data extracted and collected including but not limited to number of posted occupations, salary and wages, skills requirement, company and industry posting, education requirements, job locations, company information, industry performance, occupations hiring rate, Industry Insight, Industry Snapshot, Industry Supply Chain, Industry staffing pattern, and job posting analytics., and other information.

After extraction, the data were temporarily stored in a staging area. Files that were "ready-to-use" (such as PDF, PPTX, and some XLSX files) were directly transferred to their respective folders in the Cloud Storage. Files that needed further processing (such as XML, JSON, CSV, and some XLSX files) were kept in the staging area for the next step.

For the result demonstration in this paper, the time-period of data collection is from November 2022 – March 2023.

3.3 Transforming Data

The data that needs further processing then undergoes a data transformation process. The data transformation includes but is not limited to:

1. The removal of unnecessary variables/columns in the data.
2. Variable type conversion from a certain type to another type.
3. The removal of unnecessary strings or characters within specific variables.

All data manipulation and transformation were done in Microsoft Power Query.

3.4 Loading Data to Cloud Storage

The cleaned/transformed data is then loaded into Microsoft Excel flat table entity and stored in the cloud storage in their dedicated folder based on the previously developed database schema. The cloud storage for storing the data uses Microsoft SharePoint.

Each flat table is protected so that only person who has the authority to modify the flat table to populate the flat table with new or updated information could modify the flat table.

3.5 Connecting the Loaded Data

In order to connect both the structured (Excel flat table) and unstructured data (ppt, pdf, etc.), first, the link that redirects to the unstructured data was generated. These links were then compiled into separate Excel flat tables containing detailed information regarding the unstructured data (metadata) associated with the previously developed database schema.

Each flat table, for both structured data and unstructured metadata, was then connected to the Microsoft Power BI. The connection and relationship between the entity/flat tables were then created based on the database schema previously developed.

3.6 Presenting the Data

The connected entity/flat table were then visualized in the Microsoft Power BI environment. Several pages of the dashboard with different visualizations were created to provide specific information to the students, such as Top Occupation, Top Skills, Top Company Hiring, Skill Deep Dive, Industry Overview, Industry, and Occupations Report. Some examples of the visualizations used were bar charts, line plots, scatter plots, tables, maps, tree maps, and other visualizations.

4 Results and Discussion

The resulting interactive and user-friendly visualizations, created in Microsoft Power BI, were two dashboards providing students with current information on the job market landscape: Analytics Career Prospect Overview and Analytics Job Market Consultation dashboard, which will be discussed further below.

4.1 Dashboard 1: Analytics Career Prospect Overview

The first dashboard created was named Analytics Career Prospect Overview. The Analytics Career Prospect dashboard offers data on top occupations, salary and wage information, job posting trends, required skills information, hiring industries and companies' information, education information, and job location.

Students can browse topics of interest at their whim. For example, what are the top occupations for analytics graduates? What are the top skills required by employers? Which companies are actively hiring and posting the most job openings? Which industry hosts the most job openings? What is the job posting trends? What are the mean and median salaries? Where are the job locations?

By using this dashboard, students can jumpstart their initial career research by finding opportunities as analytics graduates, understanding salary expectations, identifying gaps in their skillset compared to industry requirements, determining the best time to apply for jobs, exploring job openings in their preferred cities, and identifying certifications that will set them apart from others in the field. An example of a page in the dashboard is available in Appendix 2, Sample Page from the Dashboard 1 Analytics Career Prospect Overview.

Some interesting takeaways from the first dashboard were:

1. Data Scientists tops the posted occupations for analytics graduates, followed by Management Analysts and Market Research Analyst.
2. The most frequently reported range of posted analytics graduate salaries was around $67,000–$75,999.
3. The job posting for analytics graduates increased over the last three years, with most jobs posted in Q1-Q2 each year.
4. The most sought-after technical skills for an analytics graduate were Structured Query Language (SQL), Python, Microsoft Excel, Tableau, and R.

4.2 Dashboard 2: Analytics Job Market Consultation

The second dashboard created was named Analytics Job Market Consultation. This dashboard provides a more in-depth analysis of required skills, industry performance and description, and specific job information reports such as Industry Insight, Industry Snapshot, Industry Supply Chain, Industry staffing pattern, and job posting analytics.

After students perform overview research about their career aspirations in the first dashboard, they can then use this dashboard to conduct in-depth research on critical items they found in the first dashboard. For example, if students want to work in the finance industry, they can see how well it performs by looking at the Industry Performance page in the second dashboard. If students know specific skill gaps, they can visit the Skills page to find detailed explanations about those skills and where to start improving them. Suppose students want to delve deeper into each career prospect within a specific industry. In that case, they can go to the Documents page and download the in-depth reports provided by Lightcast to prepare accordingly (Fig. 2).

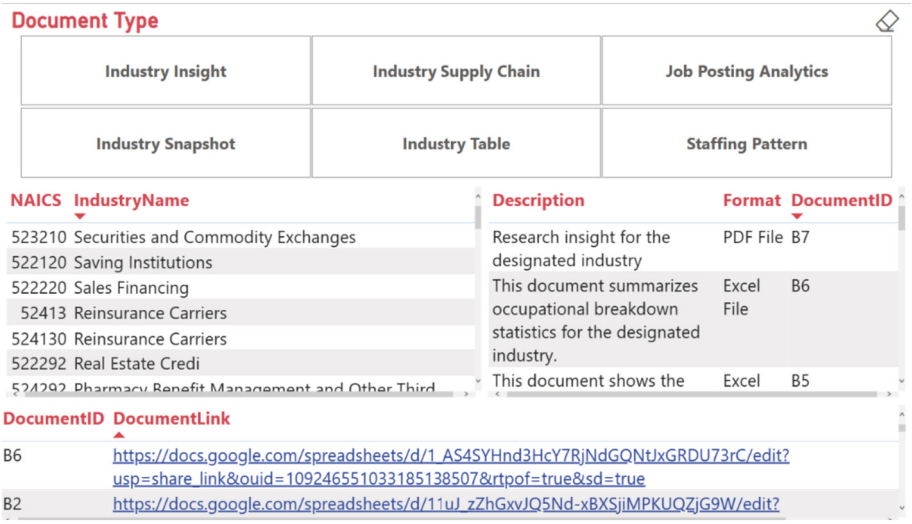


Fig. 2. Sample Page from Dashboard 2 Analytics Job Market Consultation

Some interesting takeaways from the second dashboard were:

1. There is one hire for every three unique job postings for Data Scientists, two for every individual job posting for Management Analyst, and three for every particular job posting for Market Research Analyst. Hence, the job demand for a Market Research Analyst/Marketing Specialist was relatively high in the United States.
2. Every US industry (20 NAICS categories) has had overall positive gross-output growth trends for the last 20 years. Manufacturing generally outperforms every other industry's gross output by roughly ten times.

5 Summary and Recommendations

5.1 Summary: Interactive Job Information Dashboard to Navigate the Complex Job Search Process

The dashboards, as the output from the research, are now available for students and graduates to access on the BU MET ABA Career Service Website. The dashboards were designed to simplify job research for ABA students and graduates by providing them with relevant and up-to-date information. Through the dashboards, users can easily search and navigate the job market landscape based on various parameters such as location, job title, and employer.

The dashboards provide users with critical job market insights, including industry trends, job market statistics, and salary data. The dashboards help users make informed decisions about their job search strategies by presenting this information in a clear and concise format. Overall, the dashboards are an indispensable tool for ABA students and graduates, assisting them in achieving their career goals and staying informed about the latest job market trends.

5.2 Recommendation for Next Work

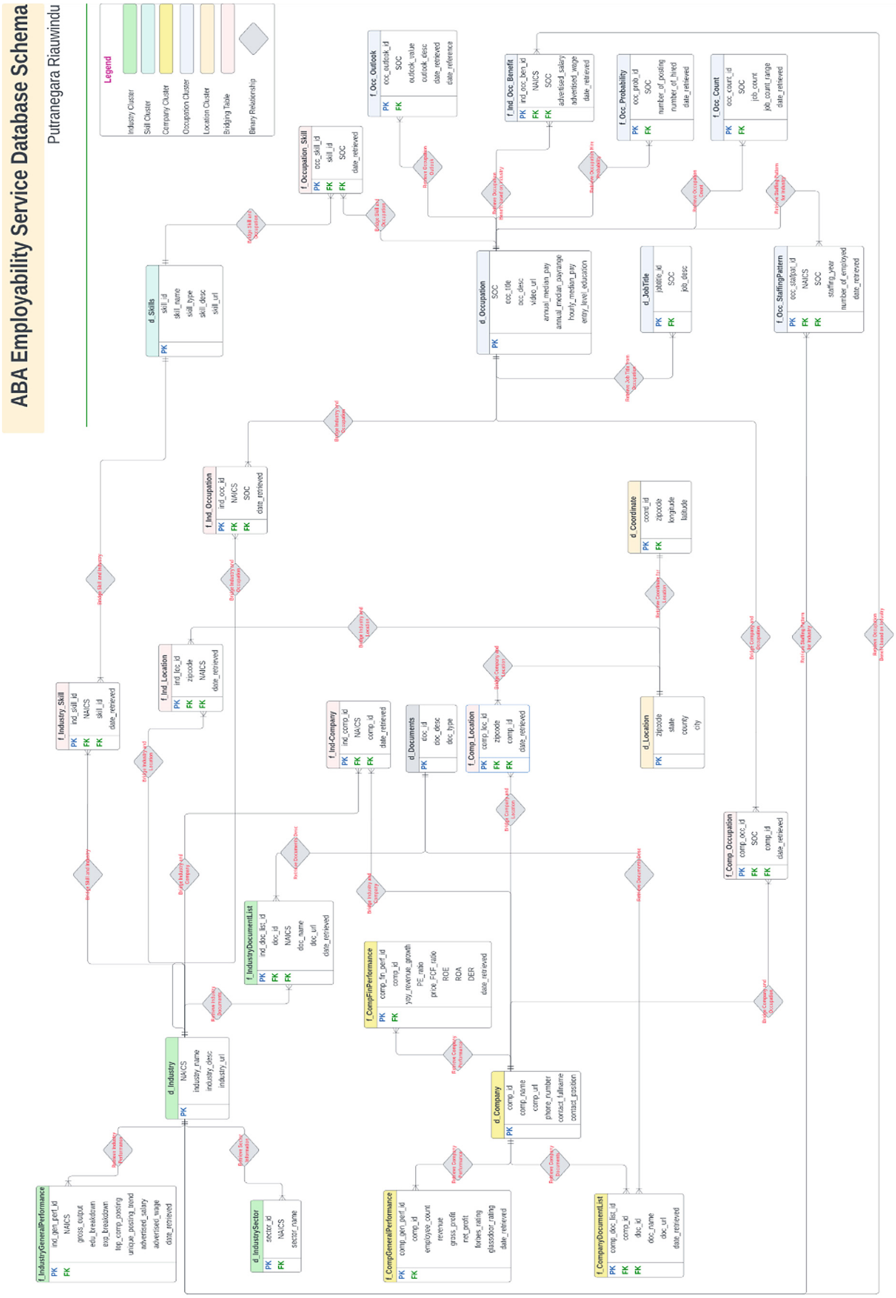
Given the wide range of industries in the US, the next step in developing the dashboard is to continue populating it with more industry and occupation information and make comprehensive comparison for the built and existing products.

Another possible next step is to perform an Analytics Occupation Deep Dive, conducting deeper analysis on each possible analytics occupation. It includes examining specific job posting trends, determining the best time to apply, identifying factors that increase a candidate's likelihood of being hired in a particular occupation, and understanding the typical profile of individuals working in those occupations within specific industries.

Another possibility is to conduct Analytics Occupation Data Mining Analysis, specifically clustering and classification. This analysis aims to identify which industries and occupations are more "analytics-graduate" friendly based on their characteristics and similarities. It could give students and graduates more information on which enterprises to focus on.

Appendix

1. Database Schema Career Services for Analytics Master's Program Student


ABA Employability Service Database Schema
Putranegara Riauwindu
Legend
Industry Cluster
Skill Cluster
Company Cluster
Occupation Cluster
Location Cluster
Bridging Table
Binary Relationship
d_Skills
f_Occupation_Skill
f_Occ_Outlook
f_Ind_Occ_Benefit
f_Occ_Probability
f_Occ_Count
d_Occupation
d_JobTitle
f_Occ_StaffingPattern
f_Ind_Occupation
d_Coordinate
f_Industry_Skill
f_Ind_Location
f_Ind-Company
d_Documents
f_Comp_Location
d_Location
f_Comp_Occupation
f_IndustryDocumentList
f_CompFinPerformance
d_Industry
d_Company
f_IndustryGeneralPerformance
d_IndustrySector
f_CompGeneralPerformance
f_CompanyDocumentList

2. Sample Page from the Dashboard 1 Analytics Career Prospect Overview

References

1. Davenport, T.H., Patil, D.J.: Data scientist: the sexiest job of the 21st century. Harv. Bus. Rev. **90**(10), 70–76 (2012)
2. Manyika, J., et al.: Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Inst. (2011). https://doi.org/10.1007/978-1-4614-4134-9_1
3. Hudson, D.: Tips for navigating the academic job market. Health Promot. Pract. **22**(1), 21–23 (2021). https://doi.org/10.1177/1524839920938802
4. Stanton, W.W., Stanton, A.D.: Helping business students acquire the skills needed for a career in analytics: a comprehensive industry assessment of entry-level requirements. Decis. Sci. J. Innov. Educ. **18**(1), 138–165 (2020). https://doi-org.ezproxy.bu.edu/10.1111/dsji.12199
5. Wilkins, D.: An in-depth analysis of the data analytics job market. Master's thesis. University of New Hampshire, Peter T. Paul College of Business & Economics (2021)
6. Kimball, R., Caserta, J.: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley, Hoboken (2004)
7. Edjlali, R., Beyer, M.A.: Magic quadrant for data warehouse and data management solutions for analytics, Gartner (2016)
8. Mukherjee, R., Kar, P.: A comparative review of data warehousing ETL tools with new trends and industry insight. In 2017 IEEE 7th International Advance Computing Conference (IACC), pp. 371–375. IEEE (2017)
9. Vassiliadis, P.: A survey of extract-transform-load technology. Int. J. Data Warehouse. Min. **5**, 1–27 (2009)
10. Mishra, S., Misra, A.: Structured and unstructured big data analytics. In: 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India, pp. 740–746 (2017). https://doi.org/10.1109/CTCEEC.2017.8454999
11. Derclaye, E.: What is a database? J. World Intellect. Property. **5**, 981–1011 (2005). https://doi.org/10.1111/j.1747-1796.2002.tb00189.x
12. Executive Office of the President, Office of Management and Budget. North American Industry Classification System. United States (2022)
13. Executive Office of the President, Office of Management and Budget. Standard occupational classification manual. United States (2018)
14. US Census Bureau. Statistics of US Businesses (SUSB) (2019). https://www.census.gov/data/tables/2019/econ/susb/2019-susb-annual.html. Accessed 12 May 12 2023
15. Lightcast. (n.d.). Open Skills Library - Frequently Asked Questions. https://lightcast.io/open-skills/faqs. Accessed 12 May 2023