Internship Report

Nikola DOBRICIC
Ecole Polytechnique Bachelor Program
July 1st - August 31st
INRIA, CEDAR team

1 Executive Summary of the Internship

The internship at the CEDAR Team of INRIA was a journey that spanned multiple aspects of data science, particularly focusing on graph databases, user-interface (UI) design, and query optimization. The team's flagship product, ConnectionStudio, serves as a robust platform for transforming structured and semi-structured data into graph databases. While it excels in data storage and transformation, ConnectionStudio presented a challenge in the form of complex SQL query formulation, particularly for users lacking technical expertise, such as journalists, who are generally the target demographic.

The goal of the internship was to bridge this gap between complex data structures and user accessibility. This was to be accomplished through the development of an interactive Graphical User Interface (GUI) tool that allows users to articulate their query intents in a user-friendly manner. These intents would then be automatically translated into SQL queries that could be executed on the underlying databases.

During the course of the internship, I engaged in a deep dive into the existing data integration architecture of ConnectionStudio to understand the intricacies involved. The task was complicated by the need to make the system compatible with Windows OS, a feature previously unexplored by the team. The challenge was not just to develop a new feature but also to ensure that it was seamlessly integrated into the existing ecosystem, irrespective of the operating system.

By the end of the internship, a fully functional interactive query tool was developed and integrated into ConnectionStudio. This tool not only simplifies the query process for existing users but also makes the system more approachable for potential users who may not have extensive technical expertise. The successful completion of this project also led to the identification of further areas of improvement and set the stage for future developments.

Contents

1 Executive Summary of the Internship **1**

2 Description of the Company/Institution **4**

2.1 Company Name: 4

2.2 Brief History: 4

2.3 Description of the Internship Company/Institution 4

2.3.1 Field of Activity: 4

2.3.2 Main Products and Services: 5

2.3.3 Unique Characteristics: 5

2.3.4 Company/Institution's Strategy and Operational Overview: 5

2.4 Why Did You Choose This Company/Institution? 5

2.4.1 Multi-disciplinary Learning: 6

2.4.2 Compatibility with Skill Set: 6

3 Internship activities and position **6**

3.1 Work Environment and Tasks 6

3.2 Analysis of One Important Task 6

4 Assessment of the Internship **8**

4.1 Skills and Qualifications Gained from the Internship 8

4.1.1 Technical Skills: 8

4.1.2 Soft Skills: 8

4.2 Influence on Future Career Plans 9

4.3 Relationship Between Internship Activities and Classroom Knowledge 9

5 Conclusions of the Report **10**

5.1 Summary of Conclusions Derived from the Internship Experience 10

5.2 General Observations About the Sector 10

5.2.1 Increasing Relevance: 10
5.2.2 Interdisciplinary Nature: 11
5.2.3 User-Centric Shift: 11

6 References **11**

2 Description of the Company/Institution

2.1 Company Name:

INRIA (Institut national de recherche en sciences et technologies du numeriques), CEDAR Team

2.2 Brief History:

CEDAR is a joint project-team of Inria Saclay and LIX, focused on rich data analytics at cloud scale.

The description from the INRIA website describes it best:

> Our research is organized along two main themes:
>
> 1. We seek to exploit parallel data processing infrastructures for highly scalable, parallel Big Data storage and processing tools. Topics we study in this area include: highly efficient query processing in polystores, efficient query answering in the presence of ontologies, and algorithms for scalable, fast data analytics.
> 2. To enhance the usefulness of Big Data, we study new paradigms of user interaction with Big Data, based on exploratory querying, analytics for semantic graphs, and intuitive query tools over highly heterogeneous data.

For more information, you can visit the CEDAR Team Website and their Research Publications.

The company has been recognized for its innovative solutions. With a multidisciplinary team comprising data scientists, engineers, and UI/UX designers, CEDAR has been successful in creating a niche for itself in the ever-evolving landscape of data science.

2.3 Description of the Internship Company/Institution

2.3.1 Field of Activity:

The CEDAR Team primarily operates in the domain of data science. Their expertise lies in data integration, graph databases, and information extraction.

2.3.2 Main Products and Services:

The flagship product is ConnectionLens (and ConnectionStudio that's built on top of it), a system designed to transform structured and semi-structured data into graph databases. The platform offers a range of services including data storage, transformation, and query optimization.

2.3.3 Unique Characteristics:

What sets the CEDAR Team apart is their multidisciplinary approach to solving complex data problems. The team comprises experts from diverse backgrounds such as computer science, data analytics, and design, fostering a culture of innovation and collaboration.

2.3.4 Company/Institution's Strategy and Operational Overview:

The overarching strategy of the CEDAR Team is to democratize access to complex data structures. They aim to simplify the data storage and query process, making it more accessible to both technical and non-technical users. This is accomplished by investing in research to develop more intuitive data abstraction methods and user interfaces.

Operational implementation of this strategy involves a continuous cycle of research, development, and user feedback. Prototypes are developed based on cutting-edge research and are then subjected to rigorous testing. Feedback from these tests is used to refine the products, ensuring they meet user needs and industry standards. This iterative process allows the company to stay ahead in the fast-paced world of data science, continually innovating and adapting to new challenges and opportunities.

2.4 Why Did You Choose This Company/Institution?

My decision to intern with the CEDAR Team was influenced by a combination of factors, each aligning with both my academic pursuits and career aspirations.

2.4.1 Multi-disciplinary Learning:

The team's multi-disciplinary approach was another major draw. With experts in data science, engineering, and UI/UX design, the CEDAR Team offered a rich learning environment. I was not only able to hone my technical skills but also gained insights into user experience design and project management.

2.4.2 Compatibility with Skill Set:

My background in both data science and front-end development made me confident that I could contribute meaningfully to the team's goals.

3 Internship activities and position

3.1 Work Environment and Tasks

I worked roughly 35 hours per week, from Monday to Friday, with a flexible schedule. I was given the freedom to choose my working hours, as long as I was available for meetings and discussions. I had regular meetings with my supervisor, as well as the other team members, to discuss the progress of the project and plan the next steps.

3.2 Analysis of One Important Task

Arguably the most important part of my task was the SQL query generation mechanism. The goal was to develop an algorithm that could translate user intents into SQL queries. This involved a deep dive into the existing data integration architecture of ConnectionStudio, as well as an understanding of the intricacies of SQL query formulation.

The goal of the task was to translate user intents from UI into an SQL query that is executed on a graph database. This proved to be very challenging. I had multiple meetings with my supervisor in an effort to understand the links between entities shown in the UI and the underlying data repre-

sentation. There were a lot of things to take into consideration, such as properly joining attributes to corresponding entities, avoiding duplicate results, and ensuring that the query is optimized for efficiency.

In the end, I achieved the solution by traversing through the entire path from the entity, down to the attribute itself, while taking care that each node only gets queried once. This was accomplished by using a `HashMap` to keep track of the nodes that have already been queried. The algorithm was also optimized to avoid unnecessary joins and ensure that the query is as efficient as possible.

Along with the UI, the end product is shown on Figure 1:

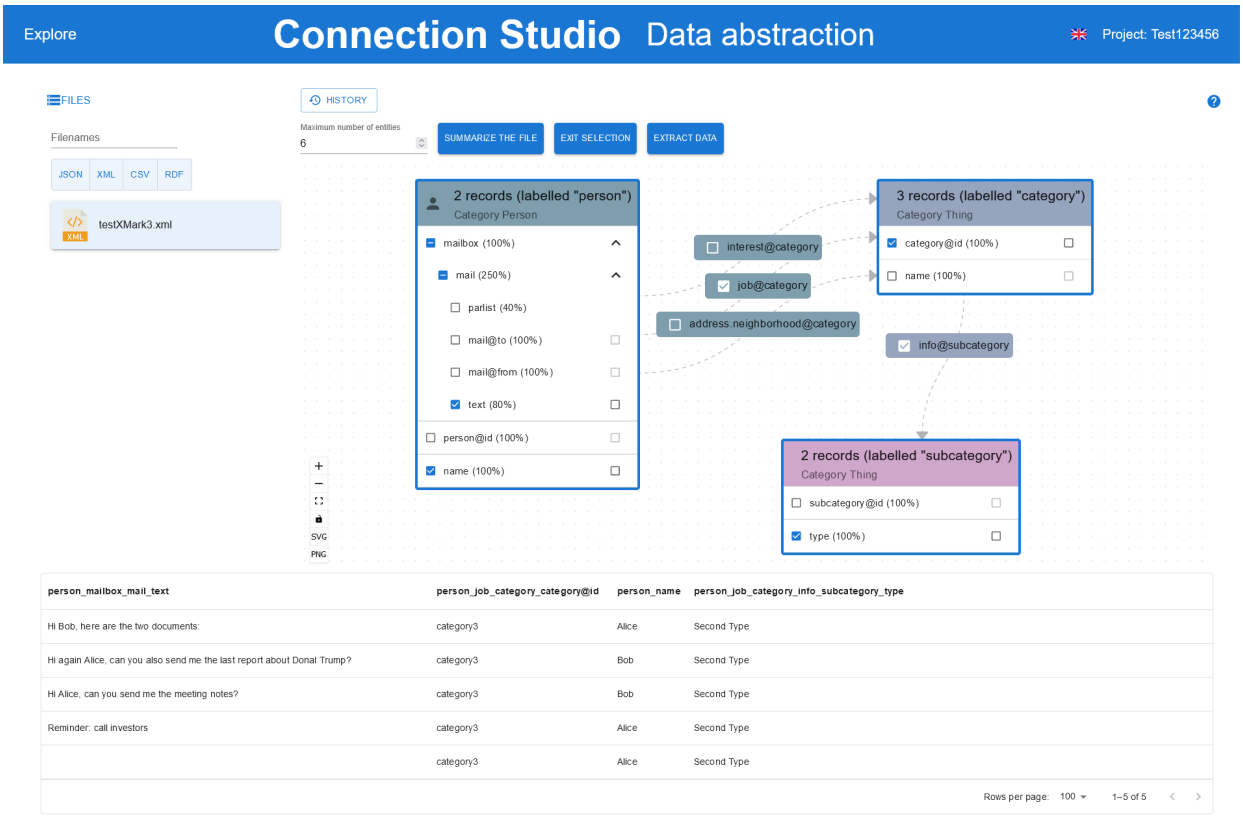


Figure 1: ConnectionStudio UI for extracting data from abstracted entities

4 Assessment of the Internship

4.1 Skills and Qualifications Gained from the Internship

Here are the main skills and qualifications gained from the internship:

4.1.1 Technical Skills:

1. **SQL Proficiency:** The project involved extensive use of SQL for data manipulation and query formulation. I gained a deep understanding of advanced SQL techniques, as well as commands specific to PostgreSQL.
2. **Database Architecture:** Understanding the architecture of ConnectionStudio's graph database was crucial for my tasks. I gained a comprehensive understanding of how data is stored, accessed, and manipulated in complex databases.
3. **Front-end Development:** My involvement in developing the interactive GUI tool allowed me to refine my front-end development skills, particularly in JavaScript, React, and Material UI.
4. **Algorithm Design:** Translating user intents into SQL queries required the development of specialized algorithms. This improved my problem-solving skills and understanding of algorithmic efficiency.
5. **Java Programming:** The internship also involved back-end development in Java. This allowed me to apply my Java programming skills in a real-world setting.

4.1.2 Soft Skills:

1. **Project Management:** Managing the development cycle of the GUI tool, from the design phase to implementation and testing, improved my project management skills.
2. **Team Collaboration:** Working closely with a multidisciplinary team taught me the impor-

tance of effective communication and teamwork.

3. **Critical Thinking:** The project presented numerous challenges that required innovative solutions. This honed my critical thinking and decision-making skills.

4. **User-Centric Design:** Gathering user feedback and iterating on the design of the GUI tool instilled in me the importance of user-centered design principles.

4.2 Influence on Future Career Plans

My internship at INRIA was a very enjoyable experience. However, it helped me realize that front-end development is not my primary area of interest. While I enjoyed the work, I found myself more drawn to the data integration and query optimization aspects of the project. My passion lay in more technical work, and front-end development was not the best fit for my career goals.

Having said that, I'm actively working on a side-project that involves a very similar technology stack, and my time at INRIA helped me immensely in that regard. I feel much more comfortable interacting with relational databases and writing SQL queries, as well as writing clear and optimized React code. I also learned a lot about the importance of user-centric design and how to incorporate it into my projects.

4.3 Relationship Between Internship Activities and Classroom Knowledge

Unfortunately, my classroom knowledge from the Bachelor Program was not directly applicable to the internship. The internship involved a lot of front-end development, which is not a part of the Bachelor Program curriculum. However, I was able to apply my knowledge of data science and relational databases to the project. I also gained a deeper understanding of SQL and its applications, which will be useful in future projects.

Although, during my high school years, we had a course on relational database design and SQL programming which definitely had a major impact on my internship.

Having said that, most of the skills required for the internship were acquired through self-learning and personal projects.

5 Conclusions of the Report

5.1 Summary of Conclusions Derived from the Internship Experience

The internship was a great learning experience. I was able to apply my technical skills in a real-world setting, while also gaining valuable insights into the world of data science. I learned a lot about the challenges involved in data integration and query optimization, as well as the importance of user-centric design. I also gained a deeper understanding of the data science sector and its future outlook.

The internship also helped me identify my strengths and weaknesses. I realized that I'm more interested in technical work than front-end development. I also learned that I enjoy working in a collaborative environment, where I can learn from and contribute to a multidisciplinary team.

The internship was a very enriching experience. It helped me grow both professionally and personally, and I'm grateful for the opportunity.

5.2 General Observations About the Sector

The sector in which the CEDAR Team operates—data science, with a focus on data integration and graph databases—is a rapidly evolving field marked by constant innovation and growing importance across industries. Here are some general observations about this sector:

5.2.1 Increasing Relevance:

In an age where data is often referred to as the "new oil", the ability to manage, integrate, and utilize data is becoming increasingly critical. Organizations across the board, from healthcare to finance to journalism, are seeking effective ways to handle their ever-expanding data sets, making

this sector highly relevant.

5.2.2 Interdisciplinary Nature:

One striking feature is the interdisciplinary nature of the work. The best solutions are often born from collaborative efforts involving not just data scientists but also software engineers, UI/UX designers, and even domain experts from the fields the data is sourced from.

5.2.3 User-Centric Shift:

There is a noticeable shift towards making data tools more user-friendly. As data becomes more integral to business decisions, there is a growing need to make these tools accessible to non-technical users, a trend clearly reflected in CEDAR Team's mission and projects.

In summary, the sector is a dynamic and evolving field, presenting a plethora of opportunities and challenges. Its interdisciplinary, user-centric, and data-driven nature makes it an exciting sector to work in.

6 References

[1] Inria Official Website: https://www.inria.fr/en

[2] CEDAR Team Website: https://team.inria.fr/cedar/

[3] CEDAR Team Publications: https://team.inria.fr/cedar/publications/

[4] Angelos-Christos Anadiotis, Oana Balalau, Catarina Conceicao, Helena Galhardas, Mhd Yamen Haddad, Ioana Manolescu, Tayeb Merabti, Jingmao You. *Graph integration of structured, semistructured and unstructured data for data journalism*. In Elsevier Journal of Information Systems, 104:101846, 2022

[5] Nelly Barret, Ioana Manolescu, Prajna Upadhyay. *Computing Generic Abstractions from Application Datasets*. EDBT 2024 (research paper).