'It is a capital mistake to theorize before one has data. Insensibily one begins to twist facts to suit theories, instead of theories to suit facts.'

Sherlock Holmes Quote

-A Scandal in Bohemia
Longitudinal Data Integration for a Tracking System for Health Professionals

by

Raafey Ahmed Mohammed

Bachelor of Engineering in Computer Science (Osmania University, 2012)

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Masters in Computer Science

in the Graduate Academic Unit of Computer Science

Supervisor:             Janet Light, PhD, Computer Science

Examing Board:  Owen Kaser, PhD, Computer Science
                 Liwen Zou, PhD, Statistics
                 Suprio Ray, PhD, Computer Science

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

February, 2016

©Raafey Ahmed Mohammed, 2016
ABSTRACT

Increase in the number of programs offered and awareness on the benefits of data collection and analysis have prompted universities across the globe to develop tracking systems that monitor the socio-economic impact of their graduates on the community. A data warehouse is required to store the data for analysis and reporting. In this thesis, an approach for building a longitudinal data warehouse is proposed for such tracking systems, in which data acquired from several sources is cleaned, integrated and structured for analytical reporting. The challenges encountered during the development of this tracking system are addressed with proposed solutions. A front-end application is also designed and developed that displays the queried data in the form of reports and charts.
DEDICATION

I would like to dedicate this thesis to my parents, siblings and wife whose support and words of encouragement have forever given me the strength to strive, achieve and conquer.
ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my supervisor, Dr. Janet Light, whose motivation and ideas have helped me complete this work. I would like to thank Dr. Owen Kaser who gave a course in winter 2015 on Data Analytics and Bitmaps that helped in acquiring a profound understanding of data warehouses. I am also thankful to the entire team of the LOCATED project at Dalhousie Medicine New Brunswick (DMNB) and Dalhousie Medical University – Halifax with whom I worked between September 2012 and February 2015. A special mention goes to Dr. John Steeves, Principal Investigator of the LOCATED project and former Associate Dean at DMNB, whose energy and enthusiasm propelled me towards success and lastly my mentor Dr. Silvane Paixão for her contributions to the direction and richness of this research.
# Table of Contents

ABSTRACT ....................................................................................................................... ii
DEDICATION ....................................................................................................................... iii
ACKNOWLEDGEMENTS .................................................................................................... iv
Table of Contents .............................................................................................................. v
List of Tables ................................................................................................................... ix
List of Figures .................................................................................................................. x
List of Symbols, Nomenclature or Abbreviations ............................................................ xii

1 Introduction and Motivation ......................................................................................... 1
   1.1 Longitudinal Data .................................................................................................. 2
   1.2 Student Tracking Systems .................................................................................... 2
   1.3 LOCATED Project ................................................................................................. 4
   1.4 Objective of the Thesis ........................................................................................ 4
   1.5 Organization of the Thesis ................................................................................... 4

2 Background .................................................................................................................. 6
   2.1 Longitudinal Data ................................................................................................. 6
   2.2 Real World Applications for Longitudinal Data .................................................... 8
   2.3 Types of Longitudinal Data .................................................................................. 8
       2.3.1 Person-Level Data ....................................................................................... 9
       2.3.2 Person-Period Data ................................................................................... 10
   2.4 Retrospective Cohort Study ................................................................................ 12

3 Related work ............................................................................................................... 14

4 Challenges with Longitudinal Datasets ..................................................................... 17
4.1 Lack of Administrative Policies ................................................................. 17
4.2 Data Standardization Challenge ............................................................... 17
4.3 Data Quality Challenges with Inconsistent Data ......................................... 18
4.4 Linking Static and Dynamic Data .............................................................. 19
4.5 Technology Variance ............................................................................... 20
4.6 Waiting For Data to Mature ..................................................................... 20
4.7 Confidentiality of Data ............................................................................ 21

5 Data Warehouse ......................................................................................... 22
5.1 Data Warehouse framework ..................................................................... 22
5.2 Data Acquisition ..................................................................................... 23
5.3 Data Cleaning ......................................................................................... 25
5.4 Data Integration ...................................................................................... 28
5.5 Data Structuring ....................................................................................... 30
5.6 Analysis/Representation ......................................................................... 31
5.7 Annual Updates ....................................................................................... 33

6 Longitudinal Data warehouse Architecture .............................................. 34
6.1 Typical Data Warehouse Architecture ...................................................... 34
6.2 Dimensional Systems ............................................................................. 35
6.2.1 Advantages and Limitations of a Dimensional System ....................... 36
6.3 Dimensional or Relational Data Model - Factors to Consider ................... 37
6.3.1 What is the Depth of Analysis? ................................................................. 37
6.3.2 How important is Performance? ................................................................. 38
6.3.3 How dynamic is the Data? ................................................................. 38
6.3.4 How dense is the Data? ................................................................. 39
6.3.5 Who are the Users? ................................................................. 39
6.4 Selection of Data Model ........................................................................ 40
6.5 Adjusted Design ........................................................................ 40

7 Implementing a Longitudinal Data Warehouse ........................................ 42
7.1 Extraction ........................................................................ 42
    7.1.1 Identifying the Research Question .......................................................... 42
    7.1.2 Identifying Data Sources ....................................................................... 44
    7.1.3 Data Acquisition .................................................................................. 45
    7.1.4 Challenges of Extraction from Different Sources .................................. 47
7.2 Transformation ........................................................................ 51
    7.2.1 Data Cleaning .................................................................................... 51
    7.2.2 Data Integration .................................................................................. 54
    7.2.3 Data Structuring .................................................................................. 58
7.3 Loading .................................................................................. 60

8 Results ............................................................................................... 61
8.1 Increase in Amount of Analyzable Data ................................................. 61
8.1.1 Quantifiable Results ................................................................. 61
8.1.2 Non-quantifiable or Qualitative Results ............................... 62
8.2 Front End Application for Analysis and Reporting .................. 62
  8.2.1 Application Architecture ...................................................... 63
  8.2.2 Design of the Application .................................................... 63
  8.2.3 Multiple Perspective Approach ........................................... 67
  8.2.4 Need for a New GUI In Spite of Existing BI Tools ............... 72
  8.2.5 Applying Filters ................................................................. 74
  8.2.6 Security ............................................................................. 76
  8.2.7 Mining Custom Data ........................................................... 76
  8.2.8 Validation and Update ......................................................... 77

8 Conclusion and Future Work ......................................................... 79
  9.1 Future Work ........................................................................... 80

Curriculum Vitae
List of Tables

Table 2-1 Data collection points in a longitudinal student tracking system .................. 7
Table 2-2: Person-level data set showing locations of subjects over 4 years ............... 10
Table 2-3: Person-period data set showing locations of subjects over 4 years .......... 11
Table 4-1 Data collected from entry source with high school field not mandatory .... 19
Table 5-1: Inconsistent data with anomalies ....................................................... 26
Table 5-2: Data cleaned of inconsistencies and anomalies ....................................... 26
Table 5-3 Example of Data collected in 2013 ........................................................ 30
Table 5-4 Example of Data collected in 2014 ....................................................... 30
Table 7-1: Table from internal data showing city and province ............................... 49
Table 7-2 Table from external source showing city and province ............................. 49
Table 7-3: Table from Census Canada showing city and province ........................... 49
Table 7-4: A table with student data from ‘Source A’ ......................................... 55
Table 7-5: Table with student data from ‘Source B’ ............................................ 55
Table 7-6: An excerpt from the province helper table .......................................... 56
Table 7-7 Dataset showing physician practice locations with timestamps ............... 58
Table 7-8 Data in a person-level dataset ............................................................. 59
List of Figures

Figure 2-1: Data collection points in a longitudinal student tracking system.............. 7
Figure 5-1 Data Warehouse Framework ........................................................................ 24
Figure 5-2: Screenshot of the LOCATED Application ............................................. 32
Figure 6-1: Typical Data Warehouse Architecture ...................................................... 35
Figure 6-2: A 3-dimensional cube in a dimensional system ..................................... 36
Figure 6-3: Adjusted Data Warehouse Architecture ................................................... 41
Figure 7-1: Chart showing time spent in various stages during this project lifecycle .... 46
Figure 7-2 Comparison of dirty and clean data for ‘Information-A’ within a dataset .... 53
Figure 7-3 Comparison of dirty and clean data for ‘Information-B’ within a dataset .... 53
Figure 8-1 Front-end Application Architecture .......................................................... 63
Figure 8-2 Screenshot of the LOCATED Windows forms Application ..................... 65
Figure 8-3 Screenshot showing the multiple perspective approach adapted in the application .................................................................................................................................. 66
Figure 8-4 Demographic Characteristic tab in the LOCATED application ............... 67
Figure 8-5 Medical Training tab in the LOCATED application .................................. 68
Figure 8-6 Residency and Fellowship tab in the LOCATED application ................. 69
Figure 8-7 Practice Location and Career Choice tab in the LOCATED application ...... 70
Figure 8-8 Quick Report tab in the LOCATED application ....................................... 71
Figure 8-9 Screenshot showing filters implemented on the LOCATED Application ..... 73
Figure 8-10 University location experience of all students from Class of 2006 .......... 74
Figure 8-11 University location experience of female students from Class of 2006...... 75
Figure 8-12 University location experience of male students from Class of 2006^6 ....... 75
Figure 8-13 Application Secure login page ................................................................. 76
Figure 8-14 Mining Custom data from the LOCATED application......................... 77
Figure 9-1 Graph database model showing subjects, professions and locations as nodes and relationships as edges ................................................................. 82
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPER</td>
<td>Canadian Post M.D. Education Registry</td>
</tr>
<tr>
<td>CaRMS</td>
<td>Canadian Resident Matching Service</td>
</tr>
<tr>
<td>DBMS</td>
<td>Database Management System</td>
</tr>
<tr>
<td>E</td>
<td>Event</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract Transform Load</td>
</tr>
<tr>
<td>$E_{ts}$</td>
<td>Event at time ‘t’ for subject ‘s’</td>
</tr>
<tr>
<td>IDE</td>
<td>Integrated Development Environment</td>
</tr>
<tr>
<td>L</td>
<td>location</td>
</tr>
<tr>
<td>LOCATED</td>
<td>Location Of Clinicians And Trainee Education Dalhousie</td>
</tr>
<tr>
<td>$L_{pl}$</td>
<td>Length of Person-level dataset</td>
</tr>
<tr>
<td>$L_{pp}$</td>
<td>Length of Person-period dataset</td>
</tr>
<tr>
<td>MDX</td>
<td>Multidimensional Expressions</td>
</tr>
<tr>
<td>$N_s$</td>
<td>Number of subjects</td>
</tr>
<tr>
<td>$N_v$</td>
<td>Number of variables</td>
</tr>
<tr>
<td>OLAP</td>
<td>Online Analytical Processing</td>
</tr>
<tr>
<td>OLTP</td>
<td>Online Transactional Processing</td>
</tr>
<tr>
<td>S</td>
<td>Subject</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>T</td>
<td>Time</td>
</tr>
<tr>
<td>$V_{st}$</td>
<td>Outcome variable V for subject s at time t</td>
</tr>
<tr>
<td>$W_{pl}$</td>
<td>Width of Person-level dataset</td>
</tr>
<tr>
<td>$W_{pp}$</td>
<td>Width of Person-period dataset</td>
</tr>
</tbody>
</table>
1 Introduction and Motivation

As universities across the globe produce graduates for their countries’ workforce each year, they have taken to gathering data and performing analysis to meet their social accountability mandates. According to State of the Nation 2012 report by Science, Technology and Innovation Council of Canada [1], the rise in credentials offered by Canadian universities has triggered a rise in the number of degrees granted, which went up by 5.4% over the four year period 2006-2010 with almost 70,000 graduating from universities in 2010 alone. With shifting demands from various sectors of employment, a comprehensive analysis is required to study the effects of demographics and socioeconomic factors on the number of degrees granted for various fields of study. The Ministry of Health and Wellness is keen on observing recruitment and retention [2]. As per Statistics Canada [3], around 46% of the population has a university degree of which approximately 81.7% are employed. From this employed population, in New Brunswick, 78.2% and 21.8% work for the service-producing sector and goods-producing sector respectively. Trade along with Health care and Social Assistance are the major industries with 16.5% and 14.75% distribution respectively. While Construction, Manufacturing and Educational service industries have 8.39%, 7.77% and 7.17% respectively, 4.33% of the employed population works in the Professional, Scientific and Technical services industry.
The swelling numbers of graduates and developing technologies have encouraged institutions to develop business intelligence technology that help in making data driven decisions for program evaluation and career tracking systems for professional programs.

1.1 Longitudinal Data

Data collected as a result from the observations of subjects which are measured repeatedly over time is called as longitudinal data [4]. The United States Bureau of Labor Statistics [5] defines longitudinal data as tracking the same sample of information consisting of individuals, households, establishments, and so on at different points in time. Also known as panel data, it can furthermore be defined as an outcome variable $V_{st}$, observed for subjects $s = 1,\ldots,n$ at a time $t = 1,\ldots,j$. The United States Bureau of Labor Statistics further states, “Longitudinal data have a number of advantages over repeated cross-sectional data. Longitudinal data allow for the measurement of within-sample change over time, enable the measurement of the duration of events, and record the timing of various events. For example, suppose the unemployment rate remained high for a long period of time. One can use longitudinal data to see if the same group of individuals stay unemployed over the entire period or if different groups of individuals move in and out of unemployment over the time period.” Longitudinal data is discussed in detail in Chapter 2.

1.2 Student Tracking Systems
Student tracking systems enable universities and colleges to record the progress of their student cohorts in early stages and towards completion of education before entering into the labor market. Data collected and stored using such systems can be used by experts to perform robust statistical analyses which aid the management in valuable data-driven decision making process. Tracking systems designed with longitudinal data can assist decision makers by allowing them to:

- Monitor academic advances and outcomes
- Identify factors influencing student success rates
- Evaluate programs
- Study the socioeconomic effects of certain programs
- Predict the elements prompting career decisions from students

Tracking systems are useful for developing predictive models that identify factors influencing the behavior of a population group. Statistical analysis of these factors can help in achieving higher productivity and benefits by providing an aggressive edge in business from strong data-driven decision making.

An ideal tracking system must consider factors including but not limited to acquisition, cleaning, integration, designing and analysis of data. These will be explained in later chapters.
1.3 LOCATED Project

In 2010, Dalhousie University – Faculty of Medicine, following the footsteps of Memorial University, initiated the LOCATED (Location of Clinician and Trainee Education Dalhousie) project to develop a tracking system for their graduates [6]. In 2013, a user-friendly application was designed to represent and analyze data integrated from various sources; internal, external and national, which is part of this thesis work.

1.4 Objective of the Thesis

The objective of this thesis is to establish an approach for developing longitudinal tracking systems for health professionals graduating from a university. This approach focuses on maximizing the amount of analyzable data by addressing issues related to data quality and data integration for the development of the longitudinal data warehouse.

The tracking system is developed over a longitudinal data warehouse which uses historical data of over 370 students spread across four academic graduation years (2006 to 2009). The data is stored locally on a PostgreSQL object-relational database system and queried using pgAdmin, a rich administration and development platform for PostgreSQL. A Windows Form application is designed and developed in C# using SharpDevelop IDE to perform statistical analysis over the remotely hosted PostgreSQL database server.

1.5 Organization of the Thesis
Chapter 2 provides some background information on longitudinal data, some of its real world applications and the different types of longitudinal data. Chapter 3 describes related work. Chapter 4 illustrates the challenges associated with longitudinal datasets. Chapter 5 describes the data warehouse framework. Chapter 6 explains the longitudinal data warehouse architecture, both typical and adjusted. Chapter 7 gives the overall implementation of the longitudinal data integration warehouse. In Chapter 8, the design of the tracking system is explained with results. The last chapter presents the conclusion and suggestions for future work.
2 Background

In this chapter, the basic concepts and terminology used in this study are described.

2.1 Longitudinal Data

With increased computational power and huge datasets available, researchers around the world are being encouraged to support their findings with strong data-driven statistical analysis. Scientists and researchers have access to sources with rich historical and present data. Appropriate statistical analysis of such datasets for a study involving events that occur over a defined period in time will help researchers identify valuable factors for an efficient system. An example of a longitudinal study can be the study of effects of different combinations in diet and medicines on patients over weeks, months or even years. The benefit of such a study is that it can help analyze datasets from multiple dimensions.

Institutions across the globe are interested in analyzing the factors influencing career decisions among their graduates. The tracking systems are intended to store data of a cohort\(^1\) over an extended and well-defined period in time. Figure 2-1 is a pictorial representation and Table 2-1 illustrates data collection points in a longitudinal student tracking system.

---

\(^1\) In education, cohort is a group of students who are educated at the same period of time.
Figure 2-1: Data collection points in a longitudinal student tracking system

<table>
<thead>
<tr>
<th>Subject id</th>
<th>Event</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>$E_{ts}$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>High school</td>
<td>2003</td>
</tr>
<tr>
<td>2</td>
<td>High school</td>
<td>2005</td>
</tr>
<tr>
<td>1</td>
<td>University</td>
<td>2006</td>
</tr>
<tr>
<td>1</td>
<td>Employment1</td>
<td>2014</td>
</tr>
<tr>
<td>2</td>
<td>University</td>
<td>2007</td>
</tr>
<tr>
<td>1</td>
<td>Employment2</td>
<td>2015</td>
</tr>
</tbody>
</table>

Table 2-1 Data collection points in a longitudinal student tracking system

Longitudinal data sets comprise an event variable $E_{ts}$, where $t$ is an instant in time and $s$ represents the subject whose data is collected. The first row in Table 2-1 represents an
event ‘E = High School’ in the life of subject ‘s = 1’ at time ‘t = 2003’. The second row is for subject 2 at time 2005, the event E being High School.

2.2 Real World Applications for Longitudinal Data

Apart from tracking systems, longitudinal data sets can be used for several other studies such as

   i. Unemployment rates within a province or a country that can be documented to check if the subjects are a group of same or fluctuating individuals.
   ii. Monitoring recovery phases for patients, post ailment.
   iii. Documenting industry inventory data to identify peak seasons for certain sales
   iv. Analyzing health data of population over time
   v. Analyzing crime statistics of regions to study the changing demographics of offenders
   vi. Storing results from surveys to identify changing trends in customer preferences
   vii. Examining employee data to perform demographic analysis of employees over years
   viii. Gathering periodic sales data to predict inventory requirement for peak seasons

2.3 Types of Longitudinal Data
Storing longitudinal data in precise format is extremely crucial. Singer and Willet [7] describe two traditional methods for storing longitudinal data, namely person-level dataset and person-period dataset. Selection of either is completely dependent upon the type of analysis to be done and the kinds of queries that are to be asked from the data. These are discussed in detail in Chapter 5 of this thesis.

### 2.3.1 Person-Level Data

The person-level dataset method for storing data is more traditional, where all of the information for a subject is stored within a single row. Hence, a count of the number of rows in a table will give the number of unique subjects. An example of person-level data can be seen in Table 2-2 where a distinct record is maintained for each person.

Person level datasets can use the subject identifiers as primary key. This method of storing data is wider and shorter i.e., the width of the table is directly proportional to the number of variables being analyzed whereas the length is equal to the number of subjects. The person period data sets can be represented as shown in Equation 2-1 where $W_{pl}$ and $L_{pl}$ are the width and length of ‘Person-level dataset’ respectively. $N_v$ is the number of variables and $N_s$ is the number of subjects.

<table>
<thead>
<tr>
<th>Id</th>
<th>Location_year1</th>
<th>Location_year2</th>
<th>Location_year3</th>
<th>Location_year4</th>
</tr>
</thead>
</table>

9
Table 2-2: Person-level data set showing locations of subjects over 4 years

<table>
<thead>
<tr>
<th></th>
<th>Toronto</th>
<th>Kingston</th>
<th>Kingston</th>
<th>Montreal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Halifax</td>
<td>Dartmouth</td>
<td>Halifax</td>
<td>Calgary</td>
</tr>
<tr>
<td>3</td>
<td>Vancouver</td>
<td>Vancouver</td>
<td>Vancouver</td>
<td>Kitchener</td>
</tr>
<tr>
<td>4</td>
<td>Ottawa</td>
<td>Bathurst</td>
<td>London</td>
<td>Ottawa</td>
</tr>
<tr>
<td>5</td>
<td>Ottawa</td>
<td>Ottawa</td>
<td>Ottawa</td>
<td>Ottawa</td>
</tr>
<tr>
<td>6</td>
<td>Saint John</td>
<td>Fredericton</td>
<td>Moncton</td>
<td>Moncton</td>
</tr>
</tbody>
</table>

\[
W_{pl} \propto N_v \\
L_{pl} = N_s
\]

Equation 2-1 : Person level data set relations

2.3.2 Person-Period Data

The person-period data sets contain a discrete row for each variable to be analyzed. This results in longer tables for person-period data sets even with fewer subjects. As can be observed from Table 2-3, the variable ‘id’ is used for identifying each subject, the variable ‘t’ is used to identify the time of the event and L the location. Thus, person-period datasets require a composite primary key, comprising the identifier id and time variable ‘t’.
Table 2-3: Person-period data set showing locations of subjects over 4 years

\[
W_{pp} = N_v \\
L_{pp} \propto N_p
\]

Equation 2-2 Person-period data set relations

Consequently, the person-period datasets are longer and narrower i.e., with more rows but fewer columns. The width of the table is equal to the number of variables whereas the length is directly proportional to the number of subjects in the study. It can be represented using the equations below where \( W_{pp}, L_{pp} \) are the width and lengths of ‘Person-period data sets’ respectively, \( N_v \) is the number of variables and \( N_p \) is the number of subjects.
Structuring of the longitudinal dataset must be initiated by understanding the aspect of data to be studied. Change in subject behavior can be assessed and analyzed best over a ‘Person-level data set’. However the effect of time on subject behavior can best be assessed and analyzed over a ‘Person-period dataset’. For the purpose of this thesis, which requires analyzing the effect of time on population behavior, ‘Person-period dataset’ is best suited. According to Singer and Willet [7], “All person-period datasets contain four types of variables: (1) a subject identifier; (2) a time indicator; (3) outcome variable(s); and (4) predictor variable(s).” These variables are included in the dataset for this study.

### 2.4 Retrospective Cohort Study

As mentioned earlier in this chapter, longitudinal study requires storing and analyzing data from periods in time which are both historical and present. These data sets are mostly created by compiling together a group of subjects to create a cohort. For example university graduates can be divided into cohorts based on the year they graduated. The analysis of longitudinal data that studies a cohort (cohort study), looking back in time (retrospective study), can be termed as *retrospective cohort study*; where the historical data of a defined cohort is gathered from various sources and integrated for an in-depth analysis. Statistical analysis of retrospective cohorts can help develop predictive models for prospective cohorts.
As part of this study, research was conducted to identify suitable longitudinal databases and based on the inferences listed above, the database was designed using the PostgreSQL database management system. The reasons for choosing PostgreSQL are as follows [8]:

- PostgreSQL’s reputation for being the world’s most advanced open source database and its free licensing.
- Supports for multi-version concurrency control which helps with operations such as re-indexing, adding or dropping columns and recreating views.
- Support for referential integrity which allows for proper foreign key management.
- Support for procedures and triggers
- Support for user authentication resulting in proper data governance
3 Related work

Mazza and Dimitrova [9] use a course management system to generate web log data to graphically render complex multidimensional student tracking data. In their paper, importance is given largely to the design guidelines for their visualization tool. Since the data used for visualization is extracted from a well-designed database, data integration and data cleaning are overlooked altogether. Haertel [10] in his work at Stanford University for the California State Department of Education illustrates designs for using test data in school accountability systems. Once again, the tracking system is developed over a pre-defined database with no explanation about the database design. Four universities in Florida led by Florida State University [11] have developed a student tracking system that uses internal administrative data to track student progress and manage enrollment data. These tracking systems are used with only internal administrative data. Once again, the challenge of designing, developing and maintaining longitudinal data sets is not addressed.

Koys, in his unit-level longitudinal study [12], used longitudinal data collected from a restaurant chain via employee, manager, customer surveys and organizational records to perform cross-lagged regression analysis to show a relationship between organizational effectiveness and employee attitudes and behaviors. Harackiewicz et al. [13] in their longitudinal study for predicting success in college examined the role of achievement goals, ability and high school performance in predicting academic success over students’ college careers. The authors collected student data until graduation to examine continued interest in psychology and performance in subsequent classes. The National Center for
Analysis of Longitudinal Data in Education Research (CALDER) located at Washington D.C., USA, is a joint effort of the American Institutes for Research and scholars at several universities [14]. CALDER uses state and district individual level longitudinal data on students and teachers to answer questions such as “How stable is teacher performance over time? How stable is teacher performance over settings, e.g., high/low poverty schools? What are the teacher mobility patterns over time and across school settings? Do the patterns differ for high performing and low performing teachers? How does student mobility affect student performance? What are the effects of selection and training on creating a productive teacher workforce? What are the pros and cons of different ways to evaluate teacher performance?”

On a more similar related front to this study, Memorial University in 2008 [15] started working on a ‘Learners and Locations Project’ to examine medical students’ geographical backgrounds, educational placements, and eventual practice locations, and allows for analysis of links among backgrounds and practice locations. Although developed with a similar objective, this project overlooks some of the major challenges that are encountered during development of a longitudinal data set such as loss of information due to missing or incomplete data, a deterioration of data over time. This study had challenges towards data integration due to variations in standards among internal and external data sources. Moreover, the project did not have a front-end tool for representation of analyzed data and hence it required that a technically skilled person be assigned every time to conduct analysis.
Data in blocks is like pieces of a huge puzzle that merely exist to give a partial image of what upon integration will be an analytical masterpiece. The integrated blocks of data form a complete picture that presents the analysts with an opportunity to assess the integrated data for predicting changes or even identifying trends.
4 Challenges with Longitudinal Datasets

While collecting and integrating longitudinal student data, several challenges were faced. Following is a list of these challenges and their brief explanations. Solutions towards these challenges are discussed in Chapter 7.

4.1 Lack of Administrative Policies

Even with the enormous advances in technology and ease of availability for specialized software, personal experience suggests that many educational institutions do not have a designated database to store and maintain administrative data especially of their graduates. It may so happen that the details get stored in simple Excel files by temporary administrative staff resulting in a lot of data either going missing or not getting entered at all. Thus working on such data for the purpose of performing analysis can become tedious for analysts. Due to lack of regulations and administrative policies, the format of saved data on systems changes every year thereby causing problems for integrating historical data with present data. In some cases, the information required may be have been recorded on scratch papers or even worse, sketched in the memories of workers without any documentation whatsoever. Consequently a simple task of collecting data which on some project proposals would not be listed as taking more than a week, ends up taking several weeks or months. There is still no assurance of getting quality data due to lack of standards.

4.2 Data Standardization Challenge
While there are several “best practices” about how to store data, there are no accepted standards for data input. Every data entry specialist and organization follow data input standards that best suit their needs and are convenient for them. Therefore, the data input standards for no two organizations are alike thereby making data integration extremely difficult and laborious.

Following are some examples:

- Organizations save the names of our provinces differently, such as ‘New Brunswick’, ‘NB’, ‘N.B.’ etc.
- Some may follow language accents while others may prefer not to; ‘Montreal’ or ‘Montréal’
- Some may observe abbreviations while others may not; ‘Saint John’ or ‘St. John’, ‘M’ or ‘Male’

### 4.3 Data Quality Challenges with Inconsistent Data

Longitudinal data sets house very large collections of data. It would be naïve to expect the data to be free from empty fields with missing information or typing inaccuracies. Many applicants tend to leave blank spaces in administrative forms that do not list all fields as mandatory. This results in inconsistent data leading to inaccurate predictions and analysis. People also tend to change their names after being married which will result in erratic
integration over years. Table 4-1 shows an example of inconsistent data with missing values for high school resulting from students leaving high school information blank since it was not a mandatory field on the form.

<table>
<thead>
<tr>
<th>Subject_id</th>
<th>High_school</th>
<th>High_school_city</th>
<th>High_school_province</th>
</tr>
</thead>
<tbody>
<tr>
<td>235</td>
<td>Halifax High School</td>
<td>Halifax</td>
<td>NS</td>
</tr>
<tr>
<td>236</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>237</td>
<td>Harbour View High School</td>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>238</td>
<td>Queen Elizabeth High School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>239</td>
<td>St. Francis High School</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4-1 Data collected from entry source with high school field not mandatory

4.4 Linking Static and Dynamic Data

Over the years, certain properties of data may change in dynamic data while others may remain constant. A person’s date and place of birth, locations of high school and universities attended are static data. On the other hand, the rural or urban categorization of any city is bound to change over time based on its population and other dynamics. For example, a subject ‘S’ attending high school at location ‘L’ is static data, whereas the ‘Rural or Urban’ categorization of ‘L’ is dynamic data. Thus the ‘Rural or Urban’ property of ‘L’ is completely dependent upon the time of Event ‘E’ i.e., ‘S’ attending high school.
It would be a flawed approach to assign a static value to a property that is dynamic in nature.

4.5 Technology Variance

The most important property of a longitudinal data set is the wide range of time periods that the data is saved for. Changes or upgrades in technology are a part of modern development. Thus, historical data which was saved on an application, service or platform that has undergone substantial updates or has been discontinued by the developers, is under severe threat of being lost forever. Thus it is of utmost importance to keep abreast with changes or updates in technology and industrial trends.

4.6 Waiting For Data to Mature

Longitudinal data sets that have been recently developed may be immature and hence inadequate for developing predictive models. These data sets need to mature enough for reliable and accurate predictions and trends to be derived from them. Results obtained from analysis on immature data sets may be of little or no use and misleading.
4.7 Confidentiality of Data

Tracking systems are home to extremely personal and sensitive data whose confidentiality should be preserved at all costs. The system at no point should reveal personal information that may be used to locate an individual or to obtain personal information about its subjects. Therefore the data warehouse should be hosted from a secure remote server that logs user accesses to help in data governance.
5 Data Warehouse

The purpose of this study is to establish an approach for longitudinal data integration for developing tracking systems for health care professionals graduating from a university. This is implemented with a set of students spread over multiple graduating years.

A data warehouse is different from a traditional relational database in the sense that it is designed to facilitate query and analysis rather than transactional operations. Mostly, a data warehouse is designed to contain historical data from an OLTP (On Line Transaction Processing) system but it may also contain data from various operational and external systems [16].

The purpose of this chapter is to describe the methodology, explain the sample sets, describe the procedure used for collecting and integrating data and provide explanation of procedures used to develop this tracking system.

5.1 Data Warehouse framework

A data warehouse needs to be updated to add new pieces of information as they are generated. Similarly, the data warehouse for a graduating student tracking system is updated annually to accommodate new information as it is made available. Therefore, a
cyclic approach is adapted for developing this tracking system. Figure 5-1 is a diagrammatic representation of the methodology used.

Note: This research does not list user requirements since it is out of the scope of this study and we assume that the user requirements were gathered and listed as needed.

Creation of a data warehouse typically requires extraction, transformation and loading (ETL) of data into the warehouse.

*Extraction* involves identifying the data sources and extracting data, which is nothing but data acquisition

*Transformation* involves data cleaning, integration and categorization

*Loading* involves moving the structured data to a data warehouse

### 5.2 Data Acquisition

Selection of sources for data depends upon various factors, including but not limited to, the hypothesis or the questions being asked from the data, the kind of data that is available for analysis and its respective sources. Based upon the source of origin, data can be classified as internal or external.
**Internal data:** Any data that can be obtained from sources internal to the organization is classified as internal data. The internal data is relatively rich in information and since it is from within the organization, it is most likely to follow standards same as the project being developed. For example, data pertaining to the background information of students such as their date of birth, place of permanent residence, high school and prior universities attended etc. can be obtained from the administrative department.
**External data:** Any piece of data that is not available within the organization and is attained from sources outside the organization is classified as external data. Unfortunately, the external data may have very limited information since the sources sharing data are bound by legal contracts that prohibit them from sharing fine details. Each of these external data sources follow standards that are deemed fit for their organization but that may not necessarily be an industrial best practice standard. This creates a conflict of data standards between datasets. For this study, the practice locations of graduates was requested from Canadian Post M.D. Education Registry (CAPER) – a central repository for information on postgraduate medical education in Canada [17] and Scott’s Directories (Canadian Medical Directories) [18]. CAPER gives data on graduates in the form of access database files whereas Scott’s Directories gives data in Comma Separated Values (CSV) files.

### 5.3 Data Cleaning

Before integration, it is essential to get rid of any anomalies in data, both external and internal. Hypothetically, external data is more prone to anomalies since it may be extracted from sources that are most likely following different standards. These anomalies may include spelling errors, deviation from standards, missing information, repeated information, etc. In the case of this project, these anomalies were corrected manually within spreadsheets using Microsoft Excel.
The first step towards cleaning anomalies is to correct any spelling errors in the data. This includes usage of abbreviations or short forms to represent information. Consistency should be maintained throughout in text spellings and usage of abbreviations if any. Table 5-1 shows city and province data where the values for province are entered inconsistently. Table 5-2 shows the city and province data which has been cleared of any inconsistencies and anomalies.

<table>
<thead>
<tr>
<th>City</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint John</td>
<td>New Brunswick</td>
</tr>
<tr>
<td>Fredericton</td>
<td>N.B.</td>
</tr>
<tr>
<td>Toronto</td>
<td>Ont.</td>
</tr>
<tr>
<td>Sussex</td>
<td>NB</td>
</tr>
<tr>
<td>Ottawa</td>
<td>Ontario</td>
</tr>
</tbody>
</table>

Table 5-1: Inconsistent data with anomalies

<table>
<thead>
<tr>
<th>City</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>Fredericton</td>
<td>NB</td>
</tr>
<tr>
<td>Toronto</td>
<td>ON</td>
</tr>
<tr>
<td>Sussex</td>
<td>NB</td>
</tr>
<tr>
<td>Ottawa</td>
<td>ON</td>
</tr>
</tbody>
</table>

Table 5-2: Data cleaned of inconsistencies and anomalies
Any piece of data that is deviating from the standards laid down for the data warehouse should be dealt with now. These may include representation of provincial names as ‘NB’ or ‘N.B.’ or ‘New Brunswick’ etc., usage of accented letters and other characters such as ‘Montréal’ or ‘Cap-aux-Meules’, usage of abbreviations or short forms such as ‘M’ or ‘F’ for male and female respectively and any other anomalies that can be found on the data.

Any piece of missing information that cannot be collected due to several reasons, such as failure to reach the subject, subject refusing to share information, fields that were not mandatory at the time of filling forms, errors during manual entry etc. is classified as missing information. Luengo [19] lists the following three problems associated with missing data:

- loss of efficiency
- complications in handling and analyzing the data
- bias resulting from differences between missing and complete data

Higgins and Green [20] mention 4 principal options for dealing with missing data:

1. analyzing only the available data (i.e. ignoring the missing data)
2. imputing missing data with replacement values, and operating as if these values were observed
3. imputing the missing data and accounting for the fact that these were imputed with uncertainty
4. using statistical models to allow for missing data, making assumptions about their relationships with the available data.
They further explain that option 1 is appropriate for cases when data is missing at random. Options 2 and 4 are more appropriate when data is not missing at random, option 2 being more practical. Options 3 and 4 require the knowledge of an experienced statistician.

Since the dataset used for this research is immature, missing data is dealt with using the first choice listed above since using mean or median values on a small dataset would have led to inappropriate results.

Another major part of data cleaning is to verify the authenticity information obtained from external sources. It is essential to make sure that the data shared by external source is authentic. This can be done by closely reading the policies and procedures followed by the external source to understand if their method of data acquisition is acceptable.

5.4 Data Integration

Integrating multiple internal data files is easier if done using a proper unique identifier. However, integrating internal data with external data can be a daunting task, considering that external sources are highly unlikely to have the same identifiers for individuals as those of internal sources.
During the study, it was especially difficult to integrate internal administrative data with the external data. The most reliable way of integrating internal data with external data would be to integrate the subjects using their first and last names; date of births could not be used since external data sources did not store/share the subjects’ date of births. To integrate internal administrative data with external data we used the subjects’ first and last names along with their years of graduation. Although quite common, it is often overlooked that sometimes subjects end up changing their names for various reasons, the most common being, changing last names after getting married. One such instance is shown in Table 5-3 and Table 5-4 where two subjects with same first name and having graduated in the same academic year, had different last names the following year thereby creating uncertainties in their identity. This resulted in the loss of ability to track their practice locations. Although the probability of such a thing happening is relatively low, it cannot be ruled out and certainly cannot be overlooked since elimination of uncertain subjects from the study group affects its population size.

<table>
<thead>
<tr>
<th>First_Name</th>
<th>Last_Name</th>
<th>Graduation_Year</th>
<th>Practice_City</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>Arthur</td>
<td>2008</td>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>Mary</td>
<td>Wayne</td>
<td>2008</td>
<td>Sussex</td>
<td>NB</td>
</tr>
</tbody>
</table>
Table 5-3 Example of Data collected in 2013

<table>
<thead>
<tr>
<th>First_Name</th>
<th>Last_Name</th>
<th>Graduation_Year</th>
<th>Practice_City</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>Renner</td>
<td>2008</td>
<td>Halifax</td>
<td>NS</td>
</tr>
<tr>
<td>Mary</td>
<td>Shepherd</td>
<td>2008</td>
<td>Fredericton</td>
<td>NS</td>
</tr>
</tbody>
</table>

Table 5-4 Example of Data collected in 2014

5.5 Data Structuring

Once data is integrated, the next important step is to model data into tables by structuring it. Data can be categorized into Person-Level or Person-Period as mentioned in section 2.3.

In a longitudinal study, the research question could be subject centric or time centric. Subject centric study questions are easier to query over a table that has been structured as a Person-level Dataset while time or event centric questions are easier to be queried over a Person-period Dataset structured table. Subject centric study questions focus upon the behavioral transitions of individuals over time from subject point of view whereas time acts as a focal point in a time centric study.

---

2 Names used in the examples are fictitious and have no relevance to any real person.

3 Names used in the examples are fictitious and have no relevance to any real person.
Once the data has been extracted and transformed, it is then ready to be loaded onto the data warehouse. Chapter 6 has detailed explanation about the architecture and implementation of the data warehouse.

5.6 Analysis/Representation

Analysis of data is done through a specially designed desktop application as shown in Figure 5-2 in the C# programming language. This application establishes a connection with PostgreSQL and queries the data to mine relevant information and then visually represents it in the form of reports and charts. A detailed explanation of this application is given in Section 8.2.
Figure 5-2: Screenshot of the LOCATED Application
5.7 Annual Updates

The data warehouse needs to be constantly updated to include new data for analysis. New student data needs to be added annually once it is available from both internal as well as external sources such as CAPER and Scott’s Directories. The update of the data warehouse is instigated by verifying whether the research question and data sources have changed or not. If the research question is the same, updates can be carried out with data acquisition. In case the research question or the data sources have changed, defining the research question and identifying data sources precedes data acquisition. It is highly recommended to wait for the data to be available from all sources before initiating the update.
6 Longitudinal Data warehouse Architecture

A data warehouse can be constructed by performing an ETL process on existing databases and data files. The ETL process is mostly performed over a database management system, in our case, PostgreSQL. The typical architecture of a data warehouse is explained and discussed in the next section.

6.1 Typical Data Warehouse Architecture

A typical data warehouse architecture design comprises several tiers as shown in Figure 6-1. Data collected from multiple sources is loaded on the back-end tier with ETL tools where a database management system is used to perform cleaning, transformation and integration of data. The data warehouse tier comprises the integrated data and its metadata. The OLAP (On Line Analytical Processing) tier houses the OLAP server which stores analytical data in the form of cubes. Finally, the front-end tier is used for data analysis and visualization such as front-end application, reporting tools, statistical tools or data mining tools.
A summary for typical data warehouse architecture is essential since the model adopted for the graduating student tracking system is an adapted version of a typical data warehouse. The data warehouse architecture used for developing the tracking system is discussed in detail in Section 6.5.

### 6.2 Dimensional Systems

Dimensional systems are used in cases that require multi-dimensional analysis of data. This is achieved by presentation of data in the form of hypercube or multidimensional arrays with values stored in cells accessible by multiple indexes. Figure 6-2 shows data stored in a three-dimensional system with the dimensions name, location and event.
Dimensional systems are usually built over relational database management systems and require queries to be written in Multidimensional Expressions (MDX) over them.

Dimensional systems can comprise one or more fact tables with measures surrounded by dimensional tables without measures. Measures are pre-calculated aggregations across dimensions.

![Diagram of a 3-dimensional cube in a dimensional system](image)

**Figure 6-2: A 3-dimensional cube in a dimensional system**

6.2.1 **Advantages and Limitations of a Dimensional System**

Data stored in dimensional systems has an intuitive spreadsheet like view which is easier to understand. When compared to relational systems, dimensional systems are easier to maintain since the data is stored in the same way as it is viewed. Unlike relational systems,
dimensional systems are more accommodating to analytical queries and there is no additional computational overhead required for queries compared to the complex indexing and joins in relational systems.

However, the biggest drawback for a dimensional system is that the dimensions of analysis must be known before designing the system. This is challenging for longitudinal-dataset development where the dimensions of analysis may vary heavily during the course of the project. The dimensional system design will have to be altered heavily every time someone asks for an added or altered dimension to be analyzed.

6.3 Dimensional or Relational Data Model - Factors to Consider

Both relational and dimensional systems have their sets of advantages and limitations. However, the choice of a relational or dimensional system for housing a longitudinal dataset is purely case specific. In my opinion, some of the most important questions to be answered before finalizing a system are listed below:

6.3.1 What is the Depth of Analysis?

For university graduates, longitudinal datasets traditionally involve analysis of data over a few dimensions such as age, gender, location or life-events such as career-choices made. Since the number of dimensions required for analysis is relatively small, relational systems
are advantageous since they work well in cases requiring analysis over a few tables with less complex joins. In cases where the number of dimensions is very high, dimensional systems are a good option.

6.3.2 How important is Performance?

Since complex analytical queries involving multiple joins are very time consuming, it is important to gauge the importance of performance required from the system. Longitudinal data-sets are seldom used for real-time analysis of data and are almost never used to analyze data on fly. Longitudinal datasets are analyzed once in a while for creating reports and summaries. Thus it can be safe to say that the query execution time spent on relational systems will be of no or very little significance. However, writing complex queries requires personnel with expertise in structured query languages.

6.3.3 How dynamic is the Data?

Another important factor to assess is to understand how dynamic the data is: how frequently does the information in it change? The answer for longitudinal data-sets is ‘not so frequently’. Longitudinal data-sets are updated yearly, unlike data marts storing sales information for a retail outlet. Since these updates are so less frequent, a relational system is well equipped to handle them by itself.
6.3.4 How dense is the Data?

Dimensional systems are not normalized and therefore contain a lot of duplicated data. In addition, a database with empty cells results in a bulky dataset. For example in a longitudinal dataset that has locations and persons as dimensions, remote places with very little population are bound to have empty cells in their cube. This will result in a sparsely populated dataset consuming more space than required. Relational systems, on the other hand being highly normalized reduce the amount of duplicated data and store locations as tuples and not dimensions.

6.3.5 Who are the Users?

As explained earlier, relational systems require personnel with expert knowledge of structured query languages to write complex queries that involve multiple joins. Analysis done over longitudinal dataset is highly valuable for human resources and management who may not be equipped with skills to write such queries. Multidimensional systems are highly advantageous in this regard since many business intelligence tools available require simple drag-drop mechanisms. However a well-designed front-end analysis tool over a relational system can equally serve the purpose.

The factors to consider for selecting relational or dimensional data model for longitudinal data can be summarized as shown in
6.4 Selection of Data Model

The selection of relational or multidimensional data model systems for longitudinal dataset analysis is purely case (project) specific and requires in-depth analysis of project requirements before being implemented. Although multidimensional systems offer more flexibility in terms of the dimensions to be analyzed, it should be understood that multidimensional database technology is a complementary technology to relational database technology and not a direct replacement. A longitudinal dataset is historically rich and has a potential to extend for a very long period of time. Due to the time flexibility of such a project, it is bound to have changing research questions resulting in changing dimensions (not be confused with slowly changing dimensions) which may involve adding or dropping of an entire dimension. This may be hard to implement on a dimensional system which has pre-defined dimensions. A relational system with a well-designed user-interface on the other hand is far more flexible and can be equally beneficial if not more.

6.5 Adjusted Design

Since a dimensional system is less flexible in dealing with changing dimensions, for this research, we have dropped the OLAP tier from the typical data warehouse architecture to give way to an adjusted design without the OLAP tier. For the LOCATED application that was introduced in Section 1.3, the data warehouse tier design stored data in a relational
model instead of a dimensional model to provide more flexibility. Figure 6-3 shows the adjusted data warehouse design that has been adapted for this research.

Data files gathered from multiple sources need to be cleaned, normalized and standardized before transferring them to a database management system to perform integration. As mentioned earlier, PostgreSQL is used to host the data warehouse. The challenges encountered in bringing together data from multiple sources are discussed in detail in Section 7.1.4.
7 Implementing a Longitudinal Data Warehouse

The purpose of this chapter is to explain in detail the ETL process involved in the development of the longitudinal data warehouse for the health care professional’s tracking system. The extraction process involves identification of research question and the data sources followed by data acquisition. The transformation process involves data cleaning, data integration and data structuring. Lastly, the structured data is loaded onto the data warehouse during the loading process. This chapter is divided into sub topics to address each component elaborately.

7.1 Extraction

The first process in the implementation of the longitudinal data warehouse is extraction. The extraction process can be broken down to three steps, which are defining the research question, identifying data sources and data acquisition.

7.1.1 Identifying the Research Question

Once the data is stored on the data warehouse, the type of analysis that can be performed on it is inexhaustible. Thus, for the success of any research analysis, it is imperative that the research question be identified. This helps in recognizing the data that will answer the research question. The research question may require either subject-centric or time-centric
analysis of data. For the LOCATED project [6], the research objective is to perform program evaluation to “verify if students are practicing in the nearby provinces of both main and regional campuses (the Maritimes) and/or in rural settlements.” In this study and for the LOCATED project [21], student data from the graduating classes of 2006 to 2009 and 2014 is used to develop the tracking system.

7.1.1.1 Subject-centric Analysis

For subject-centric analysis, the focal point of research is a subject or a group of subjects. Analysis questions that focus on the behavioral changes of any individual or a cohort over an extended period of time can be deemed as subject-centric. An example of this analysis would be to evaluate the performance of a cohort belonging to one or more graduation years, where the focus of study is the subjects within the cohort. For example, consider the SQL query to find the average age at the time of admission for students. Age_at_admission is a pre-calculated field in the administrative_data table.

SQL:

```
SELECT AVG(age_at_admission) FROM administrative_data;
```

7.1.1.2 Time-centric Analysis

Time-centric analysis of data focuses on the time of the event. The behavior of any subject or cohort will be analyzed based upon time stamps or events that may have occurred. For
example, consider calculating the average age at the time of admission for students from graduating years 2007. Here the cohort of 2007 graduating year is being analyzed:

**SQL:**

```
SELECT AVG(age_at_admission) FROM administrative_data WHERE graduation_year = 2007;
```

7.1.2 Identifying Data Sources

Once the research question is identified, the next step would be to identify the data that is needed for analysis. This data can be either administrative, demographic, academic, financial, human resources etc., and may be stored at one or several departments within or outside the organization.

7.1.2.1 Internal Sources of Data

Theoretically, internal data is meant to be easier to deal with since it supposedly follows the organizational standards for data storage. However, personal experience suggests that this may be far from true, since most of the data even in large organizations may sit on spreadsheets being created from various systems of entry which include application generated data or manually entered data. Data entered through applications and interfaces is highly likely to be normalized and structured since it resides on a Database Management System (DBMS) which is designed by a professional database administrator. Conversely,
data entered manually is prone to being inconsistent and dirty as was shown in section 4.3. A proposed solution towards handling missing information during the LOCATED project was to suggest the administrative department to implement strict data entry policies at the time of data collection by making all fields mandatory.

7.1.2.2 External sources of data

Some of the data required for analysis may not be available within the organization. For example, population related analysis may require census data to be gathered from Statistics Canada or the Canadian Medical Directory (Scott’s Directories) and CAPER can be requested for practice locations of physicians. Acquiring data from these external sources can be a tedious task considering they are governed by a completely different authority and are bound by restrictions on the granularity of data they can share. These external sources also need funds for purchase and it is time consuming with a lot of paperwork and legal formalities to be completed before access is gained to the data. Although external data is structured and designed professionally, the standards and regulations followed by the external source may not comply with the internal organizational standards and thus the external data may require some tweaking before being integration ready. Section 7.1.3 throws some more light on the challenges associated with data acquisition.

7.1.3 Data Acquisition
Once the data sources have been identified, the next stage is acquiring data. Data acquisition can be time consuming since sloppy data acquisition can lead to troubles with the law. Primarily, the sensitivity of data being used for research analysis needs to be understood and considered throughout the study. Secondarily, governance rules regarding the usage of external data should be made clear and a consensus should be reached regarding its implementation for the research. At this juncture, it is vital to mention the time spent on waiting for authorizations for data access, which lead to loss of productive time. Planning ahead is required to access external data. For instance, in our implementation, approximately 11% of the time i.e., 6 out of 42 weeks was spent waiting for authorizations to access data from various departments and authorities within the organization as can be seen from Figure 7-1.

![Figure 7-1: Chart showing time spent in various stages during this project lifecycle](image)
7.1.3.1 Governance Policies

Data governance is a crucial aspect of any project involving data analysis. Having appropriate governance placed ensures smooth flowing of the project without any unnecessary abruptions. As mentioned earlier, a huge chunk of time spent on the project can be wasted waiting for authorizations for access to pieces of data that can form a core of analysis. Thus, as soon as the data sources have been identified, proper governance policies should be laid with higher authorities having privilege to grant access to data. Proper governance laws laid down at the start of the project will ensure its smooth operation when the data needs to be updated to the warehouse.

7.1.4 Challenges of Extraction from Different Sources

External data, once acquired, needs to be transformed to suit the organizational standards and make it research ready. However, once data is obtained from the external source, the transformation is challenging and is faced by many difficulties as discussed below.

7.1.4.1 Lack of global identifiers

The unique identifiers assigned to subjects on internal data are mostly locally assigned variables that help identify individuals within the organization. Social insurance numbers can help identify individuals across a country but are an extremely sensitive piece of
information and thus not advisable. Lack of global identifiers across internal and external organizations make integration of datasets extremely tedious since it requires matching subjects on the names and possibly birthdates of subjects. During the research, integration of internal and external data is achieved by matching first and last names of subjects and their years of graduation; birth dates could not be used since the external dataset lacked this information.

An interesting situation encountered while trying to integrate data using the aforementioned values is, subjects changing their names mostly their last names after getting married as discussed in Section 5.4. This was not perceived as a major concern initially considering the relatively smaller size of the dataset (379) and the even smaller number of subjects who would change their names (9); it was decided to match these subjects manually if the query failed. However, during one particular instance, two subjects with the same first name and graduation year changed their last names and it was impossible to link the subjects even manually. Since it is impossible to link such data, the research dataset eventually degenerates resulting in fewer subjects over time.

7.1.4.2 Varying standards

The standards for storing data vary across organizations. The format of storing information differs between internal and external sources as shown in Table 7-1, Table 7-2 and Table 7-3.
### Internal Data

<table>
<thead>
<tr>
<th>city</th>
<th>province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>Montréal</td>
<td>QC</td>
</tr>
</tbody>
</table>

Table 7-1: Table from internal data showing city and province

### External Data

<table>
<thead>
<tr>
<th>city</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint John</td>
<td>N.B.</td>
</tr>
<tr>
<td>Montreal</td>
<td>Q.C.</td>
</tr>
</tbody>
</table>

Table 7-2 Table from external source showing city and province

### Statistics Canada

<table>
<thead>
<tr>
<th>city</th>
<th>province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint John</td>
<td>New Brunswick</td>
</tr>
<tr>
<td>Montréal</td>
<td>Quebec</td>
</tr>
</tbody>
</table>

Table 7-3: Table from Census Canada showing city and province

As can be seen, internally province names are stored as NB, QC etc., while external source stores it as N.B., Q.C., etc. and Statistics Canada as New Brunswick, Quebec, etc. There is also a discrepancy with regards to usage of accents in French names such as Montréal and
Montreal. These discrepancies across datasets make SQL query joining challenging. A solution in regards to this is achieved and is discussed in detail in Section 7.2.2.

7.1.4.3 Technological variance

While collecting historical data for analysis, data created on earlier versions of Excel when opened with newer versions of Excel corrupted the file by adding undesirable characters in cells. For example, during our implementation, when historical data was requested, the files received were created on Excel 2003 which when opened with Excel 2010 would add undesirable characters throughout the file. This is caused due to the difference in encoding formats between the Excel versions. However, this was not an irreparable damage since Excel 2010 had ways of recovering files created in earlier versions of Excel. Since longitudinal data spreads across time, developing technology and newer versions of software will always be a concern. This should be tackled by continuously updating the technology and saving historical data on newer versions of software as and when a new version is made available.

7.1.4.4 Questioning the veracity of data from external source

There may be limited information concerning the origin of information on external sources. Therefore the veracity of data acquired from external sources must be verified by discussing with the concerned authority and understanding their procedure for collecting
data. For example, CAPER releases physician practice locations every 2, 5, and 10 years after the end of their residency and the Canadian Medical Directory which gathers weekly data updates releases physician practice locations annually upon request [6].

7.2 Transformation

The objective of transformation is to alter the format and structure of the acquired data to suit the research demands and organization standards. This transformation is subject to project standards and research objective. Transformation of data can be achieved through data cleaning, integration and structuring.

7.2.1 Data Cleaning

It is considered an industry practice to assign unique identifiers for all subjects in a dataset to help identify each record. The first approach towards data cleaning should be to assign unique identifiers after eliminating duplicates. This assigned unique identifiers can either be the ones already in use within the organization or an especially assigned surrogate key specific for the project. Since universities assign student id numbers to all the students, it is highly recommended to use the student id number as a unique identifier. However, due to the sensitive nature of student ids, surrogate keys were assigned and used for the LOCATED project. Any acquired dataset will have information that may or may not be useful from the research analysis point of view. Any attributes within data that are
irrelevant to the project should be eliminated to decrease the space used by tables in the database.

Manually entered data is prone to human error and is extremely corrupt and dirty due to incomplete or missing information, typing errors, duplicate information, and inappropriate use of acronyms. During the analysis of a dataset that had student information collected at the time of enrollment, an attribute say ‘Information-A’, which is not listed as mandatory for filling, had a lot of blank fields (missing information). Figure 7-2 is a pie chart representing the dirty data (missing and incomplete information) from our project and it has alarming numbers: a field which is identified as an important attribute for answering the research question has only 47% clean data fit for analysis. From the 53% dirty data, 50% is missing information and 3% is incomplete or obscure information. In such cases, 47% of clean data inevitably means 47% analysis reliability for the results.

However, within the same dataset, when analyzing another attribute say ‘Information-B’, which is also not listed as mandatory on the form, the amount of missing data is relatively less. Figure 7-3 shows a pie chart with dirty and clean data for attribute ‘Information-B’ where 76% clean and 20% missing data and 4% erroneous data i.e., misspelled data is observed. This suggests that the ratio of clean to dirty data within any dataset varies hugely and is also a major factor for identifying the reliability of any dataset.
Figure 7-2 Comparison of dirty and clean data for ‘Information-A’ within a dataset

Figure 7-3 Comparison of dirty and clean data for ‘Information-B’ within a dataset
Data entered manually on spreadsheets is prone to human errors such as typographical errors, data duplicated due to ignorance, inconsistent data entry formats etc. Typographical errors can be troublesome when the integration of datasets is heavily dependent upon text matching. In the same data set mentioned earlier, resolving typing errors increased the data matching efficiency by 13.72% (52 entries from 379). It is also observed that approximately 4.75% of data on the spreadsheet is duplicate data as a result of manual entry errors. These typographical errors need to be verified and changed.

It is worth mentioning here that if the proposed longitudinal dataset is being developed to support accent characters, then the encoding standards across all software and the database management system should be UTF-8. Excel by default uses ANSI encoding which will replace any accent characters with undesirable symbols.

A proposed solution made in the LOCATED project to reduce the cost of data cleaning in future is to replace manual forms with electronic forms that contain combo box fields to avoid misspellings and deviations from standards.

7.2.2 Data Integration

Data cleaning is followed by data integration. Internal datasets can be integrated using organizational identifiers such as student id across tables. The challenge however is to integrate internal and external data in the absence of common identifiers. As mentioned
earlier, in the absence of common identifiers between internal and external data, the table joining gets heavily dependent upon textual matching in SQL query on first and last names and for this research also on the year of graduation. Typographical errors are a major concern with regards to textual matching.

Table 7-4 and Table 7-5 show data about same set of students from two different source with typographical errors. To the human eye, Edward and Ed or Phillip and Phil or Randell J. and Randell may be the same person but an SQL query performing a text join on the data will not return accurate results.

<table>
<thead>
<tr>
<th>first_Name</th>
<th>last_Name</th>
<th>City</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edward</td>
<td>Thomas</td>
<td>Montréal</td>
<td>QC</td>
</tr>
<tr>
<td>Phillip</td>
<td>Turner</td>
<td>Montréal</td>
<td>QC</td>
</tr>
<tr>
<td>Randell J</td>
<td>Renner</td>
<td>Toronto</td>
<td>ON</td>
</tr>
</tbody>
</table>

*Table 7-4: A table with student data from ‘Source A’*

<table>
<thead>
<tr>
<th>first_Name</th>
<th>last_Name</th>
<th>City</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ed</td>
<td>Thomas</td>
<td>Montreal</td>
<td>O.N.</td>
</tr>
<tr>
<td>Phil</td>
<td>Turner</td>
<td>Kingston</td>
<td>O.N.</td>
</tr>
<tr>
<td>Randell</td>
<td>Renner</td>
<td>Saint John</td>
<td>N.B.</td>
</tr>
</tbody>
</table>

*Table 7-5: Table with student data from ‘Source B’*
An SQL trim function on a column will ease text matching by eliminating irregular spacing and a lower or upper function will assist with matching texts irrespective of lowercase and uppercase characters. However, a text match on NB, N.B. or New Brunswick will return a false value.

The common text matching conflicts encountered during the course of this research were matching the locations of various events with Statistics Canada data. There were conflicts with standards across datasets with some of them using the French characters and the others not; or each dataset storing the province names contrarily. Making spell checks during data cleaning can be tedious and time consuming. A partial solution to this is reached during the course of this study by the use of what we might call helper tables.

<table>
<thead>
<tr>
<th>pruid</th>
<th>prname</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>PEI</td>
</tr>
<tr>
<td>11</td>
<td>Prince Edward Island</td>
</tr>
<tr>
<td>12</td>
<td>N.S.</td>
</tr>
<tr>
<td>12</td>
<td>N.-É.</td>
</tr>
<tr>
<td>12</td>
<td>NS</td>
</tr>
<tr>
<td>12</td>
<td>Nova Scotia</td>
</tr>
<tr>
<td>13</td>
<td>N.B.</td>
</tr>
<tr>
<td>13</td>
<td>N.-B.</td>
</tr>
<tr>
<td>13</td>
<td>NB</td>
</tr>
<tr>
<td>13</td>
<td>New Brunswick</td>
</tr>
<tr>
<td>24</td>
<td>Que.</td>
</tr>
<tr>
<td>24</td>
<td>QC</td>
</tr>
<tr>
<td>24</td>
<td>Québec</td>
</tr>
<tr>
<td>24</td>
<td>Quebec</td>
</tr>
<tr>
<td>35</td>
<td>Ont.</td>
</tr>
</tbody>
</table>

Table 7-6: An excerpt from the province helper table
Helper tables are created to support many to one relationship and can have multiple records pointing to one value to assist with integration of data with possible conflicting values across datasets. Table 7-6 shows an excerpt from the province helper table created on the database to aide with province integration. As can be seen from the table, a column ‘prname’ holds values for all possible conflicting values of province names found across multiple datasets. The column ‘pruid’ has its respective province’s unique identifier. Any dataset that requires a join on province name refers to the province helper table first to extract its respective province’s unique identifier which is then used to join Statistics Canada table that has assigned province unique identifiers.

It is important to evaluate the cost of having a helper table on the database. Since Canada has only 13 provinces/territories the cost of having a province helper table is very small in terms of both memory and query processing time. On the other hand having a helper table that stores all possible conflicting values for cities (E.g. Montreal and Montréal) across Canada (approximately 6,761) can be expensive in terms of both memory and query processing time. An easier workaround solution used here is to return the unmatched city names through an SQL query and make changes to the city names manually.

Since longitudinal data spreads across a large period in time, some of the properties are subject to change while the others remain unchanged. Longitudinal datasets should always incorporate timestamps on their data and keep join events based on the timestamps. For example, Table 7-7 shows data that has timestamps for the practice locations of physicians.
<table>
<thead>
<tr>
<th>Id</th>
<th>Event_time</th>
<th>Practice_Location</th>
<th>Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>177</td>
<td>2013</td>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>177</td>
<td>2014</td>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>177</td>
<td>2015</td>
<td>Fredericton</td>
<td>NB</td>
</tr>
<tr>
<td>179</td>
<td>2014</td>
<td>Sussex</td>
<td>NB</td>
</tr>
<tr>
<td>179</td>
<td>2015</td>
<td>Sussex</td>
<td>NB</td>
</tr>
</tbody>
</table>

Table 7-7 Dataset showing physician practice locations with timestamps

### 7.2.3 Data Structuring

Data structuring is an important process of data transformation and should be done on internal data before integrating with external data so that the integrated result is one or more tables with all the relevant columns in proper structure. Data for subject-centric analysis should be structured in person-period datasets whereas data for time-centric analysis should be structured in person-level datasets. For LOCATED project, since the study requires time centric analysis for GIS application and comparison of characteristics exhibited by graduates from different cohort i.e. time-centric analysis, data is structured in a person-level dataset as shown in Table 7-8.
<table>
<thead>
<tr>
<th>Id</th>
<th>Gender</th>
<th>Age_at_admission</th>
<th>High_school</th>
<th>High_school_city</th>
<th>High_school_province</th>
<th>Prior_university</th>
<th>Prior_university_city</th>
<th>Prior_university_province</th>
</tr>
</thead>
<tbody>
<tr>
<td>177</td>
<td>F</td>
<td>24</td>
<td>Halifax Regional</td>
<td>Halifax</td>
<td>NS</td>
<td>Dalhousie University</td>
<td>Halifax</td>
<td>NS</td>
</tr>
<tr>
<td>178</td>
<td>F</td>
<td>27</td>
<td>Saint John High School</td>
<td>Saint John</td>
<td>NB</td>
<td>University of New Brunswick</td>
<td>Fredericton</td>
<td>NB</td>
</tr>
<tr>
<td>179</td>
<td>M</td>
<td>34</td>
<td>Fredericton High School</td>
<td>Fredericton</td>
<td>NB</td>
<td>University of New Brunswick</td>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>180</td>
<td>F</td>
<td>23</td>
<td>Moncton High School</td>
<td>Moncton</td>
<td>NB</td>
<td>University of New Brunswick</td>
<td>Saint John</td>
<td>NB</td>
</tr>
<tr>
<td>181</td>
<td>M</td>
<td>28</td>
<td>Saint John High School</td>
<td>Saint John</td>
<td>NB</td>
<td>University of New Brunswick</td>
<td>Saint John</td>
<td>NB</td>
</tr>
</tbody>
</table>

Table 7-8 Data in a person-level dataset
7.3 Loading

After extraction and transformation, data is now ready to be loaded to the data warehouse. Since we have advocated the use of relational database management system, the data warehouse is hosted on PostgreSQL. This data warehouse is hosted from a remote server, the front-end application which runs queries over it for analytical reporting is discussed in detail in Section 8.2.
8 Results

The purpose of this chapter is to showcase the results obtained in terms of increasing the analyzable data within the longitudinal data warehouse. Further, this chapter explains the implementation and working of the front-end application built in this study over the data warehouse for analysis and reporting of data.

8.1 Increase in Amount of Analyzable Data

The results obtained in terms of increase in the amount of analyzable data can be classified into quantifiable and non-quantifiable or qualitative. Quantifiable results are immediate and can be measured. Non-quantifiable or qualitative results are gradual and cannot be measured immediately.

8.1.1 Quantifiable Results

The quantifiable increase in amount of analyzable data was obtained through data cleaning as mentioned in Section 7.2.1; where a 13.72% increase in the amount of matching efficiency was observed. Another quantifiable result obtained during this study was a 2.3% increase in the amount of analyzable data when the first and last names of the subjects could not be matched due to them changing their last name.
8.1.2 Non-quantifiable or Qualitative Results

The non-quantifiable results include solutions proposed for the challenges encountered during this study which will help increase the amount of analyzable data gradually as the dataset matures. These proposed solutions are listed below:

- Implementation of administrative policies to monitor data storage for improved data integration efficiency as discussed in Section 4.1
- Standardizing the process of storing data to improve data quality
- Making data fields mandatory to increase the amount of data available for analysis as discussed in Section 4.3.
- Incorporating timestamps in datasets to help identify various points in dynamic data
- Using helper tables as discussed in Section 7.2.2 to help increase data matching efficiency

8.2 Front End Application for Analysis and Reporting

A graphical user interface was developed in the front-end tier to query over the data warehouse architecture. The stakeholders can use this application to generate analytical reports in the form of charts and maps. The application is further explained in the subsequent sections.
8.2.1 Application Architecture

The application is designed and developed in C#. The architecture of the application is shown in Figure 8-1 where the Windows Form Application is written in C# and managed by Common Language Runtime (CLR) over the .NET Framework. The application queries data from the data warehouse located at a remote server on a Windows operating systems and displays the results in the form of charts and reports for analysis.

![Figure 8-1 Front-end Application Architecture](image)

8.2.2 Design of the Application
The application developed as shown in Figure 8-2, is designed to accommodate a multiple perspective analytical approach. Multiple perspective approach is achieved through implementation of several tabs on the form as shown in Figure 8-3, one for each broad category of analytical questions pertaining to the research question. These tabs are explained in the next section.
Figure 8-2 Screenshot of the LOCATED Windows forms Application
Figure 8-3 Screenshot showing the multiple perspective approach adapted in
8.2.3 Multiple Perspective Approach

The LOCATED application categorizes the analytical questions into categories such as ‘Demographic Characteristics’, ‘Medical Training’, ‘Residency and Fellowship’, ‘Practice Location and Career Choice’, ‘Quick Report’ and ‘Regional Analysis.’ The ‘Demographic Characteristics’ tab shown in Figure 8-4 answers questions related to the gender, ages at admission and graduation, high school locations and prior university locations. The ‘Medical Training’ tab shown in addresses question related to the time spent in communities during medical training.

![Figure 8-4 Demographic Characteristic tab in the LOCATED application](image)
The ‘Residency and Fellowship’ tab shown in Figure 8-6 shows the Canadian Resident Matching Service (CaRMS) programs students matched to and the training experiences gained.
Figure 8-6 Residency and Fellowship tab in the LOCATED application
Figure 8-7 Practice Location and Career Choice tab in the LOCATED
Figure 8-7 shows the ‘Practice Location and Career Choice’ tab that can display analysis about the practice locations and career choices of graduates. The ‘Quick Report’ tab can help display basic descriptive statistical reports of graduates belonging to any cohort as shown in Figure 8-8.

Figure 8-8 Quick Report tab in the LOCATED application⁴

⁴ Dummy data used to preserve sensitive information
8.2.4 Need for a New GUI In Spite of Existing BI Tools

Although several data analysis applications are available, a distinctive application is developed for the LOCATED project to address the analytical questions specific for the project such as demographic characteristics, rural/urban background, career choices and practice location assessment. The LOCATED application is designed such that users can break down the cohorts with respect to characteristics such as gender, age group, locations of permanent residence, career choices and high school, university and practice locations. This is beneficial in analyzing behavior of population exhibiting certain traits. Users can ‘apply filters’ as shown in Figure 8-9, to perform analysis on populations exhibiting certain characteristics. Section 8.2.5 explains the filters with examples.
Figure 8-9 Screenshot showing filters implemented on the LOCATED Application
8.2.5 Applying Filters

The filters provided in the LOCATED application allow users to analyze the characteristics of population exhibiting certain traits. For example, Figure 8-10 shows the high school location experience of a cohort belonging to graduating Class of 2006 whereas Figure 8-11 and Figure 8-12 show the university location experience of females and males belonging to the same cohort respectively. This kind of analysis helps answer the question asking whether from the Class of 2006, more males than females have had prior university experience in large centers.

![Pie chart](image)

**Figure 8-10 University location experience of all students from Class of 2006**

---

5 Dummy data used to preserve sensitive information
Figure 8-11 University location experience of female students from Class of 2006

Figure 8-12 University location experience of male students from Class of 2006

6 Dummy data used to preserve sensitive information
8.2.6 Security

Due to the sensitivity of information within the data, the data warehouse is hosted from a remote server to which the application establishes a connection. The application itself has a secure login feature that allows access to only registered users as can be seen from Figure 8-13.

Figure 8-13 Application Secure login page

8.2.7 Mining Custom Data

The application also allows users to query custom data from the data warehouse and export the spreadsheet in Microsoft Excel for further reporting using GIS tools, also developed as part of this study. Figure 8-14 shows a screenshot of the custom data mining feature within the LOCATED application. This feature allows users to select different attributes from the
dataset for selected cohorts to view in a data grid. This can then be exported to Microsoft Excel to be used for GIS analysis using ArcGIS or for other types of custom analysis.

Figure 8-14 Mining Custom data from the LOCATED application

8.2.8 Validation and Update

7 Dummy data used to preserve sensitive information
The application back-end queries, written in SQL, are designed such that the application will display any additional data added to the data warehouse during annual updates without the need for the code to be rewritten. This is possible since the graduation year selections are done by users from the application through textbox values.
9 Conclusion and Future Work

In this thesis, we integrated longitudinal data for a tracking system for graduating health professionals. We considered development of the tracking system over a longitudinal database to overcome the technical challenges associated with integrating, storing and maintaining historical as well as current data. We mentioned the types of longitudinal data and the criteria for data structuring. We highlighted the challenges associated with longitudinal datasets and methods are developed to tackle these challenges. We introduced helper tables to address issues associated with lack of data standards across different datasets.

We addressed the steps for building the data warehouse i.e., data acquisition, data cleaning, data integration, data structuring and the challenges associated with each of these steps. We compared dimensional and relational data modeling and advocated the use of the relational data model to encounter the changing complexities of longitudinal data.

We designed and developed a distinctive front-end application to query the data warehouse for analysis and representation of information in the form of charts and reports. We introduced a multiple perspective design for the application in the form of tabs.

In conclusion, tracking systems can be vital for universities to understand the impact of their graduates on the community to help present a case to the government for future funding support. They can also help make better decisions with respect to program
evaluation and geographic distribution of training or co-op locations in their programs. The proposed approach can help universities in identifying and overcoming the challenges in the development of a tracking system for their graduates. The front-end application can help in developing reports of extreme value to aid the stakeholders in making strong data driven decisions.

9.1 Future Work

The longitudinal data warehouse which has been implemented in a relational model on PostgreSQL can be applied on a graph database instead. Neo4j [22] defines a graph database as “an online database management system with Create, Read, Update and Delete (CRUD) operations working on a graph data model.” Graph databases are different from other databases due to the high priority given to relationships in graph databases. Entities in a graph database are represented using nodes and the relationship between nodes is represented using the edges between the nodes. Due to this functionality of graph databases, an “application doesn’t have to infer data connections using things like foreign keys or out-of-band processing, such as MapReduce” [22]. Cypher is the query language used by Neo4j. Graph databases are used by Google search engine and more recently by Facebook for their natural language graph search engine, which allows users to search for connections using statements such as “people who live in Saint John, NB and went to University of New Brunswick.”
By implementing longitudinal data for tracking systems in a graph database, high priority can be given to relationships between nodes. Figure 9-1 shows a graph database model with subjects, professions and locations as nodes and their respective relationships as edges. This database design may allow for easier query processing in terms of searching for students originating from a certain community or practicing at a specific location.
Figure 9-1 Graph database model showing subjects, professions and locations as nodes and relationships as edges
Bibliography


[15] C. Bethune, I. Robbe and S. Mulay, "Reflecting on 5 Years of FMEC MD Innovations," Memorial University Faculty of Medicine, 2015.


[22] Neo4j, "Why Graph Databases?," Neo4j, 2015. [Online]. Available:
Curriculum Vitae

Raafey Ahmed Mohammed

Universities attended:

Deccan College of Engineering and Technology
Bachelor of Engineering (Computer Science)
Hyderabad, India
Sep’2008 – June 2012

Publications:


- Paixao, S., Mohammed, R. & Steeves, J. (2014). Can you Track the Long Term outcomes of your Graduates. *Global Community Engaged Medical Education Muster*


Oral Presentation: Technical Challenges for Longitudinal Data Integration. *Canadian Conference on Medical Education (CCME) 2014*

Poster Presentation: Interactive Wireless Site Planning Simulator for Hand-off Parameter Analysis. *Computer Science Research Exposition 2015, University of New Brunswick*