RMIT UNIVERSITY

Project Proposal
for
Real World Health Benefits

Date: 20/07/2017
Sponsor: NostraData Pty Ltd
Author: Mike da Gama

Commercial - in – Confidence

Preface

The purpose of this document is to provide sufficient information for the RMIT School of Computer Science to determine the feasibility / eligibility of the project for the Software Engineering project course.

Students will then review the proposal to determine if it is to be selected for their course project.

Table of Contents

1 Executive Summary

NostraData is a data analytics company that has access to the deidentified dispensary information of over 4,000 of the 5,500 community pharmacies in Australia, with over 1.6 billion prescriptions spanning 6 years. We are looking to integrate this information with other, publicly available datasets through our existing Hadoop/Hortonworks instance. We need a team of talented database and software engineers to find a way to extract this information from the desired sources (e.g. PBS, ABS, BoM, AIHW) and map it to our data source at the lowest granularity possible. You have access to SQL Server, Informatica, Vertica, R, Hadoop, Hortonworks and MicroStrategy to create an optimised big data pool that can be used for data mining and analytics. Ideally, you will make the interface MicroStrategy but we are open to alternative solutions. The final step, time permitting, is to create specialised machine learning techniques to 'uncover' patients that may be at risk of a worsening health condition. Using a platform that we have already built – CarePro - we can proactively alert the pharmacy where the patient usually visits to take preventative action. A **real world** health benefit that we can measure.

There is enormous potential in a working system capable of proactively identifying patients that may benefit from pharmacist counselling, some of which include:

1. Better health outcomes for the patient (e.g. prevention of stroke, heart attack, osteoporotic fracture, glaucoma);
2. Reduced burden of disease (health economic benefit);
3. Reduced hospital admission/readmission;
4. Increased loyalty of patients to their pharmacy and increased value of patient to pharmacy;
5. More accurate predictions over time through feedback systems;
6. Earlier identification of potential health issues;
7. Savings to government and health insurers.

2 Background

NostraData has traditionally relied upon just its own data to identify useful information for pharmacy and its stakeholders. As demands for greater insights place pressure on the limitations of the existing dataset, we have recognised the potential benefits of combining our data with that which is publicly available to enrich our insights and consulting.

Strategically, NostraData sees an opportunity for pharmacies to take a proactive position with patient health to protect their role as primary care providers in our community. As NostraData is funded by pharmacies our sustainability relies somewhat on the current pharmacy model.

This opportunity to create a proactive care approach is of increasing importance and the timing to do something is within the next 12 months. In addition, with NostraData's recent development of a platform for pharmacies that enables them to identify, invite and counsel patients around their medicines and general wellbeing, the development of algorithms that can be tested on the platform is highly desired.

3 Operational Fit / Initiative

This initiative is relatively new in that our current instance of Hadoop and R is underdeveloped and in proof of concept stage. Also, the import of new datasets will augment our current import of pharmacy data and the Pharmaceutical Benefits Scheme product data (pbs.gov.au).

This will then add-on to our existing development in MicroStrategy and a proprietary platform known as CarePro. The new business capabilities will centre around our ability to overlay multiple datasets to perform geodemographic risk analysis based on medicine use and other data. This in turn can be developed into preventative health initiatives at a community pharmacy level.

Our plan is to create a development environment that contains a subset of the existing data and the technical layers available to the project team. Currently, we have proprietary software that extracts the data from the various dispensary and point-of-sale systems in community pharmacy, which in turn feeds into SQL Server, then via Informatica PowerCentre and IDQ into Vertica/R, then finally into MicroStrategy. We utilise the HDFS system from Hadoop for some data management and in the early stages of looking into Hive and other toolkits. In addition, we are on a fully scalable AWS instance in Australia. Some of the legacy systems (SQL Server) still reside on a private cloud in a data centre in Sydney. The development will be isolated from production but once tested and QA passed, could be fed into production for field testing.

4 Deliverables

The following deliverables are desired:

1. Automated import on refresh cycle of:
    a. Pharmaceutical Benefits Scheme drug data *http://www.pbs.gov.au/browse/downloads*
    b. ABS stats on standard demographics/risk factors including SES, Age, Sex, Family, Smoking, Wealth: *http://www.abs.gov.au/ausstats/abs@.nsf/ViewContent?readform&view=productsbytopic&Action=Expand&Num=5.7*)
    c. Australian Institute of Health and Welfare *http://www.aihw.gov.au/data-linking/*
    d. Bureau of Meteorology *http://www.bom.gov.au/catalogue/data-feeds.shtml*
    e. AMT *https://www.digitalhealth.gov.au/get-started-with-digital-health/what-is-digital-health/clinical-terminology*
    f. Other health data sources *http://libguides.library.usyd.edu.au/c.php?g=508094&p=3475693*
2. Assimilating the data sources in a data pool or lake for easy access and linkage
3. Data dictionaries and glossaries for reference (using online NostraData wiki)
4. Reference guides and maintenance for data import, source, update and management

Optional:

5. Access to the data pool from MicroStrategy and other analytics tools in a highly flexible format (for this item R will be required)
6. Integration of machine learning to enable the unsupervised identification of patients with various conditions that match risk criteria
7. Creation of algorithms to make the above mentioned identification process repeatable per pharmacy
8. Integration of the above algorithms into existing platform to provide data to CarePro

5 Stakeholders

Director – Mike da Gama *md@nostradata.com.au*

Head of Technology – Mark Evarts *me@nostradata.com.au*

Project Manager – Cingdy Cingdy

Database Administrator – Adrian Oprea

ETL and Data Quality Manager – Unnee Udayakumar

Project coordinator – Ann Southall

Business Analyst - Dhanya Nair

Subject Matter Expert - Ahmer Siddiqui

PMO - Tanya Brown

QA & Test - Vani Vendra

Solution Architect - Stephen Carter

All stakeholders will be available to the team much in the same manner as if they were employees. That is, we have a business to run and this would be treated as part of the ongoing business, with scheduled meetings, scrum and sprint sessions where needed.

6 Dependencies

Any increase in IT resources required for the project will need to be properly scoped out and costed. This would be a joint exercise between the project team and NostraData. The company is currently running at capacity and so any increased reliance on the NostraData team to accommodate the project shall be planned into the schedule by the PMO.

Ad hoc access to NostraData employees is discouraged as it may impact productivity. Instead, scheduled planning and clear requirements will ensure the project is given adequate attention.

7 Resource / Skill Set Requirements

The estimated resources required for this project will be students skilled in SQL, R (preferable), a basic knowledge of Linux and knowing the basic programming principles of Java/Python is a must. In addition, some working knowledge of source-to-target mapping of data sources in various file formats (XML being the favourite of government) and web scraping would assist. The complexity of some data sources will require domain expertise and so we will provide that assistance where needed.

Programming Java/Python and database skills are required.

It is acceptable if the students are not familiar with R and Hadoop, as the corresponding parts of the deliverables are optional. A learning curve for R and Hadoop is also acceptable.

8 Scholarship

The software development project is proposed in a bundle with a short research project that will be conducted within mid-semester break and will be paid (approx. AUD 1500 per student). The research topic will be focused on strategies for data import, source, update and management strategies (cf. point 4 of the deliverables).
Only the students working on their development project within the SEP course would be allowed to participate in the corresponding research project.

9 Key Issues & Risks

The key risks to this project include:

1. Lack of knowledge of the datasets to be extracted
2. Inability to extract datasets from desired sources
3. Datasets are at different granularity and cannot be properly mapped together
4. Complexity prevents development of functioning system
5. Stakeholder resource constraints (time, IT)
6. Insufficient working knowledge of technology by project team
7. Base solution not ready for deployment
8. Poorly defined requirements
9. Scope creep
10. Inadequate design

10 Other Considerations

Students are not expected to have a working knowledge of the datasets prior to the project initiation. Students are expected to be proficient in the programming/coding/database skills that they bring to the project, as support at this technical level will be limited.