# Research Proposal

*Degree (e.g., Master by Research or PhD)*

**Area of the Research Proposal:**
*(e.g., Data Science and Engineering)*

**Title of your Ph.D. Research Proposal:**

**…….**

**Complete Name of Candidate:**

**…….**

**email address of candidate:**

**……..**

**Advisor:**
Prof. Amin Beheshti
Head of Data Science Research Lab
School of Computing
Faculty of Science and Engineering
**Macquarie University**

**Planned Semester to Start:**

**…….**

# Abstract

The quality of the services any organization provides largely depends on the quality of their processes. Banking and finance companies are increasingly grasping this concept and moving towards data-driven process-oriented enterprises. This strategy is important as the world is currently awash with massive volumes of data (Big Data) generated from open (e.g., news and Web), private persona/business (e.g., emails and files), social (e.g., Twitter and Facebook) and IoT (e.g., sensors and CTVs) sources. Understanding and analysing this data, in the context of banking processes, will enable the banks revolutions their data-driven and knowledge-intensive processes. At the same time, these tasks are challenging. To address this challenge, in this proposal, we focus on novel data science and engineering techniques to understand the banking data and prepare it for big data analytics. The proposed approach has the potential to become the driving force behind critical decisions for banking and financial sector by turning massive volumes of data into actionable knowledge.

# 1- Introduction

The quality of the services any organization provides largely depends on the quality of their processes. Improving business processes is critical to any banking and finance corporation. Process improvement requires banking data curations as its first basic step. Before curating the banking big process data, there is a need to capture and organize the big data from open, private, social and IoT sources. This is important as executions of process steps, in modern enterprises, leave temporary/permanent traces in various systems and organizations. In order to analyze process data, it is possible to collect the data into a data lake [1], using extract, transform, and load (ETL) tools, and then leverage an OLAP tool to slice and dice data along different dimensions [2].

Banking and finance companies are increasingly grasping this concept and moving towards data-driven process-oriented enterprises. This strategy is important as the world is currently awash with massive volumes of data (Big Data) generated from open (e.g., news and Web), private persona/business (e.g., emails and files), social (e.g., Twitter and Facebook) and IoT (e.g., sensors and CTVs) sources. In this context, the big data problem can be seen as a massive number of data islands that need to be organized, curated, processed, analysed and visualized to reveal unknown insight and knowledge.

Understanding and analysing this data, in the context of banking processes, will enable the banks revolutions their data-driven and knowledge-intensive processes. At the same time, these tasks are challenging. To address this challenge, in this proposal, we focus on novel data science and engineering techniques to understand the banking data and prepare it for big data analytics. The proposed approach has the potential to become the driving force behind critical decisions for banking and financial sector by turning massive volumes of data into actionable knowledge. The unique contributions of this PhD proposal includes:

- Significant scientific advancement in understanding the practical problems of understanding the banking data. We will understand and analyse the features important for analysing the banking data from different points of view and based on various business need.
- A platform for organizing and curating the banking data and meta-data. We focus on an especial type of meta-data (i.e., provenance [3]) and present a novel banking provenance model to enable business/data analysts predict the customer needs in an easy way.
- A digital dashboard to facilitate the analysis of banking entities (e.g., customers and products). The dashboard will benefit from Artificial Intelligence (AI) and provide cognitive assistants to business/data analysts.

# 2- State of the Art

## *2-1 From Data Lakes to Knowledge Lakes.*

With data science continuing to emerge as a powerful differentiator across industries, almost every organization is now focused on understanding their business and transforming data into actionable insights. The notion of a Data Lake [1] has been coined to address this challenge and to convey the concept of a centralized repository containing limitless amounts of raw (or minimally curated) data stored in various data islands. The rationale behind a Data Lake is to store raw data and let the data analyst decide how to cook/curate them later. While Data Lakes do a great job in organizing Big Data and providing answers on known questions, the main challenges are to understand the potentially interconnected data stored in various data islands and to prepare them for analytics.

The notion of Knowledge Lake [4], i.e. a contextualized Data Lake, introduced to automatically transform the raw (process) data into contextualized data and knowledge. The term Knowledge here refers to a set of facts, information, and insights extracted from the raw data using data curation techniques such as extraction, linking, summarization, annotation, enrichment, classification and more. In particular, a Knowledge Lake is a centralized repository containing virtually inexhaustible amounts of both data and contextualized data that is readily made available anytime to anyone authorized to perform analytical activities. Knowledge Lakes provide the foundation for Big Data analytics by automatically curating the raw data in Data Lakes and preparing them for deriving insights.

## *2-2 Graph Modeling.*

Graphs are essential modeling and analytical objects for representing information networks. Several graph querying techniques [5] such as pattern match query, reachability query, shortest path query, and subgraph search, have been proposed for querying and analyzing graphs. These methods rely on constructing some indices to prune the search space of each vertex to reduce the whole search space. In [2] authors discusse a number of data models and query languages for graph data. Many of these models use RDF[1] (Resource Description Framework), an official W3C recommendation for semantic Web data models, to model graphs and use SPARQL[2], an official W3C recommendation for querying RDF graphs. SPARQL queries are pattern

[1] https://www.w3.org/RDF/

[2] https://www.w3.org/TR/rdf-sparql-query/

matching queries on triples that constitute an RDF data graph, where RDF is a data model for schema-free structured information. Several research efforts have been proposed to address efficient and scalable management of RDF data. SPARQ is a declarative query language, an W3C standard, for querying and extracting information from directed labeled RDF graphs. SPARQL supports queries consisting of triple patterns, conjunctions, disjunctions, and other optional patterns. However, there is no support for querying grouped entities. Paths are not first class objects in SPARQL. PSPARQL [6] extends SPARQL with regular expressions patterns allowing path queries; and supports folder and path nodes as first class entities that can be defined at several levels of abstractions and queried.

### *2-3- Knowledge-Intensive Processes.*

Case-managed processes are primarily referred to as semistructured processes, since they often require the ongoing intervention of skilled and knowledgeable workers. Such Knowledge-Intensive Processes involve operations that rely on professional knowledge. For these reasons, it is considered that human knowledge workers are responsible to drive the process, which cannot otherwise be automated as in workflow systems [7]. Knowledge-intensive processes often involve the collection and presentation of a diverse set of artifacts and human activities around artifacts. This emphasizes the artifact-centric nature of such processes.

### *2-4- Process Data Analytics.*

A recent book [7], provided an overview of the state-of-the-art in the area of business process management in general and process data analytics in particular. This book provides defrayals on: (i) technologies, applications and practices used to provide process analytics from querying to analyzing process data; (ii) a wide spectrum of business process paradigms that have been presented in the literature from structured to unstructured processes; (iii) the state-of-the-art technologies and the concepts, abstractions and methods in structured and unstructured BPM including activity-based, rule-based, artefact-based, and case-based processes; and (iv) the emerging trend in the business process management area such as: process spaces, big-data for processes, crowdsourcing, social BPM, and process management in the cloud. BPM in the Cloud solutions offer visibility and management of business processes, low start up costs and fast return on investment. Crowdsourcing can help organizations increase productivity by discovering and exploiting informal knowledge and relationships in order to improve activity execution. Crowdsourcing can also enable the socialBPM to assign an activity to a broader set of performers or to find appropriate contributors for its execution. Social BPMs inevitably require advanced crowd-management capabilities in future social computing platforms.

# 3- Research Objectives and Approach

The anticipated outcome of this project is the first generic framework and set of techniques for systematically contextualizing the banking data and resolving important gaps:

- **A unified framework banking data curation.** We provide the foundation for banking analytics by automatically curating the raw bank data in the Data Lake and to prepare them for deriving insights. We present a generic and unified model to formalize entities and relationships among them in the Knowledge Lake and to construct the Knowledge Graph, i.e. a Knowledge Base to enhance the discovery of connected events and entities.
- **Novel and fine-grained analytical services for banking predictive analytics.** The proposed techniques will isolate the process analyst from the process of explicitly linking an analytical content and enable the analyst to use interactive insight generation to select and sequence useful patterns. To achieve this, a set of innovative, fine-grained and intuitive analytical services to curate the data, and enrich them with complex data structures (e.g. timeseries, hierarchies, patterns and subgraphs) will be presented.
- **Novel Digital Dashboard.** We present novel techniques in the field of Human-Computer Interaction and Visualization to enable the business and data analysts explores the Knowledge Graph and to support interactive visualizations in an easy way. A set of components- for summarizing, annotating, linking, enriching and visualizing - in the dashboard will enable predictive analytics in an easy way.

# Conclusions

The continuous demand for the business process improvement and excellence has prompted the need for business process analysis in the enterprise. Recently, the business world has begun to become increasingly dynamic as various technologies such as the Internet and email have made dynamic processes more prevalent. Following this, the problem of understanding business process execution has become a priority in rapidly changing and knowledge-intensive organizations. To address this challenge, in this proposal, we focus on novel data science and engineering techniques to understand the banking data and prepare it for big data analytics.

# References

1. Amin Beheshti, Boualem Benatallah, Reza Nouri, Van Munin Chhieng, HuangTao Xiong, Xu Zhao: CoreDB: a Data Lake Service. CIKM 2017: 2451-2454
2. Amin Beheshti, Boualem Benatallah, Hamid Reza Motahari-Nezhad: ProcessAtlas: A scalable and extensible platform for business process analytics. Softw., Pract. Exper. 48(4): 842-866 (2018)
3. Amin Beheshti, Boualem Benatallah, Hamid R. Motahari Nezhad: Enabling the Analysis of Cross-Cutting Aspects in Ad-Hoc Processes. CAiSE 2013: 51-67
4. Amin Beheshti, Boualem Benatallah, Reza Nouri, Alireza Tabebordbar: CoreKG: a Knowledge Lake Service. PVLDB 11(12): 1942-1945 (2018)
5. Omar Batarfi, Radwa El Shawi, Ayman G. Fayoumi, Reza Nouri, Seyed-Mehdi-Reza Beheshti, Ahmed Barnawi, Sherif Sakr: Large scale graph processing systems: survey and an experimental evaluation. Cluster Computing 18(3): 1189-1213 (2015)
6. Amin Beheshti, Boualem Benatallah, Hamid R. Motahari Nezhad, Sherif Sakr: A Query Language for Analyzing Business Processes Execution. BPM 2011: 281-297
7. Amin Beheshti, Boualem Benatallah, Sherif Sakr, Daniela Grigori, Hamid Reza Motahari-Nezhad, Moshe Chai Barukh, Ahmed Gater, Seung Hwan Ryu: Process Analytics - Concepts and Techniques for Querying and Analyzing Process Data. Springer 2016, ISBN 978-3-319-25036-6, pp. 1-178