

GRAPHMK – AN INTEGRATED FRAMEWORK FOR GRAPH COMPUTATION

Nageswaran Keshan

(189330G)

Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science

Department of Computer Science

University of Moratuwa
Sri Lanka

April 2021

Declaration

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Full Name of Student: Nageswaran Keshan

Registration No: 189330G

Signature: *UOM Verified Signature*

Date: 11/11/2021

The supervisor/s should certify the thesis/dissertation with the following declaration.

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the CS6997 MSc Research Project.

Supervisor Name: Dr. Indika Perera

Signature:

Date: 11/11/2021

Abstract

The rapid change in technological innovation and end user expectation during the past decade resulted in data being an important area for new research and development. In addition, evolvement of internet of things, social graph and communication networks resulted in creation of large-scale real time, rapidly changing datasets. This led to the creation of graph systems supporting various graph analytics to become popular. Meanwhile, relational databases are used in storing and managing data including most advanced systems where graph analytics was never explored on top of a relational schema-based system.

This led to the questions whether graph analytics can be done in a relational environment whilst it is still being blindsided or does relational databases even have limitations to execute graph computations. The relational model where data needs to be imagined in a tabular format rather than a graph format with edges and nodes are inefficient in executing iterative graph analysis. Relational systems will end up creating expensive joints in order to execute such computations. Structured Query Language (SQL) queries are difficult in nature to express graph analysis, though we have downsided, relational systems that are composed of great functionalities which includes fault tolerance, integrity constraints, secure transaction and query optimization etc.

This thesis writeup reflects an integrated framework for graph analytics that comprise of three main components. Firstly, a data model that is capable of executing graph computation within a relational environment. Secondly, a query language which can be considered as a data log that helps to execute graph specific computations. Finally, a bolt-on solution which comprise of the above two, that executes sitting within a relational query engine and executes queries created from introduced data log language. The tests performed are evident that bolt-on solution introduced achieved better performances in provided scenarios.

Subject Descriptors

- **Graph Theory - Graph Problems, Graph specific algorithms**
- **Logical Design – Models & Schema**
- **Languages – Data log & Query**

Keywords

Bolt-on systems, Graph computation, Graph Algorithms and Iterative computation

Table of Contents

Declaration.....	i
Abstract.....	ii
1 Introduction.....	1
1.1 Overview of the Chapter	1
1.2 Motivation & Background Context.....	1
1.3 Identified Problem Scope.....	2
1.4 Details of Previous Research.....	4
1.5 Details on Aim of the Project.....	5
1.6 Details on Purpose of the Project	5
1.7 Requirements of Resources	7
1.8 Details on Expected Features	7
1.9 Summary of the Chapter	8
2 Literature Review.....	9
2.1 Overview of the Chapter	31
2.2 Background on different Analytical Systems.....	31
2.3 Graph Analytical Framework Performance Factors.....	31
2.4 Techniques for Graph Analytical Systems Analysis.....	32
2.4.1 Graph Computation.....	32
2.4.2 Relational Computation	33
2.4.3 Hybrid Computation	34
2.5 Graph System Approaches and Review	35
2.5.1 Review on Vertex-centric	35
2.5.2 Review on Neighborhood-centric	37

2.6	Graph Analytical Algorithms and Review	39
2.6.1	Analysis on PageRank	39
2.6.2	Analysis on Clustering.....	39
2.6.3	Analysis on Path	40
2.7	Graph Databases and Review.....	41
2.7.1	Property graphs in Information model.....	41
2.7.2	Triple stores	42
2.8	Relational Systems and Review	43
2.9	The Integrated Graph Analytical System's suitability.....	44
2.10	System Analysis – Pros & Cons	46
2.11	Deep dive of Existing Systems.....	47
2.12	Summary of the Chapter.....	49
3	Design & Architecture	51
3.1	Overview of the Chapter	47
3.2	GraphMk - High level design description	47
3.2.1	GraphMk Framework description on rich picture	47
3.2.2	GraphMk – Description on High Level Architecture	48
3.3	Description on solution and design goals.....	49
3.4	Details on introduced integrated graph model	50
3.4.1	Relational Core	50
3.4.2	Graphical Views.....	51
3.4.3	Relational Mapper views for Graph.....	52
3.5	Detail on introduced SQL functions.....	53
3.5.1	Selection of Design Methodology	53
3.6	GraphMk – Detail on Data flow diagram level 1.....	55

3.7	GraphMk – Details on Class Diagram	56
3.8	GraphMk – Details on Sequence Diagram.....	57
3.9	GraphMk – Details on Architecture refinement.....	57
3.10	GraphMk – Details on dependencies & packages	58
3.11	GraphMk – Details on ER Diagram	58
3.12	Summary of the Chapter.....	58
4	Implementation	60
4.1	Overview of the Chapter	59
4.2	Graphmk selection of Technology	59
4.2.1	Selection of operating system	59
4.2.2	Programming Language selection.....	59
4.2.3	Selection of an IDE and a development environment	60
4.2.4	Selection of Storage System	60
4.3	Description on implementation of GraphMk	60
4.3.1	GUI implementation	61
4.3.2	Query processor implementation	64
4.3.3	Query Parser implementation	66
4.3.4	Implementation of Query Optimizer.....	67
4.3.5	Details on different graph executors of graphmk	68
4.3.6	Details on Plan Executor implementation.....	70
4.3.7	Summary on process involved in Graphmk.....	71
4.4	Summary of the Chapter	72
5	Testing.....	73
5.1	Overview of the Chapter	73
5.2	Purpose of testing within GraphMk	73

5.3	Criteria for testing within GraphMk.....	73
5.3.1	Graphmk functional assessment and software quality.....	73
5.3.2	Graphmk structure assessment and quality.....	74
5.4	Testing stages within GraphMk	74
5.4.1	GraphMk testing	74
5.4.2	GraphMk integration.....	74
5.5	Description on environment to conduct test.....	74
5.5.1	Hardware Information.....	74
5.5.2	Software Information	74
5.6	Testing functionalities within GraphMk	75
5.6.1	Functional testing method.....	75
5.7	Integration testing perspectives for GraphMk.....	76
5.8	Testing non-functionalities within GraphMk.....	77
5.8.1	RMV Engine evaluation	77
5.9	Details on testing limitation	84
5.10	Summary of the Chapter.....	84
6	Evaluation	85
6.1	Overview of the Chapter	85
6.2	GraphMk project criteria for Evaluation.....	85
6.3	GraphMk – Evaluator selection process.....	86
6.4	GraphMk – Details on methodology for Evaluation.....	87
6.5	GraphMk – Details on evaluation questionnaire and format	87
6.6	GraphMk – Details on findings from survey	88
6.6.1	Opinions on Concept.....	88
6.6.2	Opinions on Scope	89

6.7	GraphMk – Details on self-assessment	90
6.7.1	Implementation of functional requirement and Status.....	90
6.7.2	Details on challenges	91
6.8	Summary of the Chapter	91
7	Conclusion	93
7.1	Overview of the Chapter	92
7.2	Details on Purpose & Aim achieved	92
7.2.1	Aim	92
7.2.2	Purposes	92
7.3	Application of Knowledge from course modules	94
7.4	Enhanced Existing Skills.....	95
7.5	Experiences from Project	96
7.6	Details on Challenges encountered	96
7.6.1	Scope Definition	96
7.6.2	Availability of Academic Publications	97
7.6.3	Support materials to learn key concepts and technologies	97
7.6.4	Constraints related to time	97
7.7	Details on Limitations from the research	98
7.7.1	Restrictions on a graph analytical system's main success factor	98
7.8	Possible Future work for GraphMk.....	98
7.9	Contributions from GraphMk.....	99
7.10	Closing comments	100
	References.....	i
8	Appendixes	v
8.1	Design Appendix.....	v

8.1.1	Data flow representation	v
8.1.2	Package Diagram	vii
8.1.3	ER Diagram	vii
8.2	Implementation Appendix.....	viii
8.2.1	Benchmarking GraphMk	viii