School of Informatics

Informatics Project Proposal
Benchmarking Graph Processing Systems

B138641
March 2019

Abstract

The aim of this project is to examine state-of-the-art graph processing systems and compare their performance and scalability in both local and parallel settings. More specifically, the project focuses on showing that often times scalability comes at a cost, and in many cases such an approach is not worth the time and effort. In our research we will prove this by conducting several experiments that will demonstrate that a single-core set-up outperforms executions on multi-core environments due to the parallelization overhead that incurs.

Date: Thursday 28th March, 2019

Tutor: Pablo León-Villagrá
Supervisor: Dr Milos Nikolic

Contents

1 Introduction **1**

1.1 Problem Statement 2

1.2 Research Hypothesis and Objectives 2

1.3 Timeliness and Novelty 2

1.4 Significance 2

1.5 Feasibility 3

1.6 Beneficiaries 3

2 Background **3**

3 Methodology **3**

4 Evaluation **3**

5 Expected Outcomes **3**

6 Research Plan, Milestones and Deliverables **4**

1 Introduction

Modern computer applications and services have evolved to an extent that they have become difficult to manage. The amount of data handled in such applications is only growing while the industry and the research community have built new systems and infrastructures to be able to keep up with this growth rate. Additionally, applications and services now feature relational data, which are often expressed as graph structures, fact that makes things even more demanding and complicated.

To extract useful information and perform data analytics tasks, many different graph data processing systems have been developed. Although these systems have been extensively used in various applications, many people argue that their performance is questionable and that there is a massive overhead that incurs due to the way these systems scale.

Moreover, there is no clear indication of which system is appropriate for each occasion. For this reason, there have been many researchers focusing on creating benchmarking systems in order to evaluate the performance of such systems and observe their strengths and weaknesses.

While research on benchmarking graph processing systems has provided information about the performance of several graph processing systems, it tends to be quite specific and concerns certain aspects of these frameworks. This project was motivated by the need for a more thorough investigation that covers a broader range of cases and systems. Our research aims at showing that scalability is not always the best approach by performing various experiments under different conditions, that is, measuring performance using many different algorithms, systems and other parameters. The main goal is to demonstrate how in many cases running such tasks in single node settings is more efficient performance-wise than distributed and multi-node settings.

Afterwards, if it is feasible time-wise, a second goal is to indicate how some tasks can be carried out with the use of higher-level systems, such as relational database management systems.

We will now break down the problem and briefly mention some of its aspects along with how this project contributes to the domain of graph processing and why it is important that such an investigation is conducted. In Section 2 we will discuss necessary and relevant knowledge to the problem's domain. Afterwards, we consider specific parts of our research and analyse steps in Section 3. We then focus on the evaluation of our results which will be detailed in Section 4 and highlight what we expect to see in our results in Section 5. Lastly, we state a detailed report of the project's plan, its objectives and the deliverables in Section 6.

1.1 Problem Statement

While there is research on individual graph processing systems that measures performance under certain circumstances, there hasn't been extensive work that exposes a more objective benchmark for these systems' performance, such as one that features the use of a wide variety of algorithms, multiple set-ups that range from one to many cores, as well as more realistic use cases. Our work focuses on these aspects and will try to provide a more broad benchmark that will be able to better indicate if such systems are worth using and, if so, under which settings.

1.2 Research Hypothesis and Objectives

Our hypothesis is that graph processing systems do not scale as well as they claim to do for common tasks. We were able to identify these common tasks and other useful information form a survey [1]. Moreover, this survey revealed how the graph research community and the industry perform their analysis tasks, as well as what data they use, what kind of algorithms they run, and so forth; all these were very helpful for formulating our objectives.

We aim to support our hypothesis by conducting experiments that will demonstrate how the selected graph processing systems perform in single-core settings, as well as when scaled up, and hope to show through our results that for common tasks executions in single-core settings perform better.

1.3 Timeliness and Novelty

Advances in technology nowadays have become more sophisticated and involve an extensive amount of data processing, a lot of which is expressed via graph representations. The popularity of this domain is constantly growing, given that more and more researchers and companies analyse graph data and perform graph processing tasks. Consequently, carrying out this research at this point is crucial since it will provide insights and indications as to how analyses and graph processing tasks can be performed, which will greatly impact the way researchers and developers approach their problems.

1.4 Significance

Having an indication as to what resources an individual may need before performing their graph data analysis tasks or research is of great significance. The reason for this is that a thorough experimentation that covers many use cases, some of which will probably be similar

to the intended task, will help in the process of the decision making with regards to the chosen technologies that will be used, the resources required, and others. This will result to better approaching a problem and providing more timely and efficient solutions.

1.5 Feasibility

While our aim seems to be quite feasible, we will nonetheless present some factors that may constitute an obstacle to our research or may limit it to an extent in some of its aspects.

Firstly, ...

Afterwards, ...

1.6 Beneficiaries

The outcome of this work will have an impact on the graph research community, as well as on the industry, where developers and analysts use such systems to perform their data analysis tasks. In detail, the results will provide insight as to whether an approach that involves distributed graph processing systems is needed, or if a more simple, single-core, implementation is more preferable. This will greatly benefit researchers and small to medium sized companies, since having such insight can potentially result to them saving time, effort and money.

2 Background

Modern applications and services handle a vast amount of The rise of interest in graph data has led the Lately there has been a lot of research around graph data processing due to the fact that

3 Methodology

We will now present the way by which our research will be conducted; that is, how we plan to approach the problem stated in Section 1.1, as well as the steps that will be followed in order to fulfill our project's purpose.

4 Evaluation

After having completed all steps and tasks mentioned in Section 3,we will carry on to analysing and interpreting our results.

5 Expected Outcomes

We will now review what we hope to achieve through our research and discuss how our contribution will impact the graph research community, as well as companies or individual developers that perform graph data processing tasks.

6 Research Plan, Milestones and Deliverables

The project will be completed once the following tasks have been accomplished:

- Determining all systems that will be examined
- Selecting necessary data sets/algorithms/workloads to perform experiments
- Conducting experiments in a single-core setting, as well as in multi-core settings (eventually scale up)
- Gathering results and comparing performances

Some concerns of the project are the following:

- Will the experiments produce the desired figures?
- Will there be an obvious difference in performance between the two examined cases?
- Will we have access to more resources (e.g. a cluster)?

References

[1] Siddhartha Sahu, Amine Mhedhbi, Semih Salihoglu, Jimmy Lin, and M. Tamer Özsu. The ubiquity of large graphs and surprising challenges of graph processing. *Proc. VLDB Endow.*, 11(4):420–431, December 2017.