School of Informatics

Informatics Project Proposal
Benchmarking Data 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. The claim that in many cases scalability is not the best option will be supported 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: Wednesday 27$^{\text{th}}$ March, 2019

Tutor: Pablo León-Villagrá
Supervisor: Dr Milos Nikolic

Contents

1 Motivation

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 in order to be able to choose the appropriate framework for a specific set of tasks.

This project aims to demonstrate that scalability is not always the best approach by performing various experiments under different conditions. 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

1.2 Research Hypothesis and Objectives

1.3 Timeliness and Novelty

1.4 Significance

1.5 Feasibility

1.6 Beneficiaries

The outcome of this work will have an impact on the graph research community, as well as on developers/analysts who use such systems to perform their (graph) data analysis tasks. In detail, the results will provide insight as to whether an approach that involves distributed processing systems is needed, or if a more simple single-core implementation is more preferable.

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

4 Evaluation

5 Expected Outcomes

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