UNIVERSITY OF BIRMINGHAM

Parallelised Data Processing

An investigation into the development of a tool for processing queries across a cluster of nodes.

Oliver Little

2011802

D.A. B.Sc Computer Science with Digital Technology Partnership (PwC)

Supervisor: Vincent Rahli

School of Computer Science

College of Engineering and Physical Sciences

April 2023

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

During my placement year from September 2021 - August 2022, the majority of my time was spent performing data processing using SQL. As part of this role, there were two common use-cases which are relevant to this project:

- Writing scripts to perform an ETL *(extract-transform-load)* workflow.

  This involves loading data from raw files (typically csv or some other kind of flat file), performing some transformation or processing steps, then exporting the data elsewhere for further analysis.

- Investigation of data after transformation, using one-off queries.

The datasets I worked with were usually 100GB in size or more, distributed over a number of source files. For many computers, this is too much data to keep in RAM all at once. In my role, the solution was to have a small number powerful SQL servers with a large amount of RAM which a number of users could connect to and perform their work on. This solution worked reasonably well for my use case, but executing scripts on the largest datasets could take upwards of a day. Furthermore, while a robust backup solution was in place for existing data, there was a risk of data loss while the execution was ongoing in the event of a server failure.

1.1.1 Datasets

A number of features

1.2 Problem Definition

Based on the above

1.3 Problem Definition

My aim is to build a framework on which data processing operations can be performed on a distributed cluster of nodes. These operations will include typical SQL commands, including:

- Filters
- Joins
- Group Bys

As part of the role I performed on my placement year, I had to spend a significant time actually executing the models we had written, as the data regularly changed format, slowing down the data ingestion process. Therefore, I would also like to build some tools into the framework to automate the following workflow:

- Ingesting new data
- Performing data processing
- Providing alerts for any execution issues
- Outputting the results

Chapter 2

Literature Review

Chapter 3

Design

Chapter 4

Implementation

Chapter 5

Testing

Chapter 6

Evaluation