# Comparative review of cloud computing platforms for data science workflows

---

Mohammad A. Rehman

*Supervisor:*
Dr Hairong Wang

A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

13 June 2022

# Declaration

I, Mohammad A. Rehman, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Mohammad A. Rehman

13 June 2022

# *Abstract*

The proposed framework for the comparative review of cloud computing platforms for data science workflows uses an amalgamation of the analytical hierarchy process, Saaty's fundamental scale of absolute numbers, and a selection of relevant evaluation criteria (namely: automation, error handling, fault tolerance, performance quality, unit testing, data encryption, monitoring, role based access, security, availability, ease of use, integration and interoperability) to enable users to evaluate criteria pertaining to cloud platforms for data science workflows. Three cloud providers were chosen as alternatives for selection by the user to determine which cloud platform would be most suitable for the user. The analytical hierarchy process used two levels, namely criteria and sub-criteria to ensure that attributes selected as sub-criteria were relevant and had a material impact on alternative selection. Experiments conducted and subsequent evaluation of criteria revealed that out of the three anonymised cloud service providers (CSP1, CSP2 and CSP3), CSP3 was most suitable for the user based on their valuation of the criteria. Out of the three alternate providers, CSP3 had an importance weighting of 0.53, CSP1 had 0.28 and CSP2 had 0.19, indicating that CSP3 would be most suitable for the evaluator in terms of the criteria considered. Thus, the proposed framework is successfully able to recommend which cloud platform would be suitable for the user based on the relative importance of the above criteria. Evaluations of the criteria are shown to be consistent and thus the weighting of criteria against the goal of cloud platform selection are sensible. The proposed framework is robust enough to accommodate for changes in criteria and alternatives, depending on user cloud platform requirements and scope of cloud platform selection.

# Acknowledgements

I acknowledge my supervisor, Dr Wang, for guiding me through conversations on the considerations that a review based research project needs to include. The additional perspective has increased the breadth of my investigations into cloud computing platforms for data science workflows.

Additionally, I extend a deep gratitude to my sponsors, the National e-Science Postgraduate Teaching and Training Platform.

# Contents

**Declaration** i

**Abstract** ii

**Acknowledgements** iii

**List of Figures** vi

**List of Tables** vii

**1 Introduction** 1
- 1.1 Literature Review . . . 2
- 1.2 Gaps in Literature . . . 4
- 1.3 Research Area . . . 4
- 1.4 Problem Statement . . . 4
- 1.5 Research Question . . . 5
- 1.6 Research Aims and Objectives . . . 6
  - 1.6.1 Research Aims . . . 6
  - 1.6.2 Objectives . . . 7
- 1.7 Limitations . . . 7
- 1.8 Assumptions and Definitions . . . 8
- 1.9 Overview . . . 8

**2 Research Methodology** 10
- 2.1 Research Design . . . 11
- 2.2 Data . . . 13
- 2.3 Methods . . . 14
  - 2.3.1 Creation of Data Science Workflows . . . 14
  - 2.3.2 Analytical Hierarchy Process . . . 21

2.4 Analysis . . . . . 28
2.5 Limitations . . . . . 29
2.6 Ethical Considerations . . . . . 30

**3 Results and Discussion** **31**
3.1 Results . . . . . 31
3.2 Discussion . . . . . 41
3.3 Summary . . . . . 43

**4 Conclusions and Future Work** **45**
4.1 Conclusions . . . . . 45
4.2 Future Work . . . . . 46

**A Terms of Service Review** **47**
A.1 Review of Cloud Platform Terms of Service . . . . . 47

**Bibliography** **49**

# List of Figures

2.1 AHP model for determining cloud platform based on data science workflow requirements . . . . . 25

# List of Tables

2.1 Saaty's fundamental scale of absolute numbers [17] . . . . . 12
2.2 Matrix of pairwise comparison of criteria . . . . . 25
2.3 Random consistency index parameters based on number of criteria [30] 29

3.1 Pairwise comparison matrix for selected criteria . . . . . 32
3.2 Criteria weights indicating relative importance to the goal . . . . . 32
3.3 Pairwise comparison matrix for selected sub-criteria under the functionality criteria . . . . . 33
3.4 Sub-criteria weights indicating relative importance to functionality . 34
3.5 Pairwise comparison matrix for selected sub-criteria under the governance criteria . . . . . 34
3.6 Sub-criteria weights indicating relative importance to governance . . 35
3.7 Pairwise comparison matrix for selected sub-criteria under the accessibility criteria . . . . . 36
3.8 Sub-criteria weights indicating relative importance to accessibility . . 36
3.9 Global sub-criteria weightings for the analytical hierarchy process . . 37
3.10 Pairwise comparison matrix and weights of alternatives against automation sub-criteria . . . . . 38
3.11 Pairwise comparison matrix and weights of alternatives against error handling sub-criteria . . . . . 38
3.12 Pairwise comparison matrix and weights of alternatives against fault tolerance sub-criteria . . . . . 38
3.13 Pairwise comparison matrix and weights of alternatives against performance quality sub-criteria . . . . . 38
3.14 Pairwise comparison matrix and weights of alternatives against unit testing sub-criteria . . . . . 39
3.15 Pairwise comparison matrix and weights of alternatives against data encryption sub-criteria . . . . . 39

3.16 Pairwise comparison matrix and weights of alternatives against monitoring sub-criteria . . . . . 39
3.17 Pairwise comparison matrix and weights of alternatives against security sub-criteria . . . . . 39
3.18 Pairwise comparison matrix and weights of alternatives against role based access sub-criteria . . . . . 39
3.19 Pairwise comparison matrix and weights of alternatives against availability sub-criteria . . . . . 40
3.20 Pairwise comparison matrix and weights of alternatives against ease of use sub-criteria . . . . . 40
3.21 Pairwise comparison matrix and weights of alternatives against integration sub-criteria . . . . . 40
3.22 Pairwise comparison matrix and weights of alternatives against interoperability sub-criteria . . . . . 40
3.23 Overall weightings of alternatives against the goal . . . . . 41

# Chapter 1

# Introduction

Cloud computing refers to the provisioning of infrastructure, platforms and computational services over the internet. Integrated provisioning is done by cloud service providers for both public and private cloud environments. To date, cloud computing has enabled users access to digital platforms and services remotely and on-demand without needing to have the necessary infrastructure, platform or service locally - as long as the user is connected to the internet. The accessibility of cloud platforms has been a major proponent for the adoption of cloud computing as users are no longer required to invest in physical storage infrastructure, platforms or services to carry out their requirements. It is important to note that there is a dual responsibility between users and providers when it comes to using cloud computing. Cloud service providers are responsible for the continuous operation and maintenance of the underlying infrastructure, platforms and services being provisioned, whereas the end user is responsible for aspects such as connectivity, middleware and end-user security while making use of the provisioned cloud computing services [1].

With the advantages that cloud computing offers in terms of platform as a service, software as a service, and infrastructure as a service, data engineers and data scientists are able to leverage cloud computing for their ETL/ELT (extract, transform and load) and ML (machine learning) requirements and deployments. An ETL/ELT process in the cloud context typically entails establishing data pipelines between sources and sinks. Sources include locations such as on-premises databases, representational state transfer (REST) endpoints, or even supervisory control and data acquisition (SCADA) systems that record and collect real-time data (for example equipment or fleet operating information). Sinks are the desired, destined location for the data, such as in storage accounts (data lakes) or SQL/NoSQL servers.

The purpose of the pipeline is to collect data for either transformation or loading to other systems, depending on requirements. Once data has been collected in a centralised location, it enables users to analyse the data or create models for investigation or business intelligence purposes.

Machine learning is quite useful in this regard as data scientists could make use of operational and financial data (in the context of a mine) to forecast and eventually predict the recovery of the commodities that they mine and process. Similarly, planning and maintenance forecasts and optimisations become possible by consolidating various data sources using ETL principles and leveraging big data for further analysis and modeling using machine learning. As can be seen, ETL and ML in the cloud context can provide significant benefit to users, and thus users need to be aware of the capabilities and usability of various platforms that are currently available.

## 1.1 Literature Review

To provide a background into research that has presently been done in the realm of cloud computing, specifically for deployment of workflows and cloud governance, a literature survey of cloud computing subject matter has been conducted.
In as early as 2008, solution architects had considered the use of cloud computing infrastructure for building out effective workflows [2].The importance and need of on-demand as well as reserve-based provisioning of computational services, automation and security was highlighted as a measure of effectiveness, and this is true to-date. Integration of services, appropriate governance and security, and automation are all critical to having efficient and effective workflows within any cloud computing platform.

Chen *et al.* have presented the need for workflow scheduling on cloud platforms and how it can be implemented [3], however, this was based on using external workflow scheduling platforms. Cloud computing has evolved in a manner whereby scheduling of activities within a workflow has become inherent within a cloud platform, for example, users can trigger data pipelines to run at predetermined times, over a desired time frame [4]. Sun *et al.* have investigated this further by incorporating parallelism into cloud-based workflows and have shown that external workflow engines can be used (together with a runtime service) to leverage

parallelism in order to achieve improvements when scaling workflows [5].

The generally accepted criteria for evaluating cloud platforms and services are: security, performance, accessibility and useability, scalability, and adaptability which can be further explored thanks to a report consolidating past cloud computing evaluations [6]. With the aid of these criteria, users can evaluate which cloud service provider is best suited for their workflow requirements. Abdel-Baset *et al.* have additionally provided individuals with a "neutrosophic multi-criteria decision analysis" methodology to determine the quality of service that a cloud provider offers [6]. A methodology that has been used in many frameworks that aim to evaluate various aspects of cloud computing is the analytical hierarchy process, which aims to decompose multi-criteria decision problems into hierarchical segments for evaluation (both qualitatively and quantitatively) [7] [8] [9]. For cloud security specifically, a "Trust Framework" has been proposed for evaluation of cloud computing technologies [10]. Additionally, the National Institute of Standards and Technology (U.S. Department of Commerce) have developed the "Cloud Usability Framework" that enables users to be informed of what cloud service providers need to adhere to and ensure is available within their platforms and services [11].

In order to improve efficiency of product cycles as well as scalability, serverless computing has enabled cloud services to provide big data focused architecture to users. The architectures are made up of various sub-services (and microservices) that facilitate the end-to-end ETL and ML workflows that a user would need to accomplish their tasks [12].

A significant number of criteria for the general evaluation of cloud services have been considered, however, they remain broadly within the realm of accountability, agility, management, performance, security and usability [13]. Metrics for evaluating cloud-based machine learning services were proposed and reviewed in a separate study [14]. The study included, but was not limited to, metrics such as availability, data capacity, compatibility, documentation, functionality, integration, performance, pricing and usability. In a recent study looking at the selection of cloud services adequate for big data, the proposed important criteria were identified as: architecture, cost, functionality, performance, usability and vendor reputation [15]. In a quality of service evaluation approach study pertaining to managing cloud service evaluation and selection, a similar range of criteria were considered, namely: accountability, agility, assurance, cost, performance and security [16].

## 1.2 Gaps in Literature

Although a copious amount of research has been done in relation to cloud assessment frameworks; scheduling and automation of cloud-based tasks; and offerings by cloud service providers, there is little to be seen regarding studies pertaining to the evaluation of cloud-based platforms for data science workflows. As such, the study will specifically investigate the experience of a user when developing end-to-end data science workflows using three popular cloud computing platforms. As automation is a desirable (if not necessary) element for efficiency and scalability of workflows, the investigation will also look at how cloud-based platforms provide users the capability to build pipelines once-off that are dynamically able to extract, transform, and load the necessary information for reporting going forward.

## 1.3 Research Area

Over the past decade, the has been a rapid adoption of cloud computing services for software, platforms, and infrastructure. This is due to the economic, physical and computational advantages that can be realised. As such, individuals working with data (and the computation activities and requirements thereof) are finding it incredibly beneficial to learn and become well-versed in cloud computing technologies. Thus, the research area will be relevant to the domains of cloud computing, data engineering, data science and analytics. Specifically, the focus of this study will be centered around the extract, transform, and load (ETL) workflows available to a data scientist or engineer.

## 1.4 Problem Statement

With an increase in migration to cloud-based ETL, analytics, ML and reporting solutions, data scientists and analysts need to become comfortable with moving away from tools such as Microsoft Excel and on-premises databases for their data storage and compute requirements. The transition will require individuals to gain confidence in storage and compute technologies offered by cloud service providers. Although there are the giants among the cloud computing service providers (CSPs),

for example Microsoft[1], Amazon[2] and Google[3], consumers can still opt to partner with many other CSPs such as Alibaba[4], Oracle[5], IBM[6], SAP[7] and Workday[8].

As the migration to cloud platforms is a journey that could span many months to years, consumers need to be sure that the services offered by their desired CSP are a good fit for their requirements, and thus need to know the advantages and disadvantages of the different CSPs based on factors such as compute, security, governance, scalability, pricing, availability, etc. Therefore, this study intends to bring clarity to data related service offerings by the cloud computing giants, and to equip data analysts, scientists and engineers with an understanding of the data science workflow capabilities of cloud computing platforms such as Microsoft Azure, Amazon Web Services, and Google Cloud Platform, and shed light on the experience of an entry level cloud technology consumer for data science workflows. The study will further investigate the feasibility of using existing evaluation frameworks for data science workflows, to determine if there is a need for a novel framework in this regard.

## 1.5 Research Question

Data scientists who transition from traditional, static, and on-premises data science platforms and solutions to cloud-based solutions need to be aware of the major types of cloud service providers as well as which cloud platforms would be most suitable for their requirements. As such, the research question that will be answered by investigation and experimentation of various cloud computing platforms is: Do the current evaluation frameworks for cloud computing hold for cloud-based data science workflows? To answer this research question, the subsequent research will look at the end-to-end view of a typical workflow which looks at linking data sources and performing an ETL opereation (i.e., executing activities inherent to a

[1]https://portal.azure.com
[2]https://aws.amazon.com/console
[3]https://console.cloud.google.com
[4]https://home-intl.console.aliyun.com
[5]https://cloud.oracle.com
[6]https://cloud.ibm.com
[7]https://www.sap.com/products/business-technology-platform.html
[8]https://www.workday.com/en-us/products/analytics-reporting/data-hub.html

data science workflow); integration into a ML service; and linking outputs to monitoring or reporting platforms. A crucial aspect to be considered is the level of end-to-end automation that can be achieved for each of the cloud platforms, as this is one of the drivers for adopting cloud-based suites for data related activities.

## 1.6 Research Aims and Objectives

### 1.6.1 Research Aims

The overarching goal of this study is to conduct a review of the data science workflow functionalities that a user would have made available to them. This can be done through grading their experience when executing data science workflows based on the platforms offered by the chosen cloud service providers. In order to achieve the overarching goal, the study will aim to shed light on three critical categories that would affect the experience of a data scientist using cloud platforms from an entry level perspective:

1. Functionality: the technical capabilities inherent to a cloud platform.

2. Governance: the security and administrative capabilities inherent to a cloud platform, for both data flows and pipelines that are developed, as well as for the data that is stored in cloud storage.

3. Accessibility: the availability of service offerings by cloud service providers based on account tiering, and how well they integrate to build end-to-end workflows.

Additionally, the study will include an exploration into the automation capability of the three cloud platforms: as mentioned earlier, this is an integral proponent to migrating to cloud. The study will enlighten readers on the considerations that need to be taken into account when selecting a cloud service provider, and bring familiarity to the automation capabilities (or the platform specific services that facilitate automation) inherent to each of the three cloud platforms.

### 1.6.2 Objectives

In order to achieve the aims of this study and answer the research questions, I have set several chronological objectives that can be treated as milestones for this study. They are summarised as follows:

1. Conduct a practical experiment using the major cloud platforms to establish data science workflows:
   - i Explore the services offered by the chosen CSPs.
   - ii Execute an end-to-end workflow (from source to reporting) on each of the cloud platforms.
   - iii Automate each of the workflows as much as possible with time-based triggers and scheduled activities.
2. Evaluate the development experience on each of the cloud platforms based on functionality, governance and accessibility for workflow activities and data.
3. Conduct a comparative review of the outcomes of the investigation and experiments. using existing evaluation frameworks and recommend whether a novel evaluation framework is required.
4. Develop a framework for evaluating data science workflows on cloud platforms based on components such as functionality, security, governance and accessibility.

## 1.7 Limitations

The scope of this study is limited to the ETL capabilities and subsequent integration into the ML services of the cloud platforms to be reviewed, in terms of functionality, security, governance, and accessibility; and how well these types of workflows can be automated. Aspects such as compute specifications, pricing tiers and subscriptions are not core characteristics for this study and thus have not been focused on as multiple researchers have already conducted in-depth reviews comparing these parameters. Additionally, I would like to keep the focus of this study on the capabilities of the cloud platforms for ETL production tasks, and the above considerations would fall in the realm of a cloud administrator.

## 1.8 Assumptions and Definitions

It is assumed that readers of this study will be data scientists or individuals who are transitioning to cloud based data science platforms with moderate to high experience levels in analytics that have traditionally been done using Excel, python, SQL scripts, etc. This means that they are acutely aware of the hassles of creating daily, weekly, and monthly reports on a manual basis, often having to redo the same type of data transformations from source and the same type of analytics every time a new report is required (in addition to having to use multiple types of software to achieve a final report, for example SSMS $\rightarrow$ Excel $\rightarrow$ PowerPoint). A further assumption is made that such users who are interested in transitioning their data workflows to a cloud platform will likely opt to use one of the major CSPs: Microsoft, Amazon Web Services or Google. Therefore, the cloud platforms provided by these companies will be focused on within the scope of this study. Ultimately, the proposed framework will use these platforms for demonstrating the evaluation process however, users of the framework will be able to use their own choice of platforms to evaluate.

## 1.9 Overview

In summary, this study has investigated, evaluated and reviewed the ETL capabilities and ML service integration of the cloud platforms offered by three popular cloud service providers: Microsoft Azure, Amazon Web Services, and Google Cloud Platform. This study has additionally looked at the extent of automation that can be achieved among these platforms, with the narrative being focused on the experience of an entry level data scientist who is transitioning into cloud-based platforms to execute their workflows. Finally, the analytical hierarchy process was assessed to determine its feasibility for building a framework for evaluating criteria pertaining to cloud based platforms specifically for data science workflows, in order to determine if a novel framework could be established for identifying suitable cloud computing platforms for data science workflows.

The following chapters in this report delve into the approach used to address the research question, the experimentation conducted during the research process, the results obtained from the experiments and their subsequent review and evaluation,

a discussion on the implications of the results obtained, and finally, a conclusion in terms of the overall outcomes of this study and how they have answered the research question. Specifically, Chapter 2 looks at the research methodology: providing detail on the data used, data science workflows built for experimentation purposes, the criteria used for evaluation, a step-by-step outline of the evaluation framework method used (i.e. the analytical hierarchy process), and an overall approach towards how the analysis of the results was conducted. Chapter 3 identifies the results obtained through experimentation, review and evaluation, and provides commentary in the form of a discussion to elaborate on the meaning of the results obtained within the context of the research question and finally explains how the research question was answered. Chapter 4 provides an overview of this study and its outcomes through means of a conclusion, and presents guidance on future work that could be done based on the learnings and outcomes of this research.

# Chapter 2

# Research Methodology

A qualitative review of criteria pertaining to the top cloud computing platforms required an in-depth understanding of the various ETL processes available across the cloud platforms. Therefore, the study necessitated the exploration of the services related to data science workflows prior to building out the workflows to gain familiarity with their functionality, governance and accessibility related capabilities. The exploration yielded metrics that were present among all platforms being explored, thus making these the criteria that were used within the evaluation framework. As such, each of the cloud computing platforms that were utilised explored the following capabilities, i.e. their sub-criteria:

**Functionality:** Automation, error handling, fault tolerance, performance quality and unit testing.
**Governance:** Data encryption, monitoring, security and role based access.
**Accessibility:** Availability, ease of use, integration, interoperability.

A focus was placed on services that enabled data engineering and analytics in order to remain in scope of the proposed research. The accessibility criteria were included in order to supplement the evaluation as a major driver of cloud adoption is the ability to learn and use the platform relatively easily to promote its adoption, as well as to have certainty that it is able to integrate well with applications and tools external to the platform, thus mitigating vendor lock-in. A similar workflow was developed using the various platforms and the experience gained while building out the end-to-end workflows was used as a basis for reviewing each of the criteria and sub-criteria from an entry level data scientist perspective, thus informing the vaulation of each of the attributes within the framework.

The analytical hierarchy process was subsequently used to develop a framework that evaluated the criteria pertaining to cloud platforms based on the valuations of the criteria and sub-criteria previously mentioned. Although Microsoft Azure, Amazon Web Services and Google Cloud Platform were used as the cloud platforms in this study, the actual development of the framework used anonymised platform providers, i.e. CSP1, CSP2 and CSP3, as the intention of this study is to ultimately propose a feasible, dynamic evaluation framework to identify suitable cloud computing platforms for data science workflows rather than to compare Azure, AWS and GCP in an official and categoric manner.

## 2.1 Research Design

In order to successfully achieve the aims and objectives of this study, an organised, method-based approach was used: firstly, it was necessary to identify which criteria and sub-criteria would be beneficial to the evaluation of a cloud computing platform specifically for data science workflows. Literature was surveyed [13-16] and research was conducted into factors of importance that a data scientist needs to be wary of when it comes to ETL workflows, and these factors were subsequently grouped logically into the following categories: functionality, governance and accessibility. Technical factors were grouped into the functionality criteria, security and administrative factors were grouped into the governance criteria, and factors relating to availability and workflow integration were grouped into the accessibility criteria.

Next, a firm understanding of the functional, governance and accessibility capabilities pertaining to each of the cloud computing platforms used in the experimental phase was required. For this, accounts were set up on the respective cloud platforms to begin building the workflows using the ETL services offered by the CSPs, which is detailed further in the Methods section of this chapter. This was possible thanks to each of the platforms' complementary credits (albeit temporary) that could be used for services not within the free tiering spectrum.

Once the workflows were established and a familiarity with the cloud platforms had been gained, it was possible to begin the evaluation process of the platforms based on the importance of each of the established criteria and sub-criteria, keeping in mind that the perspective of an entry level data scientist was being used. In

the context of this study, an entry level data scientist is defined as an individual who is responsible for the technical components of data science workflows with a focus on the development of pipelines and models, ensuring that they work and can be automated so that manual intervention within the workflow is not required. The expectation at this level is that the individual would require guidance on best practices and translation of technical aspects of the workflow into outcomes that are consumable by users in a business-friendly manner. In contrast, an advanced level (or senior) data scientist would be expected to understand the business and stakeholder requirements, thus planning and tailoring solutions that are able to fulfill requirements from inception. Advanced data scientists are experienced with methodologies such as agile, DevOps, source control, and continuous integration and continuous delivery (CI/CD) and can thus approach the development of solutions from a technical, business and operational efficiency perspective. As the experimentation is focusing on the use of cloud computing platforms for the general development of data science workflows, the perspective of an entry level data scientist suffices.
To deduce the relative importance of criteria, Saaty's fundamental scale of absolute numbers [17] was used to attribute a numeric rating to the linguistic importance of each criteria (and sub-criteria) compared to another, as shown in Table 2.1. As a note, even values of importance indicate there is no absolute designation of importance between a pairwise comparison, for example if the importance of A-B is 2, then it means that A is valued moderate to equally as important as B, but not absolutely equally or moderately more important than B.

TABLE 2.1: Saaty's fundamental scale of absolute numbers [17]

| Intensity of Importance | Definition |
|---|---|
| 1 | Equal importance |
| 3 | Moderate importance |
| 5 | Strong importance |
| 7 | Very strong importance |
| 9 | Extreme importance |

Understanding the relative importance of each sub-criteria against another within a given criteria, as well as the relative importance of one criteria over another, made

it possible to use the analytical hierachy process to develop a framework for evaluating the cloud computing platforms for data science workflows based on the experience of an entry level data scientist. The analytical hierarchy process used in this study is further detailed in the Methods section of this chapter. Ultimately, following the analytical hierarchy process, the goal of the framework is set to identify which cloud service provider would fit the needs of the consumer based on the criteria and sub-criteria, i.e., which CSP should the consumer choose, with each alternative for selection being the cloud computing platforms, the framework is able to generate the overall importance weightings for each of the CSPs, with the highest weighting indicating the appropriate platform based on the consumer's requirements.

## 2.2 Data

The data that was used for this study is contained in the Northwind database [18], which is a set of open-source product and sales related tables provided by Microsoft. Specifically, the Northwind database contains customer (91 records), employee (9 records), order (830 records), product (77 records), region (4 regional records and 53 territorial records), supplier (29 records) and shipping (3 records) information for a hypothetical business, stored across a set of 13 tables. The information contained in this database was used in the ETL portion of the experimentation, which focused on developing data pipelines between the source server and a cloud-based database, upon which the data science workflows were created to perform data engineering and analytics before presenting the final information in a report (demonstrating the integration to a BI tool such as PowerBI).

The goal of the ETL process was to understand how the cloud platforms enable users to address and make use of each of the sub-criteria when building data science workflows. For example, when an ETL process was set up to extract data from a source database that contained normalised tables such as Products, Categories and Orders; consolidated the source data into subsequent datasets such as "Products by category" or "Sales by category" for analytical purposes; and loaded the transformed data into an analytical database, what considerations can be made in terms of each of the sub-criteria? Similar questions include:

- Can performance quality and speed be tuned?

- Is there a possibility of adding fault tolerance to a workflow in the case of a failed run (e.g. changes in source schema)?
- Can debugging of workflows be done on the fly to ensure they are working as intended before publication?
- Can customised or parameterised triggers be established for workflows to avoid manual intervention in the running of workflows?

## 2.3 Methods

Prior to the development of the data science workflows on the various cloud platforms, there were several pre-requisite activities that needed to be accomplished. The first was to establish a database that would host the Northwind database, which contained the source data that was used in the ETL workflows. A SQL server was created to host this database, and the data was imported using a SQL script accessible from the Microsoft website[1]. It is an open source, templated script that recreates the Northwind database and its assets within the SQL server. Once this was done, it was necessary to create profiles on Azure, AWS and GCP for criteria testing - a student account was created on Azure which offers $100 in credits for 12 months and free access to selected services; a standard account was created on AWS with access to only free tier services; and a standard account was created on GCP with $300 credits available for a 3 month period. Due to the reasonable amount of resource provisioning within the free tier of these platforms, and the small size of the source data, it was possible to explore the various services within these platforms and develop the workflows in order to gain familiarity with the ETL services and interact with the functionalities associated to the sub-criteria of interest.

### 2.3.1 Creation of Data Science Workflows

Once the prerequisites were established, data science workflows were created using the cloud platforms. This section of the report details how each of the workflows

[1]https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/sql/linq/downloading-sample-databases

were set up in order to investigate the sub-criteria. This experience was necessary when making use of the analytic hierarchy process to enable the ranking of relative weightings of criteria and sub-criteria, and to obtain a framework for cloud platform selection specifically for data science workflows.

**Microsoft Azure**

The data science workflow for Microsoft Azure was created using the following set of services:

- A SQL server that was used to host the databases being used within the MS Azure workflow.
- Two SQL databases, one that contained the source information (i.e. the Northwind database) and another that was used as the target database when loading the transformed datasets to be used for subsequent analytics (i.e. BI reporting source).
- An instance of data factory (v2) to carry out the ETL processes as well as set triggers for automated movement and transformation of data.

For the workflow itself, the focus is on Azure data factory (v2) (ADF) as it is Microsoft Azure's go-to service for ETL. Within data factory, a linked service[2] was created to establish a connection between the data factory and the database containing the Northwind datasets (in structured, tabular format). This allowed the data factory instance to access the tables within the Northwind database, thus enabling it to copy entire tables or portions of tables (using SQL queries) from the source to a sink (the target database in this case) using a "Copy to" activity, as well as to read tables within the Northwind database and apply a number of transformations depending on user requirements to the data, and then sink it to a target using "data flow" functionality. The distinction here is that the "Copy to" activity is a single activity block whereby users can copy entire tables as they are or use SQL scripts or stored procedures, to carry out transformations on the source before copying it to the destination, while the dataflow provides a canvas to carry out various transformation step-by-step in a graphical manner which is beneficial for individuals who

[2]https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services?tabs=data-factory

are not well versed in scripting using SQL.

The experimental workflow used the data flow functionality to apply several transformations to the source data before sinking it to the target database. The reason a data flow was used instead of the "Copy to" activity was that the transformations available within the data flow allow for users to ensure that both inserts and updates are possible instead of simply writing all the information from the source into the target database, as well as to visualise the transformation process in a step-by-step manner. Mapping data flows are especially helpful in excluding duplicate entries when the source does not contain primary keys or time intelligence fields (e.g. to and from dates). Once this data flow was created, a pipeline was established to invoke the data flow, and a daily trigger was set up to automatically run the data flow so that the ETL process can be carried out from source to sink without manual intervention.

Azure's Machine Learning Studio (ML Studio) can be leveraged for subsequent machine learning requirements. Azure's ML Studio offers three methods of authoring ML models: through Jupyter notebooks that allow for building of machine learning models using python scrips; through the studio's automated machine learning service (AutoML) which offers a predefined workflow that users can follow to set up thier ML model - which involves selecting a dataset, identifying the compute instance or cluster to be used by the service, selecting the type of machine learning task (regression, classification or time series forecasting) and assigning a validation type; and the final method is using the designer functionality to create user defined machine learning models using an interface, similar to how ETL processes are created using data flows in ADF. ML Studio is able to source data from local storage, Azure blob or data lake storage, and Azure SQL databases, thus the analytical database used to sink data post ETL was used as a source. Although additional environments can be installed, ML Studio comes stock with environments such as pytorch, tensorflow, sklearn, and xgboost. Using autoML, once the ML model has been created and validated, Azure's ML studio will run a number of experiments using different machine learning methods depending on the type of task selected (however, designer and notebook implementations of machine learning models will create a ML model specific to the model specified by the author e.g. a long-short term memory model for forecasting) and returns the results of each of the experiments as models, of which the user can decide to deploy. For a production

deployment, compute using a real-time inference is necessary for deployment[3], as it enables consumers of the model to access the deployed model by sending inputs to the created application programming interface (API) endpoint associated to the ML model, and receiving outputs in real time.

For the final step of the workflow, Power BI was used to create a dashboard based on the data contained in the analytical database. The Power BI service has the functionality to schedule automated refreshes[4], which were set up to run shortly after the data factory trigger. The establishment of the scheduled refresh functionality ensures that if any new data is brought into the analytical database using the ETL process, then it would show up in the dashboard once the refresh is complete. As PowerBI is a Microsoft product, it has in-built compatibility with Azure ML Studio, thus models that are deployed and running within the studio can be called and applied to datasets within PowerBI.

Thus, an end-to-end data science workflow was established using services offered within Microsoft Azure. A key note here is that once the initial effort has been put to establishing the workflow, as long as all the services within the workflow are running, no additional manual intervention is required.

#### Amazon Web Services

In AWS, the Amazon Glue Console is used to establish ETL workflows. Simple "copy to" activities or complex code-based transformations can be scripted within the Glue Console to run on demand or according to designated schedules, or the Glue Studio can be used as a canvas to establish the transformation logic of the workflow graphically using "nodes" whereby each step in the ETL process is represented by a node (each with their assigned node types such as data source, transform, or data target) that makes up the entire ETL workflow. However, before an ETL workflow can be set up, the data source required within the workflow needs to be added to the AWS data catalogue in Glue – this ensures that the schema attributed to a data source is known when performing the job. Glue is able to do this using its "crawler" functionality[5], which connects to a data store to identify the

[3]https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/ml-deployment-inference

[4]https://docs.microsoft.com/en-us/power-bi/connect-data/refresh-scheduled-refresh

[5]https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html

schema and subsequently creates metadata tables within the data catalogue for the source data.

Thus, when creating the ETL workflow in AWS, the following process was followed: A crawler was established to read the Northwind dataset tables using a Java Database Connectivity (JDBC) connection to the server that hosts the source database. Appropriate identity and access management (IAM) role policies needed to be attributed to the crawler in order for it to have permission to access the source; AWS has a standard managed policy that was used to give read access, namely AWSGlueServiceRole, however, it is possible to create a custom IAM role for the crawler with the desired permissions. Once access had been granted to the crawler, the scheduled run frequency for the crawler was designated: for this experiment, a daily schedule was selected. Next an output destination for the crawler needed to be identified in order to store the metadata tables generated by the crawler, which was done by creating a database within Glue to store the outputs of the crawler. This database was then used in the ETL process when creating the job that is responsible for the desired workflow.

As the crawler had been created, it became possible to use the source data to create an automated ETL workflow within Glue. However, for this to be done, it was necessary to create a job for the process. Once again, an appropriate IAM policy was required in order to give the job access to the assets being used within the workflow (read permission from source, write permission to targets, access to the data catalogue as well as the ETL scripts that are generated by Glue). Once the job has access to the source and target, the mapping between the two can be seen visually in order to verify that the schema have been identified correctly by the crawler, as well as to add additional columns that contain metadata such as the time of the run or name of the source, for example. Since this experiment involved creating a new workflow, an existing script was not used as a template for the job; rather, Glue was advised to generate a python script based on the ETL logic that would make up the ETL workflow. As the intention for this workflow was not to be once-off, the script needed to be stored for reuse. An Amazon S3 bucket is the default service for storing and accessing the ETL scripts that are generated by Glue. Upon creation of the job, a trigger was assigned to the job for it to run daily without manual intervention.

Creating the crawler, job and trigger ensured that the building blocks for the ETL

workflow were established. Using the Workflows functionality[6] in the Glue Console, a logical flow was orchestrated to define the relationship of how the crawler, job and trigger need to interact, i.e., when the trigger is set off, the crawler should run in order to update the metadata tables to reflect the latest information at source before running the job to extract information, transform it, and then load it into the target database. This ensured that the latest data was being extracted from the source at the time of the trigger.

For ML workflow requirements, Amazon's Sagemaker is the go-to machine learning service for data scientists, business analysts and machine learning engineers. As with other AWS services, establishing appropriate IAM roles is essential to ensure that the necessary access to resources is available. Specific users can be created within the environment with customised execution IAM roles to manage activity within Sagemaker. For machine learning deployments, users need to ensure that Sagemaker has access to the required S3 bucket(s). Amazon Sagemaker provides two distinct methods of developing and deploying ML models: using jupyter notebooks (referred to as Studio) with preset python kernels such as base python, spark, MXNet, pytorch and tensorflow (there is also an iteration called RStudio which uses R instead of python); as well as a no-code interface for users who want to make use of a graphical user interface (referred to as Canvas). Canvas is limited to data sources such as local storage, S3 buckets, snowflake and Amazon redshift, therefore, the studio was used to import *sqlalchemy* which is a python SQL toolkit to access the analytical database and capture data into a pandas dataframe. In Sagemaker, once a model has been created, trained and validated, it can be deployed to an inference endpoint[7] for real-time consumption.

Finally, in order to have an end-to-end workflow using AWS, Power BI was once again used to access the target database to visualise the transformed data contained within the analytical database. In order to incorporate the output of Sagemaker's machine learning model into the workflow, the python script used to generate the model outputs can be run within PowerBI, with the caveat that the outputs are contained within a dataframe[8].

[6] https://docs.aws.amazon.com/glue/latest/dg/workflows_overview.html
[7] https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html
[8] https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-python-scripts

**Google Cloud Platform**

Cloud data fusion is the go-to service for graphical user interface based ETL and ELT workflows in Google Cloud Platform (GCP). It is a comprehensive platform that allows users to establish pipelines without the need of writing scripting to perform ETL operations, and rather, provides a canvas to create workflows based on step-by-step nodes that are designated as sources, transformations, analytics, sinks, conditions and actions, as well as error handlers and alerts. This allows the user to create an entire workflow on a single canvas.

In order to use cloud data fusion, the IAM role for cloud data fusion Admin needed to be assigned to the service account for the cloud data fusion instance in GCP, which allows read and write access to the instance, enables it to perform the required ETL operations, and grants access to the resource manager for projects. Once the required IAM role was assigned, the cloud data fusion hub was used to install the appropriate drivers and plug-ins that data fusion requires to access sources and sinks, in this case, it would be the SQL server containing the Northwind database, as well as the server containing the analytical database that contained the final outputs for use in BI reporting. In order to access the Microsoft SQL server, a JDBC driver plug-in was deployed, specifically version 6.0 of the java runtime environment. The JDBC driver plug-in made it possible for the wrangler in data fusion, which enables user to explore, transform, cleanse and enrich data using a code-free environment, to access the required sources.

The cloud data fusion studio was then used to establish the workflow, using a SQL server source node that read the Northwind database, the wrangler node to enable the required transformations, the analytical join node to combine various sources, and the SQL server sink node to write the final outputs to the analytical database. Once the design of the workflow was established, the scheduler was used to create a daily trigger for the pipeline, and then deployed for operation. As this was the first time creating a workflow, the nodes needed to be designed from scratch, however cloud data fusion allows users to create templated pipelines or share pipelines using the import and export feature. Jobs that are successfully created can then be monitored with cloud data fusion monitor which provides operational details pertaining to a specific job.

For machine learning modelling and deployment, the Vertex AI service hosted by GCP can be used. First, the Vertex API needs to be enabled to ensure that the

user has accepted the terms and conditions of the service[9]. Once the API was enabled, a standard Jupyter notebook was created with numpy, scipy and sklearn and pandas environments, however it is possible to establish notebooks with environments such as tensorflow, pytorch and apache beam. The interface allows users to install additional packages subsequent to creating a notebook with specific environments using the terminal (similar to how pip install commands are run on command prompt or anaconda prompt). As tabular datasets in vertex AI are limited to local csv files, cloud csv files or Google BigQuery tables, the Models functionality[10] which allows for autoML deployments could not be used as the analytical database was on a SQL server. Thus, it was once again required to use the *sqlalchemy* package within the Jupyter notebook to access the analytical database and load the required datasets into a dataframe so that they could be used for machine learning.

For the final step of the workflow, i.e. BI reporting, GCP offers data studio which is a visualisation tool that can be used to create dashboards, charts and reports; albeit not as robustly as PowerBI. Due to the desire to test the interoperability criteria, PowerBI was also used as the BI reporting tool for this experimental workflow rather than Google's data studio, which used the outputs from cloud data fusion's ETL workflow that were stored in the analytical database to populate the visuals. The link between Vertex AI models and PowerBI is similar to the interoperability of Sagemaker and PowerBI in that a python script can be used to load ML model outputs into PowerBI.

### 2.3.2 Analytical Hierarchy Process

The context of this study is within the domain of a multi-criteria decision making problem, where a selection of alternatives need to be evaluated based on a set of criteria. It is required that the evaluation of criteria should aid the user in deciding which alternative is the most favourable for a given requirement. The analytical hierarchy process can be used to develop a framework for decision making pertaining to the comparative review of a multitude of criteria that pertain to a set of cloud platforms for data science workflows. To ensure that the goal is bound to the user's requirement, the perspective of an entry level data scientist who has the

[9]https://support.google.com/googleapi/answer/6158841?hl=en
[10]https://cloud.google.com/vertex-ai

goal of deciding which platform would be most suitable for the user's data science requirements is set.

The analytical hierarchy process (AHP) provides decision-makers with a framework to attribute mathematical valuations to an individual's subjective opinion pertaining to a set of criteria in an overall evaluation, thereby enabling users to mathematically deduce which alternative is most suitable for the user against a defined goal [17]. It does this by associating relative measures among a pair of criteria that are relevant to the user, thus propagating a pairwise matrix of relative importance for each of the criteria. This can be developed further by means of using sub-criteria, which provide a granular numeric representation of importance among the hierarchy of attributes relevant to the user. As the process is based on subjective beliefs, aggregated inputs from a myriad of users is beneficial, especially for important decision making. This study will explore the use of the analytical hierarchy process to develop a framework for evaluating cloud computing platforms for data science workflows. Upon successful development of a robust framework, further work can be done to bolster the usability of the framework through aggregated weightings (based on multiple inputs) as well as dynamically changing attributes based on user preferences and requirements.

There is a logical, systematic approach that is taken when using the analytical hierarchy process, which will be detailed in this portion of the report. The aim is to convert linguistic representations of importance between a pair of criteria into numerical equivalents, and subsequently derive the weightings of each of the sub-criteria in relation to the pool of sub-criteria. The weightings are used to calculate which platform (alternative) has the highest score, ultimately indicating that based on the chosen criteria, the platform with the highest score would be the most appropriate for the user. The method of using the analytical hierarchy process is as follows:

**Step 1: Establish a goal, criteria and alternatives**

In order for the analytical hierachy process to be used, three key concepts needed to be established [17]. The first was a goal, i.e. what is the targeted outcome that needs to be addressed? The goal was set to be the cloud computing platform that would be most suitable for an entry level data scientists needed to construct and

automate production-level data science workflows. Next, the criteria needed to be selected. These were the attributes that would be used to review the cloud platforms in terms of the desired outcome, namely: functionality, governance and accessibility. In order to get a deeper valuation of criteria pertaining to the outcome, these criteria were broken up into sub-criteria that spoke to the components to be evaluated within the context of each of the criteria:

**Functionality:**

- Automation: the capability of ensuring that the ETL workflows are able to be performed independently as per designated rules [19].
- Error handling: the capability of assigning specific operations when exceptions to the norm are encountered within a workflow [20].
- Fault tolerance: the ability of the workflow to continue operating when encountering failure at a component within the workflow, for example, at a specific activity within the workflow [21].
- Performance quality: the ability to optimise workflow operations to increase performance through partitioning, broadcasting joins, or choosing specific compute types for operational clusters [22] [23].
- Unit testing: the inherent capability to debug and trial workflows without affecting external systems or services [24].

**Governance:**

- Data encryption: the ability to encrypt data at rest, as well as offer masking of sensitive information that needs to be present within a workflow [25].
- Monitoring: the ability to monitor jobs and operational metrics of workflows to have a summative overview of performance and successful operation of workflows.
- Security: the compliance to international standards and regulatory assurances such as SOC certification (systems and organisational controls), ISO 27xxx, etc. [26].

- Role based access: access controls that can be assigned and managed by platform owners or administrators to ensure that users have minimum required access necessary to perform their tasks [27].

**Accessibility:**

- Availability: ensuring that the platform is available at anytime and is accessible from anywhere, and is backed by a service level agreement by the platform provider.
- Ease of use: the ability for new users to intuitively make use of services when establishing data science workflows.
- Integration: the extent in ability of an ETL platform to be able to connect to external services, platforms and tools to create an end-to-end workflow.
- Interoperability: the ability for the ETL platform to have real time connections for efficient data movement across independent platforms [28].

The final concept that was established before using the analytical hierarchy process was that of the alternatives available for answering the question contained in the goal. These are the cloud platforms that could potentially be used by the data scientists to conduct their tasks. Because this study intends to develop a framework for comparative reviews of cloud platforms for data science workflows rather than provide a definitive comparison of existing cloud platforms, the alternatives used in this AHP model were classified as CSP1, CSP2 and CSP3. Thus, the preliminary model can be illustrated as shown in Fig.2.1. As can be seen, the criteria bear a weighting on the overall goal that is desired, however, each sub-criteria set also bears a weighting in terms of its parent criteria. Thus, the total weighting of each sub-criteria group is bounded by the weighting of its parent criteria [17]. Finally, each alternative will have a weighting towards the goal based on how the review of the services and capabilities determined during the practical construction of the ETL workflows scored. The alternative with the highest score will indicate that based on these attributes, the data scientist should choose the platform attributed to the cloud service provider with the highest final weighting.

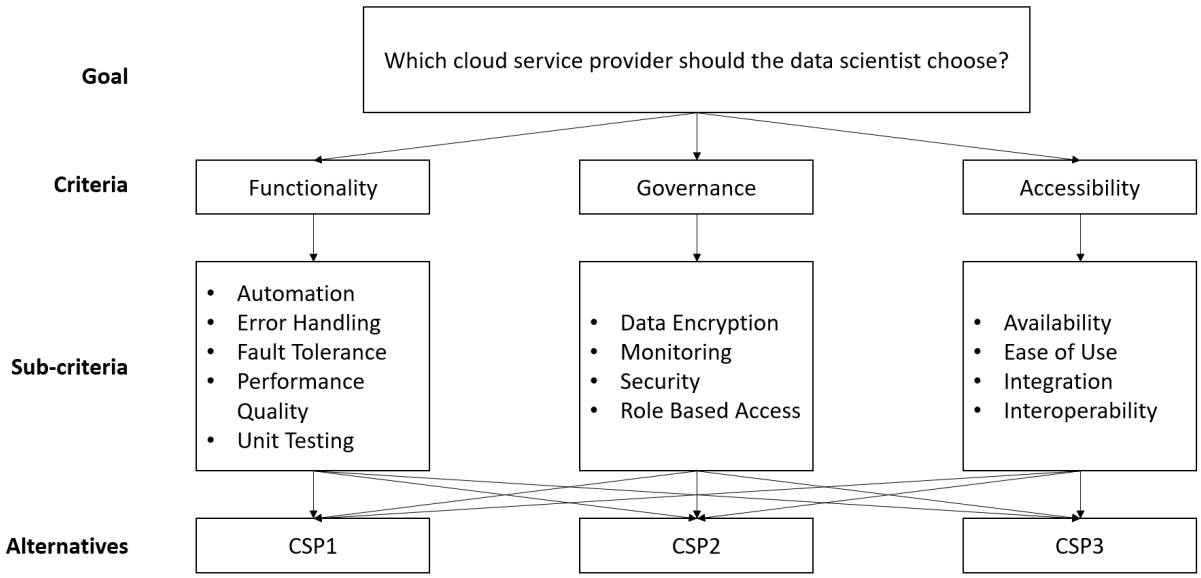


FIGURE 2.1: AHP model for determining cloud platform based on data science workflow requirements

**Step 2: Conduct pairwise comparison between criteria**

Based on Saaty's fundamental scale of absolute numbers, a matrix is established to gather the values of importance between a pair of criteria, as shown in Table 2.2.

TABLE 2.2: Matrix of pairwise comparison of criteria

| | Functionality | Governance | Accessibility |
|---|---|---|---|
| Functionality | 1 | X | Y |
| Governance | 1/X | 1 | Z |
| Accessibility | 1/Y | 1/Z | 1 |

As explained in Section 2.1 of the report, and illustrated in Table 2.1, the parameters $X$, $Y$ and $Z$ are the numerical representations of the relative importance of one criterion compared to another. Here, based on the scale shown in Table 2.1, $X$ would indicate the intensity of importance of functionality compared to governance, $Y$ would indicate the intensity of importance of functionality compared to accessibility, and $Z$ would indicate the intensity of importance of governance compared to accessibility. As a corollary, the intensity of importance of governance compared to functionality would be the inverse of the intensity of importance of functionality compared to governance, hence it is calculated as $\frac{1}{X}$. It can be noted

that the number of pairwise comparisons, $N$, that are required for each criteria or sub-criteria matrix can be shown as:

$$N = \frac{n(n-1)}{2} \tag{2.1}$$

where n is the number of criteria or sub-criteria within a set.

**Step 3: Normalise the pairwise matrix**

The pairwise matrix is normalised by obtaining the column-wise sum of the numeric importance of each of the criteria calculated in the previous set and dividing each element $v$ within a column by the column-sum, as shown by:

$$\bar{v}_{ij} = \frac{v_{ij}}{\sum_{j=1}^{n} v_j}. \tag{2.2}$$

**Step 4: Obtain criteria weights**

The criteria weights are obtained by aggregating the row-wise normalised values. This yields the relative weighting of a criteria in relation to the goal. The highest weight will indicate that the criteria is relatively the most important whereas the lowest weight is relatively the least important. It is noted that this does not negate the absolute importance of a criteria, as the importance of criteria are being evaluated based on their importance against other criteria. The formula for obtaining criteria weights ($w$) is given by:

$$w_i = \frac{\sum_{j=1}^{n} \bar{v}_{ij}}{n}. \tag{2.3}$$

**Step 5: Obtain sub-criteria weights**

The weights of the sub-criteria can be obtained in the same way that the weights of the criteria are obtained, keeping in mind that pairwise comparisons are done for sub-criteria within a set of a given criteria. For example, a pairwise comparison of relative importance is done for automation, error handling, fault tolerance, performance quality and unit testing relative to each other only, as they fall under the functionality criteria. Similar comparisons are done for the sub-criteria pertaining

to governance and accessibility. Thus, to obtain the sub-criteria weightings, steps 2-4 are repeated for each sub-criteria set.

**Step 6: Obtain global weights**

Global weights refer to the relative weighting of a sub-criteria to the overall goal [17]. A global weight is a dot product of the weight attributed to the parent criteria and the weight attributed to the sub-criteria. Thus, having 13 sub-criteria means that the analytical hierarchy process should have 13 global weights indicating how important a given sub-criteria is to the overall goal.

**Step 7: Review each alternative against each sub-criteria using the pairwise matrix**

Based on the experience gained during the creation of the data science workflows, each of the sub-criteria are given an alternative-based weighting, which indicates that for that specific sub-criteria, this is how a given CSP's platform was rated compared to the others. This was done by repeating steps 2-4 based on the alternatives for a given sub-criteria; where Equation 2.3 yielded the relative weight for each alternative, for a given sub-criteria. Once again, the number of matrices that will be generated in this step are equal to the number of sub-criteria being used to evaluate the platforms from a data science workflow perspective.

**Step 8: Obtain overall weightings of each alternative to select recommended CSP**

As each of the weightings related to the alternatives for each sub-criteria are normalised, the sum of the dot product of a weighting per sub-criteria and the global weighting of the sub-criteria yields the overall rating for each of the alternatives. The output is a $m \times 1$ matrix where m is the number of alternatives, and the sum of the values in the matrix are equal to 1. Thus, using the analytical hierarchy process, the most feasible cloud platform can be recommended for an entry level data scientist to use for data science workflows based on the sub-criteria chosen.

## 2.4 Analysis

Upon completion of the data science workflows, the experience gained while creating the end-to-end workflows on each of the platforms was reviewed in terms of the sub-criteria to be used in the analytical hierarchy process. Although each of the platforms have their own ways of building the workflow and making use of the functionality sub-criteria, for example, the useability of those sub-criteria (setting up triggers for automation, error handling and fault tolerance processes, performance tuning mechanisms and debugging of workflows) was a key driver in the review and evaluation of the platforms. Indeed, the perspective of this study is from that of an entry level data scientist, thus useability of features in an intuitive manner is quite important.

The base multi-criteria evaluation framework, analytic hierarchy process was used to weigh the desirability and importance of the different criteria and sub-criteria for each of the cloud computing platforms in the scope of the investigation as the methodology for developing an evalution framework using AHP worked well with the requirements of this study's requirement. As per the name of the framework, each of the evaluation categories (functionality, governance and accessibility) were set to be the parent nodes of the sub-categories that fall below them, thus creating a logical hierarchy that could be systematically assessed, with the initial goal of converting linguistic significance of a criteria and sub-criteria into a numeric value in a pairwise fashion. A suitable comparison scale was used to attribute a relative importance rating to each criteria and sub-criteria pair, thereby creating a comparison matrix that represents the relative importance of each element in the hierarchy in relation to another element.

Once the comparison matrix was established, an eigenvector was used to attribute the significance of each criteria to the overarching requirement (i.e. which platform has a higher desirability for data science workflows based on the criteria in question). This eigenvector is known as the priority vector and is a key metric in the AHP as it represents the relative weights of each criteria, enabling users to have a relative rating of each criteria and sub-criteria for each of the three cloud computing platforms. Furthermore, the priority vector was used to measure another key metric, namely, inconsistency which needed to be minimised. An inconsistent

weight attributed to the criteria or sub-criteria would mean that the weightings calculated are flawed, and the evaluation process has been done incorrectly. Thus it is vital to ensure that inconsistency is within acceptable bounds (which is measured by the consistency ratio) before using the framework to attain global weightings for the sub-criteria and the final weighting for the alternatives.
For example, inconsistency would occur when the functionality criteria has a higher priority than the governance criteria (F>G) and the governance criteria has a higher priority than the accessibility criteria (G>A), but it is discovered that the accessibility criteria has a higher priority than the functionality criteria (A>F). Using the sum of the dot product of the pairwise comparison matrix and the criteria weight matrix, $\lambda_{max}$ (the largest eigenvalue [29]) is calculated, which is used to obtain the consistency index attributed to the priority vector:

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{2.4}$$

where $n$ is the number of evaluated criteria or sub-criteria.

The consistency index is deemed acceptable when its ratio to the random consistency index is less than 10% [30], where the random consistency index per number of criteria is given as:

TABLE 2.3: Random consistency index parameters based on number of criteria [30]

| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| RI | 0 | 0 | 0.58 | 0.90 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.49 |

## 2.5 Limitations

The entirety of services and microservices was not taken into account during the course of this study as a majority of services in each of the cloud-based platforms have niche applications to certain types of user requirements. The general method of investigation was Source > ETL > Reporting on each of the cloud platforms, in order to have a fair basis of consideration for the subsequent review. The scope of

this study remained on data science workflows and the direct cloud-based services that are required to establish and automate end-to-end workflows from a data science and engineering perspective.

For this study, the establisher of the data science workflows (the developer) is responsible for evaluating the relative importance of criteria against one another, as it is based on their experiences while developing the workflows among the different cloud platforms. The developer in this study was able to do the evaluations as he is familiar with the platforms, having two years of project experience using cloud computing platforms in industy, as well as having done cloud practitioner courses in Azure, AWS and GCP. It was necessary to ensure that the developer focused the evaluations to the established goal of selecting the platform that was most suitable for the workflow requirements in terms of automated ETL $\rightarrow$ ML $\rightarrow$ BI Reporting. The analyical hierarchy process iterates evaluations across the hierarchy of criteria, that is, weights are generated among sub-criteria, criteria, and alternatives. Therefore, a single AHP evaluation was able to yield weights indicating the most favourable alternative (cloud plaform) to achieve the goal of cloud platform selection for data science workflows.

## 2.6 Ethical Considerations

From a data perspective, there were no ethical considerations that needed to be made as the data that was used is generic, open-source procurement information that can be obtained and used by any public or private entity, thus the data did not require ethical clearance. The terms of service for the cloud platforms within the scope of this study have been reviewed, and there is no requirement for initial approval of platform use for evaluation or sharing of research methodology and content, as decisive statements or conclusions pertaining to the evaluation of specific cloud service proviers are not drawn from the research experiments or outcomes (see Appendix A). Furthermore, the alternatives (CSPs) required for the framework have been anonymised. In terms of this study as a whole, an unbiased view was maintained throughout the investigation as any bias would risk the legitimacy of the review and evaluation process.

# Chapter 3

# Results and Discussion

The analytical hierarchy process was used to develop a framework for reviewing criteria pertaining to cloud computing platforms for data science workflows. The process entailed the construction of ETL pipelines using the cloud-based ETL platforms of various (anonymised) cloud service providers to investigate attributes such as automation, error handling, fault tolerance, performance quality, unit testing, data encryption, monitoring, security, role based access, availability, ease of use, integration and interoperability. The attributes were subsequently used to review and evaluate the ETL platforms relative to one another based on the experience of an entry level data scientist, to determine which cloud platform would be the most suitable in terms of selection for carrying out data science workflows. In the evaluation process, Saaty's fundamental scale of absolute numbers was used to attribute numerical values to linguistic associations of importance to a pair of relevant attributes. This conversion of linguistic valuation to numeric valuation ensured that matrices containing the relative importance of criteria and sub-criteria were able to be constructed, which aided in assigning global weights to sub-criteria in terms of their overall importance. Subsequently, a similar approach made it possible to evaluate each of the cloud platforms relative to one another in order to determine the platform with the highest overall weighting in terms of the chosen sub-criteria, thus serving as a recommendation for the platform that would likely be most suitable for the data scientist.

## 3.1 Results

Initially, the analytical hierarchy process was used to determine the weighting of each criteria in terms of its importance to the goal of cloud platform selection. First,

a pairwise comparison matrix was created to illustrate the relative importance of each criteria, as shown in Table 3.1.

TABLE 3.1: Pairwise comparison matrix for selected criteria

| | Functionality | Governance | Accessibility |
|---|---|---|---|
| Functionality | 1.00 | 4.00 | 2.00 |
| Governance | 0.25 | 1.00 | 0.33 |
| Accessibility | 0.50 | 3.00 | 1.00 |
| **Sum** | **1.75** | **8.00** | **3.33** |

Next, each element in the pairwise matrix was normalised before performing a row-wise aggregate in order to calculate the weight of each criteria towards the goal. The criteria weights are shown in Table 3.2. It can be noted that the sum of the criteria weights is equal to 1.

TABLE 3.2: Criteria weights indicating relative importance to the goal

| | Weights |
|---|---|
| Functionality | 0.56 |
| Governance | 0.12 |
| Accessibility | 0.32 |

Before continuing, it was important to ensure that the relative evaluation of the criteria was consistent so that there would be confidence in the calculated weights. This was done by first summing the dot product of the pairwise comparison matrix and the criteria weights in order to determine $\lambda_{max}$, which was found to be 3.02. Using Equation 2.4, the consistency index was calculated, where $CI = 0.009$. As there were three criteria, the random index parameter of 0.58 was used to calculate the consistency ratio:

$$CR = \frac{CI}{RI} = \frac{0.009}{0.58} = 0.016 \tag{3.1}$$

The result of Equation 3.1 is less than 0.1, thus indicating that the evaluation of the criteria was reasonably consistent.

The process was iterated through once again for the sub-criteria under functionality, governance and accessibility. This ensured that weights for each sub-criteria

could be determined, as well as to ensure that the evaluations of relevant sub-criteria relative to one another was done consistently. For the sub-criteria under functionality (automation, error handling, fault tolerance, performance quality and unit testing), Table 3.3 shows the pairwise comparison matrix and Table 3.4 shows the sub-criteria weights under the functionality criteria. Once again, it can be noted that the sum of the weights are equal to 1.

The detailed explanation of the analytical hierarchy process in Section 2.3.2 can be referenced to see the methodology and equations used to calculate the relative weightings of sub-criteria. Specifically, Table 2.2 and Step 2 of Section 2.3.2 illustrate how the evaluator is able to determine the numerical values for relative importance of two different criteria, which are portrayed in Table 3.3, Table 3.5 and Table 3.7.

An example of the logic followed during the evaluation process in order to arrive at the values within the pairwise matrices is as follows:

For the relative importance of error handling to automation, the evaluator views error handling as **moderately more important** than automation, as successful automation is contingent on fit-for-purpose error handling, thus a numeric value of 3 is attributed (see Table 2.1 for the scale depicting intensity of importance). The relative importance of automation to error handling would be the inverse of the relative importance of error handling to automation, therefore, it gets a valuation of $\frac{1}{3}$. Each sub-criteria is compared to another sub-criteria in this manner, by the evaluator, as the evaluator has an understanding of these sub-criteria and how they rank in terms of relative importance based on the workflow experiments conducted and detailed in Section 2.3.1.

TABLE 3.3: Pairwise comparison matrix for selected sub-criteria under the functionality criteria

| | Automation | Error Handling | Fault Tolerance | Performance Quality | Unit Testing |
|---|---|---|---|---|---|
| Automation | 1.00 | 3.00 | 6.00 | 4.00 | 0.33 |
| Error Handling | 0.33 | 1.00 | 4.00 | 3.00 | 0.20 |
| Fault Tolerance | 0.17 | 0.25 | 1.00 | 0.33 | 0.11 |
| Performance Quality | 0.25 | 0.33 | 3.00 | 1.00 | 0.17 |
| Unit Testing | 3.00 | 5.00 | 9.00 | 6.00 | 1.00 |
| **Sum** | **4.75** | **9.58** | **23.00** | **14.33** | **1.81** |

TABLE 3.4: Sub-criteria weights indicating relative importance to functionality

| | Weights |
|---|---|
| Automation | 0.25 |
| Error Handling | 0.13 |
| Fault Tolerance | 0.04 |
| Performance Quality | 0.08 |
| Unit Testing | 0.50 |

The same consistency check that was done for the evaluation of criteria previously was conducted for the evaluation of the sub-criteria under functionality. Using Equation 2.4, a consistency index of 0.058 was calculated. Because there are 5 attributes in this specific evaluation, the random index parameter to be used, as shown in Table 2.3, is 1.12. The consistency of the evaluation of sub-criteria under the functionality criteria was calculated using Equation 3.2:

$$CR_{functionality} = \frac{CI}{RI} = \frac{0.058}{1.12} = 0.051 \tag{3.2}$$

As can be seen, $CR_{functionality} < 0.1$ thus the evaluation of these sub-criteria was consistent.

For the sub-criteria under governance (data encryption, monitoring, security and role based access), an iteration of the evaluation process was conducted and the consistency of the evaluation of determined. Table 3.5 shows the pairwise comparison matrix as per the evaluation of the sub-criteria under governance:

TABLE 3.5: Pairwise comparison matrix for selected sub-criteria under the governance criteria

| | Data Encryption | Monitoring | Security | Role Based Access |
|---|---|---|---|---|
| Data Encryption | 1.00 | 0.02 | 3.00 | 0.33 |
| Monitoring | 5.00 | 1.00 | 7.00 | 3.00 |
| Security | 0.33 | 0.14 | 1.00 | 0.20 |
| Role Based Access | 3.00 | 0.33 | 5.00 | 1.00 |
| **Sum** | **9.33** | **1.68** | **16.00** | **4.53** |

Normalising the pairwise comparison matrix and performing a row-wise aggregtion yielded the weights for the sub-criteria under governance, as shown in Table 3.5. Once again, it can be noted that the sum of sub-criteria weights under governance is equal to 1.

TABLE 3.6: Sub-criteria weights indicating relative importance to governance

| | Weights |
|---|---|
| Data Encryption | 0.12 |
| Monitoring | 0.56 |
| Security | 0.06 |
| Role Based Access | 0.26 |

Using Equation 2.4, a consistency index of 0.039 was calculated for the sub-criteria pertaining to governance. Since the number of sub-criteria in this set was 4, the random index parameter was set to 0.9. Equation 3.3 shows the calculation for the consistency related to the evaluation of these sub-criteria:

$$CR_{governance} = \frac{CI}{RI} = \frac{0.039}{0.90} = 0.044 \tag{3.3}$$

$CR_{governance} < 0.1$ which indicates that the evaluation of the sub-criteria under governance was done consistently.

The final set of sub-criteria that were evaluated were those under accessibility, namely; availability, ease of access, integration and interoperability. Table 3.7 shows the pairwise comparison matrix for these sub-criteria:

The weights of the sub-criteria pertaining to accessibility are given in Table 3.8. Once again, it can be noted that the sum of the weights is equal to 1.

Equation 2.4 resulted in a consistency index of 0.024 for the sub-criteria pertaining to accessibility. The number of sub-criteria in this set was also 4, thus the random index parameter was set to 0.9 for this criteria as well. Equation 3.4 shows the calculation for the consistency related to the evaluation of accessibility related

TABLE 3.7: Pairwise comparison matrix for selected sub-criteria under the accessibility criteria

| | Availability | Ease of Use | Integration | Interoperability |
|---|---|---|---|---|
| Availability | 1.00 | 0.17 | 0.50 | 0.20 |
| Ease of Use | 6.00 | 1.00 | 4.00 | 3.00 |
| Integration | 2.00 | 0.25 | 1.00 | 0.50 |
| Interoperability | 5.00 | 0.33 | 2.00 | 1.00 |
| **Sum** | **14.00** | **1.75** | **7.50** | **4.70** |

TABLE 3.8: Sub-criteria weights indicating relative importance to accessibility

| | Weights |
|---|---|
| Availability | 0.07 |
| Ease of Use | 0.54 |
| Integration | 0.13 |
| Interoperability | 0.26 |

sub-criteria:

$$CR_{accessibility} = \frac{CI}{RI} = \frac{0.024}{0.90} = 0.027 \tag{3.4}$$

$CR_{accessibility} < 0.1$, therefore, there is confidence in that the evaluation of the sub-criteria under accessibility was done consistently.

Once all criteria and sub-criteria evaluations were validated to be consistent, the weightings of the sub-criteria were converted to global weights. This was done by multiplying the weight of each sub-criteria with the weight of its parent criteria. Remembering that the sum of weights for criteria, as well as sub-criteria sets, was equal to 1, the weights can be treated as percentages, thus the multiplication of sub-criteria weight and parent weight yields the weighting of a sub-criteria in terms of the goal of the analytical hierarchy process. As such, the sum of the global weights is also equal to 1. Table 3.9 illustrates the global weights of each of the sub-criteria:

Next, using the insights gained through building the experimental data science workflows and a subsequent review of the sub-criteria associated to each platform, the platforms of each of the CSPs were evaluated against one another. In terms of the analytical hierarchy process, these platforms provided by the CSPs were the

TABLE 3.9: Global sub-criteria weightings for the analytical hierarchy process

| Criteria | Sub-criteria | Global Weight |
|---|---|---|
| Functionality | Automation | 0.14 |
| Functionality | Error Handling | 0.07 |
| Functionality | Fault Tolerance | 0.02 |
| Functionality | Performance Quality | 0.04 |
| Functionality | Unit Testing | 0.28 |
| Governance | Data Encryption | 0.015 |
| Governance | Monitoring | 0.07 |
| Governance | Security | 0.01 |
| Governance | Role Based Access | 0.03 |
| Accessibility | Availability | 0.02 |
| Accessibility | Ease of Use | 0.17 |
| Accessibility | Integration | 0.04 |
| Accessibility | Interoperability | 0.08 |

alternatives present to the data scientist in terms of achieving the goal of selecting suitable provider for data science workflow requirements. Table 3.10 to Table 3.22 below show the pairwise comparison matrices and weights associated to each of the alternatives against each of the sub-criteria that were used in the overall evaluation process:

TABLE 3.10: Pairwise comparison matrix and weights of alternatives against automation sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 3.00 | 1.00 | 0.42 |
| CSP2 | 0.33 | 1.00 | 0.25 | 0.13 |
| CSP3 | 1.00 | 4.00 | 1.00 | 0.46 |

TABLE 3.11: Pairwise comparison matrix and weights of alternatives against error handling sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 0.33 | 0.50 | 0.16 |
| CSP2 | 3.00 | 1.00 | 2.00 | 0.54 |
| CSP3 | 2.00 | 0.50 | 1.00 | 0.30 |

TABLE 3.12: Pairwise comparison matrix and weights of alternatives against fault tolerance sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 0.20 | 0.33 | 0.11 |
| CSP2 | 5.00 | 1.00 | 3.00 | 0.63 |
| CSP3 | 3.00 | 0.33 | 1.00 | 0.26 |

TABLE 3.13: Pairwise comparison matrix and weights of alternatives against performance quality sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 0.33 | 0.17 | 0.09 |
| CSP2 | 3.00 | 1.00 | 0.25 | 0.22 |
| CSP3 | 6.00 | 4.00 | 1.00 | 0.69 |

TABLE 3.14: Pairwise comparison matrix and weights of alternatives against unit testing sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 2.00 | 0.33 | 0.23 |
| CSP2 | 0.50 | 1.00 | 0.20 | 0.12 |
| CSP3 | 3.00 | 5.00 | 1.00 | 0.65 |

TABLE 3.15: Pairwise comparison matrix and weights of alternatives against data encryption sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 3.00 | 2.00 | 0.54 |
| CSP2 | 0.33 | 1.00 | 0.50 | 0.16 |
| CSP3 | 0.50 | 2.00 | 1.00 | 0.30 |

TABLE 3.16: Pairwise comparison matrix and weights of alternatives against monitoring sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 3.00 | 3.00 | 0.60 |
| CSP2 | 0.33 | 1.00 | 1.00 | 0.20 |
| CSP3 | 0.33 | 1.00 | 1.00 | 0.20 |

TABLE 3.17: Pairwise comparison matrix and weights of alternatives against security sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 0.33 | 0.25 | 0.12 |
| CSP2 | 3.00 | 1.00 | 0.50 | 0.32 |
| CSP3 | 4.00 | 2.00 | 1.00 | 0.56 |

TABLE 3.18: Pairwise comparison matrix and weights of alternatives against role based access sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 2.00 | 5.00 | 0.58 |
| CSP2 | 0.50 | 1.00 | 3.00 | 0.31 |
| CSP3 | 0.20 | 0.33 | 1.00 | 0.11 |

Having obtained the weights of each alternative against each sub-criteria ultimately allowed for the calculation of the overall weighting of the alternatives against the goal, similar to how the sub-criteria were weighed against the goal, as shown

TABLE 3.19: Pairwise comparison matrix and weights of alternatives against availability sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 1.00 | 1.00 | 0.33 |
| CSP2 | 1.00 | 1.00 | 1.00 | 0.33 |
| CSP3 | 1.00 | 1.00 | 1.00 | 0.33 |

TABLE 3.20: Pairwise comparison matrix and weights of alternatives against ease of use sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 2.00 | 0.20 | 0.19 |
| CSP2 | 0.50 | 1.00 | 0.25 | 0.13 |
| CSP3 | 5.00 | 4.00 | 1.00 | 0.68 |

TABLE 3.21: Pairwise comparison matrix and weights of alternatives against integration sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 3.00 | 0.25 | 0.23 |
| CSP2 | 0.33 | 1.00 | 0.20 | 0.10 |
| CSP3 | 4.00 | 5.00 | 1.00 | 0.67 |

TABLE 3.22: Pairwise comparison matrix and weights of alternatives against interoperability sub-criteria

| | CSP1 | CSP2 | CSP3 | Weights |
|---|---|---|---|---|
| CSP1 | 1.00 | 2.00 | 0.33 | 0.23 |
| CSP2 | 0.50 | 1.00 | 0.20 | 0.12 |
| CSP3 | 3.00 | 5.00 | 1.00 | 0.65 |

in Table 3.9. Specifically, this was done by multiplying the weight of each alternative per sub-criteria to the global weight of the sub-criteria. As a result, the final weighting of the alternatives were calculated and are demonstrated in Table 3.23:

TABLE 3.23: Overall weightings of alternatives against the goal

| | Weights | Implications |
|---|---|---|
| CSP1 | 0.28 | Suitable for user |
| CSP2 | 0.19 | Least suitable for the user |
| CSP3 | 0.53 | Most suitable for the user |

The results indicate that based on the relative importance of criteria and sub-criteria to the user, the experience obtained by making use of the cloud platform provided by **CSP3** would be most suitable for the data scientist as it has the highest weighting based on the evaluation of the criteria and sub-criteria relevant to the tasks needed to be performed on said platform. Conversely, although **CSP2** does have more desirable functionality in terms of error handling (Table 3.11) and fault tolerance (Table 3.12), overall it would not provide the user with as positive an experience as CSP3.

## 3.2 Discussion

The goal of this study was to conduct a comparative review of criteria pertaining to cloud computing platforms for data science workflows, and to develop a framework for aiding in such evaluations, in order to identify an appropriate platform for data science workflows since there are no specific evaluation frameworks for cloud computing platforms from a purely data science workflow perspective. In order to set the groundwork for the development of a framework which could be used to evaluate importance of criteria among cloud platforms, this study ensured that the experimentation was based on reviewing the services that would be required by an entry level data scientist as a basis. Furthermore, the analytical hierarchy process was used to develop and showcase the functionality of the framework as this specific multi-criteria decision making methodology was quite suitable for the requirements of this study's objectives.

Firstly, a goal was set to ensure that the framework had an outcome to produce: which cloud platform should the data scientist select for use? Next, criteria and sub-criteria were established to be specific to the requirements of the data scientist. These attributes were what the data scientist would look for when using the cloud platforms for data science workflows, thus ensuring that the criteria used to review

and evaluate each platform were suitable. Additionally, because a comparative review was desired, and not a review to determine the absolute best platform, the analytical hierarchy process was perfect for achieving this. Pairwise comparisons were done between criteria and sub-scriteria to determine the numeric representation of the relative levels of importance among attributes. As the analytical hierarchy process has a means of evaluating consistency, the evaluations made when reviewing the attributes were able to be tested for any inconsistency and as shown in the results section, the outcomes for the evaluations pertaining to the criteria and sub-criteria were indeed consistent. Finally, the alternatives were reviewed based on the experience of an entry level data scientist against each of the sub-criteria in order to weigh the various alternatives (cloud platforms provided by the cloud service providers) against the goal. Ultimately, this allowed the framework to recommend a cloud provider to the user by answering the question present in the goal of the analytical hierarchy process with the alternative that yielded the highest weight, which in the experiment conducted for this study was CSP3.

Evaluation of multi-criteria systems can become quite complex, especially when the criteria are of a linguistic nature and are dependent on subjective beliefs. This was an additional motivation of using the analytical hierachy process, along with Saaty's fundamental scale of absolute numbers, so that linguistic associations of relative importance between each of the attributes could be converted to numeric values upon which mathematical operations could be conducted to produce weightings of sub-criteria and alternatives against the goal. It enabled the framework to be both replicable and robust, as the analytical hierarchy process is a widely known methodology for multi-criteria decision analysis and thus can be followed to reproduce the experiments and results obtained in this report. Additionally, based on how the framework is set up, users can apply the framework to systems with varying criteria, sub-criteria and alternatives, depending on their requirements and accessibility to platforms.

As such, having shown the consistency in evaluation as well as the ability for the framework to recommend a cloud service provider, the outcomes of this study's efforts have been positive: a successful framework was developed using the analytical hierarchy process in order to review, compare and evaluate criteria pertaining to cloud platforms based on the required attributes, and to recommend a suitable platform for a user based on the review and subsequent evaluation of the attributes,

while maintaining confidence in the evaluation process by means of determining that the evaluations of the criteria and sub-criteria were consistent.

A comparison of this study to studies that have been conducted in the domain of cloud computing evaluation frameworks revealed that a contribution has been made in terms of being able to use sub-criteria for the evaluation of cloud computing platforms. Recent studies [6] [15] [31] have made use of multi-criteria decision making frameworks to evaluate relative importance at the criteria level, whereas this study has shown that successful frameworks can be established at deeper levels, that is, at the sub-criteria level. As such, a clearer understanding of the relative importance of attributes pertaining to a given goal can be established and used to explain why one alternative would be better than another for the goal. Furthermore, weighing the sub-criteria against one another provides a means for the evaluator to ensure that attributes that are relevant and have a material impact on the decision are being considered. An example of this is if there are sub-criteria with relatively small weights compared to other sub-criteria, then they should most likely not be included in the evaluation as their contribution to the evaluator's decision of which alternative to choose will be negligible.

Although past research into cloud computing evaluation had been conducted, it was largely done at a general level, and did not consider the application of cloud computing platforms to a specific task or operation, which is a gap identified within this study's literature survey. This study addresses that gap in literature as it focuses on the application of cloud computing platforms for data science workflows. Having a precise application to derive a goal for AHP, and being able to apply sub-criteria evaluations to the alternatives enables the framework to be robust, thus facilitating changing user requirements and platform selection.

## 3.3 Summary

Using the analytical hierarchy process along with Saaty's fundamental scale of absolute numbers, it was possible to numerically evaluate the attributes relevant to data science workflows against one another, and to subsequently evaluate how well various cloud platforms rank in terms of these attributes, keeping in mind that the evaluation is based on the experience of an entry level data scientist. The process made it possible to establish and demonstrate a framework for the comparative

review of criteria pertaining to cloud platforms for data science workflows, while proving to be robust enough to be used for varying attributes and alternatives. It has been shown that confidence in evaluation can also be demonstrated using the analytical hierachy process, using the concept of consistency. When developing the framework, the evaluations conducted for the criteria and the three sets of sub-criteria were shown to be consistent, thus there is confidence in the global weighting of the sub-criteria in terms of the overall goal of the analytical hierarchy process. The developed framework was able to recommend a cloud service provider to the user based on the ranking of relevant attributes and how they presented themselves during data science workflow experimentation with the respective cloud platforms.

# Chapter 4

# Conclusions and Future Work

## 4.1 Conclusions

Although a significant amount of research has been done regarding scheduling of data science workflows, cloud computing theory, and the comparison of cloud computing platforms in terms of concepts such as performance, pricing and storage, not much has been done with regards to evaluating or reviewing cloud based platforms for data science workflows specifically. Thus research was conducted to understand various data science workflow platforms, the requirements that a data scientist would have when using such platforms and the availability of existing frameworks or methodologies that would enable someone to conduct a comparative review of cloud platforms for data science workflows. Since no such framework was identified, a novel framework using the analytical hierarchy process, Saaty's fundamental scale of absolute numbers, and relevant attributes to a data science workflow was created. It was applied to a scenario where a data scientist needed to select a cloud platform that would be suitable for the user's requirements and the framework was successful in recommending a suitable cloud service provider to the data scientist.

The benefits of using the analytical hierarchy process were twofold: the first was that a measure, namely consistency, could be used to ensure that evaluations were done correctly, and the second was that it enabled the review and evaluations of criteria pertaining to the cloud platforms (alternatives) to be done in a relative manner depending on the experience of an entry level data scientist. In this way, decisive or categoric proposals of one cloud platform being objectively better than another could be avoided, with the additional benefit on users of the framework being able to have their own alternatives depending on which set of platforms would be feasible for them (depending on budget, vendor agreements, brand affiliation, etc.).

Overall, a successful framework was created and demonstarted for the comparative review of cloud computing platforms for data science workflows. The evaluations were deemed consistent and the framework is dynamic enough to be tailored to user requirements (attributes can be changed as needed) and platform scope (relevant alternatives to the user's circumstances can be chosen).

## 4.2 Future Work

This study has demonstrated that a multi-criteria decision analysis (MCDA) methodology can be used to review and evaluate cloud platforms for data science workflows in a robust manner. The proposed framework can, however, be further developed into a tool that is usable by individuals to compare alternatives against a set of attributes, while being bounded by numeric valuations of relative importance being consistent. Additionally, other MCDA methods such as the aggregated indices randomisation method (AIRM) or stochastic multicriteria acceptability analysis can be explored to evaluate whether they are able to yield more consistent evaluations and weights of attributes and alternatives. Lastly, a means of generating a recursive algorithm to deal with the iterative portions of the analytic hierarchy process would be a welcomed addition to the framework, and would tie in well to the intial note of developing a tool to use the proposed framework in a systematically robust manner.

# Appendix A

# Terms of Service Review

## A.1 Review of Cloud Platform Terms of Service

The experimental portion of this research leveraged existing cloud platforms, namely Amazon Web Services, Google Cloud Platform and Microsoft Azure, to gauge the experience of a data scientist when using services catered towards data science workflows in terms of ETL and ML. The selected criteria and sub-criteria aided in focusing on important aspects that have a material effect on the work done by, and user experience of, the data scientist on a selected cloud platform. As such, it was necessary to review the terms of service for the various CSPs to ensure that use of their platform and subsequent publishing of results was in line with their guidelines and policies.

Section 1.8 of AWS' terms of service[1] indicates that although allowed, benchmarking of their services need to be disclosed in a manner that allows results to be replicated. This research focuses on an evaluation of relative importance of cloud computing attributes based on a data scientist's requirements and not a categorical benchmarking of one cloud platform's service against another. As work has not been done to benchmark any of AWS' services, the clause is not applicable to this research.

Section 1.1 of GCP's terms of service[2] indicates that use and integration of services are allowed for any customer application as long as it does not breach the terms outlined in Section 3 (customer obligations). Thus no approval or disclosure was required when using the services provided by Google Cloud Platform for the experimental portion of this research.

[1] https://aws.amazon.com/service-terms/
[2] https://cloud.google.com/terms

Microsoft Azure expressly states that the Azure for Students subscription enables students to have access to all Azure services for non-commercial research[3]. As this research is an exploration of the relative importance of data science workflow criteria to a data scientist and subsequent development of an evaluatory framework, with no commercial gain attributed to the findings and outcomes of the research, this research is permitted to make use of the Azure for Students subscription to fulfill the cloud platform investigation requirements of the research.

[3]https://azure.microsoft.com/en-us/free/free-account-students-faq/

# Bibliography


[1] *Cloud Computing*. 2020. URL: https://www.investopedia.com/terms/c/cloud-computing.asp (visited on 04/14/2021).

[2] M. A. Vouk. "Cloud Computing - issues, research and implementations". In: *Journal of Computing and Information Technology* 4 (2008), pp. 235–246.

[3] C. Chen, J. Chen, J. Liu, and Y. Wen. "Research on workflow scheduling algorithms in the cloud". In: *Process-Aware Systems* (2014), pp. 35–48.

[4] *Pipeline execution and triggers in Azure Data Factory*. 2018. URL: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers (visited on 04/14/2021).

[5] C. Sun, W. Fu, J. Liu, L. Jia, and H. Tan. "Modeling, design, and implementation of a cloud workflow engine based on Aneka". In: *Journal of Applied Mathematics* 2014 (2014), pp. 1–9.

[6] M. Abdel-Basset, M. Mohamed, and V. Chang. "NMCDA: A framework for evaluating cloud computing services". In: *Future Generation Computer Systems* 86 (2018), pp. 12–29.

[7] W. Liao and W. Qiu. "Applying analytic hierarchy process to assess healthcare-oriented cloud computing service systems". In: *SpringerPlus* 5 (2016), pp. 1–9.

[8] B. M. R. Wilson, B. Khazaei, and L. Hirsch. "Cloud adoption decision support for SMEs using analytical hierarchy process (AHP)". In: (2016), pp. 1–5.

[9] N. Chawla and D. Kumar. "Decision-making model to evaluate cloud computing service nodel using analytic hierarchy process (AHP) and benefits, costs, opportunities, and risks (BCOR) analysis". In: *International Journal of Sensors, Wireless Communications and Control* 10 (2020), pp. 846–856.

[10] E. Ghazizadeh and B. Cusack. "Evaluation theory for characteristics of cloud identity trust framework". In: *Cloud Computing - Technology and Practices* (2018).

[11] B. Stanton, M. Theofanos, and K. P. Joshi. "Framework for cloud usability". In: *NIST Special Publication* 500 (2015), pp. 1–18.

[12] *Architecture for high-throughput low-latency big data pipeline on cloud*. 2020. URL: https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5 (visited on 04/14/2021).

[13] P. Costa, J. Santos, and M. Mira da Silva. "Evaluation criteria for cloud services". In: June 2013, pp. 598–605. ISBN: 978-0-7695-5028-2. DOI: 10.1109/CLOUD.2013.70.

[14] A. Tataru. "Metrics for evaluating machine learning cloud services". In: (2017), pp. 1–103.

[15] N. Afifi, H. Belhadaoui, I. Hilal, and F.E. Mdarbi. "An approach for selecting cloud service adequate to big data". In: *International Journal of Computer Science and Information Security* 16 (2018), pp. 20–32.

[16] N. Uphadyay. "Managing cloud service evaluation and selection". In: *Procedia Computer Science* 122 (2017), pp. 1061–1068.

[17] T.L. Saaty. "Decision making with the analytical hierarchy process". In: *International Journal of Services Sciences* 1 (2008), pp. 83–98.

[18] *Get the Northwind sample database for SQL Server*. 2021. URL: https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/sql/linq/downloading-sample-databases (visited on 07/13/2021).

[19] *What is workflow automation?* 2021. URL: https://www.ibm.com/cloud/blog/workflow-automation (visited on 10/11/2021).

[20] H. Mourãoand and P. Antunes. "Exception Handling Through a Workflow". In: vol. 3290. Oct. 2004, pp. 37–54. ISBN: 978-3-540-23663-4. DOI: 10.1007/978-3-540-30468-5_5.

[21] D. Poola, M. A. Salehi, K. Ramamohanarao, and R. Buyya. "A taxonomy and survey of fault-tolerant workflow management systems in cloud and distributed computing environments". In: *Software architecture for big data and the cloud*. 2017, pp. 285–320. ISBN: 978-0-12-805467-3. DOI: https://doi.org/10.1016/B978-0-12-805467-3.00015-6.

[22] *How to improve ETL performance in data integration process?* 2020. URL: https://dataintegrationinfo.com/improve-etl-performance/ (visited on 10/11/2021).

[23] S.M.F. Ali and R. Wrembel. "From conceptual design to performance optimization of ETL workflows: current state of research and open problems". In: *The VLDB Journal* 26 (2017), pp. 777–801.

[24] *ETL testing: an overview*. URL: https://www.talend.com/resources/etl-testing/ (visited on 10/11/2021).

[25] *Encryption vs masking*. URL: https://www.encryptionconsulting.com/education-center/encryption-vs-masking/ (visited on 10/11/2021).

[26] *ISO 27001 certification vs ISAE 3402 SOC 2 assurance report*. 2016. URL: https://www.cm-alliance.com/news/2016/05/us-client-want-isae-soc-2-report-i-already-iso-27001-certified-i (visited on 10/11/2021).

[27] B.S. Al-Attab. "Role-based access control's framework for cloud computing". In: Jan. 2015, pp. 1–15.

[28] *Integration vs interoperability: what's the difference?* 2017. URL: https://blog.sisfirst.com/integration-v-interoperability-what-is-the-difference (visited on 10/11/2021).

[29] Y. Tung, T. Pai, S. Lin, and L. Shih. "Analytic hierarchy process of academic scholars for promoting energy saving and carbon reduction in Taiwan". In: *Procedia Environmental Sciences* 20 (Dec. 2014), pp. 526–532. DOI: 10.1016/j.proenv.2014.03.065.

[30] T.L Saaty and L. Vargas. "The analytic network process". In: *Decision Making with the Analytic Network Process* (2006), pp. 1–26.

[31] L. Wenquan, P. Xindong, and C. Zhen Song. "Evaluation of cloud computing copyright protection based on AHP". In: *Mathematical Problems in Engineering* 2020 (Dec. 2020), pp. 1–11. DOI: 10.1155/2020/6671331.