Republic of Iraq
Ministry of Higher Education and
Scientific Research
University of Babylon
College of Information Technology
Department of Software

Reducing Cost of File Transportation in Cloud Computing by Using Vogel's Approximation Method

A Thesis
Submitted to the Council of the College of Information Technology, for
Postgraduate Studies of the University of Babylon in Partial Fulfillment
of the Requirements for the Degree of Master in Information Technology/
Information Software

By

Hawraa Mahdi Salih Hadi

Supervised by
Asst. Prof. Dr. Mahdi Saleh Neamaa Mousa

2022 A.D. **1444 A.H.**

بسم الله الرحمن الرحيم

﴿الحَمدُ لله الّذي هَدَانا لِهَذا وما كُنّا لِنَهتَدي لولا أن هَدانا اللّه﴾

صدق الله العظيم

سورة الأعراف آية 43

Declaration

I hereby declare that this thesis entitled “**Reducing Cost of File Transportation in Cloud Computing by Using Vogel’s Approximation Method**” submitted to the University of Babylon in partial fulfillment of requirements for the degree of Master in Information Technology \ Software, has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose source is appropriately cited in the references.

Signature:
Name: Hawraa Mahdi Salih

Date: / /2022

Supervisor Certification

I certify that the thesis entitled (**Reducing Cost of File Transportation in Cloud Computing by Using Vogel's Approximation Method**) was prepared under my supervision at the department of Software/ College of Information Technology/ University of Babylon as partial fulfillment of the requirements of the degree of Master of Philosophy in Information Technology-Software.

Signature:

Supervisor Name: **Asst. Prof. Dr. Mahdi Saleh Al-Mhanna**

Date: / /2022

The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled "**Reducing Cost of File Transportation in Cloud Computing by Using Vogel's Approximation Method**" for debate by the examination committee.

Signature:

Prof. Dr. Ahmed Saleem Abbas

Head of Software Department

Date: / /2022

Certification of the Examination Committee

We, the undersigned, certify that (**Hawraa Mahdi Salih**) candidate for the degree of Master in Information Technology - Software, has presented his thesis of the following title "**Reducing Cost of File Transportation in Cloud Computing by Using Vogel's Approximation Method**" as it appears on the title page and front cover of the thesis that the said thesis is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:

Signature:
Name:
Title:
Date: / / 2022
(Chairman)

Signature:
Name:
Title:
Date: / / 2022
(Member)

Signature:
Name:
Title:
Date: / / 2022
(Member)

Signature:
Name:
Title:
Date: / / 2022
(Member)

Signature:
Name:
Title:
Date: / / 2022
(Member)

Signature:
Name:
Title:
Date: / / 2022
(Member and Supervisor)

Approved by the Dean of the College of Information Technology, University of Babylon.

Signature:
Name:
Title: Professor
Date: / / 2022
(Dean of Collage of Information Technology)

Declaration Associated with this Thesis

Some of the works presented in this thesis have been published or accepted as listed below.

The published Paper:

Title: **Improving the performance of Apache Hive by using Vogel's Approximation Method**,

Author: Hawraa Mahdi Salih, Mahdi S. Almhanna

Journal: Guangdianzi Jiguang/ Journal of Optoelectronics Laser

Publisher: Tianjin Daxue Jidian Fenxiao/Tianjin Institute of Technology

Dedication

To everyone who taught me from my childhood to this day,

To the soul of my father, who is proud to bear his name and I hope that he will share my joy with me and I miss his presence

To who was her supplication the secret of my success and tenderness as a surgical balm.. my mother

To my brothers and sister,

To who gave me all the support to achieve my ambitions, Muhammad Al-Saaedi.

Acknowledgments

In the name of Allah, the Most Merciful, the Most Merciful.
My sincere thanks and appreciation to all my teachers who helped me with their knowledge in my academic are career, especially my supervisor **Asst. Prof. Dr. Mahdi Saleh Al-Muhanna** "who supports me with his valuable observations, advice, and guidance throughout his supervision period, and the detailed ideas that enlightened my insight.

I extend my sincere thanks to the members and management of the Information Technology college for their help and support.

I owe a big thanks to the head of the Department of Information Programs, **Prof. Dr. Ahmed Saleem**, who without his paternal and motivating support would not have completed my career, and heartfelt thanks to the dean of the college of information Technology, Prof. Dr. Hussein Attia Lafta.

Last but not least, I am greatly grateful to my family and my friend for their continued love and support.

Hawraa Mahdi Salih
September 2020

Abstract

Cloud computing is a technology that enables virtual integration and provision of computing resources located at remote sites. On the World Wide Web, many users request many files scattered on many servers around the world and this takes time, which leads to an increase in the cost of transferring the file within cloud computing. Reducing the time required to download a file is the same as reducing the cost of delivering products. The least time required to deliver the product to the customer at the lowest possible cost.

In this thesis, the methodology is proposed to reduce the cost of file transfer using Vogel's Approximation Method (VAM). The main idea of the system determines which server helps deliver the file quickly to the client requesting this file. The time required to download the file via the proposed system is reduced by selecting the best server that returns the file in the shortest time, and an efficient method VAM is used to get the minimum cost needed for each file to be delivered to the user within the middle server receiving client requests, and specify the server for each client to download the file from.

The proposed system consists of three parts which are the cloud computing part, server vam part, and client part and used three file which are (case1, case2, case3) with (6.5, 2, 1.3) GB respectively.The proposed system was tested using a standard deviation. The results showed that the VAM contributed to reducing the cost of file transferring. It is shown in this thesis that the ratio of time reduction was about (97- 98)% compared to the normal file transfer time.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

Symbols	Full Form
VAM	Vogel's Approximation Method
IT	Information Technology
TOCM-MT	Total Opportunity Cost Matrix – Minimal Total
IWD	Intelligent Water Drop
PSO	Particle Swarm Optimization
GA	Genetic Algorithm
MODI	Modified Distribution Method
NCM	Northwest Corner Method
LCM	Lowest Cost Method
SQL	Structured Query Language
TCP/IP	Transmission Control Protocol/Internet Protocol
PCs	Personal Computers
UDP	User Datagram Protocol
HDInsight	Hadoop Distributed insight
HDFS	Hadoop Distributed File System
HQL	Hive Query Language
DBMS	DataBase Management System
JDBC	Java™ Database Connectivity
ODBC	Open Database Connectivity
ETL	Extract/Transform/Load
LLAP	Low Latency Analytical Processing
Min.	Minute
CPU	Central Processing Unit
GB	Giga Byte
GHz	Giga Hertz
HDD	Hard Disk Drive
SSD	Solid-State Drive
RAM	Random Access Memory
STD	Standard Deviation

Chapter One

Introduction

1.1 General Introduction

In recently, business organizations seeks to provide their services and management. Its business uses new technological models, which leads to the development of the level of performance of the organization, which helps it keep abreast of what is happening around the world. Information technology has undergone a huge change in the last few years. By adapting its material or human resources, it can improve its performance and complete tasks faster and easier with the speed of delivery of services and information, as well as the ease with which beneficiaries can exchange information [1].

The huge change in information and communication technology has made it much easier for businesses to keep track of the things and people they own, both material and human [1].

As the world's largest business organizations, which need to cut costs, began to use cloud computing, this area has seen an interest in cloud computing technology in order to maintain and obtain a competitive advantage through the advantages that are based on the use of cloud computing technology, and work to confront the resulting risks. Keeping pace with changes in rapid information and communication technologies [1].

The internet has significantly impacted the computer world from its inception in the 1990s and the capabilities of ubiquitous computing of today. Parallel computing has evolved into distributed computing, grid computing, and most recently cloud computing. The term "cloud computing" refers to a computer environment where one party can outsource its processing requirements to another party, and that party can then use the computing power or resources, such as databases or emails, as needed. A new development in IT is cloud computing, which moves processing and data from desktop and portable PCs to huge data centers [2].

The National Institute of Standards and Technology's (NIST) definition of cloud computing reads as follows: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage applications, and services) that can be rapidly provisioned and released with little management effort or service provider interaction [3]." Applications may be supply as services through the Internet due to the widespread growth of the internet worldwide. This lowers the total cost as a result [2].

The focus of transportation models is to select the best way to get the goods from various manufacturers (source) and get the assets to various warehouses or clients (it is the destination) where they are needed [4].

To determine an initial transportation cost, Vogel's Approximation Method (VAM), one of the well-known transportation methods in the literature, was examined Initial Transportation Cost (ITC) [5].

Users may quickly query, summarize, and analyze Big Data using SQL-like expressions known as Hive-QL thanks to Apache Hive, data ware house infrastructure solution to process structured data in Hadoop. To import and export data to and from the storage file system, it supports a variety of file types. Hive wants to make processing petabytes of data simple and effective [6].

1.2 Motivation and Problem Statement

The widespread interest in accelerating discovery by offering researchers advanced functionality that reduces their IT costs. The work presented is on file transfer. The intended users need to download many files, of potentially large total size, from among the various network-connected sites ("endpoints"). This requires maximum ease of use and the least possible cost of time.

1.3 Thesis Objectives

1.3.1 Main Objectives

The main objective of this thesis is to reduce the cost of file transport by using Vogel's Approximation Method.

1.3.2 Specific Objectives

To accomplish the main objective:-

1- Copying the parts of the file.

2- Uploading this replica to the servers according to the speed and time via a VAM application.

3- Downloading this file with minimal time and cost.

1.4 Thesis Contribution

The contribution of this thesis is to design and implement a model for reducing the time needed to download files approximately in equal time to reduce the cost of delivering products by mail. Reducing the time is a challenge that often cannot be ignored to deliver the file to the client with the minimum cost needed. With this concept, the proposed system has been contributed by reduces the time needed to download the file which will reflect positively on reducing the money.

1.5 Challenges of Thesis

1. May need to have a fast interconnect between compute resources.
2. Some applications may need to be tweaked to take full advantage of the new model.

1.6 Related Works

Vogel’s Approximation method is a way to find the best solution for transfer problems but not much research has done yet regarding file transfer. To the best of the researcher’s knowledge, this is the first effort using the VAM method to download files at the lowest time cost in the cloud computing.

- In 2014, Rafah M. Almuttairi [7], The researcher proposed a dynamic optimization method, namely Smart Vogel's Approximation Method, to improve the performance of data transfer in Data Grids. This approach reduces the ideal time spent waiting for the slowest replica provider to be equal to or less than the predefined data transfer completion time with minimum replica prices. The results demonstrate that the Efficient Sites Technique EST has a drawback in that two files (f1, f2) cannot be downloaded because their sizes exceed the sites' downloading capacity, whereas the SVAM technique achieves acceptable results.

- In 2018, Debapriya Banik and Md. [8], used VAM to Optimize the Transportation Cost of an Online Business, At the conclusion of this program, the optimal method of transporting products in order to achieve the lowest possible cost is printed out, including which warehouse and how many units of goods need to be delivered to a specified destination. A suitable location for distributing goods in the chosen location has carefully chosen based on demand and warehouse rent. The program can output the optimal amount of supply along with the corresponding destination. The company can associate the supply unit with the destination in order to minimize transportation costs prior to each product.

- In 2019, Amaliah B, et al. [9], developed a new method called Total Opportunity Cost Matrix – Minimal Total (TOCM-MT) to determine the initial basic feasible state of the Transport Proplem TP. TOCM-MT is capable of achieving a total cost that is comparable to or less than the optimal solution. TOCM-MT outperforms VAM, JHM, and TDM1 because it uses a total opportunity cost matrix as the initial matrix, provides a more robust mechanism when multiple highest penalty HP have the same values, and when the least cost of the HP is equal to zero. Thirty-one numerical examples were also used to evaluate the proposed method, twenty-five from journals and six generated randomly.

- In 2019, M. L. Aliyu1, et al. [10], the result illustrates that VAM is the most cost effective way to determine the Initial Basic Feasible Solution, as it requires the least amount of transportation prior to optimization and needs fewer iterations than the least cost method. The transportation design assisted BUA Cement Company's logistics directors in making strategic decisions about the optimum utilization of production from the two varieties (Cement Company of Northern Nigeria CCNN and OBU) to the various customers (distributors) at the lowest possible transportation cost.

- In 2019, Sujaudeen Nannai John and T. T. Mirnalinee [11], presented a novel dynamic data replication strategy with intelligent water drop algorithm that takes into account parameters such as bandwidth, user access, storage space availability, and traffic. They evaluated this algorithm using Hadoop-like storage and analyzed the storage and retrieval of data access times. The increase in storage space is nearly 40%. They tested it against four different access scenarios in order to optimize access efficiency and obtained great promise while maintaining the cloud storage's high availability, performance, and load balancing.

- In 2019, A. Shetty, et al. [12], presented a peer-assisted parallel downloading system which uses the concept of segmented file transfer to decrease the download time. The proposed system allows the utilization of unused bandwidth while reducing the download time. Contributed towards assistance helps maintain balance in the network while reducing the overall time for download for peers with low data transfer rates. Tests show the overall download time for files can be effectively reduced by 15% for peers with poor data transfer rates.

- In 2020, Kamba, et al. [13], used transportation problem techniques to determine minimum cost of transportation and optimize of total transportation cost, the study determined minimum cost of transportation of Gimbiya Furniture Factory using online software, Modified Distribution Method (MODI). The observation made was that if Gimbiya furniture factory, this transportation model will be useful for making strategic a decision by the logistic managers of Gimbiya furniture factory, in making optimum allocation of the production from the company in Kebbi to various customers (key distributions) at a minimum transportation cost.

Table (1.1): Summary of Related Works

No.	Ref.	Author	Year	Contribution
1.	[7]	Rafah M. Almuttairi	2014	Improve the performance of data transfer in Data Grids by Reducing the ideal time spent waiting for the slowest replica provider to be equal to or less than the predefined data transfer completion time with minimum replica prices.
2.	[8]	Debapriya Banik, Md. Zahid Hasan	2018	Used Vogel's Approximation Method to Optimize the Transportation Cost of an Online Business
3.	[9]	Amaliah B, Fatichah C and Suryani	2019	Developed a new method called Total opportunity cost matrix – Minimal total to determine initial basic feasible
4.	[10]	M. L. Aliyu , U. Usman, Z. Babayaro, M. K. Aminu	2019	Showed that Vogel's Approximation method is the most efficient of all the methods in finding the Initial Basic Feasible Solution because it has the least transportation cost.
5.	[11]	Sujaudeen Nannai John & T. T. Mirnalinee	2019	Presented a novel dynamic data replication strategy with intelligent water drop algorithm.
6.	[12]	A. Shetty, S. Mhatre, N. Sinvhal and K. K. Devadkar	2019	Presented a peer-assisted parallel downloading system which uses the concept of segmented file transfer to decrease the download time and also aims to be truly decentralized
7.	[13]	Kamba, A. I., Kardi, S. M., & Dikko, Y. K. G	2020	Used transportation problem techniques to determine minimum cost of transportation and optimize of total transportation cost

As previously stated, the past studies that dealt with the VAM did not address the performance of VAM with the files or the cloud computing. Therefore, the current work will concentrate on verifying the method's performance with files within a specific approach that consists of a number of stages for specify the best server, as well as employing Apache Hive to optimize the efficiency of the proposed approach.

1.7 Thesis Outlines

The given thesis has four main chapters with chapter one includes the following:

- **Chapter Two (Theoretical Background):** The scientific previous works and backgrounds are given to show the concepts of cloud computing, Vogel’s Approximation Method, cost reduction, Apache Hive, and standard deviation.
- **Chapter Three (System Design and Implementation):** This chapter shows in detail the structure and the methodology of the presented work.
- **Chapter Four (Implementation Results):** This chapter gives the outcomes and the resulted of downloading the files and a comparison of the performance measure of the final results.
- **Chapter Five (Conclusions and Future Works):** This chapter gives the final results in summary as well as showing the main thoughts about additional future works.

Chapter Two

Theoretical Background

2.1 Overview

Provides an overview of cloud computing, followed by a discussion of the benefits and challenges of cloud computing and cost reduction, the context of topic modeling will be explained, and Apache Hive; in specific, the concentration will be on the used VAM.

Since the thesis focuses on using VAM with files in cloud computing. First, a brief introduction to VAM will be introduced and its procedure principles will be demonstrated.

2.2 Cloud Computing

Cloud computing is a technology that enables a virtual integration and provisioning of computing resources residing in remote locations [14][15]. When data is stored in a convenient location, it can be quickly retrieved and processed. However, in today's cloud computing environment, securing enough space to store a massive amount of service data is difficult. Cost-effective storage capacity at local sites is critical for the cloud computing service to perform well, and this is an active area of research at the moment [1].

With the advancement of technologies accessible via the Web, the emergence of Web 0.2 and Web 0.3, and the major increase in available Internet speeds, many companies have adopted the practice of making their software programs available for download and use via the internet, including it is now referred to as "cloud computing" because this technology enables additional benefits for its users. Figure (2.1) illustrates the cloud computing architecture [16].

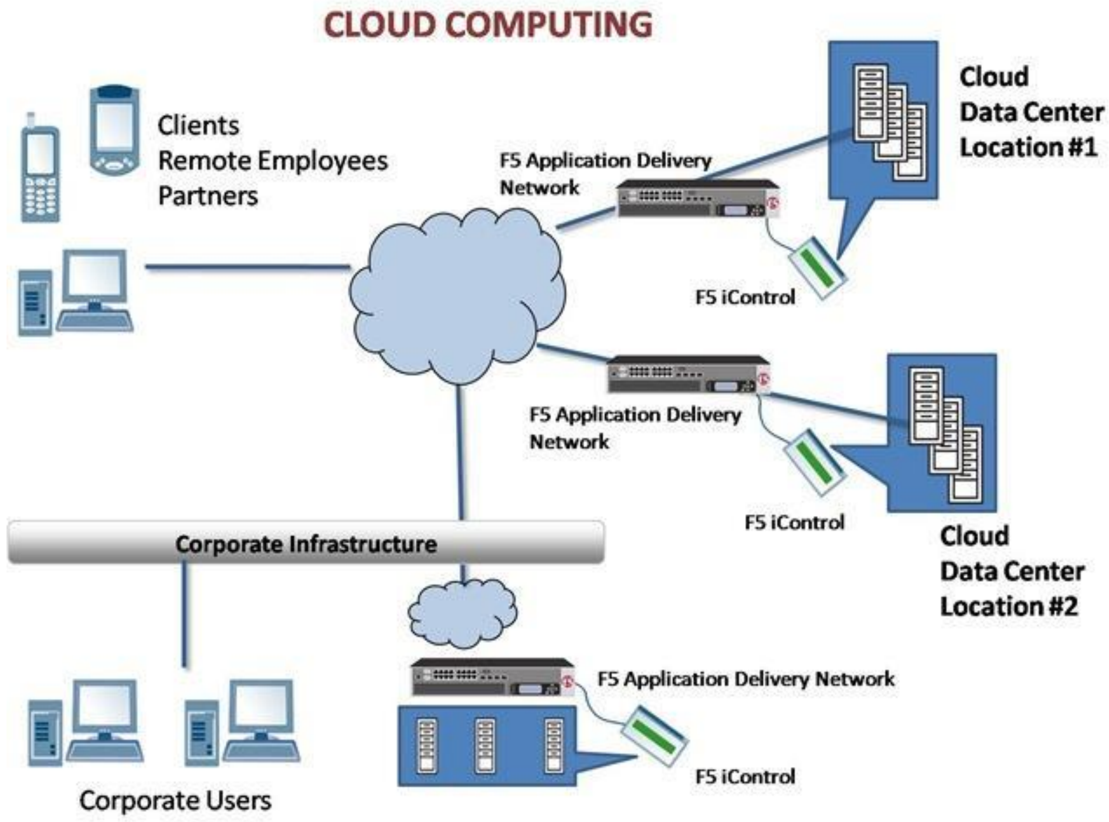

CLOUD COMPUTING
Clients
Remote Employees
Partners
F5 Application Delivery Network
Cloud Data Center Location #1
F5 iControl
F5 Application Delivery Network
F5 iControl
Cloud Data Center Location #2
Corporate Infrastructure
F5 Application Delivery Network
F5 iControl
Corporate Users

Figure (2.1): Cloud Computing Architecture

Numerous, including (reducing expenses and increasing the availability of information services to a broader segment of beneficiaries, and also providing the Without a commitment to using a personal computer, information institutions can store, process, transmit, and share data from anywhere and at any time, but all of these processes (storage, processing, transmission, and sharing) are completed once [17].

External servers are accessible via the Internet cloud while maintaining the protection of this data and preventing it from being hacked or infected by viruses [18]. Many of them resorted to subscriptions. Computing projects are made available through some institutions that dedicate their services to libraries, such as the Dura Cloud project [3].

There is a good chance to that do not give cloud computing enough credit in our personal or professional lives, even though we use a variety of cloud services. The term "cloud computing" began to be used in the early 1990s. Is the term "cloud computing" inspired? From the cloud symbol that was often used to represent the Internet in graphs and maps. Moreover, as with so many other new [19].

Undoubtedly, in light of recent trends in information technology, the nature of the Internet is constantly evolving and changing, which has led to the emergence of successive generations of development in the structure, content, and products of the network [20].

Cloud computing is a type of cloud computing where the consideration is estimated according to what each customer consumes in terms of processing capabilities, storage space, and volume. In other words, can use the computer to communicate via memory, several clients are allowed to work, etc., in other words, a network and store programs, files, etc. All these resources are stored in the cloud (i.e., data centers) and become the computer's tool for communicating with this cloud. And this is the case with the various computers in a company[20].

Instead of putting the applications online, they are working on employees' devices. These applications are installed in the cloud, and work is done as usual. Cloud computing has emerged as a significant new trend in technology, and many industry experts anticipate that this shift will result in changes to the processes involved in information technology (IT) as well as the processes involved in the information technology market. Users can access their cloud computing resources from a variety of devices when they use the technology, including desktop computers, laptops, phones, smartphones, and digital assistants, to access software, storage, and online platforms for the development of applications through the use of services offered by cloud computing providers. The use of cloud computing can result in significant cost reductions, increased availability, ease of assimilation and other benefits [21].

2.2.1 Cloud Computing Benefits

Cloud computing offers numerous advantages to the user at the receiving end. These consist of the following:

- Access to a huge selection of applications without the need to install or download [22].
- supplying educational resource, databases, e-mails, tools and educational applications to students, teachers, and clients located all over the world who are taking part in an educational program [23].
- Access to the applications is possible from any computer, in any part of the world [24].
- Users are able to save money by only purchasing the hardware and software that is necessary for their tasks [23].
- There is a central location where businesses can pool their resources [25].
- Consumption is billed as a utility with minimal upfront costs [19].
- Scalability through the utilization of resources on demand [26].

2.2.2 Cloud Computing Challenges

Despite all the justifications and benefits that have been said and written about cloud computing, some human, material and procedural defects emerge from time to time, which impede, from near or far, the ways of implementing cloud computing and which are:

1. **Environment**: There are those who believe that cloud computing is not necessarily green computing, as it causes more pressure on the Internet, and increases the number of copies of the same data on more than one cloud.

2. **The Internet**: Cloud applications need to be connected to the internet, as disconnecting from the internet will affect the inability to perform work [27], but some institutions have taken the initiative to remedy this, and thanks to some modern HTML5 and JavaScript technologies, it is possible to build web applications [28][29] to work without an internet connection, then synchronize when the connection is back.

3. **Security and security concerns**: Some are afraid to put all their information and files with companies providing cloud services. When the service is expose to hacking operations, the hacker may be able to obtain users' information, or the company may resort to selling user information or benefiting from it in one way or another. The only guarantee is in such cases, it is to resort to large companies with high reliability and good reputation in this field. (On the other hand, your private equipment and your computer are not immune from hacking, theft or loss, and cloud service companies are more secure to store and save information, but these concerns remain with some users).

4. **Where to save files**: The user does not know where his information or files should be saved. It is possible that you are, for example, in a cloud server in a hostile country, and therefore

political problems enter here, but with the cloud you can choose places to save those data or files and the provider companies try hard in this area, to avoid this defect.

5. **Cloud applications**: They are not as good as traditional desktop programs yet, and photo editing applications over the web have not reached levels comparable, for example, to the traditional Photoshop application. Applications have not yet reached the level of traditional desktop applications.

6. **Editing documents** online has reached the level of Microsoft Office, but it is gradually approaching this as the years go by.

7. **Internet speed**: The use of cloud computing is plagued by this issue in a number of less developed countries.

8. **Protection of intellectual property rights**: which raise users' concerns, there are no guarantees that these rights will not be violated.

9. **The problem of information security and privacy**: Some users are afraid of the possibility of others accessing their private information [28].

2.3 Vogel’s Approximation Method (VAM)

Vogel’s method is among of most important third approaches (Vogel’s Approximation Method (VAM), North West Corner Method (NWCM), Least Cost Method (LCM)) [30].The ability to arrive at the best solution as soon as possible and calculations take longer than necessary. The Northwest Corner Method (NCM) and the Lowest Cost Method (LCM), after ensuring that the transfer schedule is in equilibrium.

The transmission problem is a special problem for network optimization. It contains a special data structure in the solution with a transmission diagram. Transportation models play an important role in supply chain and logistics. This question mainly involves calculating the cost of carrying a single commodity from many sources to various destinations [31].

The goal is to decrease the cost of moving goods from one location to another while still meeting the needs of each new region, and each transport location operates within its capacity [32].

The purpose is to use the power of supply point m to determine the most effective means of meeting demand of point n. determining effective solutions the application of operations research to large-scale transportation problems is a critical endeavor [33].

In this thesis, the VAM is one of the transfer methods to find better elemental solutions that have been written about. Heuristic approach VAM often offers a better first solution than other methods. [33].

Applying a VAM to a particular problem does not guarantee that it will produce the best solution. However, a very good solution is always obtained with relatively little effort [34]. In fact, VAM usually provides the best or near-optimal solution to small transmission problems [35]. The notion of VAM is based on the cost of a fee or remorse[35].

The difference between the costs of the greatest and second-largest units in a row or column is the penalty cost. VAM gives as much money as feasible to the cell with the lowest penalty cost per row or column [36].

2.3.1 Vogel's Approximation Method Procedure

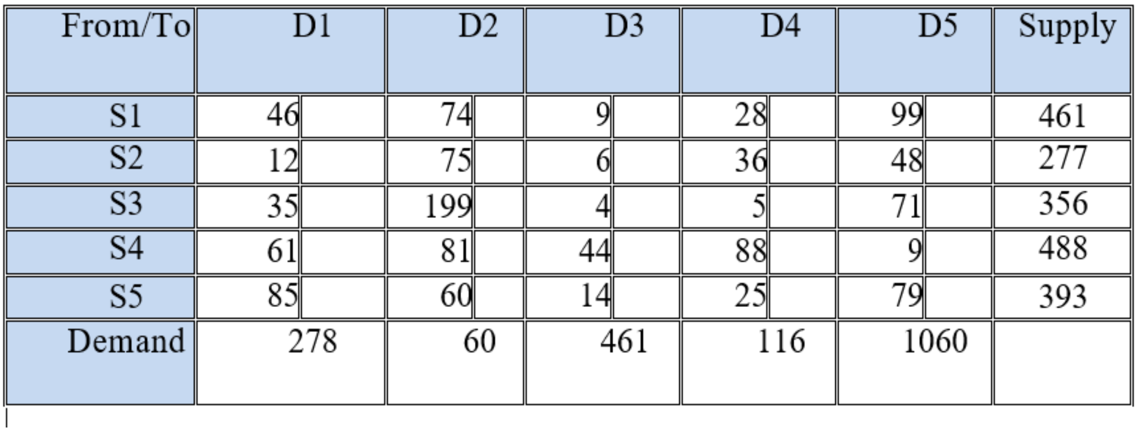
This method is regarded as the best important and efficient integrated for arriving at the best solution, as well as one of most important methods [37]. Since this method is defined by the ability to attain the ideal solution, determining the basic beginning solution is not possible. It's near to the best and uses a scientific estimation procedure that usually yields a basic solution and at this point we explain the VAM method. The detailed process of VAM is as follows [38][39] as shown in Table (2.1):

Table (2.1): Vogel’s Approximation Method Procedure

Step 1	The provided transportation issue must be balanced if either (total supply>total demand) or (total supply<total demand)
Step 2	Subtract the lowest cell cost in the row or column from the next lowest cell cost in the same row or column to get the penalty cost for each row and column.
Step 3	Break ties randomly or choose the lowest-cost the cell with the highest penalty value of cost in the column or row).
Step 4	Allocate as much as possible to the viable cell in the row or column with the highest penalty cost that has the lowest transportation cost.
Step 5	Steps 2, 3, and 4 should be repeated until all requirements have been met.
Step 6	Calculate the for each of the available allocations, the total transportation cost.

In the example, the size of the matrix is (5 x 5), S1-S5 are the source points, and D1-D5are the destination points. Each of the boxes on the left the fixed cost (cij) is represented by the column and each blank box on the right the column represents the allocated quantity (xij), which is the quantity of units sent from point I to point j [40][41].

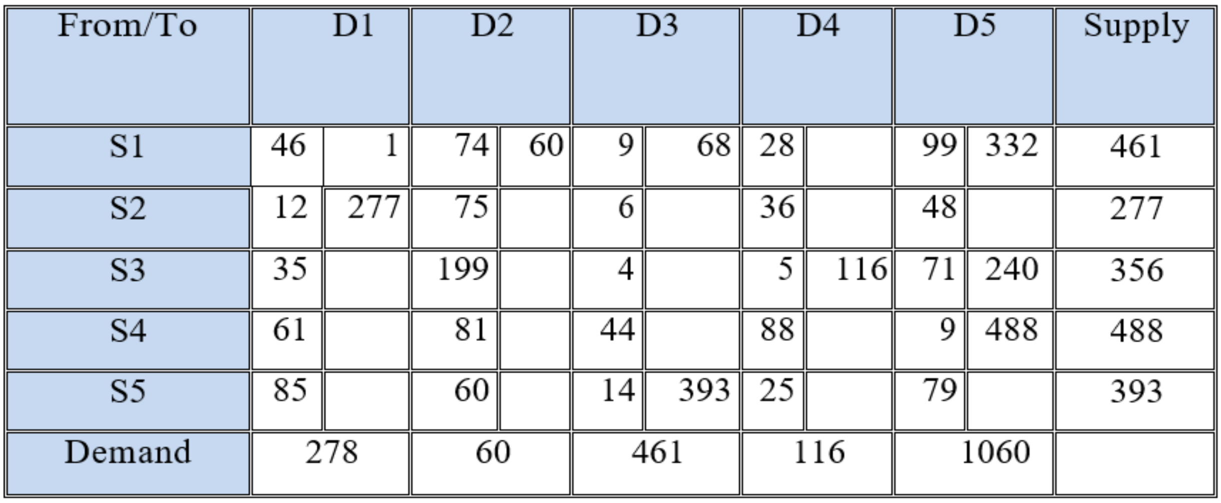
Table (2.2): An illustration of a (5x5) transportation issue



From/To	D1		D2		D3		D4		D5		Supply
S1	46		74		9		28		99		461
S2	12		75		6		36		48		277
S3	35		199		4		5		71		356
S4	61		81		44		88		9		488
S5	85		60		14		25		79		393
Demand	278		60		461		116		1060		

The initial baseline solution VAM was used to find a solution to this problem, which is presented in Table (2.3). The initial cost was determined using the data in Table (2.2) as follows: 68804.

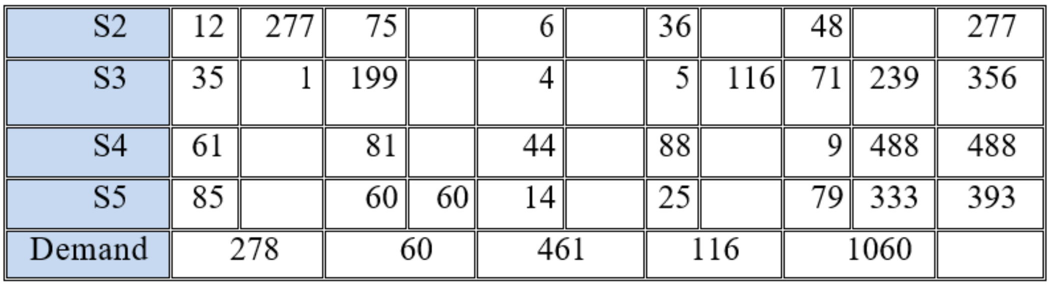
Table (2.3): Initial solution tableau of VAM



From/To	D1		D2		D3		D4		D5		Supply
S1	46	1	74	60	9	68	28		99	332	461
S2	12	277	75		6		36		48		277
S3	35		199		4		5	116	71	240	356
S4	61		81		44		88		9	488	488
S5	85		60		14	393	25		79		393
Demand	278		60		461		116		1060		

It has been reached after achieving the optimal solution using the simple transfer algorithm through five iterations, the ultimate cost was determined to be 59356. Then the optimal solution table was presented in table (2.4). In the following section, we'll go over the proposed method (IVAM) for solving the identical problem[41].

Table (2.4): Optimal solution tableau of VAM



S2	12	277	75		6		36		48		277
S3	35	1	199		4		5	116	71	239	356
S4	61		81		44		88		9	488	488
S5	85		60	60	14		25		79	333	393
Demand	278		60		461		116		1060		

Assume that there is a minor file transfer issue that can be resolved with the VAM approach. The VAM approach can be used to solve a file transmission difficulty.

Table (2.5) is called the FTT file transfer where the problem is replicated on a table. We assume that the client requested four files (f1, f2, f3, f4) of sizes (60, 40, 30 and 110) GB respectively. The list of available replica providers is (S1, S2, S3) at (120, 70, 50) Gb/s with taxa displaycosts listed in Table (2.5). For a The following approach is utilized to make a better selection decision [7].

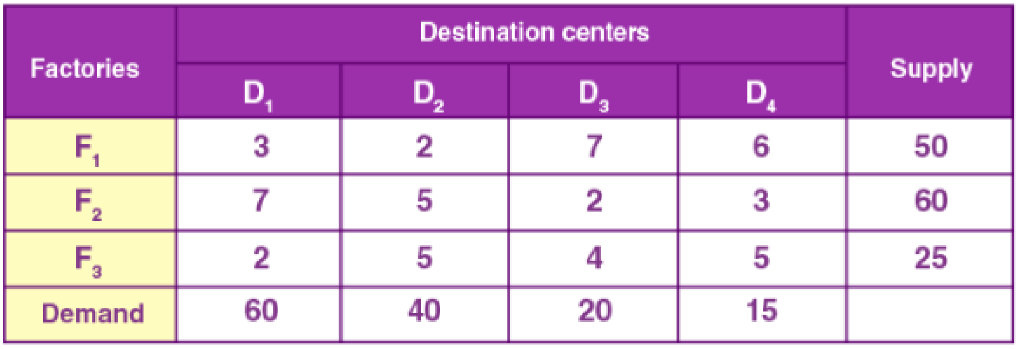
Table (2.5): File Transportation Table

Replica Providers Sites					
Requested File	f_1	f_2	f_3	f_4	Demand
S_1	× (19)	× (30)	× (50)	× (10)	7
S_2	× (70)	× (30)	× (40)	× (60)	9
S_3	× (40)	×(8)	× (70)	× (20)	18
Capability	5	8	7	14	

2.3.2 Vogel's Approximation Method Example

When attempting to determine the initial basic solution that is feasible to a transportation issue, one of the methods that is utilized is known as Vogel's Approximation Method (VAM). However, Vogel's Approximation

Method is an iterative procedure, and so at each step, we need to determine the penalties for each column and row by selecting the least expensive option and the second least expensive respectively[42]. The following example explains the VAM procedure:



Factories	Destination centers				Supply
	D_1	D_2	D_3	D_4	
F_1	3	2	7	6	50
F_2	7	5	2	3	60
F_3	2	5	4	5	25
Demand	60	40	20	15	

Solution:

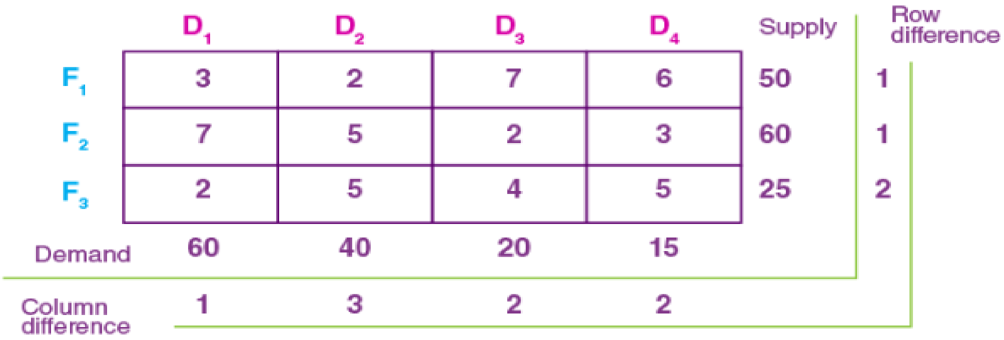
For the given matrix of the cost,

Sum of supply = 25+ 60 +50 = 135

Sum of demand = 15 + 60 + 20 +40 = 135

As a result, the presented issue is a balanced transportation issue.

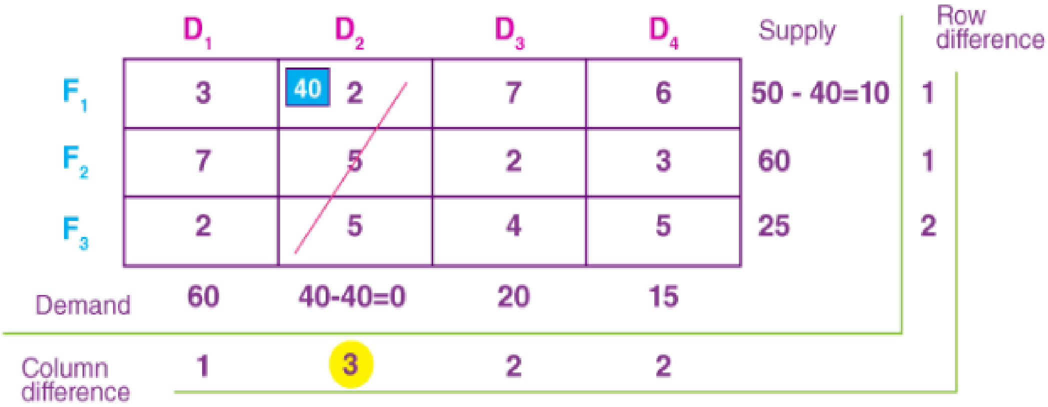
Step 1: Find the lowest and second-lowest cost in each column and row, then write down the differences between these two costs. In the first row, for example, 2 and 3 are the lowest and second-lowest numbers, and their difference is 1.



	D_1	D_2	D_3	D_4	Supply	Row difference
F_1	3	2	7	6	50	1
F_2	7	5	2	3	60	1
F_3	2	5	4	5	25	2
Demand	60	40	20	15		
Column difference	1	3	2	2		

Step 2: Now, find the worst penalty and pick lowest number in that row or column. Then, give each min value of (supply, demand).

Here, the biggest penalty is 3, and the smallest number in the same column is 2. In same cell, min(supply, demand) = min(50, 40) = 40 the put 40 in this cell and cross out the column next to it, demand will be met in this case (40-40=0)**.**



	D_1	D_2	D_3	D_4	Supply	Row difference
F_1	3	40 2	7	6	50 - 40=10	1
F_2	7	5	2	3	60	1
F_3	2	5	4	5	25	2
Demand	60	40-40=0	20	15		
Column difference	1	3	2	2		

Step 3: Find the absolute differences between the rest of the rows and columns. Then do step 2 again.

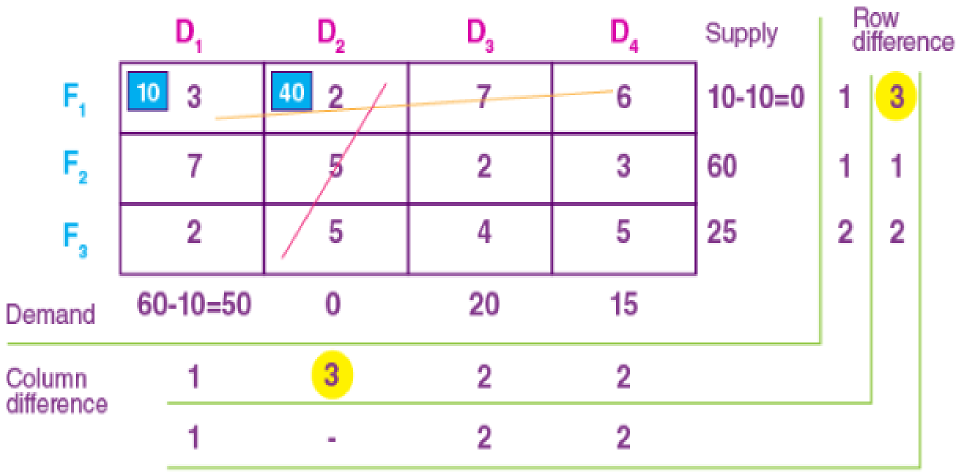
Here, the most you can be fined is 3, and the least you have to pay is also 3. Also, the minimum supply and demand is 10 and 60, which is 10.

So, put 10 into that cell and write down the new supply and demand in the row and column that go with it.

Demand = 60 – 10 = 50

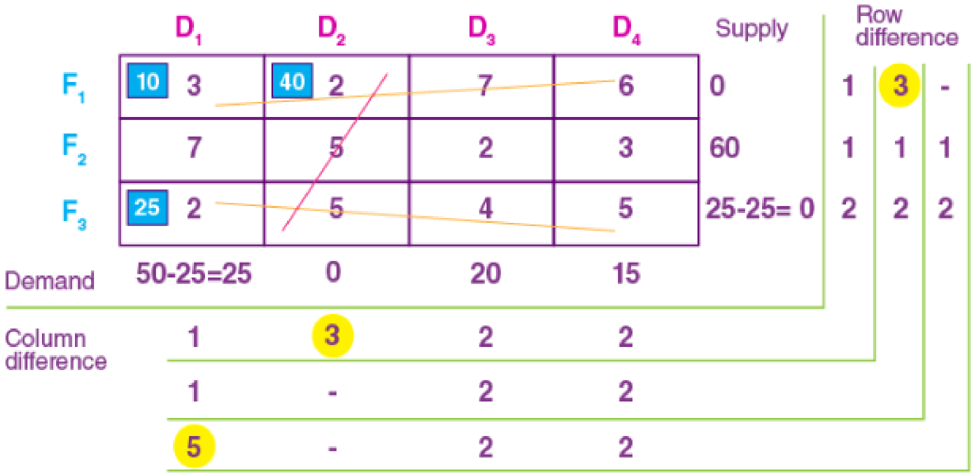
Supply = 10 – 10 = 0

As supply is 0, strike the corresponding row.



	D_1	D_2	D_3	D_4	Supply	Row difference	
F_1	10 3	40 2	7	6	10-10=0	1	3
F_2	7	5	2	3	60	1	1
F_3	2	5	4	5	25	2	2
Demand	60-10=50	0	20	15			
Column difference	1	3	2	2			
	1	-	2	2			

Step 4: Repeat the step above, which was step 3. This will lead to the following:

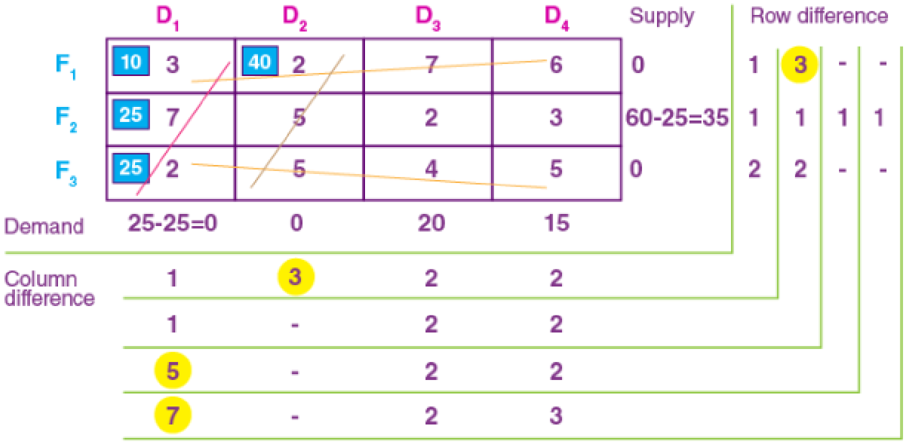


	D_1	D_2	D_3	D_4	Supply	Row difference		
F_1	10 3	40 2	7	6	0	1	3	-
F_2	7	5	2	3	60	1	1	1
F_3	25 2	5	4	5	25-25= 0	2	2	2
Demand	50-25=25	0	20	15				
Column difference	1	3	2	2				
	1	-	2	2				
	5	-	2	2				

In this step, the second column goes away, and the cell with the value 2 gets the value min (supply, demand) = 25.

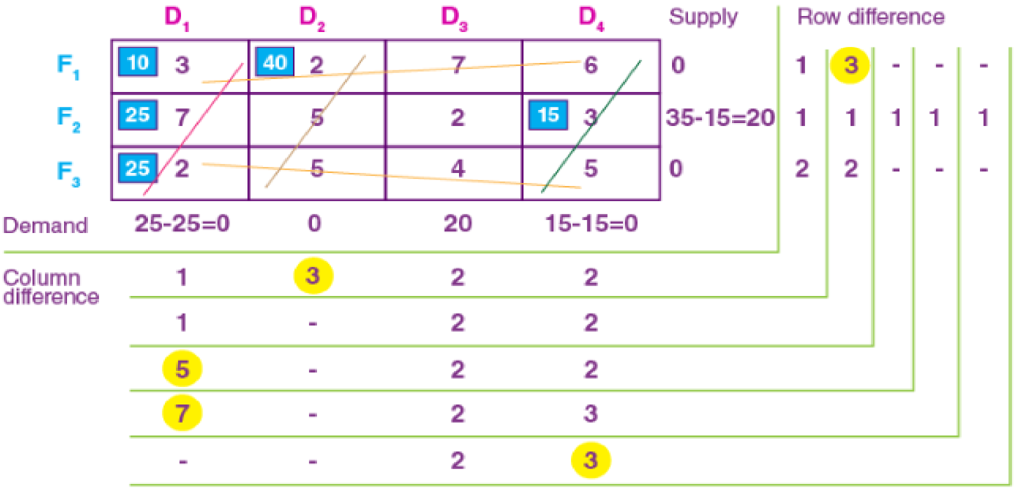
Step 5: Do step 3 again, just like we did in the last step.

In this case, the highest penalty was 7 and the lowest cost in the corresponding column was also 7.



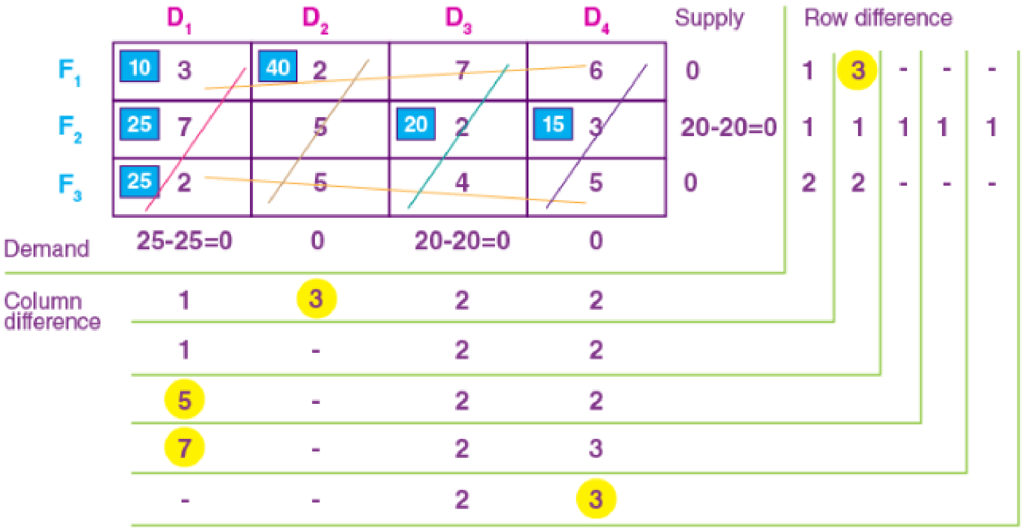
	D_1	D_2	D_3	D_4	Supply	Row difference			
F_1	10 3	40 2	7	6	0	1	3	-	-
F_2	25 7	5	2	3	60-25=35	1	1	1	1
F_3	25 2	5	4	5	0	2	2	-	-
Demand	25-25=0	0	20	15					
Column difference	1	3	2	2					
	1	-	2	2					
	5	-	2	2					
	7	-	2	3					

Step 6: Now, do step 3 again by figuring out the differences between the rest of the rows and columns.



	D_1	D_2	D_3	D_4	Supply	Row difference				
F_1	10 3	40 2	7	6	0	1	3	-	-	-
F_2	25 7	5	2	15 3	35-15=20	1	1	1	1	1
F_3	25 2	5	4	5	0	2	2	-	-	-
Demand	25-25=0	0	20	15-15=0						
Column difference	1	3	2	2						
	1	-	2	2						
	5	-	2	2						
	7	-	2	3						
	-	-	2	3						

Step 7: In the step before, every row and column except one disappeared. Now, put the remaining value of supply or demand in the right cell.



	D_1	D_2	D_3	D_4	Supply	Row difference				
F_1	$\boxed{10}$ 3	$\boxed{40}$ 2	7	6	0	1	3	-	-	-
F_2	$\boxed{25}$ 7	5	$\boxed{20}$ 2	$\boxed{15}$ 3	20-20=0	1	1	1	1	1
F_3	$\boxed{25}$ 2	5	4	5	0	2	2	-	-	-
Demand	25-25=0	0	20-20=0	0						
Column difference	1	3	2	2						
	1	-	2	2						
	5	-	2	2						
	7	-	2	3						
	-	-	2	3						

Total cost = (10 × 3) + (40 × 2) + (25 × 7) + (20 × 2) + (15 × 3)+ (25 × 2)

= 420 [42].

2.4 Cost Reduction

One of the most influential factors that must be taken into consideration is the cost/time reduction factor, as the concept of cost leadership is tohave the lowest operational level of IT costs in the industry in parallel with the quality of IT services. Especially since the financial crisis in 2008, companies have been constantly striving for economic flexibility and IT costs usually belong to the expenditure period, they are under review by top management, some research [43] has demonstrated that the potential TCO benefits of service software in cloud computing are greater compared to traditional on-premises IT, as well as several studies which has indicated the possibility of reducing pre-existing costs under cloud computing [44].

A 2013 study by Morgan and Conboy [45], cited the case for the three companies which showed that companies had reduced costs for servers, licenses, maintenance, backup, and electricity. But they also point out that there may be hidden costs such as additional training courses, therefore, an implicit assumption that has been widely cited is a near- natural occurrence of cost reduction through cloud computing adoption.

A 2012 study by Meer et al. [46], demonstrated that the cost effect is beneficial under real conditions by monitoring transportation and distribution processes. The study also indicated that the use of cloud computing helped improve the scalability of the data layer in the multi-application Internet by directing a request to the database instance thatit can process using the minimum amount of work general, capital expenditure prevention depreciation), and administrative costs for the user, but the increased cost is to the service provider, as contracts areoften long-term due to higher switching costs [47]. Moreover, in the short term the freedom of choice to repeat upgrades and downgrades depends on the economic situation.

2.5 Ways to Find the Best Solution

They are the methods for selecting the acceptable basic solutionobtained from the basic solution methods And we make sure of the solution we reached, is it an optimal solution (Solution Optimal), or we do improve it in case it is not optimal by using other methods to obtain the optimal solution that The value of the objective function for the costof transportation shall be as low as possible, and this is done accordingto one of the two methods [48]:

2.5.1 Stepping Stone Method

Once a fundamental, workable solution has been computed to check whether the answer is necessary, one must now proceed. Whether or if the result is ideal, the Stepping Stone Method seeks to solve transportation issues in the best possible way [49].

2.5.2 Method Multipliers

The improvement indicator is calculated using mathematical equations in this manner rather than by drawing paths as it was done in the past, and the indicator with highest values is picked. Negative to begin solution optimization [50].

2.6 Apache Hive

Apache Hive is an open source data warehouse software that enables users to read, write, and manage massive data set files [51]. These files can be kept directly in the Apache Hadoop Distributed File System (HDFS) or in other data storage systems such as Apache HBase [52].

Hive gives SQL developers the ability to query and analyze data using HQL statements, which are quite similar to regular SQL queries. These HQL statements are written in Hive Query Language. The programming for MapReduce will be simplified as a result, relieving us of the need to be familiar with and compose extensive Java code [53].

Instead, can write queries in HQL, which is a simpler language, and then Hive automatically generate the map and reduce the functions [54]. Can able to apply a table structure to enormous volumes of unstructured data thanks to the Hive metastore, which is included with installation of the Hive data warehouse platform. After have created a Hive table and specified its columns, rows, data types, and other attributes, all of this information is saved in the metastore and is considered to be a component of Hive architecture [55].

Other tools, such as Apache Spark and Apache Pig[52], can then access the information included in the metastore. Figure (2.2) illustrate the architecture of Apache hive [52]:

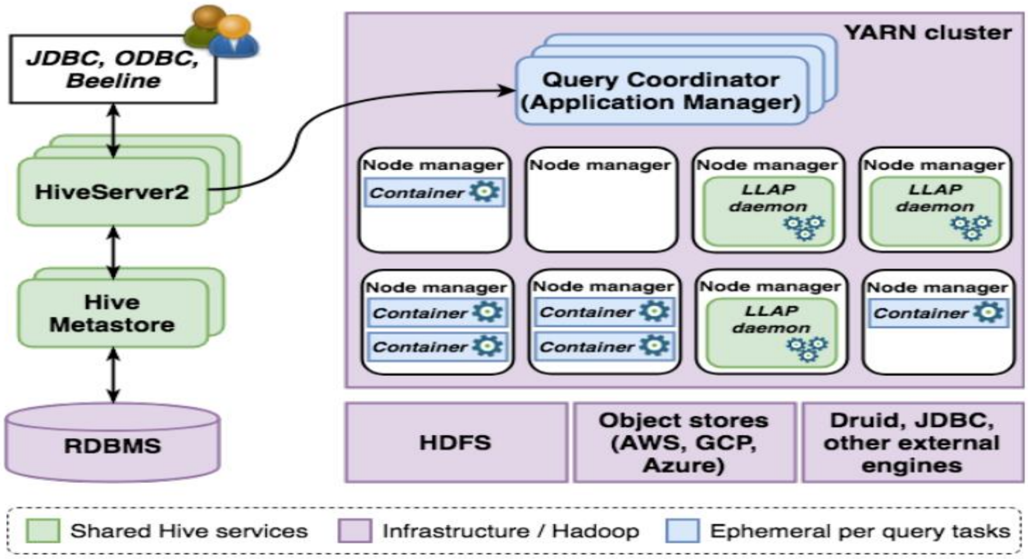


Figure (2.2): Apache Hive Architecture

As is the case with any database management system (DBMS), we are able to execute our queries on Hive from a command-line interface (also referred to as the Hive shell), from a JavaTM Database Connectivity (JDBC) application[56], or from an Open Database Connectivity (ODBC) application by utilizing the Hive JDBC/ODBC drivers. It is possible for us to execute a Hive Thrift Client inside of applications that have been written in C++, Java, PHP, Python, or Ruby [57]. This is analogous to using these client-side languages with integrated SQL in order to access a database like IBM Db2® or IBM Informix®.

Hive is designed to resemble conventional database programming and supports SQL access [56]. However, because Apache Hadoop and Hive operations serve as the foundation for Hive, there are significant distinctions between the two. First, Hadoop was designed for lengthy sequential scans [58], and because Hive is built on top of Hadoop, query latency is extremely high (several minutes) [59]. Because of this, the use of Hive is not recommended for applications that require extremely quick reaction rates. Second, because Hive is a read-based system [60], it is not suitable for transaction processing [56], which often involves a high proportion of write operations. It is better suited for data warehousing operations such as Extract, Transform, and Load (ETL), reporting, and data analysis, and it contains tools that offer easy access to data using SQL [61]. Hive gives the ability to impose structure on data that is, for the most part, unstructured [62]. Therefore is used to deal with the cost table as explain this in Chapter Four.

2.7 The Standard Deviation for Testing the Results

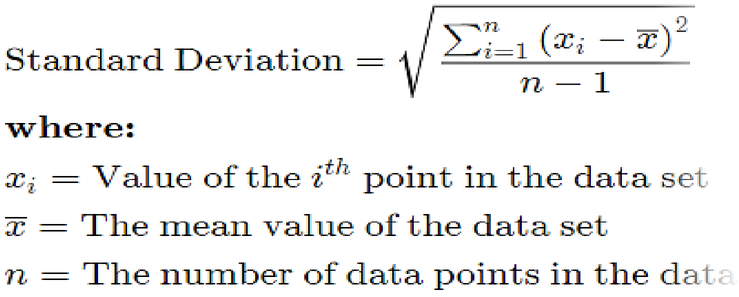
Standard deviation is a statistic that measures how spread out a set of data is in relation to its mean. It is calculated by taking the square root of the variance by determining each data point's deviation relative to the mean [63].

If the data points are more away from the mean, then there will be a greater deviation within the data set; hence, the more spread out the data are, the larger the standard deviation will be.

2.7.1 Standard Deviation Formula

The formula for calculating standard deviation involves finding the square root of a value that is obtained by comparing individual data points to the population's overall average.

The formula is [64]:



$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

where:

x_i = Value of the i^{th} point in the data set

$\overline{x}$ = The mean value of the data set

n = The number of data points in the data

2.7.2 Using Standard Deviation

The standard deviation is a crucial tool in trading and investment techniques since it helps monitor market and securities volatility and predict performance patterns. This makes the standard deviation an especially helpful instrument. When it comes to investment, for instance, the objective of an index fund is to mimic the performance of its benchmark index; hence, the fund is likely to have a low standard deviation in comparison to the benchmark index. One of the most important basic risk measurements that analysts, portfolio managers, and financial advisors utilize is the standard deviation. The standard deviation of their mutual funds and other products is something that investment companies declare publicly. A high dispersion indicates that the actual return on the fund is significantly different from the normal returns that were anticipated. This number is routinely presented to the end customers and the investors because it can easily be grasped by both of these groups [64].

Chapter Three

System Design and Implementation

3.1 Introduction

The proposed system illustrates the parts that process the implementation of the system. The main idea of the system is to specify each client-requested specific file to the server which will deliver the file more efficiently than and as fast as possible.

The steps followed to achieve the key aim of this thesis are described. This includes proposing an approach for conclude the minimum cost time needed for each file delivery to the user. The algorithm used is called Vogel’s Approximation Method (known as VAM). VAM algorithm is running inside a middle server which receives client’s files requests and determines each client which server downloads from, using the Apache Hive to distributing the CostTable (which will be large in the real world) among various Hive nodes.

3.2 Proposed System Architecture

There are three parts in the proposed system. As described in Figure (3.1), these steps are:

1- Cloud Computing Part: which contains the servers and holds the cost table and related data for VAM method.

2- VAM Server Part: which acts as middle party to run the method and deliver the links for each server to each corresponding client.

3- Clients Part: users who request files from VAM.

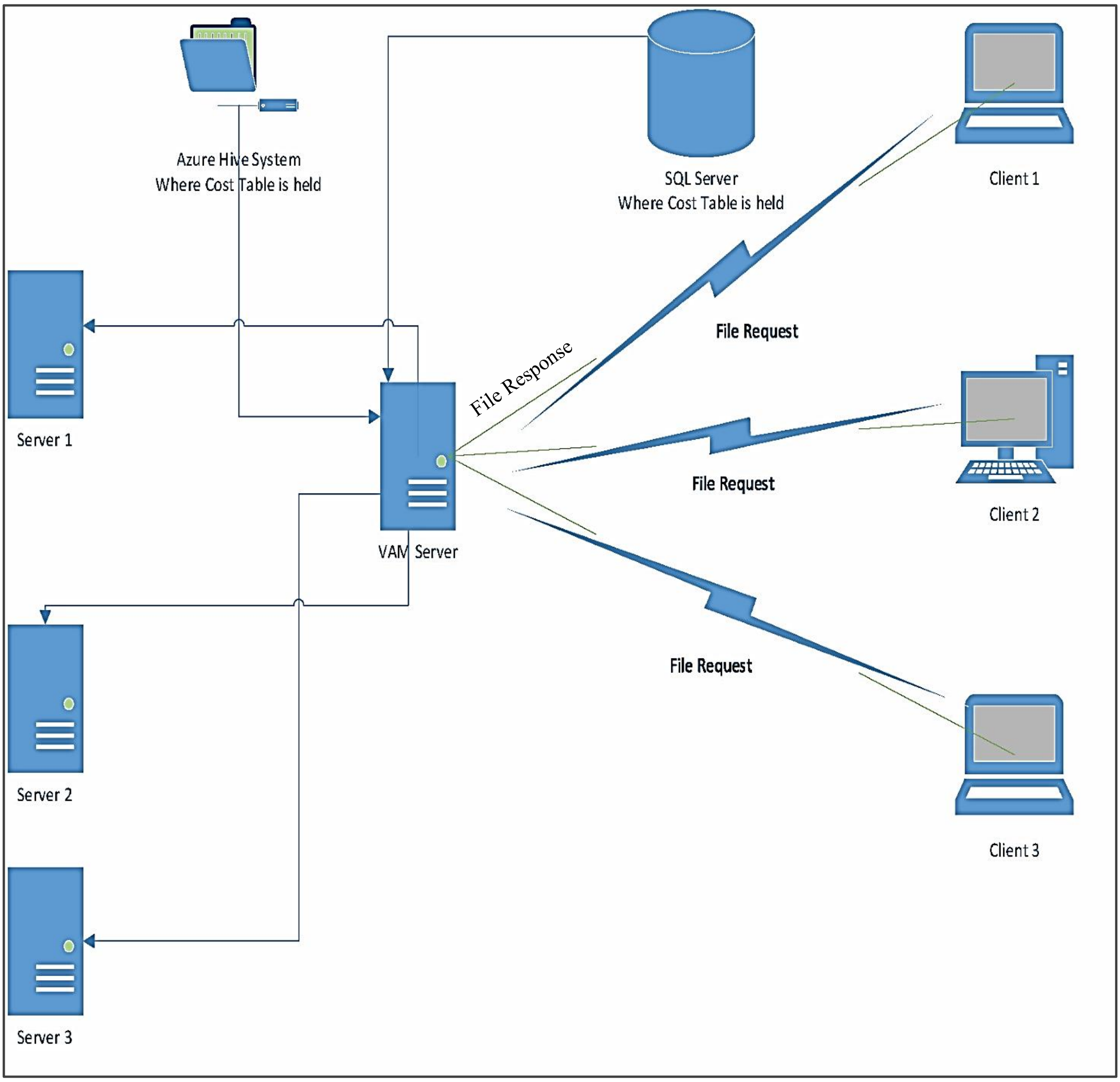


Figure (3.1): Proposed System Architecture

3.2.1 Collecting Dataset

In this thesis, three files, two of movie type (.mp4) and one of text type (.csv) collected by downloaded from the web browser. Table (3.1) lists the sources for each file and the size.

Table (3.1): The Dataset

File	Type	Size (GB)	Sources
Case1	Mp4	6.5	https://www.youtube.com/watch?v=TEsIeYLsxsk
Case2	Mp4	2	https://www.youtube.com/watch?v=ZrUYRGvKT2w&t=211s
Case3	csv	1.3	https://www.unb.ca/cic/datasets/ddos-2019.html

3.2.2 Cloud Computing Part

VAM Server uses a table called CostTable, this table is used by the VAM method to determine which client should request the file from which server. This table could be huge (according to the number of servers and clients) and it must be stored in Cloud and fetch this table whenever VAM Server needs it. To achieve this, there are two approaches:

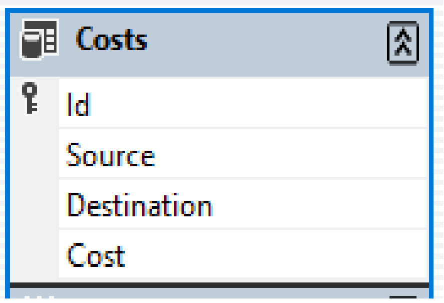
The first method, CostTable stored in Microsoft SQL Server. Microsoft SQL Server can hold structured data and manipulate these data by SQL queries via TCP/IP connection, but this approach has a problem, the problem is CostTable (as we mentioned earlier) could be huge data amount. In addition, this could cause the lack of requests this table in the time needed. Microsoft SQL Server doesn’t have the proper techniques for distributing the data and minimizing the time needed to gather the table and send it to the requested part.

So, the second method is proposed to solve this issue.

The second method is to stored CostTable in Microsoft Azure Service, this service contains Hive techniques to distribute the Table among server nodes (nodes mostly is represented as PCs) and by using special algorithms, the data could be gathered and send it to the party as fast as possible.

In the proposed system, had implemented both methods to compare these methods, in order to illustrate the power of Microsoft Azure HDInsight with dealing big data. Indeed, Azure Hive is the most efficient way to deal with CostTable as explain this in Chapter four.

The following figure describes the CostTable contents:



Id
Source
Destination
Cost

Figure (3.2): Cost Table

3.2.2.1 Microsoft Azure HDInsight

Azure HDInsight is a Microsoft service that provide Hadoop, Spark, R Server, HBase, Hive, and Storm clusters, and many distributed techniques to deal with big data [65].

CostTable has been built in a structured way (.csv file) and then uploaded to storage in Azure HDInsight. In Azure Cloud, some settings must be configured in order to distribute these data in a matter to be sent whenever requested by using queries just like SQL query. Figure(3.3) illustrate the Azure HDInsight Hive Services Structure [61].

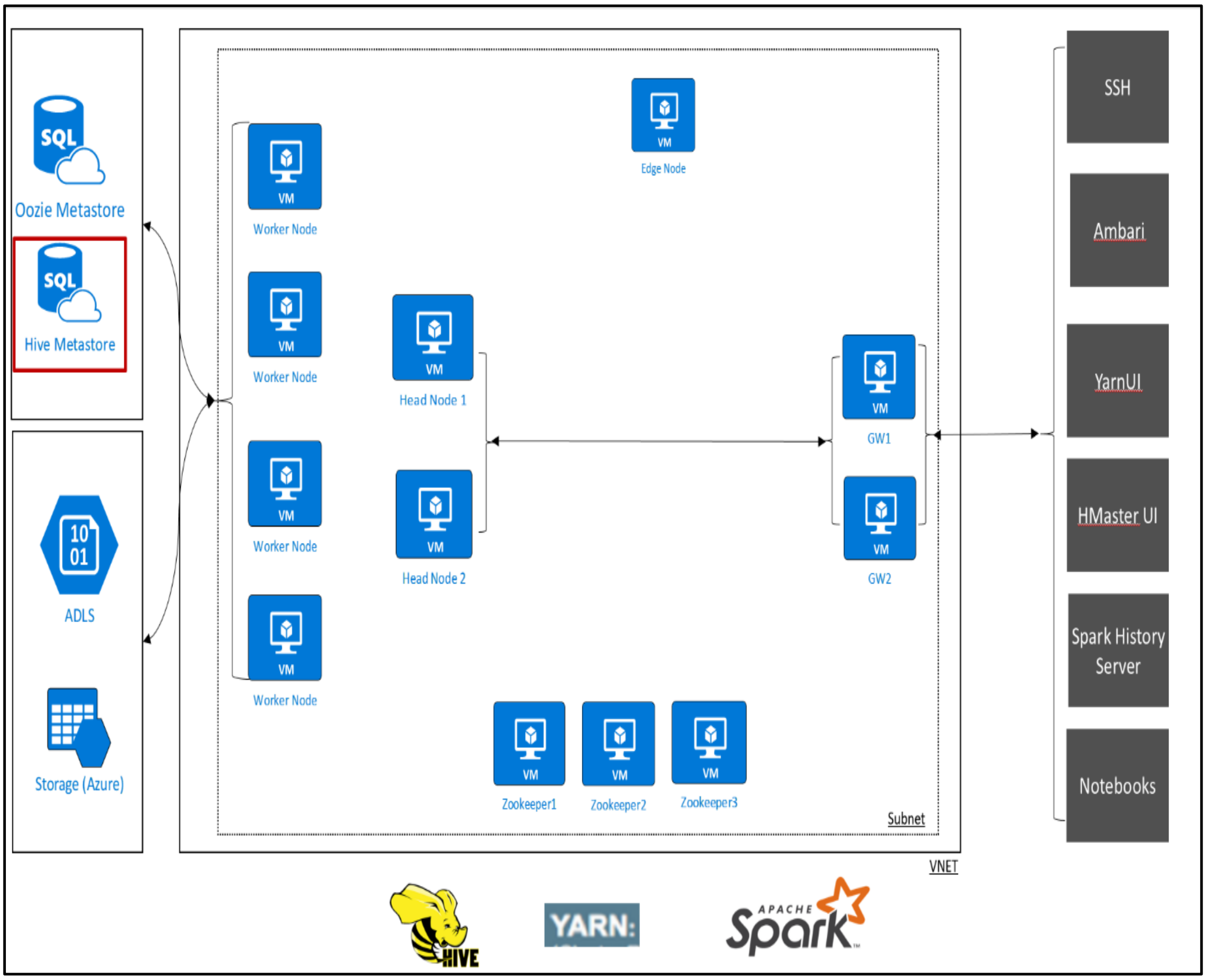


Figure (3.3): Azure HDInsight Hive Service Structure

HDInsight offers a number of different cluster types, each of which is optimized for a particular kind of workload. Query that is Interactive[66]. A Hadoop cluster that improves response times for interactive queries by providing Low Latency Analytical Processing (LLAP) functionality.

A Hadoop cluster that has been optimized for carrying out batch processing tasks. Working with Hive is made easier thanks to functionality that is built into Spark Apache. To query the data that is stored in Apache HBase, you can use either HBase or HiveQL.

3.2.2.2: Implementation Azure HDInsight:

An unprocessed CSV data file is used as a starting point (CostTable.csv), it is then loaded into Azure SQL Database using Apache Sqoop after being imported into an Azure HDInsight cluster where it is then transformed using Apache Hive.

To upload CostTable is used the following stages:

- Retrieve the information and then submit it to an HDInsight cluster.
- Utilize Apache Hive to perform the transformations on the data.
- Sqoop should be used to load the data into the Azure SQL Database.

Accordingly, an Azure Subscription must be created.

3.2.3 VAM Server Part

The core and main function of the proposed system is VAM server, which acts as an access point for clients and analyze the demands of clients to supply them with the files needed for each corresponding server. VAM Server (by it’s called) uses Vogel’s Approach Method Algorithm (shorted by VAM) as analyzing method to determine each client demand with servers that reducing the cost of time to deliver the file for each client. The following structure in figure (3.4) explains how the VAM server works:

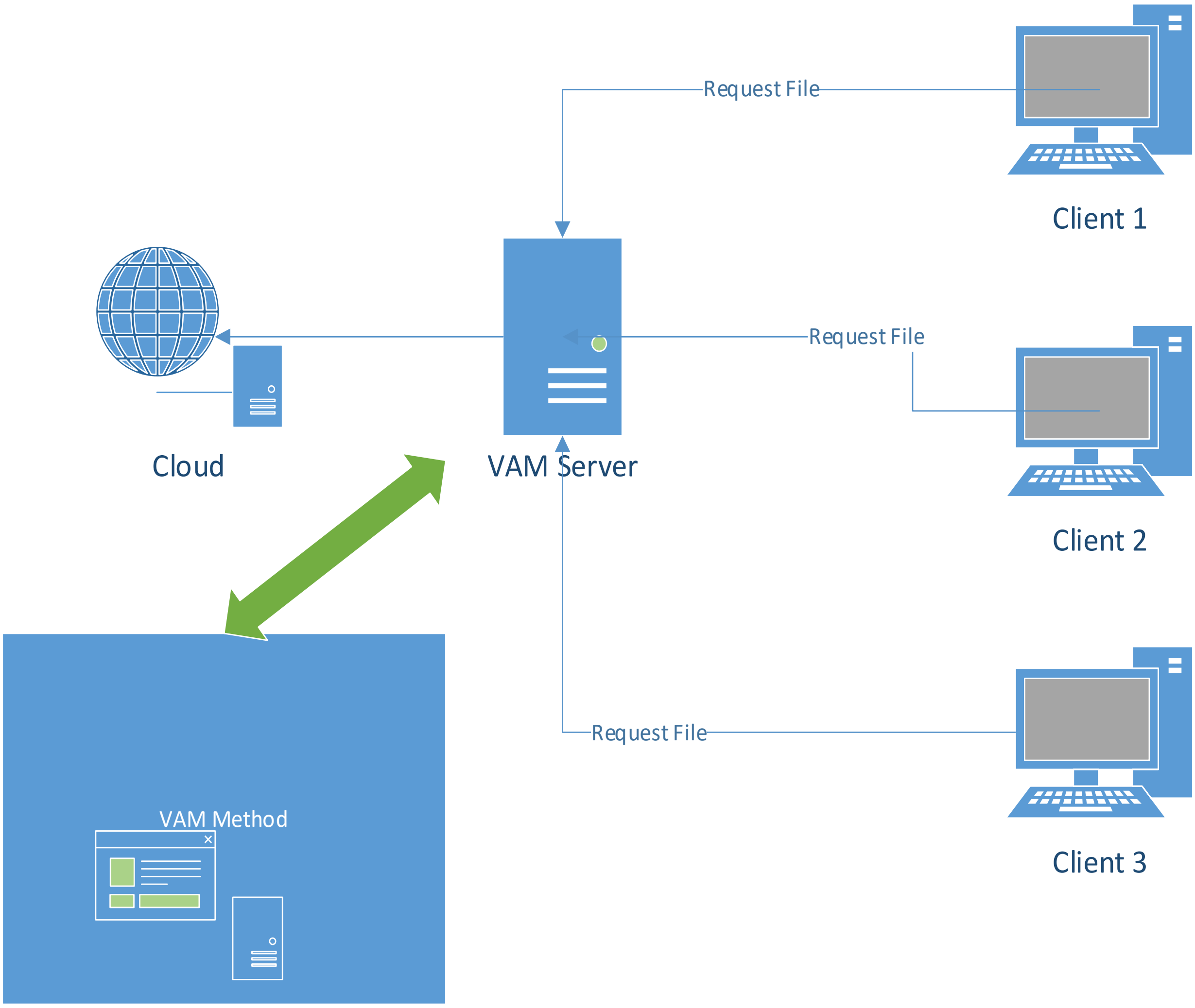


Figure (3.4): Structure of VAM Server

In the proposal system, there are three clients requesting files, in real world, there are thousands of clients requesting thousands of files at the same time. Some of these clients could request same file. In anyway, VAM Server will use VAM method to analyze the request and make decision which client should download the file from.

The following flowchart as shown in Figure (3.5) describes how the request and data flows throw out the VAM Server:

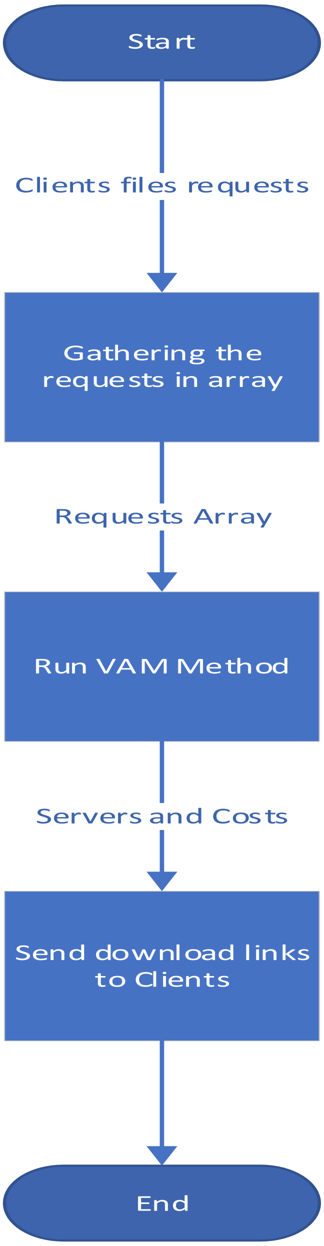


Figure (3.5): Flow Chart of request and data flows throw out the VAM Server

These Clients are communicating with VAM Server using TCP/IP communications. The requests send throw out the network using TCP/IP to VAM Server (which acts exactly as Server) and running the VAM method and returns the download links throw out the network using TCP/IP again.

3.2.3.1 Vogel's Approximation Method (VAM) Algorithm

The steps of algorithm involved in solving the transportation problem using VAM such as in the step1 the method determine a penalty cost for each row (column) by subtracting the lowest unit cell cost in the row (column) from the next lowest unit cell cost in the same row (column), then in step2 identify the row or column with the greatest penalty, then allocate as much as possible to the variable with the lowest unit cost in the selected row or column, after that adjust the supply and demand and cross out the row or column that is already satisfied and repeat until the best solution true, these steps show in algorithm (3.1) as bellow:-

```
Algorithm (3.1): Vogel's Approximation Method VAM
Notations:
  ✓ Requested file(s),$f_i$, $i=\{1,2,…,n\}$, where $n$ representes number of files
  ✓ Replica providers sites, $S_j$ , $j=\{1,2,…,m\}$, where $m$ represents number of providers sites
  ✓ Cost (time) of each file $C_{ij}$ (Cost of $f_i$ in replica provider $j$)
  ✓ Demand is to determine required each file
  ✓ Supply is to determine the ability of suppliers to send files
  ✓ Penalty $P_i$, represents the difference of the distribution costs between the two lowest unit costs (first best route and second best route)
Begin
  Step 1: For each row and column, find the Penalty $P_i$
  Step 2: Identify the row or column with the largest Penalty values
  Step 3: Assign as many demand units as possible to the lowest cost supplier that belongs to the row or column selected
  Step 4: Eliminate any row or column when its ability of the supplier becomes zero.
  Step 5: Re-compute the cost differences for the transportation table, omitting rows or columns crossed out in the preceding step
  Step 6: Return to step 2 and repeat the steps until an initial feasible solution has been obtained
End
```

3.2.3.2 VAM Server Application

Microsoft Visual Studio designs the main form of VAM Server application and the contents are the button (Apply VAM by Hive) shows the time of fetch CostTable from Azure Hive to the VAM method and the button (Apply VAM by SQL) shows the time of fetch the CostTable from SQL and the value of CostTable shows in the middle of this form as shown in Appendix B. TCP/IP protocol was used to communications between Client and VAM Server

3.2.4 Client Part

In this section, illustrate the client application that requests file from VAM Server. As mentioned client communicates with VAM Server using TCP/IP protocols. Thus, Client Side uses Client procedures to send the request file to VAM server.

The form of client application designed by Microsoft Visual Studio also and the content is button (Download using Http) used to download the file that selected in File scroll and show the time of download this file in Download Time text box as shown in Appendix B and the button (Download using VAM) used to download the file that selected by VAM method and show the time of download this file in Download Time text box.

File download step by step after receiving the link from VAM Server, thus, the progress will be displayed as progress bar in user interface.

Note that, clients can request same file or different files. In either way, VAM Server will handle these requests and communicate with these clients in order to exploit the reduction of cost time.

The speed progress of the download can be various from client personal computer (PC) to another. The factors controlling the download are:

1- Internet and network speed.
2- PC Ram, CPU, and type of Hard drive.

The test PCs properties used in the proposed system are described in Table (3.2):

Table (3.2): The test PCs properties

	RAM	CPU	Type of Hard Drive	Internet speed
PC1	4 GB	Intel Core i5 2.40 GHz	HDD	6.9 Mbps
PC2	8 GB	Intel Core i5 2.40 GHz	HDD	12.5 Mbps
VAM Server	16 GB	Intel Core i7 2Core 2.40 GHz, 2.40GHz	SSD	50 Mbps

As described above, VAM Server has the highest PC properties, which chosen like this to handle the load of clients high number of requests.

On the other hand, client's PCs does not need high RAM or CPU or either SSD hard drive. Client supposes only request file and download it, so no need for high properties. Sure, Client's PC properties could affect the download speed, but not much effect on the results. In chapter four, the results of downloading files not various from one to another so far as shown in Table (4.4) and Table (4.5).

3.3 Result Testing

In this stage, the results obtained in the same work environment are tested using standard deviation and evaluated whether they are good results or not, where the value of standard deviation lower than one that mean the result is good .

Chapter Four

Experimental Results and Discussion

4.1 Introduction

The main result of the proposed system is the rate of reduction of time cost of delivering the file to the client after requesting it. In regular approach, client requests file using normal http request, the speed of delivering the file count on the speed of Internet and the distance between the client and the server the holds the file in its storage. Thus, if the file requested is too far from client and the same file is near, but the client requested the one in the far server, the client will suffer waiting for long time more than needed if he requested in the near server.

This conclude that the need for a method to determine the better server is extreme important. This proposed system (as explained in chapter three), is to use VAM method to fit such a need. The results that will be explained in this chapter, and illustrate the difference between the proposed approach and the regular approach. The results into two sides:

1- **Cloud side**: which will improve the proposed project more by distributing the Cost Table (which will be large in real world) among various Hive nodes.

2- **VAM side**: the results of the goal of this research.

4.2 Hardware and Software Requirements

The proposed approach is implemented using the following hardware and software requirements.

- **Hardware:** three PCs as shown in Table (3.2).

- **Software:**
  - **Operating System:** Windows10 pro-64-bit.
  - **Programs:** Microsoft Visual Studio Community 2022 (64-bit) – Version 17.1.1, SQL Server Management Studio Version 18.10.
- **Programming language:** Python language.
- **Subscriptions:** SmartASP.NET, Microsoft Azure.

4.3 The Dataset

The dataset that is used to obtain the result as described in section (3.2.1) of chapter three. Which include three files as shown in Appendix B.

4.4 Cloud Side

The results of the Azure side are measured by the time spent fetching the cost table from the cloud to VAM Server. In order to determine whether this service is more suitable from another, we compared this service result with fetching the cost table from regular SQL Server database. Thus, Azure service must show results better than SQL Server.

To do that, Cost Table is created in SQL Server using Query show in Appendix B and the table resides the cost of source and destination of client and server created by SQL Server query.

The same structure also created in Azure Service storage, but this time, the structured Cost Table is created in .csv file as shown in Table (4.1).

Table (4.1) : Cost Table

7196	490	10782
1520	2195	6686
1326	1901	7285

The first access of the table is Servers, and the second is Clients, and finally the cells are the cost from which to another. This file has been uploaded and configured in Azure service.

At first, VAM Server measure the costs by downloading the files from each server and determine how much cost needed for each file, then update the CostTable with these data.

After running clients and requesting files, the following results of brining CostTable are concluded as shown in Figure (4.1):

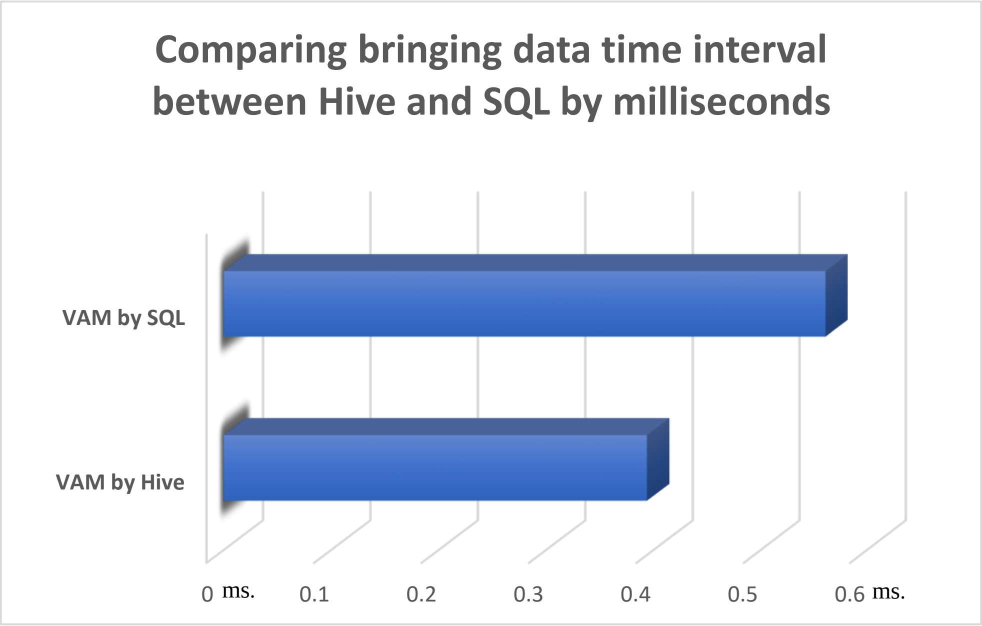


Figure (4.1): Comparing bringing data time interval between Hive and SQL by milliseconds

These results according to the following updated CostTable in Table (4.2):

Table (4.2): Time required of transport the files from each site to client

From/To	File1	File2	File3
Server1	7196 seconds (sec.)	1520 sec.	1326 sec.
Server2	490 sec.	2195 sec.	1901 sec.
Server3	10782 sec.	6686 sec.	7285 sec.

4.5 VAM Server Side

To measure the results of VAM method and its improvement, the amount of time calculated from execution of such a method with regular downloading links are compared. In other words, users can download the requested file using VAM decision.

The results of running VAM method are illustrated as below:

4.5.1 Scenario:

The scenario of the execution the system as following:

Client // request duaa(6.5G).mp4
Client // request quraan(2G).MP4
Client // request portmap(1.3G).csv

4.5.2 Cost Table and Result Table

These results according to the CostTable as shown in Table (4.1) is unbalanced because the demand does not equal to supply therefore the dummy has been added as shown in Table(4.3) and the Result table of the VAM shown in Table(4.4):-

Table (4.3): Cost Table by adding supply and demand (Unbalanced)

From/To	D1	D2	D3	Dummy	Supply
S1	7196	1520	1326	0	3
S2	490	2195	1901	0	3
S3	10782	6686	7285	0	3
Demand	1	1	1	6	9

Table (4.4): VAM Result Table

From/To	D1	D2	D3	Dummy	Supply
S1	7196	1520(1)	1326(1)	0	3
S2	490(1)	2195	1901	0	3
S3	10782	6686	7285	0	3
Demand	1	1	1	6	9

4.6 Experiments and Results

The results of downloading the three file by normal approach and the download of the same files by using the proposed system shown in Table (4.5):

Table (4.5): Comparing the time of downloading files in the usual way and the proposed system method in PC1

From/To	Time by http	Time by proposed system	Ratio of reduction
File1	922 min.	20 min.	98%
File 2	435 min.	10 min.	97.8%
File 3	165 min.	7 min.	96%

Moreover, the result of downloading the three file by normal approach and the download of the same files by using the proposed system in the PC2 as shown in Table (4.6):

Table (4.6): Comparing the time of downloading files in the usual way and the proposed system method in PC2

From/To	Time by http	Time by proposed system	Ratio of reduction
File1	632 min.	9 min.	98.5%
File 2	321 min.	4 min.	98.7%
File 3	113 min.	2 min.	98.3%

Finally, in this proposed system the execution of the system has run in same environments in order to measure the standard deviation and the result of time download the files after using VAM and the standard deviation shown in Table (4.7) and the values of STD are less than (1) that mean the time of download the files are not much different.

Table (4.7): Standard Deviation Result

Files	Time of download the file in Test1	Time of download the file in Test 2	Time of download the file in Test 3	Value of STD
Case1 douaa (6.5 GB)	9.0 Min.	10 Min.	8.5 Min.	0.7637
Case2 quraan (2 GB)	4.0 Min.	5.3 Min.	3.5 Min.	0.9292
Case3 portmap (1.3 GB)	2.02 Min.	3 Min.	1.5 Min.	0.7617

Chapter Five

Conclusion and Future Works

5.1 Conclusions

Vogel's Approximation Method play a major role in finding the best solve. The most important characteristics obtained from the results of the proposed work explained in the following:

1- The use of Vogel's Approximation Method has greatly reduced the cost of transfer time.

2- The proposed algorithm helps to use the least number of servers, which is considered the best.

3- Using the Apache Hive helps to speed the time the project where it is able to deal with large datasets easily.

5.2 Suggestions of the Future Works

To develop the proposed work, many recommendations for the future system are illustrated below:

1- It is possible to use the price cost with the cost of time to get the best results.

2- Using the Smart Vogel's Approximation Method (SVAM) for the same files and make a comparison of the obtained results.

3- Using a larger number of files and a statement of the results obtained from the application of the proposed system.

References

References

[1] V. Kundra, “State of public sector cloud computing,” *Fed. Chief Inf.*, 2010.

[2] Y. Jadeja and K. Modi, “Cloud computing-concepts, architecture and challenges,” in *2012 international conference on computing, electronics and electrical technologies (ICCEET)*, 2012, pp. 877–880.

[3] N. I. of S. and T.-C. Security and Resource Center -, September 2021, “Www.csrc.nist.gov.”

[4] H. A. Hussein and M. A. K. Shiker, “A modification to Vogel’s approximation method to Solve transportation problems,” in *Journal of Physics: Conference Series*, 2020, vol. 1591, no. 1, pp. 12029.

[5] M. W. Ullah, M. A. Uddin, and R. Kawser, “A modified Vogel’s approximation method for obtaining a good primal solution of transportation problems,” *Ann. Pure Appl. Math.*, vol. 11, no. 1, pp. 63–71, 2016.

[6] M. Gunay, M. N. Ince, and A. Cetinkaya, “Apache Hive Performance Improvement Techniques for Relational Data,” in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–6.

[7] R. M. Almuttairi, “Smart Vogel’s Approximation Method SVAM,” *Int. J. Adv. Comput. Res.*, vol. 4, no. 1, pp. 198, 2014.

[8] D. Banik and M. Zahid Hasan, “Transportation Cost Optimization of an Online Business Applying Vogel’s Approximation Method,” *World Sci. News*, vol. 96, no. April, pp. 179–190, 2018.

[9] B. Amaliah, C. Fatichah, and E. Suryani, “Total opportunity cost matrix – Minimal total: A new approach to determine initial basic feasible solution of a transportation problem,” *Egypt. Informatics J.*, vol. 20, no. 2, pp. 131–141, 2019, doi: 10.1016/j.eij.2019.01.002.

[10] M. L. Aliyu, U. Usman, Z. Babayaro, M. K. Aminu. (2019). A Minimization of the Cost of Transportation, American Journal of Operational Research, Vol. 9 No. 1, pp. 1-7. doi: 10.5923/j.ajor.20190901.01.

[11] S. Nannai John and T. T. Mirnalinee, “A novel dynamic data replication strategy to improve access efficiency of cloud storage,” *Inf. Syst. E-bus. Manag.*, vol. 18, no. 3, pp. 405–426, 2020, doi: 10.1007/s10257-019-00422-x.

[12] A. Shetty, S. Mhatre, N. Sinvhal, and K. K. Devadkar, “Peer assisted parallel downloading system,” in *2019 International Conference on Communication and Signal Processing (ICCSP)*, 2019, pp. 650–655.

[13] A. I. Kamba, S. M. Kardi, and Y. K. G. Dikko, “Optimization of total transportation cost,” *Glob. J. Pure Appl. Sci.*, vol. 26, no. 1, pp. 57–63, 2020, doi: 10.4314/gjpas.v26i1.7.

[14] M. Jeon, et al., “Dynamic data replication scheme in the cloud computing environment,” *Proceedings - IEEE 2nd Symposium on Network Cloud Computing and Applications, NCCA 2012*. pp. 40–47, 2012. doi: 10.1109/NCCA.2012.10.

[15] D. Kapil, E. S. Pilli, and R. C. Joshi, “Live virtual machine migration techniques: Survey and research challenges,” in *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013, pp. 963–969. doi: 10.1109/IAdCC.2013.6514357.

[16] M. P. Yadav and M. A. Gulati, “a Novel Approach for Cloud-Based Computing Using Replicate Data Detection,” *J. Glob. Res. Comput. Sci.*, vol. 3, no. 8, 2012, [Online]. Available: www.jgrcs.info

[17] S. M. Kallow, “Cloud Computing: Its Concept and Applications in Libraries and Information Centers,” *QScience Proc.*, 2014.

[18] E. A. Lavrov, A. L. Zolkin, T. G. Aygumov, M. S. Chistyakov, and I. V Akhmetov, “Analysis of information security issues in corporate computer networks,” in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1047, no. 1, pp. 12117.

[19] G. El, N. Moawad, G. Ebrahem, S. Ebrahem, and G. Elnabawy, “The Relationship between use of Technology and Parent- Adolescents Social Relationship,” Journal of Education and Practice, Vol.7, No.14, 2016.

[20] J. Bresnahan, K. Keahey, D. LaBissoniere, and T. Freeman, “Cumulus: an open source storage cloud for science,” in *Proceedings of the 2nd international workshop on Scientific cloud computing*, 2011, pp. 25–32.

[21] L. Qian *et al.*, “Overview of Cloud Computing,” *IOP Conference Series: Materials Science and Engineering*, vol. 677, no. 4. 2019. doi: 10.1088/1757-899X/677/4/042098.

[22] M. Vuyyuru, P. Annapurna, K. G. Babu, and A. S. K. Ratnam, “An overview of cloud computing technology,” *Int. J. Soft Comput. Eng.*, vol. 2, no. 3, pp. 244, 2012.

[23] A. Ghazizadeh, “Cloud Computing Benefits and Architecture in E-Learning,” in *2012 IEEE Seventh International Conference on Wireless, Mobile and Ubiquitous Technology in Education*, 2012, pp. 199–201. doi: 10.1109/WMUTE.2012.46.

[24] S. P. Mirashe and N. V Kalyankar, “Cloud computing,” *arXiv Prepr. arXiv1003.4074*, 2010.

[25] D. Velev and P. Zlateva, “Cloud infrastructure security,” in *International Workshop on Open Problems in Network Security*, 2010, pp. 140–148.

[26] R. L. Grossman, “The case for cloud computing,” *IT Prof.*, vol. 11, no. 2, pp. 23–27, 2009.

[27] E. Ahmed, A. Naveed, A. Gani, S. H. Ab Hamid, M. Imran, and M. Guizani, “Process state synchronization for mobility support in mobile cloud computing,” in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[28] T. Budai and M. Kuczmann, “Towards a modern, integrated virtual laboratory system,” *Acta Polytech. Hungarica*, vol. 15, no. 3, pp. 191–204, 2018.

[29] M. T. Valdez, C. M. Ferreira, and F. P. M. Barbosa, “3D virtual laboratory for teaching circuit theory—A virtual learning environment (VLE),” in *2016 51st International Universities Power Engineering Conference (UPEC)*, 2016, pp. 1–4.

[30] Soomro, A.S., Junaid, M., & Tularam, G.A. (2015). Modified Vogel’s Approximation Method For Solving Transportation Problems. Mathematical theory and modeling, 5, 32-42.

[31] H. A. Taha, “Operations research analysis of a stock market problem,” *Comput. Oper. Res.*, vol. 18, no. 7, pp. 597–602, 1991.

[32] Reeb, J.E. and S.A., Leavengood, “Transportation problem: a special case for linear programming problems”, EM8779. Corvallis: Oregon State University Extension Service, pp. 1-35, 2002.

[33] R. Dutton, G. Hinman, and C. B. Millham, “The optimal location of nuclear-power facilities in the Pacific Northwest,” *Oper. Res.*, vol. 22, no. 3, pp. 478–487, 1974.

[34] H. H. Shore, “The transportation problem and the Vogel approximation method,” *Decis. Sci.*, vol. 1, no. 3-4, pp. 441–457, 1970.

[35] H. A. Taha, “Operations Research: An Introduction". *Macmillan, New York*, 1987.

[36] B. Alsehli, “The Cost of Transporting Oil: a Case Study of Saudi Arabia,” EasyChair, 2022.

[37] Sinjini De Sarkar, Sakshi Kanchan, Shaurya Manohar, Saumya Patodia, Senthil Ramachandran. “Application of Operations Research in Supply Chain Management of FMCG Industry", , International Journal of Innovative Science and Research Technology (IJISRT), Vol. 4 Issue. 10, - 2019, www.ijisrt.com. ISSN - 2456-2165, PP :- 303-308.

[38] N. Joshi, S. S. Chauhan, and R. Raja, “A new approach to solve mixed constraint transportation problem under fuzzy environment,” *Int. J.*, vol. 16, no. 4, 2017.

[39] J. Pratihar, R. Kumar, S. A. Edalatpanah, and A. Dey, “Modified Vogel’s approximation method for transportation problem under uncertain environment,” *Complex Intell. Syst.*, vol. 7, no. 1, pp. 29–40, 2021.

[40] G. B. Dantzig, “Linear Programming and Extensions, Prince.” ton University Press, Princeton, New Jersey (1 963), 1963.

[41] S. Korukoğlu and S. Balli, “An improved vogel’s approximation method for the transportation problem,” *Math. Comput. Appl.*, vol. 16, no. 2, pp. 370–381, 2011, doi: 10.3390/mca16020370.

[42] Karthik., “Vogel’s Approximation Method (VAM),” 2022. https://byjus.com/maths/vogels-approximation-method, Sep. 2022.

[43] Benlian, A. 2009. "A transaction cost theoretical analysis of Software as a Service (SaaS) based sourcing in SMBs and enterprises". ECIS 2009 proceedings paper 4.

[44] W. Venters and E. A. Whitley, “A critical review of cloud computing: researching desires and realities,” *J. Inf. Technol.*, vol. 27, no. 3, pp. 179–197, 2012.

[45] Morgan, Lorraine and Conboy, Kieran, "Factors Affecting The Adoption Of Cloud Computing: An Exploratory Study" (2013). ECIS (2013).

[46] D. VanderMeer, K. Dutta, and A. Datta, “A cost-based database request distribution technique for online e-commerce applications,” *MIS Q.*, pp. 479–507, 2012.

[47] H. Demirkan, H. K. Cheng, and S. Bandyopadhyay, “Coordination strategies in an SaaS supply chain,” *J. Manag. Inf. Syst.*, vol. 26, no. 4, pp. 119–143, 2010.

[48] S. M. Saleh, “using the optimal solution (S.F.B.S) to plan and solve The problem of transportation for the study community,” *East. Co. Frozen Ready-made Foods, Eng. J. Technol.*, vol. 30, no. 2, 2012.

[49] B. Srinivas and G. Ganeshan, “Optimal solution for fuzzy transportation problem using stepping-stone method,” *Int. J. IT Eng.*, vol. 3, no. 3, pp. 185–198, 2015.

[50] M. Al-Amri, “Applications of Operations Research in Adequacy,” *Middle East Univ. Aya Stud. Ithra Publ. Distrib.*, 2009.

[51] J. Anuradha, “A brief introduction on Big Data 5Vs characteristics and Hadoop technology,” *Procedia Comput. Sci.*, vol. 48, pp. 319–324, 2015.

[52] J. Camacho-Rodríguez *et al., 2019. "Apache Hive: From MapReduce to Enterprise-grade Big Data Warehousing". In Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19). Association for Computing Machinery, New York, NY, USA, 1773–1786. https://doi.org/10.1145/3299869.3314045*

[53] U. R. Pol, “Big data analysis: comparison of hadoop mapreduce, pig and hive,” *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 5, no. 6, pp. 9687–9693, 2016.

[54] Abdul Ghaffar Shoro, & Tariq Rahim Soomro. (2015). "Big Data Analysis: Apache Spark Perspective". Global Journal of Computer Science and Technology, 15(C1), Retrieved from https://computerresearch.org/index.php/computer/article/view/1137

[55] A. Thusoo *et al.*, “Hive-a petabyte scale data warehouse using hadoop,” in *2010 IEEE 26th international conference on data engineering (ICDE 2010)*, 2010, pp. 996–1005.

[56] V. Chavan and R. N. Phursule, “Survey paper on big data,” *Int. J. Comput. Sci. Inf. Technol*, vol. 5, no. 6, pp. 7932–7939, 2014.

[57] T. White, *Hadoop: The definitive guide*. “ O’Reilly Media, Inc.,” 2012, ISBN: 978-1-449-31152-01327616795

[58] J. Shafer, S. Rixner, and A. L. Cox, “The hadoop distributed filesystem: Balancing portability and performance,” in *2010 IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS)*, 2010, pp. 122–133.

[59] X. Tian, R. Han, L. Wang, G. Lu, and J. Zhan, “Latency critical big data computing in finance,” *J. Financ. Data Sci.*, vol. 1, no. 1, pp. 33–41, 2015.

[60] S. G. Kolte and J. W. Bakal, “Big Data Summarization: Framework, Challenges and Possible Solutions,” *Adv. Comput. Intellegence An Int. J.*, vol. 3, no. 4, pp. 1–9, 2016.

[61] V. M. Ngo, N.-A. Le-Khac, and M. Kechadi, “Designing and implementing data warehouse for agricultural big data,” in *International Conference on Big Data*, 2019, pp. 1–17.

[62] J. Singh and V. Singla, “Big data: tools and technologies in big data,” *Int. J. Comput. Appl.*, vol. 112, no. 15, 2015.

[63] N. I. O. Education, “Course Hero”, [Online], August 2022, Available: https://www.coursehero.com/file/65877002/TASK-18pdf/, Sep. 15, 2022

[64] M. HARGRAVE, F. Bio, “Standard Deviation Formula and Uses vs. Variance.”, July 05, 2022, https://www.investopedia.com/terms/s/standarddeviation.asp

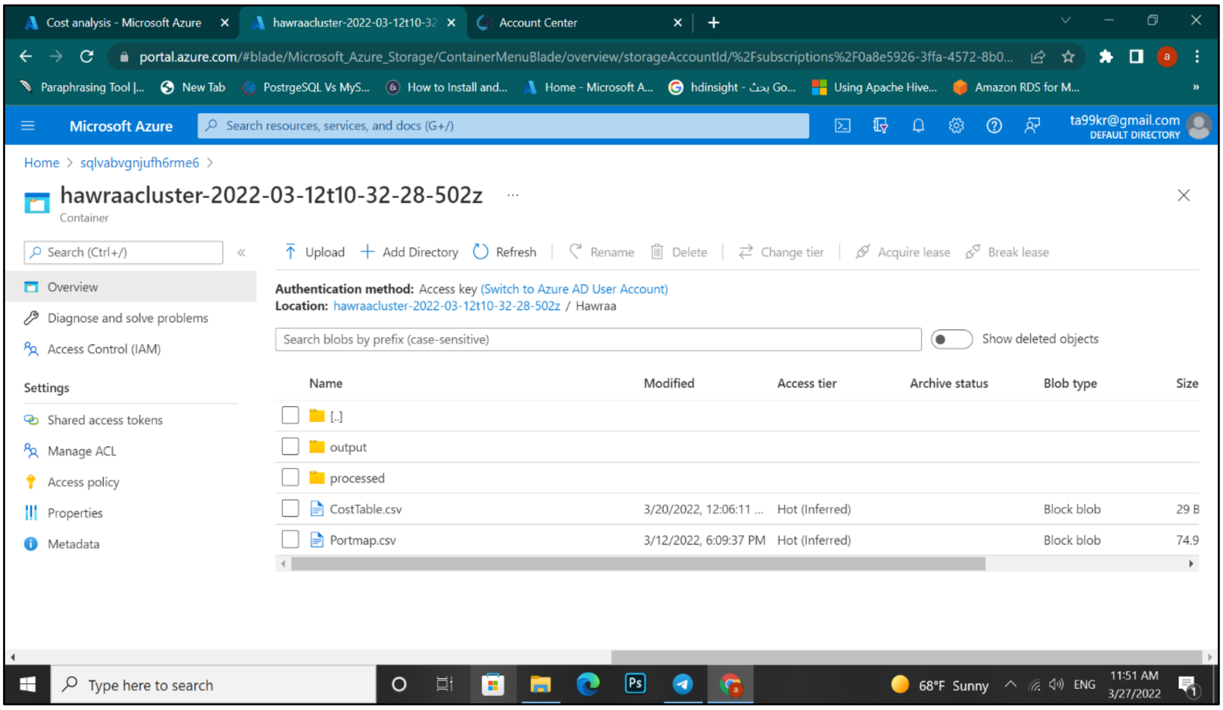
[65] V. Yadav, *Processing Big Data with Azure HDInsight*. Springer, 2017.

[66] A. Roy, A. Jindal, H. Patel, A. Gosalia, S. Krishnan, and C. Curino, “Sparkcruise: Handsfree computation reuse in spark,” *Proc. VLDB Endow.*, vol. 12, no. 12, pp. 1850–1853, 2019.

Appendix A

The proposed system using Microsoft Azure HDInsight.

1) After sign up in Microsoft Azure service, the cluster has been created in HDInsight service and the CostTable.csv is uploaded.


Cost analysis - Microsoft Azure
hawraacluster-2022-03-12t10-32
Account Center
portal.azure.com/#blade/Microsoft_Azure_Storage/ContainerMenuBlade/overview/storageAccountId/%2Fsubscriptions%2F0a8e5926-3ffa-4572-8b0...
Paraphrasing Tool |... New Tab PostgreSQL Vs MyS... How to Install and... Home - Microsoft A... hdinsight - بحث Go... Using Apache Hive... Amazon RDS for M...
Microsoft Azure
Search resources, services, and docs (G+/)
ta99kr@gmail.com
DEFAULT DIRECTORY
Home > sqlvabvgnjufh6rme6 >
Search (Ctrl+/)
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Search blobs by prefix (case-sensitive)
Show deleted objects
Type here to search
68°F Sunny
ENG
11:51 AM
3/27/2022

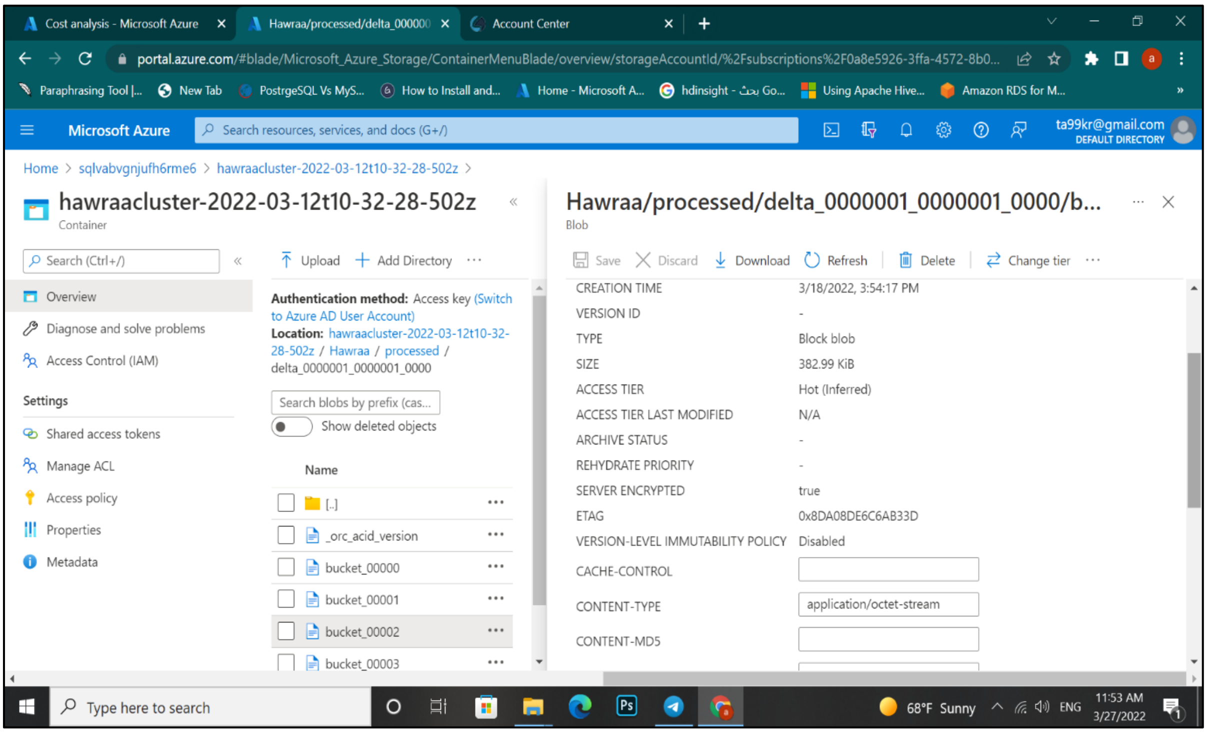
hawraacluster-2022-03-12t10-32-28-502z

Container

Authentication method: Access key (Switch to Azure AD User Account)
Location: hawraacluster-2022-03-12t10-32-28-502z / Hawraa

Name	Modified	Access tier	Archive status	Blob type	Size
[..]					
output					
processed					
CostTable.csv	3/20/2022, 12:06:11 ...	Hot (Inferred)		Block blob	29 B
Portmap.csv	3/12/2022, 6:09:37 PM	Hot (Inferred)		Block blob	74.9

2) The CostTable.csv divided in to bucket by hive.


Cost analysis - Microsoft Azure
Hawraa/processed/delta_000000
Account Center
portal.azure.com/#blade/Microsoft_Azure_Storage/ContainerMenuBlade/overview/storageAccountId/%2Fsubscriptions%2F0a8e5926-3ffa-4572-8b0...
Paraphrasing Tool |... New Tab PostgreSQL Vs MyS... How to Install and... Home - Microsoft A... hdinsight - بحث Go... Using Apache Hive... Amazon RDS for M...
Microsoft Azure
Search resources, services, and docs (G+/)
ta99kr@gmail.com
DEFAULT DIRECTORY
Home > sqlvabvgnjufh6rme6 > hawraacluster-2022-03-12t10-32-28-502z >
Search (Ctrl+/)
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Search blobs by prefix (cas...
Show deleted objects
Save
Discard
Download
Refresh
Delete
Change tier
Type here to search
68°F Sunny
ENG
11:53 AM
3/27/2022

hawraacluster-2022-03-12t10-32-28-502z

Container

Authentication method: Access key (Switch to Azure AD User Account)
Location: hawraacluster-2022-03-12t10-32-28-502z / Hawraa / processed / delta_0000001_0000001_0000

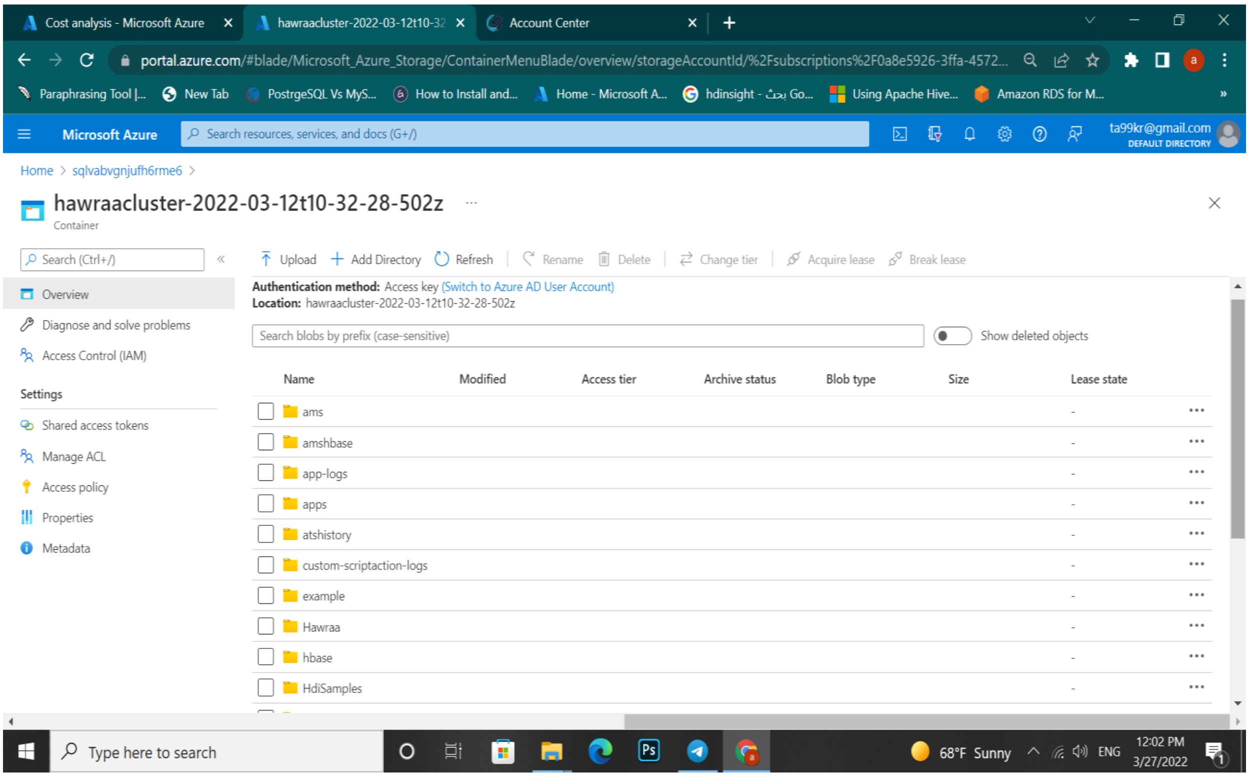
Name
[..]
_orc_acid_version
bucket_00000
bucket_00001
bucket_00002
bucket_00003

Hawraa/processed/delta_0000001_0000001_0000/b...

Blob

CREATION TIME	3/18/2022, 3:54:17 PM
VERSION ID	-
TYPE	Block blob
SIZE	382.99 KiB
ACCESS TIER	Hot (Inferred)
ACCESS TIER LAST MODIFIED	N/A
ARCHIVE STATUS	-
REHYDRATE PRIORITY	-
SERVER ENCRYPTED	true
ETAG	0x8DA08DE6C6AB33D
VERSION-LEVEL IMMUTABILITY POLICY	Disabled
CACHE-CONTROL	
CONTENT-TYPE	application/octet-stream
CONTENT-MD5	

3) Below, the pictures is shown the files has been implemented in the cluster.

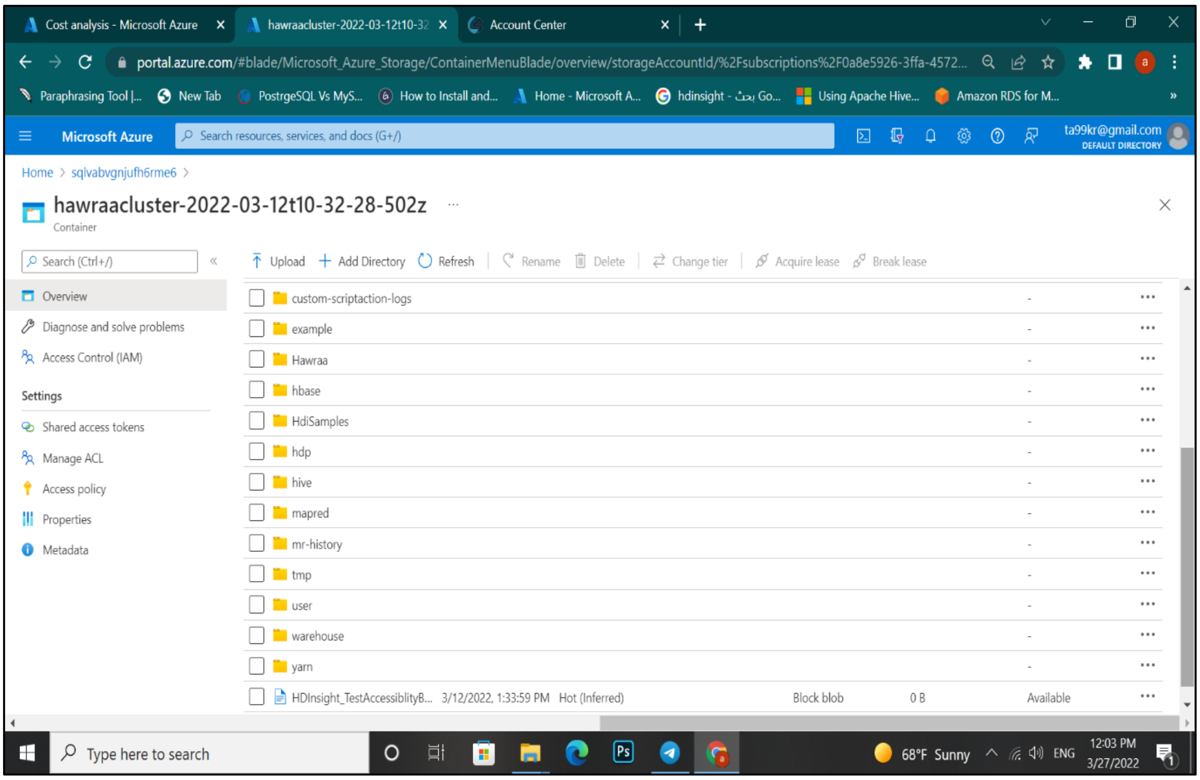


Cost analysis - Microsoft Azure
hawraacluster-2022-03-12t10-32
Account Center
portal.azure.com/#blade/Microsoft_Azure_Storage/ContainerMenuBlade/overview/storageAccountId/%2Fsubscriptions%2F0a8e5926-3ffa-4572...
Paraphrasing Tool |... New Tab PostgreSQL Vs MyS... How to Install and... Home - Microsoft A... hdinsight - بحث Go... Using Apache Hive... Amazon RDS for M...
Microsoft Azure
Search resources, services, and docs (G+/)
ta99kr@gmail.com
DEFAULT DIRECTORY
Home > sqlvabvgnjufh6rme6 >
Container
Search (Ctrl+/)
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Search blobs by prefix (case-sensitive)
Show deleted objects
Type here to search
68°F Sunny
ENG
12:02 PM
3/27/2022

hawraacluster-2022-03-12t10-32-28-502z

Authentication method: Access key (Switch to Azure AD User Account)
Location: hawraacluster-2022-03-12t10-32-28-502z

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
ams						-
amshbase						-
app-logs						-
apps						-
atshistory						-
custom-scriptaction-logs						-
example						-
Hawraa						-
hbase						-
HdiSamples						-

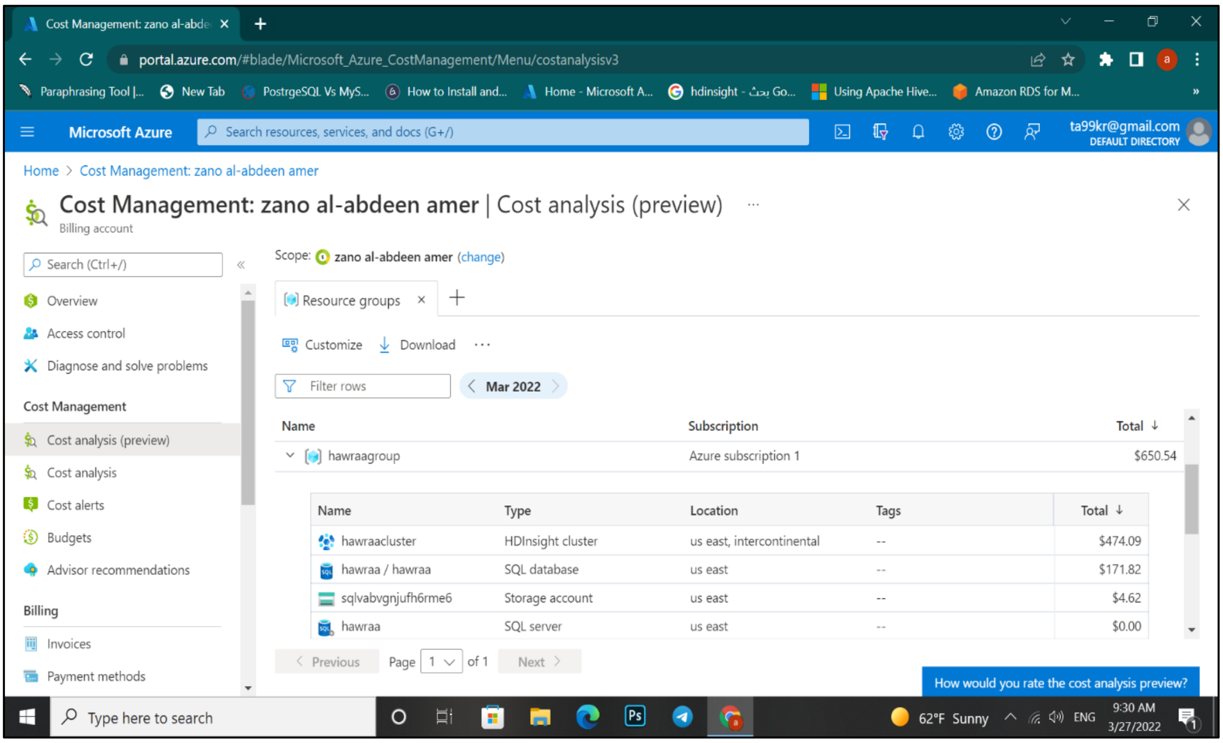


Cost analysis - Microsoft Azure
hawraacluster-2022-03-12t10-32
Account Center
portal.azure.com/#blade/Microsoft_Azure_Storage/ContainerMenuBlade/overview/storageAccountId/%2Fsubscriptions%2F0a8e5926-3ffa-4572...
Paraphrasing Tool |... New Tab PostgreSQL Vs MyS... How to Install and... Home - Microsoft A... hdinsight - بحث Go... Using Apache Hive... Amazon RDS for M...
Microsoft Azure
Search resources, services, and docs (G+/)
ta99kr@gmail.com
DEFAULT DIRECTORY
Home > sqlvabvgnjufh6rme6 >
Container
Search (Ctrl+/)
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Type here to search
68°F Sunny
ENG
12:03 PM
3/27/2022

hawraacluster-2022-03-12t10-32-28-502z

Name	Modified	Access tier	Blob type	Size	Lease state
custom-scriptaction-logs					-
example					-
Hawraa					-
hbase					-
HdiSamples					-
hdp					-
hive					-
mapred					-
mr-history					-
tmp					-
user					-
warehouse					-
yarn					-
HDInsight_TestAccessiblityB...	3/12/2022, 1:33:59 PM	Hot (Inferred)	Block blob	0 B	Available

4) Below, the pictures is shown the Cost Management/Cost Analysis.



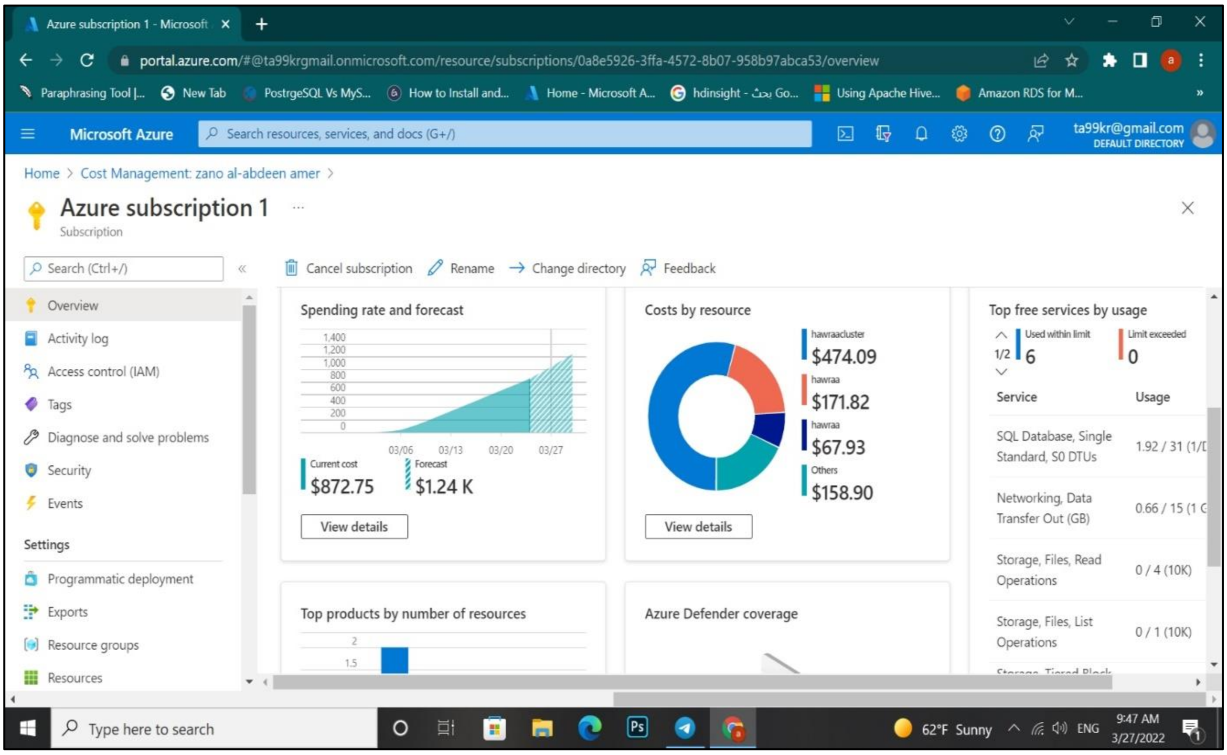
Cost Management: zano al-abdeen amer | Cost analysis (preview)

Billing account

Scope: zano al-abdeen amer (change)

Name	Subscription	Total ↓
hawraagroup	Azure subscription 1	$650.54

Name	Type	Location	Tags	Total ↓
hawraacluster	HDInsight cluster	us east, intercontinental	--	$474.09
hawraa / hawraa	SQL database	us east	--	$171.82
sqlvabvgnjufh6rme6	Storage account	us east	--	$4.62
hawraa	SQL server	us east	--	$0.00



Azure subscription 1

Subscription

Spending rate and forecast

Current cost $872.75

Forecast $1.24 K

Costs by resource

hawraacluster $474.09

hawraa $171.82

hawraa $67.93

Others $158.90

Top free services by usage

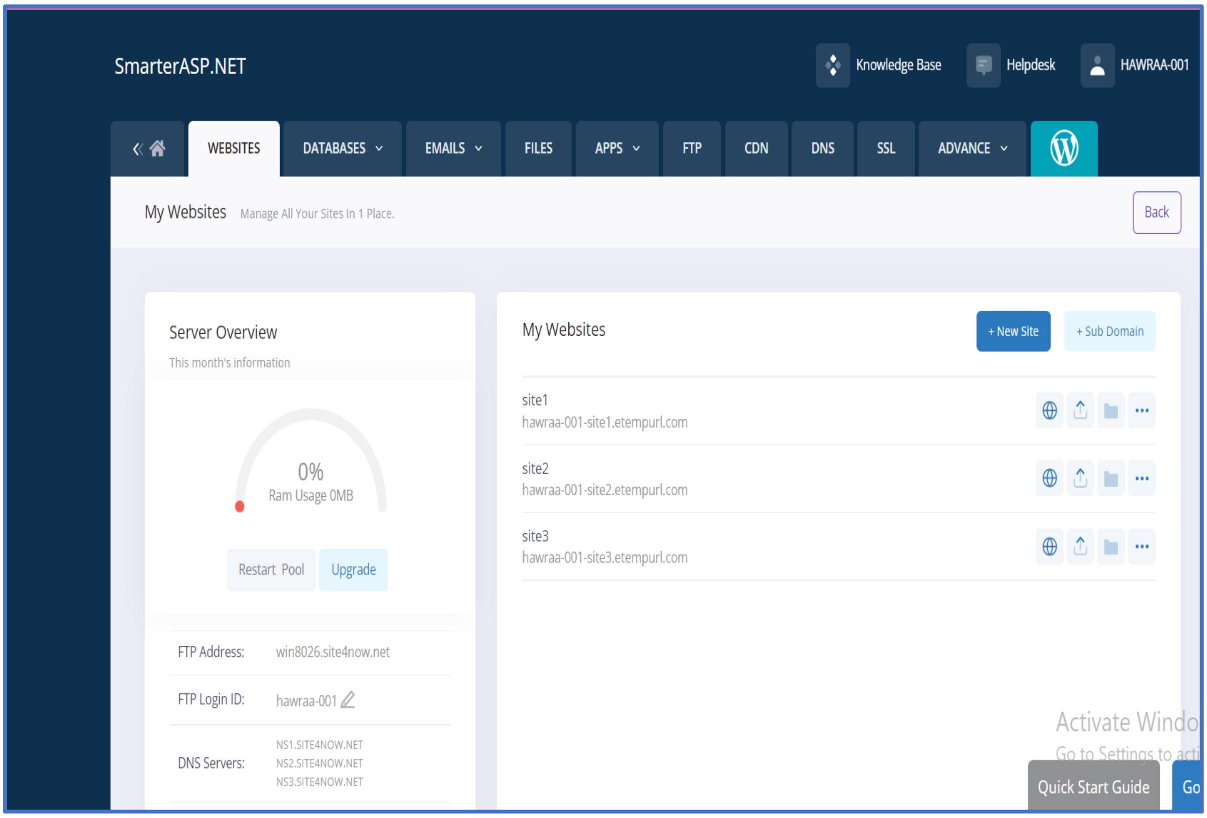
1/2 Used within limit 6 | Limit exceeded 0

Service	Usage
SQL Database, Single Standard, S0 DTUs	1.92 / 31 (1/D
Networking, Data Transfer Out (GB)	0.66 / 15 (1 C
Storage, Files, Read Operations	0 / 4 (10K)
Storage, Files, List Operations	0 / 1 (10K)

Top products by number of resources

Azure Defender coverage

5) Below, the pictures is shown the subscription of cloud computing (SmartASP.NET)



My Websites Manage All Your Sites In 1 Place.

Server Overview
This month's information

0%
Ram Usage 0MB

FTP Address: win8026.site4now.net

FTP Login ID: hawraa-001

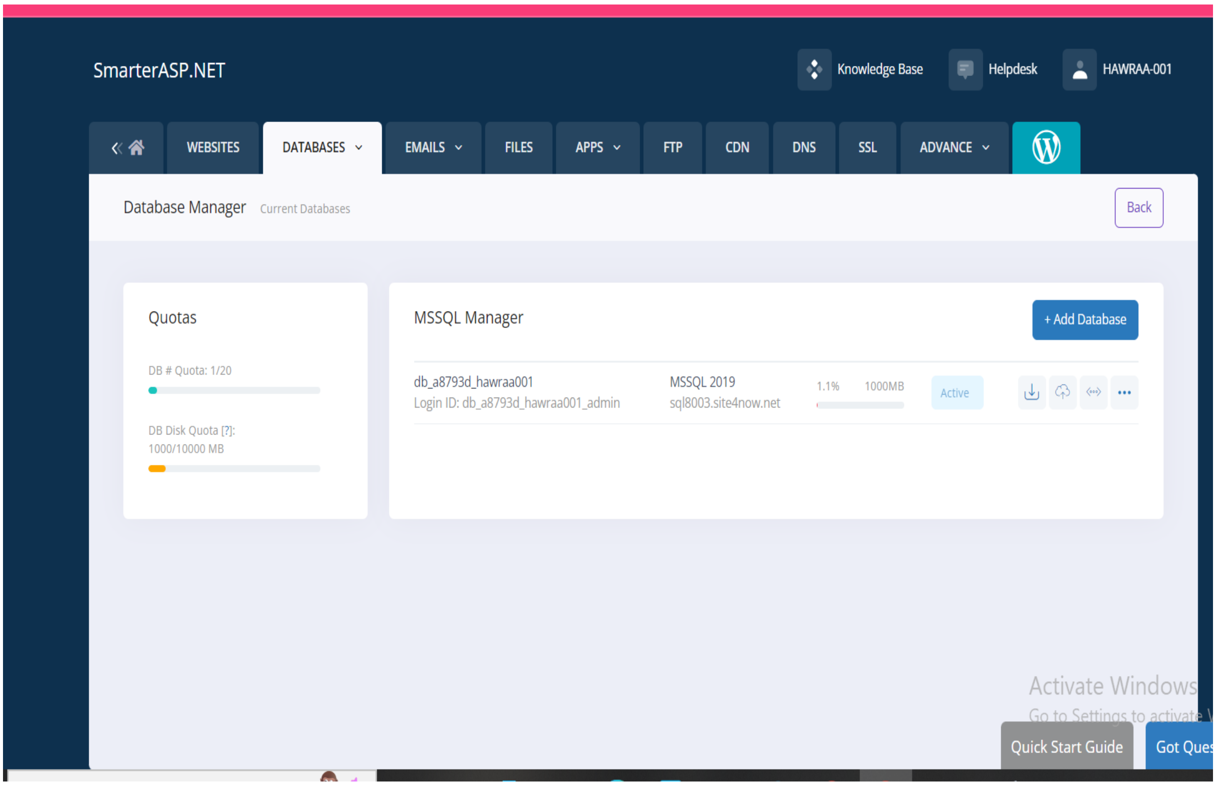
DNS Servers: NS1.SITE4NOW.NET
NS2.SITE4NOW.NET
NS3.SITE4NOW.NET

My Websites

site1
hawraa-001-site1.etempurl.com

site2
hawraa-001-site2.etempurl.com

site3
hawraa-001-site3.etempurl.com



Database Manager Current Databases

Quotas

DB # Quota: 1/20

DB Disk Quota [?]:
1000/10000 MB

MSSQL Manager

db_a8793d_hawraa001 Login ID: db_a8793d_hawraa001_admin	MSSQL 2019 sql8003.site4now.net	1.1%	1000MB

Appendix B

VAM Server application and client show in Figure (1) and Figure (2):

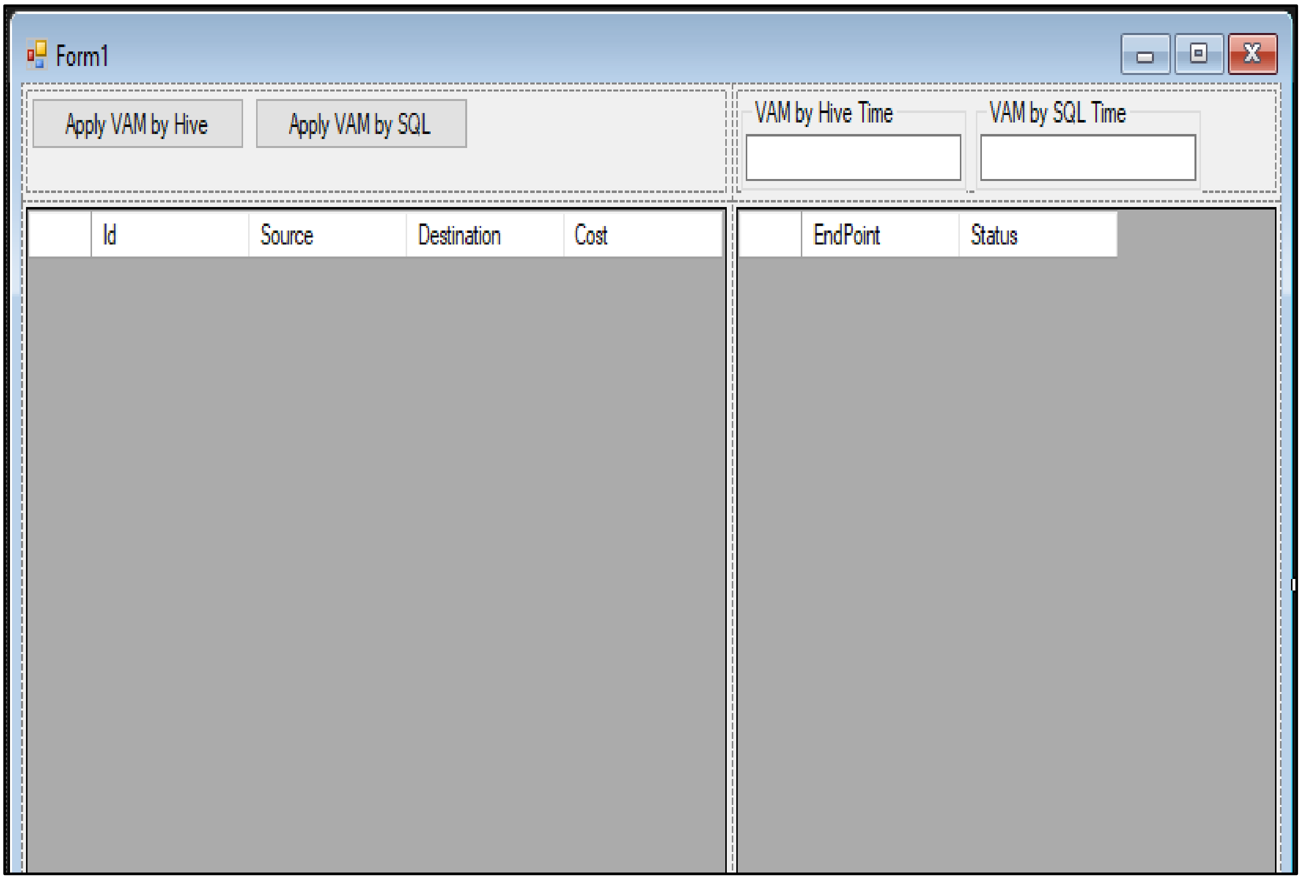


Figure (1): VAM Server application

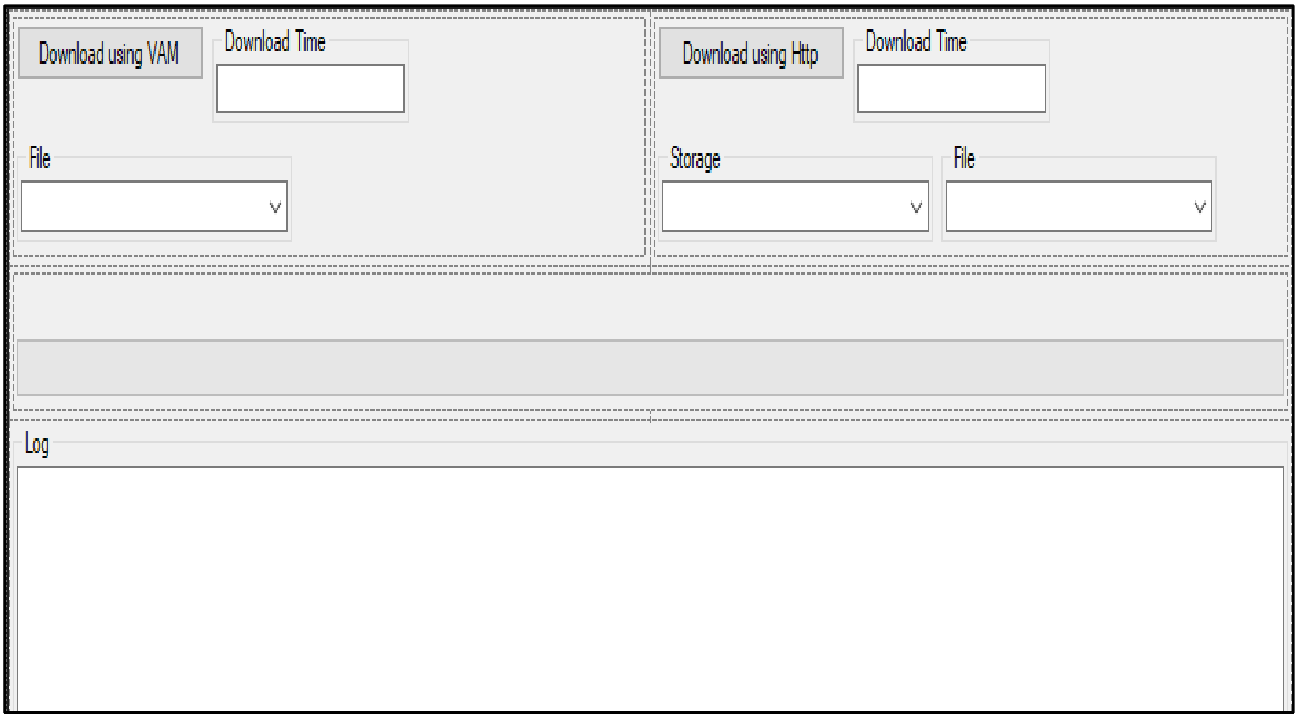
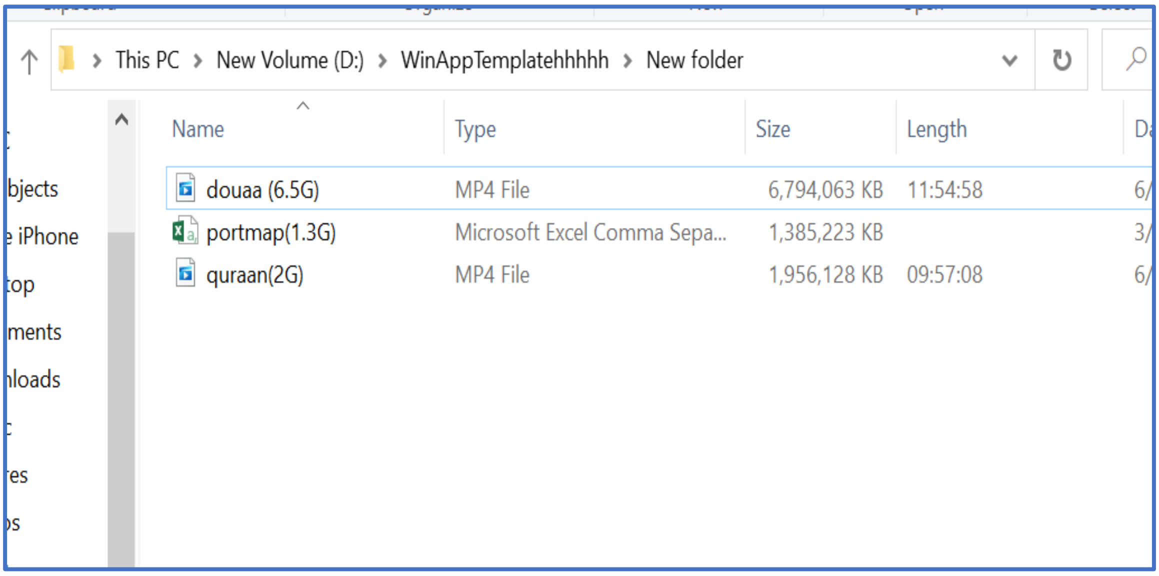


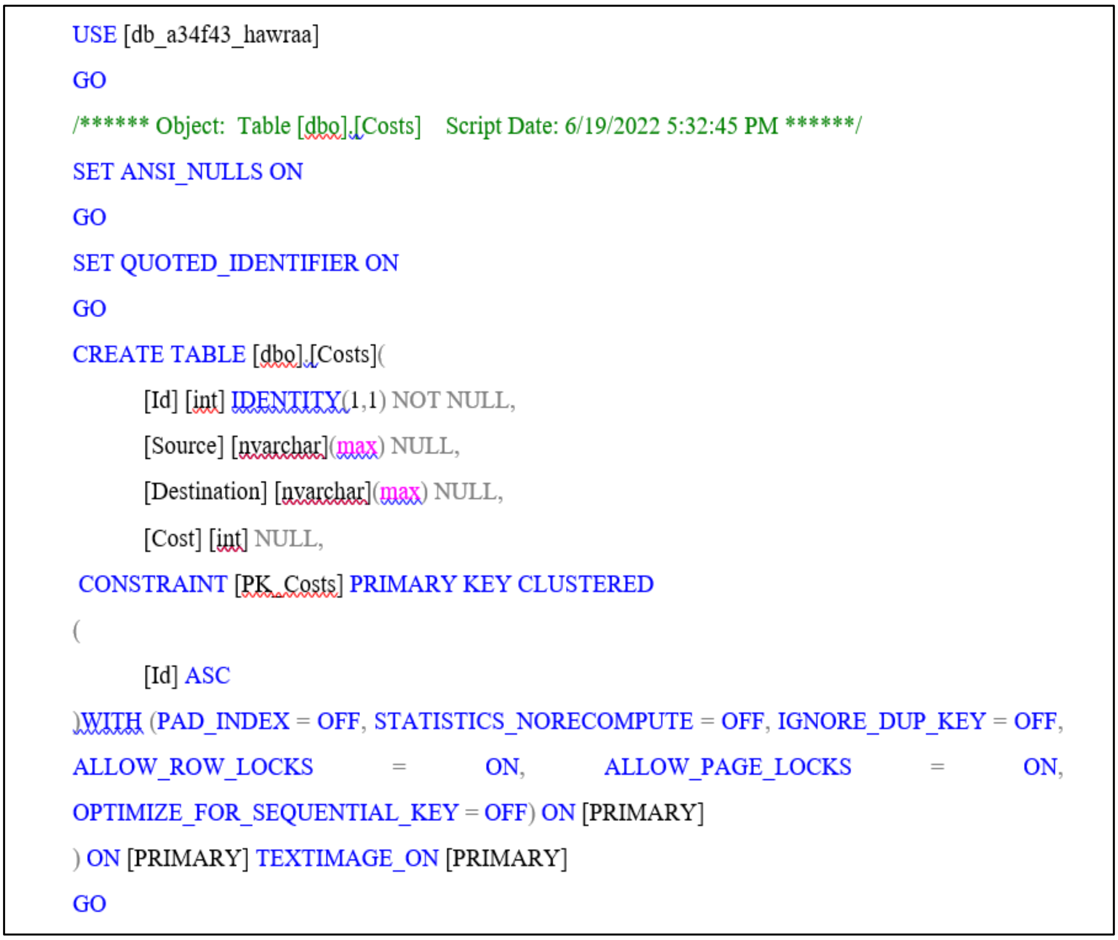
Figure (1): Client’s application

The dataset of the proposed system and the query of CostTable shown in Figure (3) and Figure (4):


This PC › New Volume (D:) › WinAppTemplatehhhhh › New folder
bjects
iPhone
top
ments
nloads
es
ps

Name	Type	Size	Length
douaa (6.5G)	MP4 File	6,794,063 KB	11:54:58
portmap(1.3G)	Microsoft Excel Comma Sepa...	1,385,223 KB	
quraan(2G)	MP4 File	1,956,128 KB	09:57:08

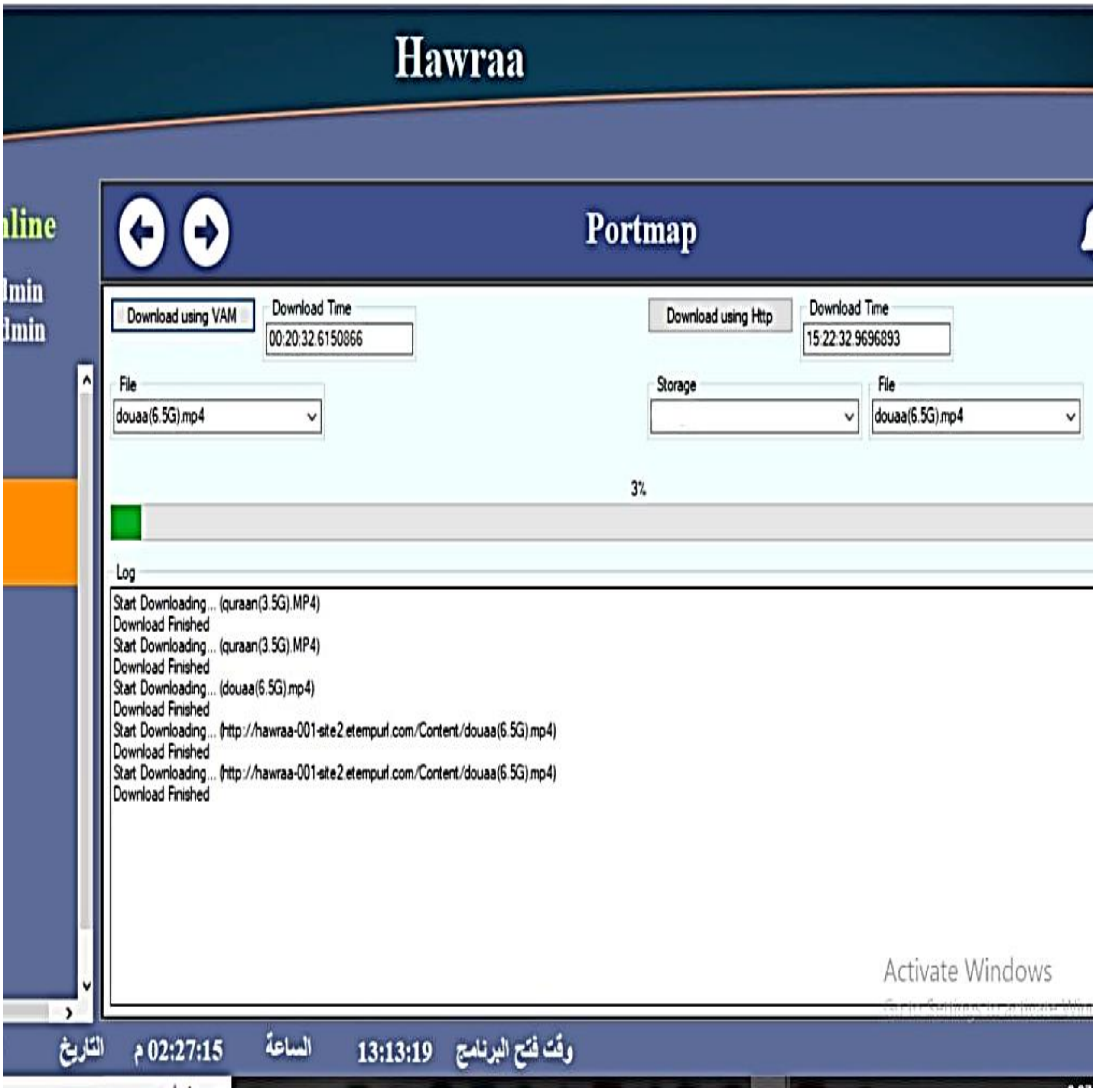
Figure (3): The data set



```
USE [db_a34f43_hawraa]
GO
/****** Object:  Table [dbo].[Costs]    Script Date: 6/19/2022 5:32:45 PM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[Costs](
        [Id] [int] IDENTITY(1,1) NOT NULL,
        [Source] [nvarchar](max) NULL,
        [Destination] [nvarchar](max) NULL,
        [Cost] [int] NULL,
 CONSTRAINT [PK_Costs] PRIMARY KEY CLUSTERED
(
        [Id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON,
OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]
GO
```

Figure (4): Query of CostTable

Took a snapshot of the client's application output in PC1 as shown in Figure (5). The button that name is "Downloading using Http" show the time of download the (case1) file in regular approach and the button that name is "Downloading using VAM: show time of download the (case1) file in the proposed system.



Download Time: 00:20:32.6150866

File: douaa(6.5G).mp4

Download Time: 15:22:32.9696893

Storage:

File: douaa(6.5G).mp4

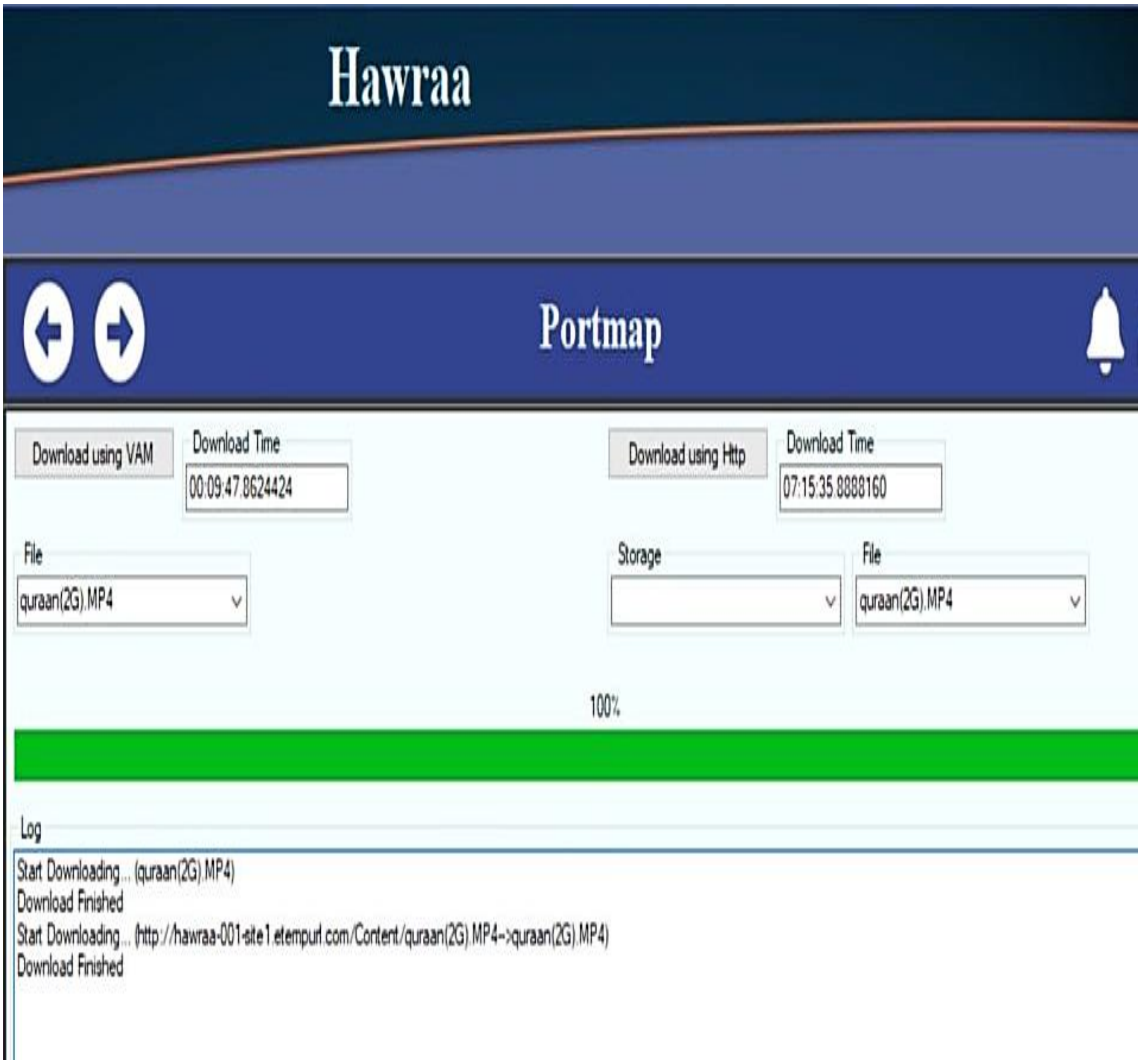
Log

```
Start Downloading... (quraan(3.5G).MP4)
Download Finished
Start Downloading... (quraan(3.5G).MP4)
Download Finished
Start Downloading... (douaa(6.5G).mp4)
Download Finished
Start Downloading... (http://hawraa-001-site2.etempurl.com/Content/douaa(6.5G).mp4)
Download Finished
Start Downloading... (http://hawraa-001-site2.etempurl.com/Content/douaa(6.5G).mp4)
Download Finished
```

التاريخ م 02:27:15 الساعة 13:13:19 وقت فتح البرنامج

Figure (5): Client's application output in PC1shows the time of download the case1 douaa(6.5G).mp4 in regular approach and proposed system

The snapshot of the client's application output in PC1 as shown in Figure (6) shows the time of download the case2 (quraan(2G).mp4) file in regular approach and the button that name is Downloading using VAM show time of download the case2 (quraan(2G).mp4) file in the proposed system.



Download Time: 00:09:47.8624424

File: quraan(2G).MP4

Download Time: 07:15:35.8888160

Storage:

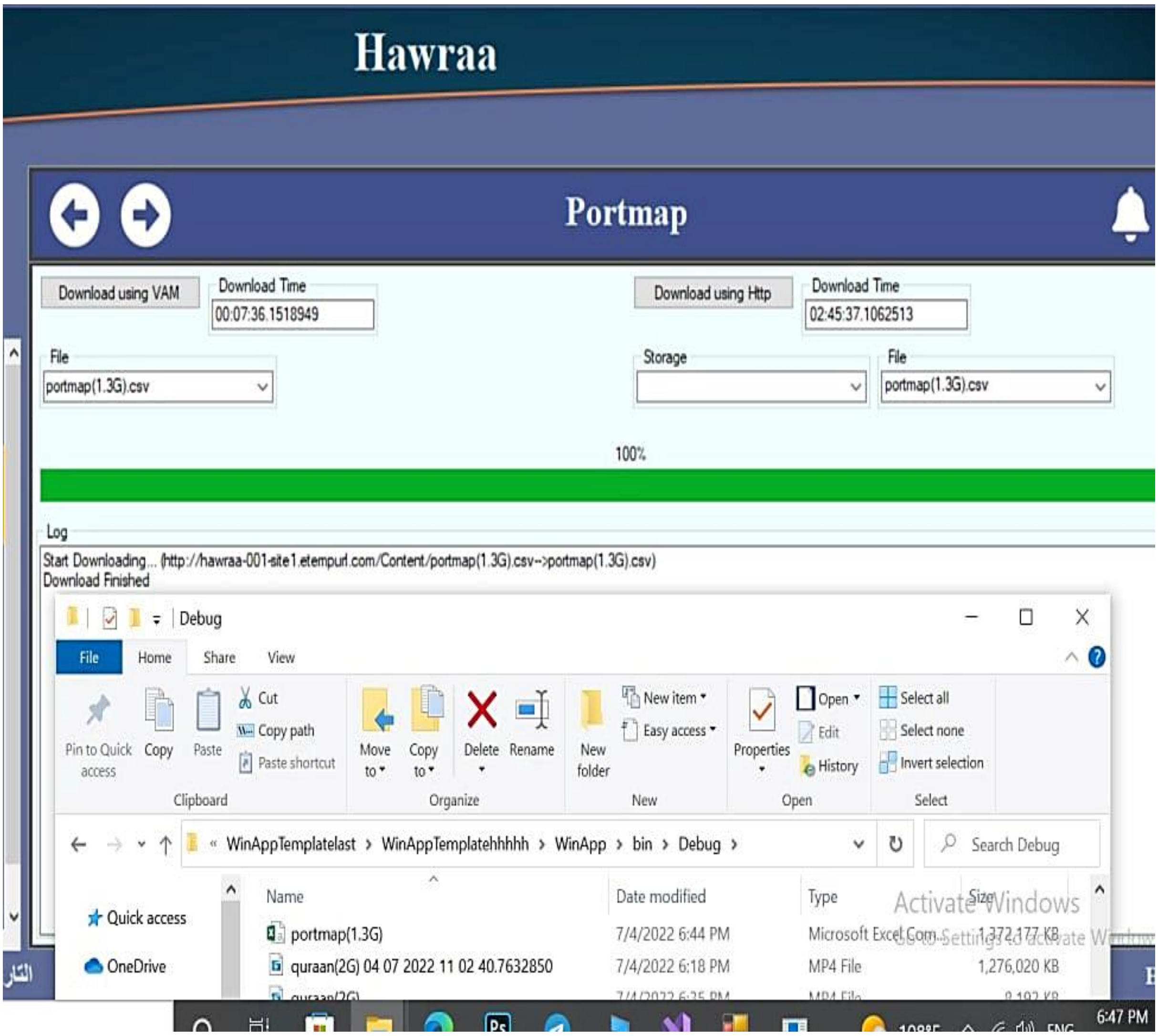
File: quraan(2G).MP4

Log

```
Start Downloading... (quraan(2G).MP4)
Download Finished
Start Downloading... (http://hawraa-001-site1.etempurl.com/Content/quraan(2G).MP4-->quraan(2G).MP4)
Download Finished
```

Figure (6): Client's application output in PC1shows the time of download the case2 quraan(2G).mp4 in regular approach and proposed system

Figure (7) shows the time of download the (portmap(1.3G).mp4) file in regular approach and the button that name is Downloading using VAM show time of download the (portmap(1.3G).mp4) file in the proposed system in PC1.



Download Time: 00:07:36.1518949

Download Time: 02:45:37.1062513

File: portmap(1.3G).csv

Storage:

File: portmap(1.3G).csv

Log

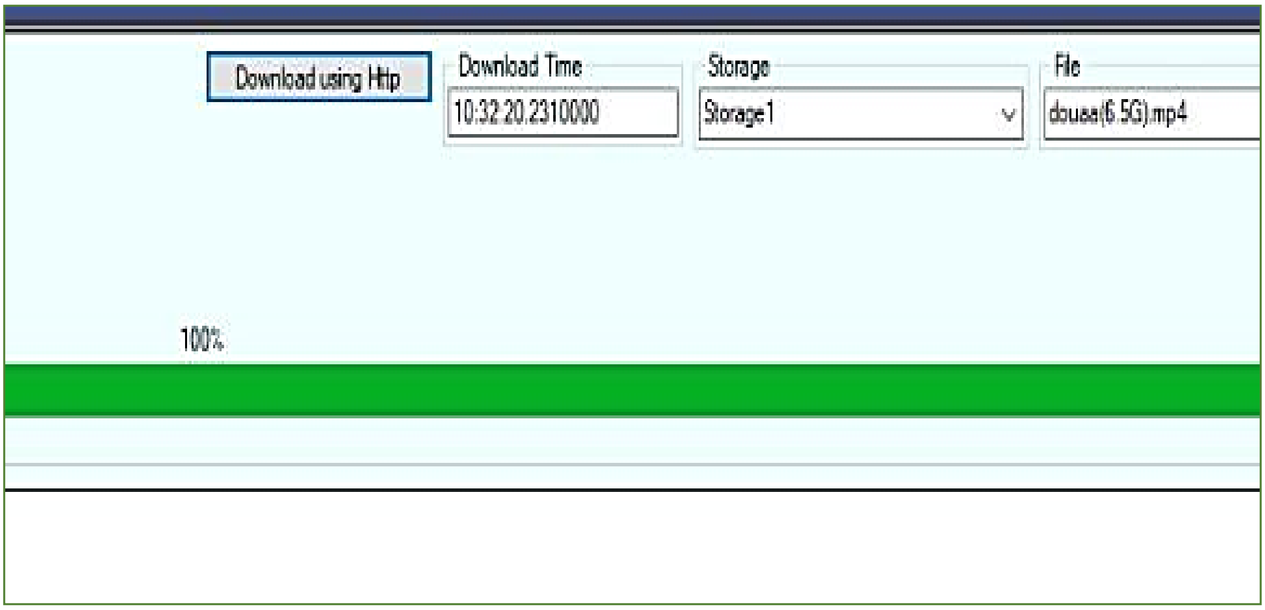
```
Start Downloading... (http://hawraa-001-site1.etempurl.com/Content/portmap(1.3G).csv-->portmap(1.3G).csv)
Download Finished
```

« WinAppTemplatelast › WinAppTemplatehhhhh › WinApp › bin › Debug ›

Name	Date modified	Type	Size
portmap(1.3G)	7/4/2022 6:44 PM	Microsoft Excel Com...	1,372,177 KB
quraan(2G) 04 07 2022 11 02 40.7632850	7/4/2022 6:18 PM	MP4 File	1,276,020 KB

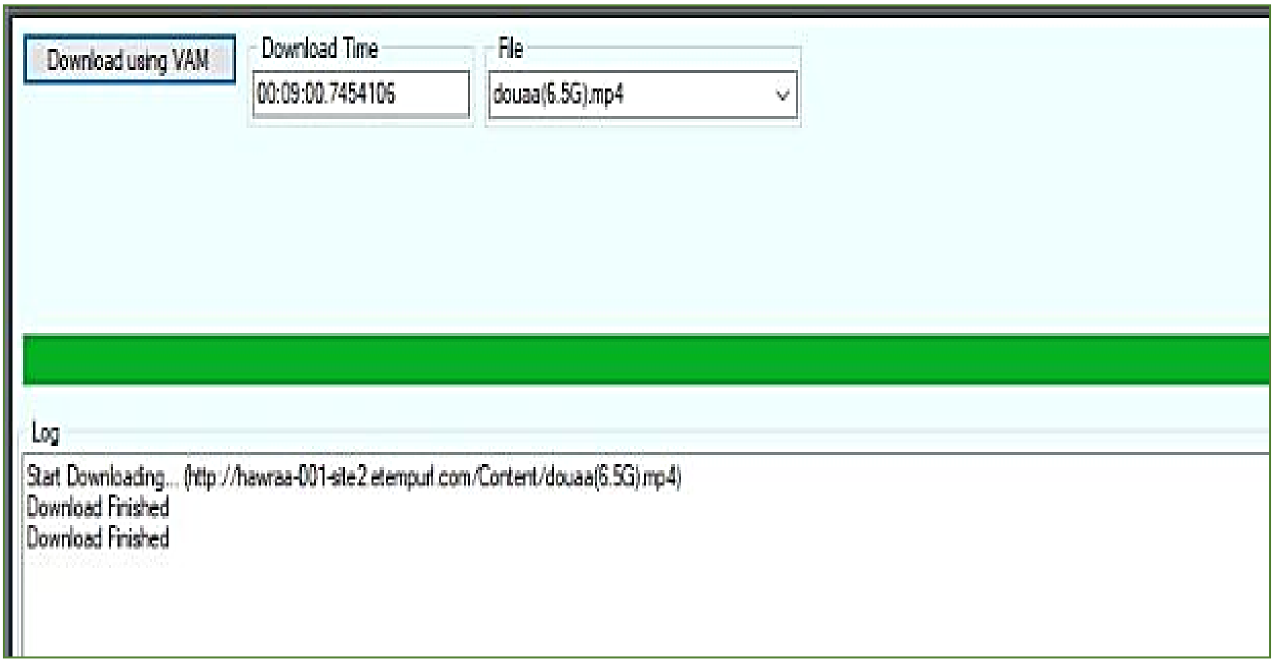
Figure (7): Client’s application output in PC1shows the time of download the case3 portmap(1.3G).mp4 in regular approach and proposed system

Figure (8) shows the time of downloading case1 (douaa(6.5G).mp4) file in regular approach and the figure(9) shows the time of downloading the case1 (douaa(6.5G).mp4) file in the proposed system in PC2.



Download Time: 10:32:20.2310000
Storage: Storage1
File: douaa(6.5G).mp4

Figure (8): Client’s application output in PC2 shows the time of download the douaa(6.5G).mp4 file in regular approach



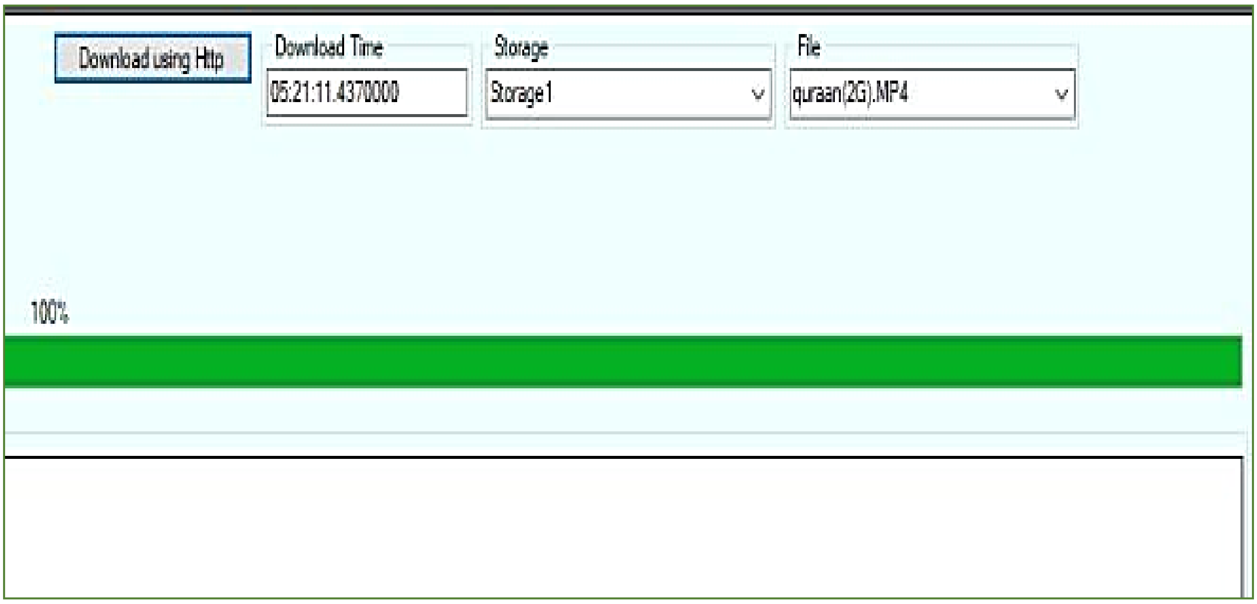
Download Time: 00:09:00.7454106
File: douaa(6.5G).mp4

Log

```
Start Downloading... (http://hawraa-001-site2.etempurl.com/Content/douaa(6.5G).mp4)
Download Finished
Download Finished
```

Figure (9): Client’s application output in PC2 shows the time of download the douaa(6.5G).mp4 file in proposed system

Figure (10) shows the time of download the case2 (quraan(2G).mp4) file in regular approach and the Figure(11) shows the time of download the case2 (quraan(2G).mp4) file in the proposed system in PC2.



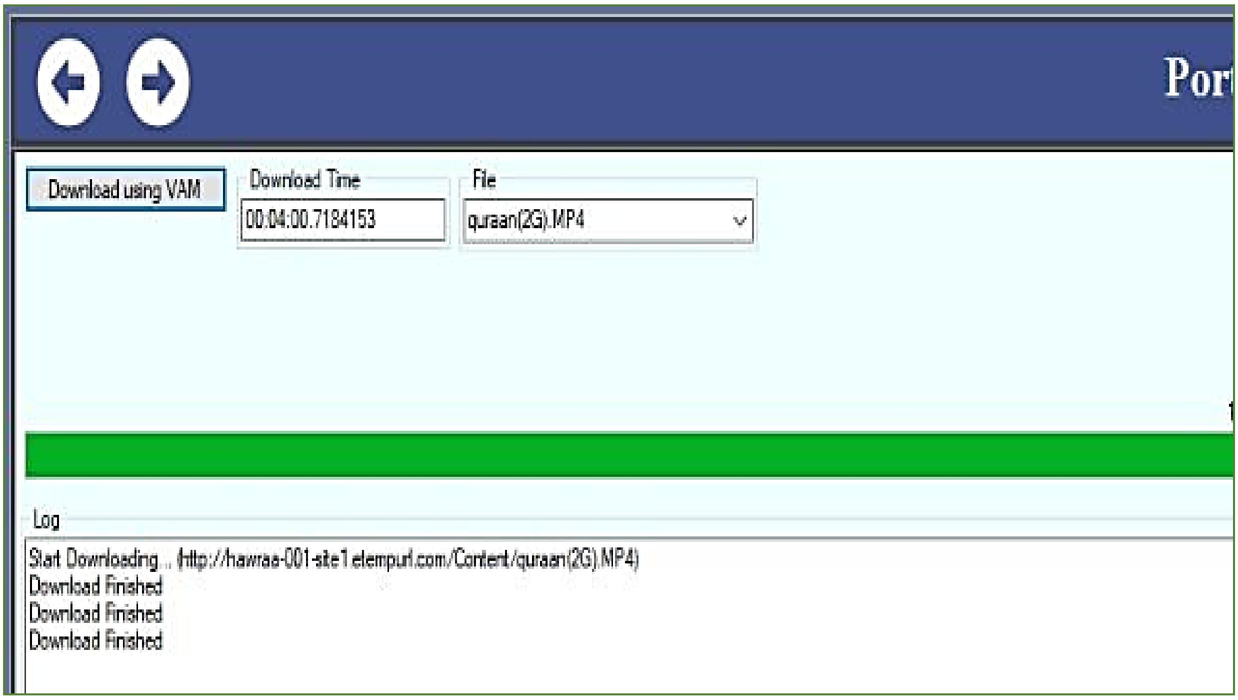
Download Time
05:21:11.4370000

Storage
Storage1

File
quraan(2G).MP4

100%

Figure (10): Client's application output in PC2 shows the time of download the quraan(2G).mp4 file in regular approach



Download Time
00:04:00.7184153

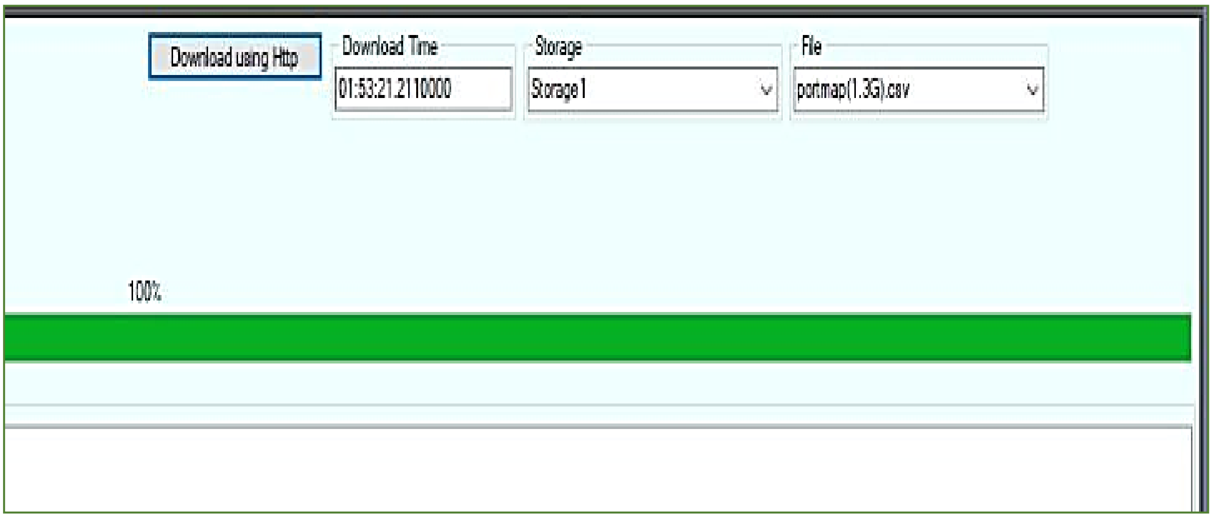
File
quraan(2G).MP4

Log

```
Start Downloading... (http://hawraa-001-site1.etempurl.com/Content/quraan(2G).MP4)
Download Finished
Download Finished
Download Finished
```

Figure (11): Client's application output in PC2 shows the time of download the quraan(2G).mp4 file in proposed system

Figure (12) shows the time of download case3 (portmap(1.3G).mp4) file in regular approach and the Figure (13) shows the time of download case3 (portmap(1.3G).mp4) file in the proposed system in PC2.



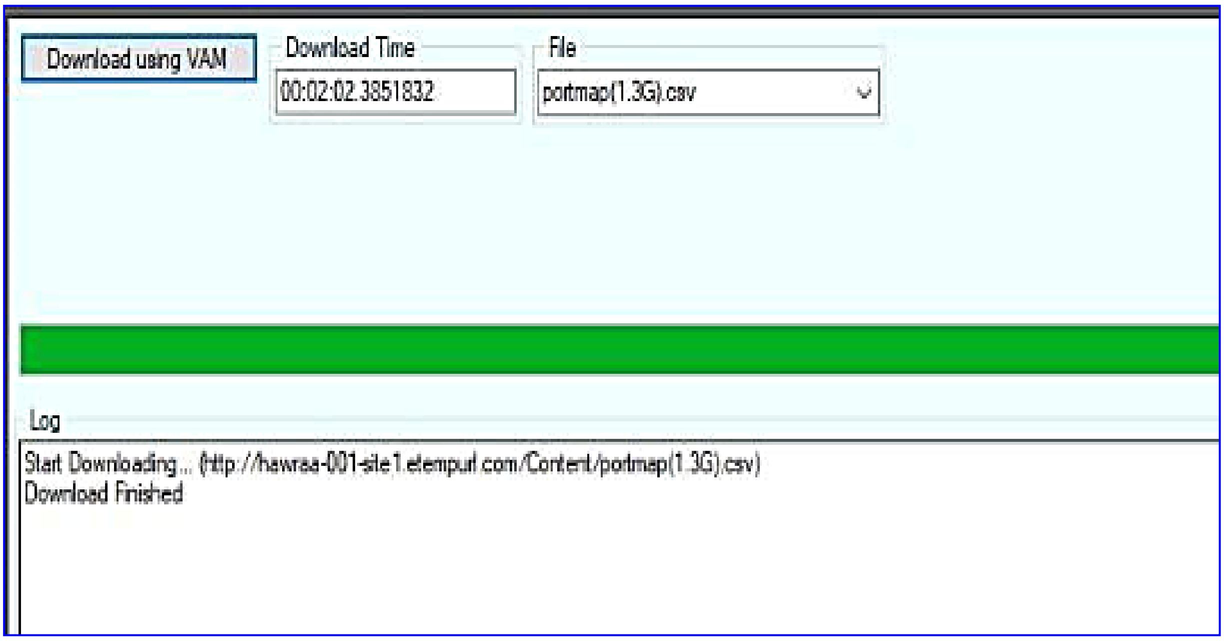
Download Time: 01:53:21.2110000

Storage: Storage1

File: portmap(1.3G).csv

100%

Figure (12): Client's application output in PC2 shows the time of download the portmap(1.3G).mp4 file in regular approach



Download Time: 00:02:02.3851832

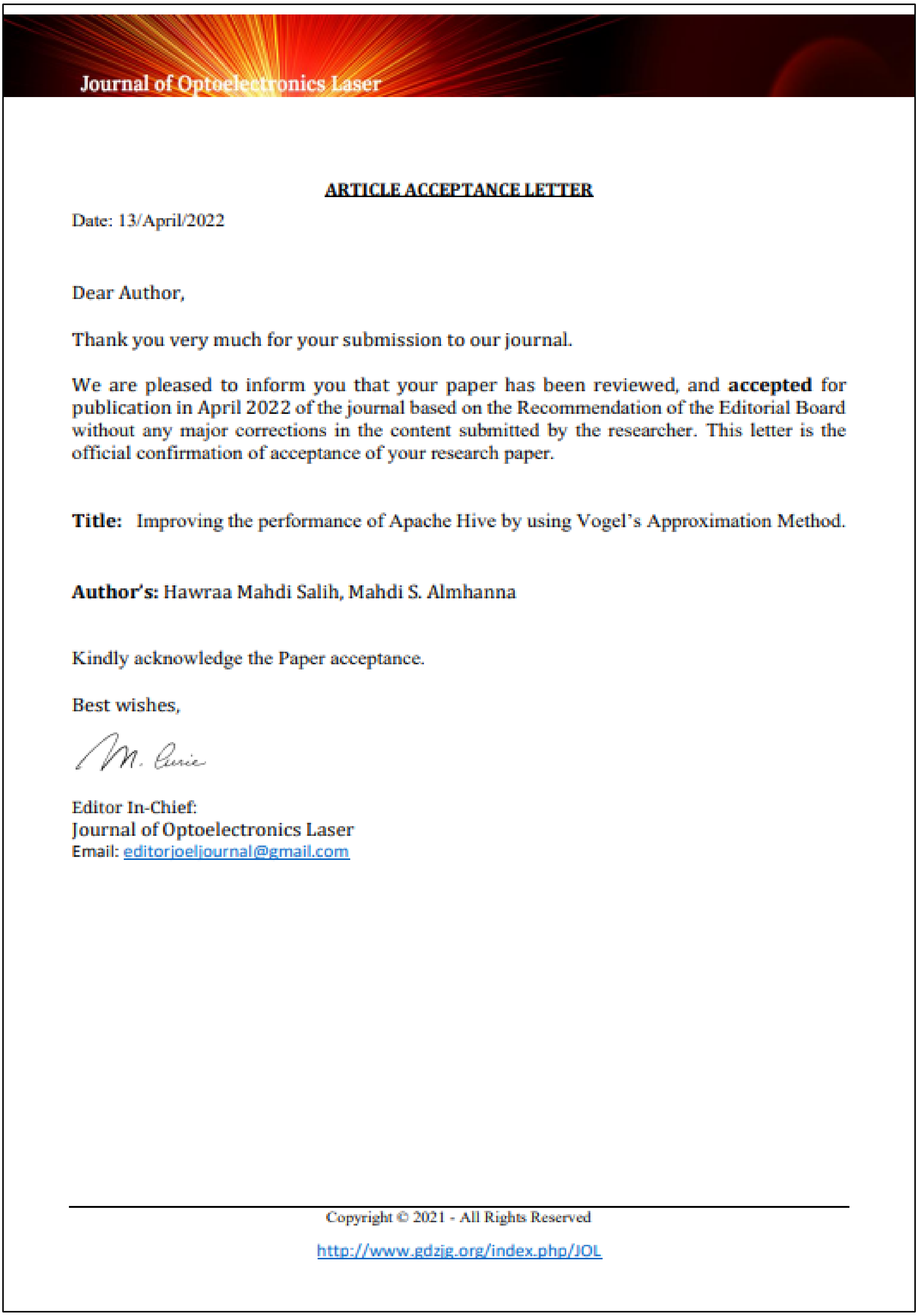
File: portmap(1.3G).csv

Log

```
Start Downloading... (http://hawraa-001-site1.etempurl.com/Content/portmap(1.3G).csv)
Download Finished
```

Figure (13): Client's application output in PC2 shows the time of download the portmap(1.3G).mp4 file in proposed system

Appendix C



Journal of Optoelectronics Laser

ARTICLE ACCEPTANCE LETTER

Date: 13/April/2022

Dear Author,

Thank you very much for your submission to our journal.

We are pleased to inform you that your paper has been reviewed, and **accepted** for publication in April 2022 of the journal based on the Recommendation of the Editorial Board without any major corrections in the content submitted by the researcher. This letter is the official confirmation of acceptance of your research paper.

Title: Improving the performance of Apache Hive by using Vogel's Approximation Method.

Author's: Hawraa Mahdi Salih, Mahdi S. Almhanna

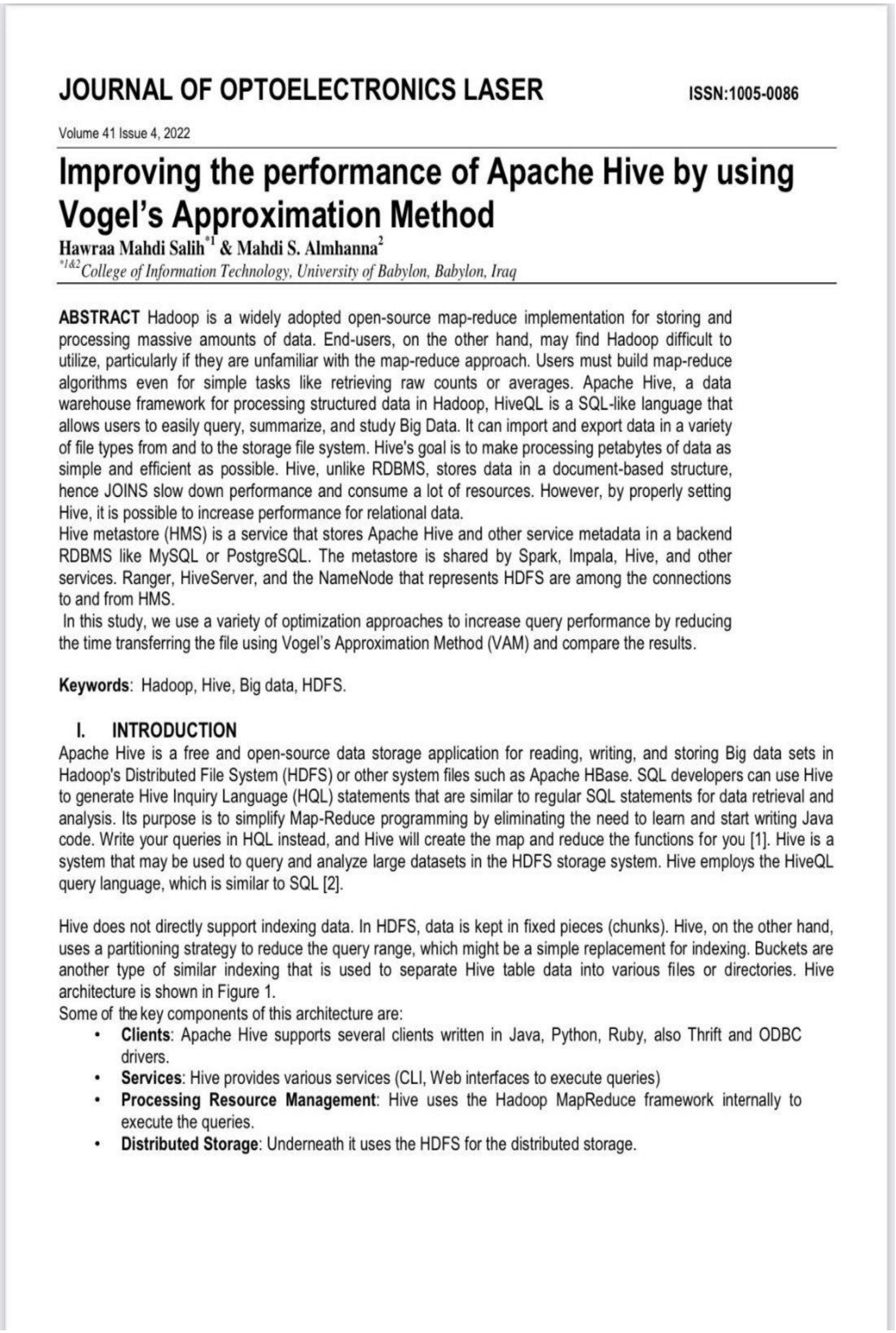
Kindly acknowledge the Paper acceptance.

Best wishes,

Editor In-Chief:
Journal of Optoelectronics Laser
Email: editorjoeljournal@gmail.com

Copyright © 2021 - All Rights Reserved

http://www.gdzjg.org/index.php/JOL



JOURNAL OF OPTOELECTRONICS LASER ISSN:1005-0086

Volume 41 Issue 4, 2022

Improving the performance of Apache Hive by using Vogel's Approximation Method

Hawraa Mahdi Salih*1 & Mahdi S. Almhanna2

*1&2 *College of Information Technology, University of Babylon, Babylon, Iraq*

ABSTRACT Hadoop is a widely adopted open-source map-reduce implementation for storing and processing massive amounts of data. End-users, on the other hand, may find Hadoop difficult to utilize, particularly if they are unfamiliar with the map-reduce approach. Users must build map-reduce algorithms even for simple tasks like retrieving raw counts or averages. Apache Hive, a data warehouse framework for processing structured data in Hadoop, HiveQL is a SQL-like language that allows users to easily query, summarize, and study Big Data. It can import and export data in a variety of file types from and to the storage file system. Hive's goal is to make processing petabytes of data as simple and efficient as possible. Hive, unlike RDBMS, stores data in a document-based structure, hence JOINS slow down performance and consume a lot of resources. However, by properly setting Hive, it is possible to increase performance for relational data.

Hive metastore (HMS) is a service that stores Apache Hive and other service metadata in a backend RDBMS like MySQL or PostgreSQL. The metastore is shared by Spark, Impala, Hive, and other services. Ranger, HiveServer, and the NameNode that represents HDFS are among the connections to and from HMS.

In this study, we use a variety of optimization approaches to increase query performance by reducing the time transferring the file using Vogel's Approximation Method (VAM) and compare the results.

Keywords: Hadoop, Hive, Big data, HDFS.

I. INTRODUCTION

Apache Hive is a free and open-source data storage application for reading, writing, and storing Big data sets in Hadoop's Distributed File System (HDFS) or other system files such as Apache HBase. SQL developers can use Hive to generate Hive Inquiry Language (HQL) statements that are similar to regular SQL statements for data retrieval and analysis. Its purpose is to simplify Map-Reduce programming by eliminating the need to learn and start writing Java code. Write your queries in HQL instead, and Hive will create the map and reduce the functions for you [1]. Hive is a system that may be used to query and analyze large datasets in the HDFS storage system. Hive employs the HiveQL query language, which is similar to SQL [2].

Hive does not directly support indexing data. In HDFS, data is kept in fixed pieces (chunks). Hive, on the other hand, uses a partitioning strategy to reduce the query range, which might be a simple replacement for indexing. Buckets are another type of similar indexing that is used to separate Hive table data into various files or directories. Hive architecture is shown in Figure 1.

Some of the key components of this architecture are:

- **Clients**: Apache Hive supports several clients written in Java, Python, Ruby, also Thrift and ODBC drivers.
- **Services**: Hive provides various services (CLI, Web interfaces to execute queries)
- **Processing Resource Management**: Hive uses the Hadoop MapReduce framework internally to execute the queries.
- **Distributed Storage**: Underneath it uses the HDFS for the distributed storage.

الخلاصة

الحوسبة السحابية هي تقنية تتيح التكامل الظاهري وتوفير موارد الحوسبة الموجودة في المواقع البعيدة. على شبكة الويب العالمية، يطلب العديد من المستخدمين العديد من الملفات المنتشرة على العديد من الخوادم حول العالم وهذا يستغرق وقتًا، مما يؤدي إلى زيادة تكلفة نقل الملف داخل الحوسبة السحابية. إن تقليل الوقت المطلوب لتنزيل ملف هو نفس تقليل تكلفة توصيل المنتجات. أقل وقت مطلوب لتسليم المنتج للعميل بأقل تكلفة ممكنة.

تم اقتراح نظام في هذه الأطروحة لتقليل تكلفة نقل الملفات باستخدام طريقة فوجل التقريبية VAM. تحدد الفكرة الرئيسية للنظام أي خادم يساعد في تسليم الملف بسرعة إلى العميل الذي يطلب هذا الملف. تم تقليل الوقت اللازم لتنزيل الملف عبر النظام المقترح عن طريق اختيار أفضل خادم يقوم بإرجاع الملف في أقصر وقت، ويتم استخدام طريقة فعالة VAM للحصول على الحد الأدنى من التكلفة اللازمة لتسليم كل ملف إلى المستخدم في غضون الخادم الأوسط الذي يتلقى طلبات العميل، وحدد الخادم لكل عميل لتنزيل الملف منه.

يتكون النظام المقترح من ثلاثة أجزاء وهي جزء الحوسبة السحابية وجزء VAM الخادم وجزء العميل ويستخدم ثلاثة ملفات وهي (case1، case2، case3) مع (6.5، 2، 1.3) GB على التوالي. وتم باستخدام الانحراف المعياري لاختبار النتائج. أظهرت النتائج أن طريقة VAM ساهمت في تقليل تكلفة نقل الملفات حيث أن نسبة تقليل الوقت كانت حوالي (97-98)٪ مقارنة بوقت نقل الملفات العادي.

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية تكنولوجيا المعلومات
قسم البرمجيات

تقليل كلفة نقل الملفات داخل الحوسبة السحابية باستخدام طريقة فوجل التقريبية

رسالة
مقدمة الى مجلس كلية تكنولوجيا المعلومات, جامعة بابل وهي جزء من
متطلبات نيل درجة الماجستير في تكنولوجيا المعلومات / برمجيات المعلومات

من قِبَل
حوراء مهدي صالح هادي

بإشراف
أ.م.د. مهدي صالح نعمة موسى

1444 هـ 2022 م