# The Use of Hybrid Clouding to Reduce the Cost of Database Management System

**Project Proposal**
**Data 225 - Database Management Systems**
**Fall 2020**

**Mabisa Gharti Chhetry** **Rudra Gandhi** **Nghi Nguyen**

# Table of Contents

Page#

# List of Tables:

# Abstract

The rapid growth in demand for big data storage leads to the development of many types of storage. As the size of the dataset is growing, storing big data in the traditional database management system is no longer efficient due to the high storage and management costs. Besides, on-premises data centers have low scalability, inflexibility, and no backup in case of disasters. We essentially focus on the cost of the management system and analyze costs from different storage methods given the same performance, scalability, flexibility, and consistency. To be more specific, the paper proposes hybrid cloud systems as the best data storage in terms of cost advantage. Cloud database management systems are now widely used due to its flexibility, mobility, quality control, disaster recovery, etc. This paper investigates the advantages and disadvantages of hybrid cloud storage and other low-cost types of database management systems. To prove the cost advantage of hybrid cloud, we build a tool that can calculate the cost of running all possible queries for a sample dataset for a specific period.

# 1. Introduction

## 1.1 Objective

The applications of Database Management Systems (DBMS) have been used in many industries. Users always want it all when working with databases: minimum cost, maximum performance, maximum scalability, maximum consistency, maximum flexibility, and so forth. However, the system covering those requirements has not existed yet, and it is nearly impossible to build such a perfect structure. With the development of technology, the cost of DBMS is rising rapidly. Different kinds of costs are the cost of hardware and software, cost of staff training, cost of data conversion, etc. Applications like Quick Base, Microsoft SQL Server, and Apple iCloud asks users to pay monthly bills. Hardware resources are no longer a one-time fixed investment, and cloud computing and Amazon Web Services have made this cost a continuous metric *. In the past, users tended to concentrate more on performance, “try to minimize the response time of each request with a fixed number of machines.” Today, the statement turns out to be, “try to minimize the number of machines with a fixed response time goal.”

## 1.2 What is the problem

This paper will essentially focus on the cost of the database system rather than other elements of the users’ requirements. The purpose of this paper is to propose a new technique to reduce/minimize the cost of the traditional database system. Specifically, we are introducing cloud computing and hybrid cloud computing and the way they can significantly cut out some inefficient costs in the DBMS. Notice that a method that can reduce most of the cost might not be the best option as it also relates to performance, scalability, consistency, and flexibility; therefore, we need to consider those factors throughout the analysis. Since cloud and hybrid

cloud systems are new trends with a variety of benefits, it is difficult to choose which one is better. Thus, we decide to analyze both systems as well as the traditional DBMS to draw a conclusion for the best approach. Because both methods are very new in the industry, we need to carefully analyze both to come to a conclusion. The chosen system will not only minimize the cost of the DBMS but also satisfies other requirements of performance levels. The theme is to clarify and explain how hybrid cloud computing is the best solution to having a cost-efficient database management system. This technique will reduce the cost of a traditional database system.

### 1.3 Why this is a project related to this class

The project relates to Data-225 course because it raises a popular issue of the traditional DBMS and proposes a solution to resolve the problem. The purpose of the class is to demonstrate the knowledge data model, relational databases design, data definition and manipulation languages, architectures of database management system, security, etc. The research might not go over all those topics, but it will give deeper knowledge for many parts of them, such as the operational and analytical databases, data models, and architectures of DBMS. Not only the cost but also many aspects of the DBMS will be investigated. The cost of the traditional DBMS will be shown and compared to the new approach. Python codes and analytical skills will be applied to test the hypothesis. Consequently, it is a good chance to show the practice of data mining programming and open source database tools. Not only outside resources, the knowledge from the class will also be used in this paper. After the analysis, students will have more solid understanding about the traditional DBMS and its current costing problem.

### 1.4 Why other approaches are not good

Other approaches are optimizing performance and using database management tools. These approaches are not as efficient. For example, optimizing performance sounds great, it can help increase the productivity of businesses and would be able to increase profits that way, however, the downside is that there will be an increase in costs and that defeats the purpose. Using database management tools is another approach. These tools can aid in lowering license costs. However, this is more expensive and the cost of restoring data and maintaining security will also be very high. Also, to be able to choose a specific tool can be very difficult because it needs to match the requirements of the companies and those can change. Requirements are always changing so this would not be a good approach. Thus, our approach seems to be most efficient and everyone would benefit from this. Hybrid clouding will help companies that want to shift to the cloud but are not quite sure. This will help them make a transition and be able to switch back if it is not right for them. The benefits that hybrid cloud bring are significant and are definitely cost-efficient.

### 1.5 Statement of the problem

Database Management System is an effective tool to manage data, but the cost it takes might be higher than the benefits it brings about. For DBMS, it is mandatory to have a high-speed processor and a large memory size, because now there is a large amount of data in every field which needs to be stored safely and securely. This becomes very costly. This research paper will investigate the current cost of DB Management System/Software/IOs analyze our approach of hybrid cloud as a new and improved solution to having a DBMS with efficient cost management.

### 1.6 Area or Scope of investigation

In this project, we are implementing a hybrid cloud architecture to reduce the cost. Further to cut costs on the cloud we will be using RavenDB requests by utilizing ‘lazy requests.’ This project is expected to help users reduce the cost of DBMS and may affect many industries inside the United States and all over the world. The paper may develop innovative thinking by introducing a new trend combining with the traditional system, which helps eliminate the hesitation of changing.

# 2. Theoretical bases and literature review

## 2.1 Definition of the problem

The database system's goal before was to minimize the response time of each request given a number of machines. However, now we need to minimize the number of machines (cost) given the response time goal. The goal of this paper is to show how hybrid cloud is the best approach for a cost-efficient database system.

## 2.2 Theoretical background of the problem

Users want everything. They want zero cost, zero response time, infinite scalability, and much more. Nothing has changed, except the priorities. Cloud computing has made cost a continuous metric.

| *Feature* | *Trad. DB* | *New DB* |
|---|---|---|
| Cost [$] | fixed | minimize |
| Performance [secs and tps] | optimize | fixed |
| Scalability [machines] | maximize | fixed |
| Predictability [$ and secs] | - | fixed |
| Consistency [%] | fixed | maximize |
| Flexibility [#variants] | - | maximize |

Table 1: Traditional vs. New DB Optimization Problem

Table 1 shows us how the traditional database system used to work and the new database system that people are wanting. A hybrid cloud will be able to give this to the users. It is good to use a hybrid cloud when the workloads are dynamic or frequently changing. Use an easily scalable public cloud for your dynamic workloads, while leaving more sensitive workloads to a

private cloud or on-premises data center. You would be able to separate critical workloads from less-sensitive workloads. Also, there will be no rush in moving to the cloud, you can move at your own pace. A hybrid cloud lets you allocate public cloud resources for short-term projects, at a lower cost than using your own data centers infrastructure. A hybrid cloud has many benefits. A hybrid cloud is a better and more efficient approach.

### 2.3 Related research and their advantages/disadvantages

To conduct this research paper, seven research papers are referred to: “An Efficient Cost Model for Data Storage with Horizontal Layout in the Cloud,” “Cloud Database Management System Architecture,” “Rethinking Cost and Performance of Database Systems,” “Towards Eco-friendly Database Management Systems,” “Development of a Database Management System Design Involving Quality Related Costs,” “Database Management Systems for Statistical and Scientific Applications: Are Commercially Available DBMS Good Enough?”, and “Impacts of Data Mining on Relational Database Management System Centric Business Environments.” Since the two first research paper talking about the same method, this section will discuss only “An Efficient Cost Model for Data Storage with Horizontal Layout in the Cloud”, the other research will be used for other section.

“An Efficient Cost Model for Data Storage with Horizontal Layout in the Cloud" proposes cloud storage as an efficient cost model for DBMS. The paper emphasizes that the horizontal layout aggregated data will help improve the query performance and reduce the cost of data storage in the cloud architecture more than the vertical layout data [1]. Accordingly, it gives a table of price tags for cloud consumption. However, the paper also indicates that the approach effectively reduces the storage cost for small data; data transfer cost and computing

cost will be added with large databases, and it might be more expensive. The analysis is quite short, and it does not show all the aspects of the proposal solution.

"Rethinking Cost and Performance of Database Systems" points out the expectation of the users to the DBMS and proposes a new architecture to minimize the cost of the system given the performance requirements of an application. The research paper is well-written with clear statements and analysis. It is specific with the comparison between the two three-tier architectures of the traditional DBMS and the new DBMS. Both good and bad sides of the two systems are also indicated. The proposed new database application architecture can significantly reduce the cost of the DBMS, but it needs time to be mature, and the process will be complicated and require high technical-level developers [4]. To be more specific, the new structure is a general idea as the authors do not give in detail of how to build it. Big players like Google still need a great deal of work to launch their products.

"Towards Eco-friendly Database Management Systems" emphasizes the tasks of managing the energy consumption when processing queries. The cost of energy consumption has increasingly gone up in recently years, and it is the third largest cost in data center, the other top two are server hardware and power distribution and cooling costs [6]. The paper is quite interesting because it attracts attention to an old energy and environment issue but in the aspect of DBMS. The paper proposes a project called ecoDB to develop energy efficient data processing techniques including "global" and "local" techniques. The task is difficult because modern motherboards are multi-layered and tapping into the components is not trivial.

"Development of a Database Management System Design Involving Quality Related Costs" reduces the cost of DBMS by minimizing the quality costs which are costs resulting from the failures of inadequate systems. The authors design an interface that can store data of quality

system from users and calculate to check whether the money spending on the systems is useful or not [8]. Some entity-relationship diagrams to model the cost DBMS is shown to better convey the idea. The biggest achievement of this paper is the proposal and process that are easy to understand. Moreover, the system introduces high-level flexibility and could be developed to incorporate the elements of production and manufacturing systems. The drawback of the research is that it might be difficult to build such a platform. The disadvantage of the calculating tool is that it is just a general approach in tracking and measuring quality cost. Different entrepreneurs in different industries might have to go over a great number of complex phases and activities to develop the operations.

"Database Management Systems for Statistical and Scientific Applications: Are Commercially Available DBMS Good Enough?" investigates if the available DBMS is commercially good enough for Statistical and Scientific Applications. Although the title is not directly about the cost of DBMS, the analysis does comprise some relevant information and statistics. It also gives some instant methods to decrease the cost of DBMS such as low-cost storage technologies like disks and high-density cassettes and telecommunication [9]. The conclusion suggests that the most flexible and versatile systems should be chosen, which is quite reasonable. In general, the research has clear structures and analysis.

"Impacts of Data Mining on Relational Database Management System Centric Business Environments" raises the importance of data mining in reducing cost in DBMS and points out some challenges in relational database management system (RDBMS). Data mining is the process of searching useful information to support the policy maker to take the strategic decision in the business [5]. This might be one of the simplest ideas for DBMS's cost problem. Nevertheless, the process is more complicated than how it looks. The disadvantage of the paper

is that data mining might be not easy and understandable to the end users. The good side of the research is that beside cost reduction, it considers other crucial issues such as scalability, availability, flexibility, security. Moreover, data mining can be applied to both traditional DBMS and the new proposed architecture.

### 2.4 Other possible solutions

Some other solutions to reduce the cost of DBMS are consolidating, optimizing performance, using database management tools, and cloud computing. The disadvantages of those strategies could be pointed out as follows.

First, consolidate servers and pay less for each instance. We can reduce licensing costs by moving multiple databases onto the same logical database instance. The challenges associated with this approach are network connectivity, transfer methodologies, and data normalization. Due to cybersecurity threats, certain percentages of the IP network necessarily need to be separated from each other, which can prevent facility management systems from taking data from sharing the Internet. Data transfer methodologies include three typical options: web services with the risk of pushing data to a destination platform that is unready to process data, flat file transfer, and open protocols which are the most basic. Data normalization leaves out critical parameters associated with data that are difficult to translate.

The second method is optimizing performance. Solely focusing on optimizing performance with the traditional database system can help increase productivity of the business and therefore get more profits, but the drawback is that using the same database structure with optimized performance will certainly increase costs. In addition, many performance problems are no longer caused by the DBMS; thus, improving on the performance might not help.

Using database management tools is another good choice. The tools can help lower license costs and lessen serious incidents. A vast array of software applications offers many forms of licenses. A full range of features license for enterprises is usually the most expensive. However, organizations might not use the full suite of features and want to use database management tools to analyze actual usage and report on which software and features are being used, then adjust the licensing levels. On the other hand, serious incidents can lead to a very high financial cost for the DBMS such as the cost of restoring data and maintaining security. Monitoring tools can be used to detect and correct problems before they become serious, discover security weaknesses, and institute regular preventative maintenance across the organizations' servers and instances. The only issue of database management tools is that it might be a little difficult to choose a tool that matches the type and requirements of the companies.

Cloud computing is separated from other methods because it strongly relates to our proposal. Cloud computing has been developed significantly in recent years due to its huge benefits. It offers good performance even for big databases since it is in company responsibility. Databases can be scaled quickly, cheaply, and efficiently. The administrative burden is reduced because the cloud-hosted system can eliminate unnecessary features that consume much of the database administrators' time and efforts. The issues of Cloud Database are internet speed, query and transactional workload, multi-tenancy, privacy, and the high cost. Cloud computing requires stainable internet connection since it is the communication between devices over long distance. Internet speed is both a supporter and a barrier of performance. Next, although we can control the transactions, we do not have the control over query workload since we do not know the number of queries. Multi-tenancy occurs when several users are served by a single software

instance at the same time, which will affect the efficiency of the database. For privacy, cloud databases are accessible through the network, which is an important place for hackers to try breaking the system. Some might advise that users just need to leave it to the cloud company and have some peace in mind, but there is no guarantee that the data will never be leaked. Finally, cloud pricing is quite expensive: enterprises need to pay a high fee when migrating a large dataset to the cloud and when using the storage. The more storage they use, the more they must pay, so enterprises might need to scarify some efficiency at scale. However, for some organizations, the total cost of the ownership with cloud solutions can be far lower than the on-going cost of maintaining on-premises hardware. Thus, it really depends on the situation and the industry of companies to evaluate whether the cost for migrating to the cloud is too high or not. The good side is that there are solutions for the above challenges and users just need to take some more steps to resolve them.

### 2.5 Our solution and why our solution is better

To take the advantages and minimize the disadvantages of the clouding system, our paper proposes hybrid clouding. This model helps an enterprise try out cloud tools before fully committing to the system. Since the cost for cloud migration is not small and companies might be unsure if clouding is a good match for their system, companies can experiment by transferring some portions of the database to cloud. When we need temporary processing capacity, a hybrid system helps save money by letting us allocate public cloud resources for short-term projects; in this way, we do not overpay for the equipment that we do not need in the future. Thanks to highly scalable cloud resources, big data analytics is possible on cloud. Additionally, the hybrid approach is a good solution for privacy because it enables companies to keep data in an on-

premises private data center. We can connect existing systems containing sensitive data that are not suitable to the public cloud. The biggest benefit of hybrid cloud models is that users are free to continue expanding their cloud presence as needed. Whenever they feel like the clouding system is no longer suitable for their purpose, they can stop anytime. On the other hand, hybrid cloud has some similar problems like cloud databases such as internet speed, query workload, and multi-tenancy. The biggest challenge lies in getting staff prepared in a way that does not compound the issue that starts the disaster recovery. For the cost issue, hybrid clouding systems could save some money for entrepreneurs since only some portions of the database are migrated to cloud. Despite some disadvantages of hybrid cloud DBMS, the benefits it brings are significant, including reducing the cost of DBMS.

# 3. Hypothesis

## 3.1 Multiple Hypothesis

There are many costs and benefits that need to be considered when incorporating a new approach. Many of the benefits that we see have to do with whether the initial investment makes sense for a company. For many companies, the main factor they consider is the dollar value associated with taking on a new concept and incorporating it into a workflow. However, it is also important for companies to think about the future and take manageable risks which could help sustain the company in the long run. Not only that but it may also make them more marketable and relevant for the future. This has to do with not just marketing products but also marketing the company so that new talent will want to join, learn and grow together with the company.

We are surrounded by data and we are generating data in large amounts every second of our life. There is so much data in the world, but it has not yet been put to good use. Companies can use the data that is out there to make their products better and more applicable. A good example of this is an Apple Watch. Maybe the first generation of the watches could not pick up the slight steps taken by the human, however as more and more data get collected the Apple Watch is able to detect it better. Data is all around us. It is not just in our phones and laptops but also the Roomba vacuum cleaner and the nest that is on our doors. All the internet of things (Iot). That is why we need good data management systems. A well thought out data management system secures and maintains data. It helps maintain crucial information within an organization.

Using cloud computing reduces cost in the long term and increases flexibility of a database that is also mobile. Even though there are many costs and benefits associated with having a cloud-based data system, it also helps companies stay relevant in a technologically advancing society.

Cloud computing differs from traditional databases because it offers flexibility. Costs related to cloud implementation would be resources that are consumed by a system, deployment method chosen, and the cloud service provider's pricing. These are all the factors that play a key role for decision makers in an organization. Transitioning to a cloud-based system has quite a few benefits as well. In terms of accounting, there is no need to purchase expensive hardware and network infrastructure. This lowers the upfront cost involved in setting up a database. It is also energy saving.

### 3.2 Positive or Negative Hypothesis

Servers consume a lot of energy to operate. That is why energy consumption is an important factor in design. There are different ways to minimize cost by investigating other energy efficient methods for general data processing. This is not just for servers but also clusters, data centers, running either a database management system or some other data processing system. We can save energy by taking a bit of penalty on response time if a project does not consider that to be an important aspect.

CPU nowadays consume the majority of the energy compared to the other rest of the components in a server, even though other components are also not that energy efficient. It could be that these components do become more energy efficient over time. However, at the moment they are not. The two biggest sources of energy consumption are the processor and memory.

Yes, cloud storage is expensive however it reduces the need for upward scaling and adding memory. Transitioning to a cloud database consolidates the number of physical machines and their replacement for servers, network devices, and other security systems. In a company, it

also saves staff because there is less need for overhead. Cloud database simplifies managing the information system, by only concentrating staff on essential activities.

In the environment that is globalized, there is a true need for flexibility that influences all of the different business processes. It is increasingly important to use all the resources available in an efficient and effective manner. Cloud computing is a technology that can help organizations achieve these objectives.

# 4. Methodology

## 4.1 How are data generated?

So, how exactly does cloud storage work? Cloud storage involves having at least one data server that users can access via the internet. Then the user is able to send files manually or in an automated fashion, to the data server. The data is not just stored in one location, rather it is replicated in vast numbers of data servers to ensure availability of the data. Users can access that data that is in the cloud from anywhere with the help of internet service. These locations where data is stored are called data centers.

## 4.2 How does it solve the problem?

There are hundreds of cloud service providers. The amount of storage being offered is also growing rapidly. Because of this increase in supply, the price of cloud storage has lowered. Usually there is also some amount of free space available for clients to test it out.

Cloud based data storage solves the monetary issues that come with traditional database management systems. Cloud computing is scalable. When servers reach their potential and there is no further growth happening, the system can get cloned. Elasticity offers some of the same computing experience that we are used to traditionally, but cloud has other resources. There can be a system set up which allows an organization a more efficient place for communicating and file sharing. Many of the times, companies use cloud storage as a backup rather than the only source. This ensures that crucial data is not lost, and it is backed up to the cloud. Cloud computing providers tend to keep the information more securely backed up.

There is lower initial investment needed with cloud. The only things that will be necessary are a computer and an internet connection. Even though having this is nice, there is no

need to invest in any new hardware, other specialized software. There is also no need to add new staff overseas. It is a much-simplified process to manage data. Having these benefits and saved investment allows companies to invest in new projects and ideas without risking a big loss.

Since there is not processing power and storage space considerations there isn't a need to understand all of the technology that is involved with traditional databases. Planning and executing this method is considerably simpler and does not require an organization to maintain and update new hardware every time something goes wrong. Cloud based systems are easy to deploy and there is not a need to plan a new system. This is something that can be set up within a few days. Yes, there is a learning curve involved in putting something like this set up, however this is something that will put the company at an advantage later in the future. Being able to accept new technology and hiring people that understand that technology enable an organization to benefit from in the long run.

Putting up a cloud-based service allows the company to be able to access these vital data from anywhere there is internet. This is help to people that travel all around the world, and work from different situations. It is not independent to any specific system or browser.

Overall, data storage and retrieval amount to a lot of money and larger upfront costs are associated with it. Relational database models are the core of most transactional systems, meaning as the data grows it will start to take more processing power.

### 4.2.1 What algorithms are used?

Cloud computing has grown and with it there are concerns regarding security. Some of these issues are ensuring a secure data transfer, secure interface, separation of data, stored data, and user access control. Cloud computing should use a method to make the transfer of data more

cryptic. This is important because of data integrity. Security Algorithms can be Private Key or Symmetric Algorithms. Using a secret key that encrypts a large amount of data with fast processing speed. Some examples of this algorithm are RC6, 3DES, Blowfish. This uses a secret key that is known to the sender and receiver.

There is also Public Key or Asymmetric Algorithms. This has key pairs, with public key for encryption and private key for decryption. Such an algorithm has high computation cost which in turn causes slow speed when compared to a single key algorithm. Some examples of public key algorithms are RSA and Diffie Hellman.

There are also Signature Algorithms. This is where the client signs in and authenticates the use of data based on a single key. The password is saved in encrypted format to secure the data.

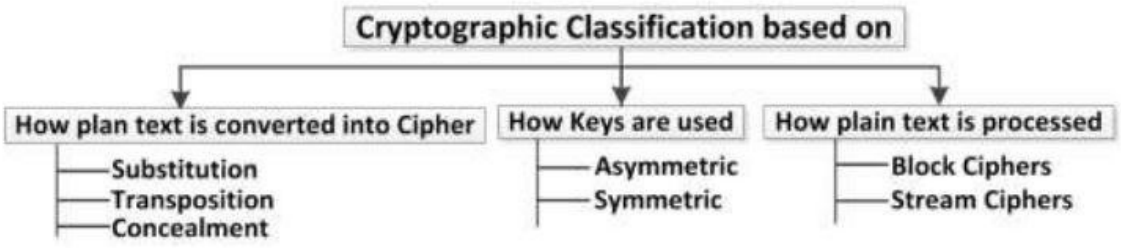



Image:1

Image 1 shows how an algorithm system, Cryptographic Classification works (*). It is based on converting a plain text into a Cipher, with the use of substitution, transportation, and concealment. The keys that are involved are either Asymmetric or symmetric. The plain text is then processed using block ciphers and stream ciphers. Having these algorithms helps make cloud computing secure and cuts down the cost of buying even more security systems.

### 4.2.2 What language is used?

The above algorithms are based on .NET and other Microsoft technologies. The languages differ based on the cloud service providers. The amount of space, pricing are all factors that companies need to consider when looking at incorporating a cloud-based infrastructure.

### 4.2.3 What tools are used?

There are many tools available to make cloud computing, more secure, cheaper, and even more flexible. Unlike transactional database management systems, most of these tools are easy to incorporate into an already existing system, such as increasing the amount of storage that is needed, and only paying for what is being used. We can look at Cloud Monitoring, which is an automated and manual tool used to manage, monitor and evaluate the cloud computing architecture, the infrastructure, and the services. Having such tools enable the client to compare the cost benefits associated with running a cloud-based system. It helps detect any outstanding issues and come up with a resolution in a timely manner. This ensures that the client's customers are satisfied and there are not many complications involved.

There are many tools to supplement a cloud database system. They make the experience even better and worth the investment. Even though there is a lot of unknown for someone to venture off into, there is much information available and it is easier to understand compared to components necessary while maintaining a traditional database management system. There is a high use of energy, high processing power, and the need to store space.

Some of the best cloud management and monitoring tools out in the market now, are Amazon Cloudwatch, Microsoft Cloud Monitoring, AppDynamics, BMC TrueSight Pulse, DX

Infrastructure Manager (IM), New Relic, Hyperic, Solarwinds, ExoPrise, Retrace, Aternity, PagerDuty, and many more. Many of these tools are used for organizations, for added security, offering intelligent monitoring, and analyzing and getting insights for the data that is stored. For example, Amazon Cloudwatch offers the client insights into the systems performance, and need for maintenance. It provides metrics, stores log files, views graphs and statistics, and monitors or takes action if there are any resource changes that need to be handled within the system.

# 5. Implementation:

## 5.1 Code

The codes are written by Python and SQL. Table 9.1 – 9.5 on appendices are used to build the tools. Since using large dataset (several GBs) could lead to technical difficulty in the programming process, we use some datasets of several KBs each.

The following files are used to implement our proposed algorithm:

- governors_county.csv
- governors_county_candidate.csv
- governors_state.csv
- house_candidate.csv
- house_state.csv

Most functions are written in Jupyter notebooks (Python), including the functions to calculate the storage cost of the clouding system, hybrid clouding system, and in-house server data center. Please refer to the code submission for those Python functions. On the other hand, function to calculate query cost needs to be done in SQL server to make the experiment more reliable.

This is the explanation of the SQL codes that calculates the query cost.

We wanted to see the storage cost, duration, and cost in dollars for every query that is executed. To do this we used Postgres SQL to create tables of our sample data and create queries on that data.

First, we created a list of queries that are most likely to be executed in a month.

**Queries**

SELECT total_votes FROM house_candidate where total_votes < 300000

INSERT INTO governors_state(state, votes) VALUES ('California', 500000);

SELECT house_candidate.district, house_candidate.total_votes FROM house_candidate INNER JOIN house_state ON house_candidate.district = house_state.district

SELECT count(*) FROM house_state

SELECT * from house_state

SELECT votes FROM governors_state where votes > 300000

Second, we created a function called query_cost that would calculate the cost and duration of each query.

```
CREATE OR REPLACE FUNCTION query_cost(
      queries text[],
      query OUT text, cost OUT float8, duration OUT float8
   ) RETURNS SETOF record LANGUAGE plpgsql STRICT AS
$$DECLARE
   i integer;
   p json;
BEGIN
   /* loop through input queries */
   FOR i IN array_lower(queries, 1)..array_upper(queries, 1) LOOP
      query := queries[i];
      /* get execution plan in JSON */
      EXECUTE 'EXPLAIN (ANALYZE, FORMAT JSON) ' || query INTO p;
      /* extract total cost and execution time */
      SELECT p->0->'Plan'->>'Total Cost',
             p->0->'Plan'->>'Actual Total Time'
         INTO cost, duration;
      /* return query, cost and duration */
      RETURN NEXT;
   END LOOP;
END;$$;
```

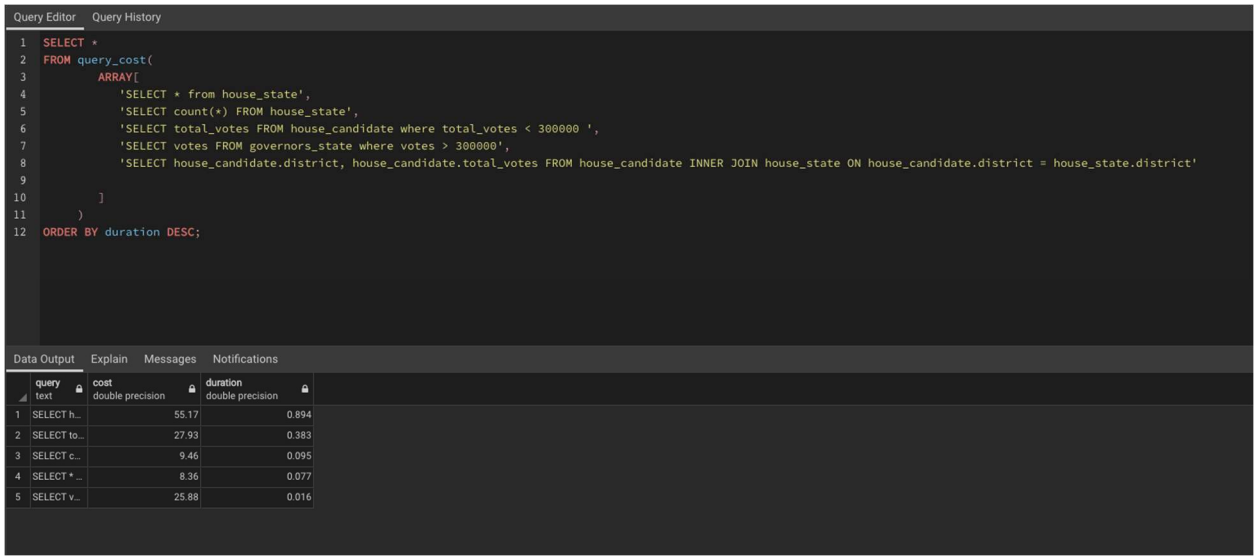
This function goes through each query and analyses the query to extract the total cost and execution time. It will then return the query, cost, and duration.

To execute this function,

```
SELECT *
FROM query_cost(
        ARRAY[
            'Your Queries HERE'
        ]
    )
ORDER BY duration DESC;
```

This will execute the queries and order by duration by calling the function we created.

Here is an example of using the function:




As we see in this figure each query is being executed and the following output is correct.

### 5.2 Design document and flowchart

Our goal is to compare the costs of the database storage and management when we store the dataset on premises and on hybrid clouding system. We build possible queries that can be used if a user or a company has a dataset, then calculate the costs if we run those queries given the dataset is stored on premises or on cloud. For cloud storage, we consider the price of BigQuery on Google Cloud in this paper and use that price level to calculate the cost of storage and the cost the queries we run. Since our proposal is hybrid cloud database, we give some sample fractions of on-premises database and cloud database and to create a hybrid system and analyze it accordingly. There are two document flows which are the analysis flow and the user flow. The analysis flow shows the analysis process of the paper's analysis and the user flow shows the process that appears in the user's end.

**Analysis flow:**

1. Start
2. Read the dataset
3. Find all possible queries for one month on the dataset.
4. Calculate the storage and management cost in case the dataset is stored on premises
5. Calculate the storage and management cost in case the dataset is stored on cloud. The cost of cloud system includes storage cost and running queries cost.
6. Give sample fractions for in-house and cloud dataset to analyze the hybrid system
7. Conclusion

**Flow Chart of Analysis:**

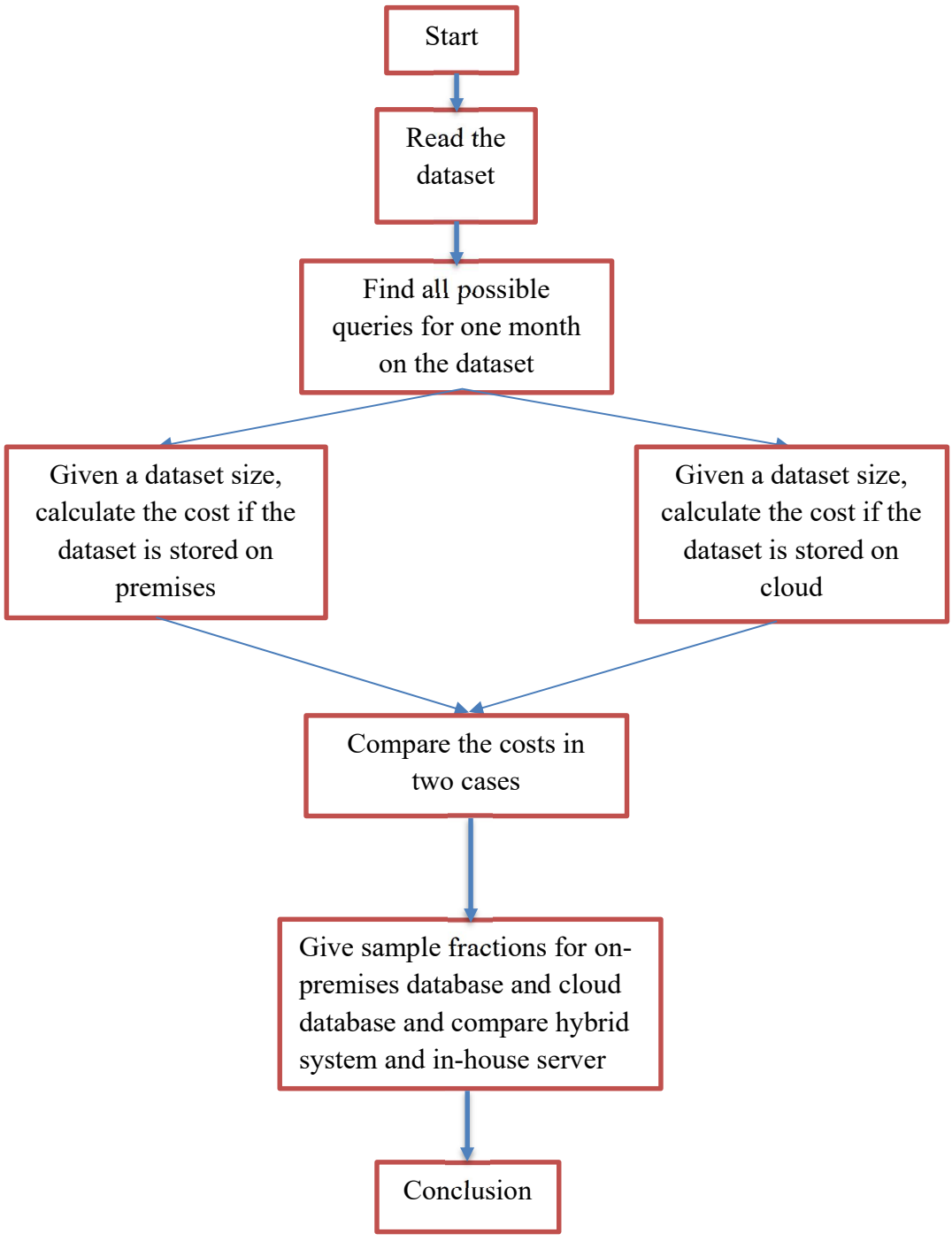



**Figure 5.1:** The steps to analyze the storage and operation cost of on-premises dataset and hybrid-cloud dataset

**Users flow:**

The first step is to ask the user to enter the storage amount he/she wants to store his/her dataset on cloud or on premises.

- On Cloud:
    a. Ask the user to enter storage amount and storage time, return the cost of cloud storage for the dataset.
    b. Ask the user to enter each query he/she will execute on a cloud database in one month. Record the cost for each query to a table.
    c. Add up all the cost and return the total cost for cloud storage on the dataset.
- On-premises: calculate the cost of the storage and return it
- Add up the cost of cloud and in-house database. In case the user wants to investigate the cost of storing data on-premises only, enter 0 in the amount stored on cloud.

**Flowchart of Users:**

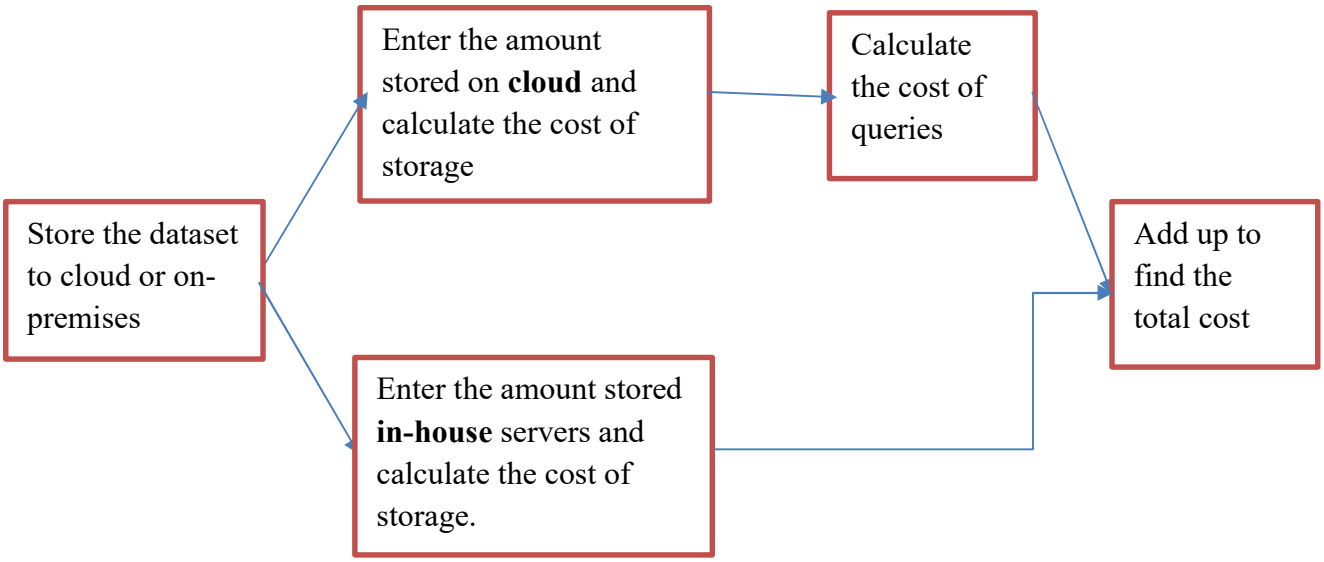



**Figure 5.2:** Flowchart of finding the cost of a hybrid system

## 6. Data analysis and discussion

The price is calculated on the Google Cloud standard storage price as of December 2020 of Los Angeles area, which is $0.023 per GB per month from table 9.2 or 2.3 cents per GB per month. Besides, there are some options of cloud storage systems on Google Cloud such as Google BigQuery and Google Cloud Storage. Our storage system is Google BigQuery since it is more suitable for managing relational databases.

Based on table 9.1, the total cost for 1 PB of on-premises storage in five year is $1,970,000. Since the table is in 2018, we consider the inflation rate of 1.81% in 2019 and 0.62% in 2020 [3] and suppose that the total cost in five year for 1 PB of data is about $2,054,000 in 2020. Then, the cost per GB per month is $0.034 per GB per month or 3.4 cents per GB per month. However, notice that this amount does not include retrieval, which means if disaster happens and we lose all the dataset, we are not able to take the data back. To make it comparable with clouding system, there are two ways for our assumption: the first option is to save a back-up version on Google Cloud Storage, and the second option is to save them in a storage hardware. In this paper, we prefer doing the latter method. Based on table 9.1, we use the storage hardware cost of $500,000 per PB per five years, or 0.87 cents per GB per month with inflation rate. Thus, the total cost for on-premises data center is $0.0427 per GB per month.

The query cost we are using also derives from Google Cloud website (table 9.2), which is $5 per TB.

## 6.1 Output generation

```
check_file_size("C:/Users/14086/Desktop/Data225-Dataset/")

File name: governors_county.csv
Size: 38861 bytes

File name: governors_county_candidate.csv
Size: 248348 bytes

File name: governors_state.csv
Size: 220 bytes

File name: house_candidate.csv
Size: 68175 bytes

File name: house_state.csv
Size: 19783 bytes

Total dataset size (in gbs):
0.000375387
```

**Figure 6.1.1:** Amounts of data storage

```
cloud_storage_cost()

Enter the storage amount in one month(in GB):  375.387
'Total storage cost in one month is: $8.403901'

premises_storage_cost()

Enter the storage amount for on-premises in one month(in GB):  375.387
'Total storage cost for one month is: $16.0290249'
```

**Figure 6.1.2:** Storage costs of clouding system and in-house server

From query function written in PostgreSQL server, we can withdraw the storage cost and duration of a random query. We want to create a table that includes all possible queries that can be run and the cost in GBs and in dollar for those queries. Since delete operations are free resources on BigQuery, we ignore those in this paper.

| Query | Storage Cost (in GB) | Duration | Cost (in a dollar) |
|---|---|---|---|
| SELECT total_votes FROM house_candidate where total_votes < 300000 | 2.793*10^(-9) | 1.685 | 0.139*10^(-9) |
| INSERT INTO governors_state(state, votes) VALUES ('California', 500000); | 1 *10^(-9) | 0.699 | 0.00005*10^(-9) |
| SELECT house_candidate.district, house_candidate.total_votes FROM house_candidate INNER JOIN house_state ON house_candidate.district = house_state.district | 5.517*10^(-9) | 0.997 | 0.2758*10^(-9) |
| SELECT count(*) FROM house_state | 9.46*10^(-9) | 0.073 | 0.0473*10^(-9) |
| SELECT * from house_state | 8.36*10^(-9) | 0.052 | 0.0418*10^(-9) |
| SELECT votes FROM governors_state where votes > 300000 | 2.588*10^(-9) | 0.007 | 0.1294*10^(-9) |

**Table 6.1.3:** List of possible queries and the cost for each (in dollar)

```
hybrid_cost()

Enter the total storage amount in one month(in GB):  375.387
Enter the storage fraction that is stored on cloud (i.e: 0.3):  0.3
'Total storage cost for one month is: $14.589527986'
```

**Figure 6.1.4:** Hybrid cloud cost for 30% cloud storage and 70% on-premises storage

## 6.2 Output Analysis

From figure 6.1.1, the total amount of the datasets is 0.000375387 GBs, which is quite small. We use the small dataset to minimize technical difficulties, but still need a large-enough dataset to deduct the 10GB free storage each month of Google Cloud BigQuerry (table 9.2). We want to make up the dataset by copy each file 1,000,000 times, then we have a total of 375.387 GBs dataset.

If we store all the dataset on solely cloud or an in-house server, the storage cost of the on-premises data center is approximately 1.9 times more expensive than that of the clouding system. Since we do not need space to run queries in the data center, the storage cost of $16.029 is also the cost for the entire in-house server database system in one month.

Now, we want to see the monthly cost of using these queries.

To calculate query data usage we need to start by estimating a few basic parameters of our service,

# of Users (per day)

# of Queries ( per User, per day)

Average Data usage per Query

Monthly_Query_Data_Usage = numUsersPerDat * numQueriesPerUser * dataPerQuery * daysPerMonth

Lets say for example, we have about 1 user per day, each running 30 queries per day, with an average data usage of 5 GB per query,

1 * 30 * 5GB * 30 = 4,500 GB = 4.5 TB Monthly_Query_Data_Usage.

Also, keep in mind that google cloud gives 1TB free per month.

So we would have 4.5(total TB) - 1(1 Free TB) = 3.5 TB (Cost we pay for)

We can simply multiply that by $5 per TB cos of BigQuery at the time we get an estimation of ~$17.50 per month for Query Data usage. Based on this method we are able to use the same technique in calculating the cost of the query.

Suppose a month has 22 business days. Notice that we assume that we have a dataset that is 1,000,000 bigger than what we are having now. Therefore, the cost for each query will be 1,000,000 times more expensive than table 6.1.3. From the information and table 6.1.3, we create the below table which shows the total query cost for one user per month, and from the new data, we can calculate the total query cost for a specific number of users accessing the dataset in one month.

| Query | Cost per query for our 375.387GBs dataset | Number of times it runs per month | Cost per month |
|---|---|---|---|
| SELECT total_votes FROM house_candidate where total_votes < 300000 | $0.139*10^(-3) | 20 times/day * 22 days = 440 times/month | $61.16*10^(-3) |
| INSERT INTO governors_state(state, votes) VALUES ('California', 500000); | $0.00005*10^(-3) | 150 times/day * 22 days = 3300 times/month | $0.165*10^(-3) |
| SELECT house_candidate.district, | $0.2758*10^(-3) | 20 times/day * 22 days = 440 times/month | $121.352*10^(-3) |

| | | | |
|---|---|---|---|
| house_candidate.total_votes FROM house_candidate INNER JOIN house_state ON house_candidate.district = house_state.district | | | |
| SELECT count(*) FROM house_state | $0.0473*10^(-3) | 20 times/day * 22 days = 440 times a month | $20.812*10^(-3) |
| SELECT * from house_state | $0.0418*10^(-3) | 20 time/day * 22 days = 440 times a month | $18.392*10^(-3) |
| SELECT votes FROM governors_state where votes > 300000 | $0.1294*10^(-3) | 20 time/day * 22 days = 440 times a month | $56.936*10^(-3) |
| | | Total cost per month per one user | $0.278817 |
| | | Total cost per month for 30 users | $8.36451 |
| | | Total cost per month for 30 users after 1TB is free ($5 deduction) | **$3.36431** |

**Table 6.2.1:** Query cost per user per month

According to the table, we see the query, the number of times it runs per month, and the cost per month. This shows how much each query is being used and the monthly cost. Lastly, we calculated the cost per month, and we also calculated the cost after knowing that 1TB is free which is a $5 deduction. We see that for the queries that we have created the total cost per month for 30 users would be $8.36451 and the final total cost after 1TB free would be $3.36431.

At this point, the cost of the clouding system can be determined as the sum of the cloud storage cost and the cloud query cost. Since our proposal is hybrid clouding system, assume that

we have a dataset of size 375.387 GBs, we want some portion of that be stored on the cloud, and the remaining portion is stored in the in-house data center.

In figure 6.1.4, the dataset size is 375.387 GBs with 30 percent of that is are stored in cloud and 70 percent of the data stored in in-house server. In this case, the total cost for storage and management of the hybrid clouding system is $14.590, which is lower than the in-house data center. To make it more general, it is necessary to show the cost of the hybrid clouding system with different fractions of cloud storage.

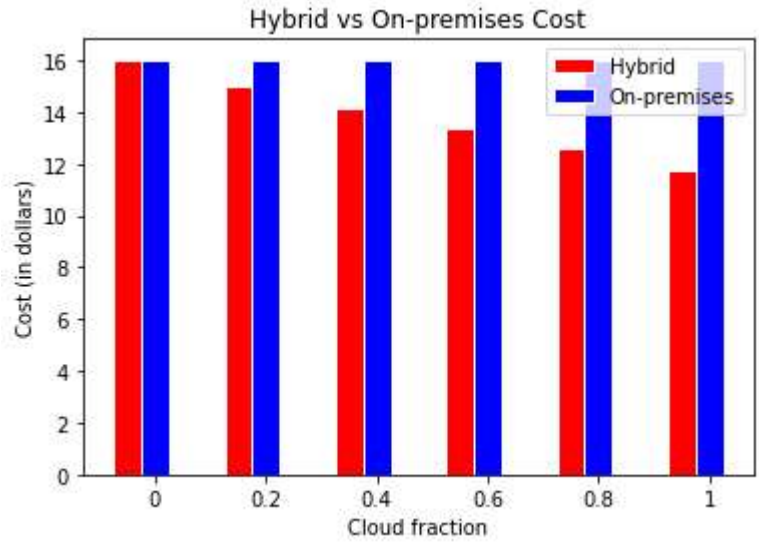



**Figure 6.2.2:** Hybrid vs On-premises system cost

The above bar graph shows the cost of hybrid database management versus solely on-premises data center. The red bars represent the hybrid system cost, which corresponds to different storage fraction cloud/on-premises levels, such as 0, 20, 40, 60, 80, and 100 percent of the dataset stored on the cloud, and the remaining portions stored on premises. The blue bars indicate the total cost if we solely save all the dataset in the data center. Observe that as the fraction of cloud storage increases, the cost of hybrid system decreases. Notice that it just

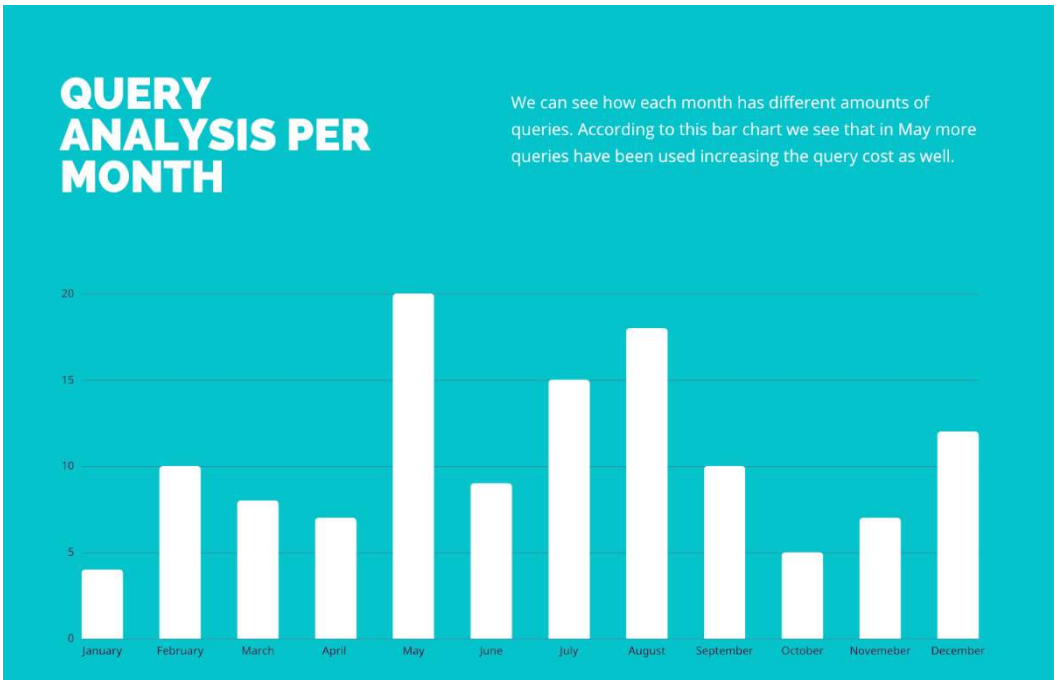
illustrates the cost for one month. In the long run, like one year or one decade, the cost of hybrid clouding system will be much lower.

### 6.3 Compare output against hypothesis

Our hypothesis is that the average cost of hybrid system is lower than the average cost of solely saving the dataset in the in-house data center. Through our experiments, we can conclude that we are failed to reject our null hypothesis.

### 6.4 Abnormal case explanation

The abnormal case is when the query cost is too high. For example, a data analyst in one specific month runs much more queries than a normal month. If there is a huge difference between the number of queries towards the end of the year compared to the beginning of the year, this would differ in the monthly costs and affect the companies or teams budget. For example, if the amount of queries in one month is 100 and in the next month is twice that amount the query cost would become very high.




This figure illustrates a month in which query costs are significantly high. We can see how each month has different amounts of queries. According to this bar chart we see that in May more queries have been used increasing the query cost as well. This would be considered an abnormal case.

### 6.5 Discussion

Our goal is to show how a hybrid cloud is much more beneficial and can be achieved at a lower cost than on-premises. Hybrid cloud computing offers numerous benefits and advantages to enterprise organizations. The primary advantage of a hybrid cloud is agility. The need to adapt and change direction quickly. Hybrid cloud has the benefit of being cost-effective as organizations pay for the public cloud position of their infrastructure only when it is needed. Also, with a hybrid cloud, you are able to get a centralized private infrastructure on-premises. While researching the cost of queries and analyzing the cost differences between on-premise and cloud we were able to see how much more hybrid cloud is efficient. There are also a few

disadvantages to having a hybrid cloud. Although the long-term cost savings are one of the many benefits, the initial deploying cost of a hybrid cloud exceeds as compared to the setup cost incurred in the case of a public cloud. While creating a hybrid cloud environment, specific hardware is required to deploy on-premises, and that's what shaves off a large chunk of the budget. Also, if not picked correctly, cloud compatibility can become a real nuisance for Hybrid Cloud environments. A fast performing on-premise infrastructure may not be able to successfully perform in coherence with a slow-performing public infrastructure resulting in a sluggish performance of the Hybrid Cloud. However, the hybrid cloud does have a high number of advantages and it is one of the most stable cloud environments.

During this project, we did face a few challenges. While we did do research on hybrid versus on-premises and were able to understand the advantages. We needed to show factual data to prove our hypothesis, which is that the hybrid cloud is more cost-efficient than on-premises. To do this, we needed to do more research on how we should code these functions and show the data. After talking to the professor, we realized that we need to take a sample dataset and create and execute our functions around that data. Thus, we were able to overcome this hurdle and start to execute the functions using the sample dataset that we retrieved from Kaggle.com.

# 7. Conclusions and recommendations

## 7.1 Summary

The results of the analysis say that cloud database management may be costly if we continuously run the queries in a specific period. However, if we just save the data for back-up or not run queries so many times, the cloud database is a much cheaper system. It also allows enterprises to take advantage of cloud services and free up personnel to concentrate on other assignments. Although cloud databases seem to have many cost advantages, we need to consider its drawbacks, especially privacy issues. Based on the experiment and the logical analysis, we still propose a hybrid clouding system with a low portion of private database stored on-premises data center and higher portion of public database stored on cloud.

## 7.2 Recommendations for future studies

Our present implementation of the queries is limited because there might be more complicated queries that need to be executed in the reality. The price chart also varies with different clouding systems and storage devices. This research focuses on the cost of cloud databases and on-premises databases. To compare the cost of a cloud database management system with other proposed systems, it might need more work and effort since some proposals are incomplete, such as building new architecture for the management system, or building a new tool to calculate the quality related cost. Besides, depending on the requirements of companies and users, other factors need to be considered. For example, privacy, scalability, consistency, etc.

## 8. Bibliography

[1] Brintha, S., and Nalini, C. (2014). An Efficient Cost Model for Data Storage with Horizontal Layout in the Cloud. Retrieved from https://www.researchgate.net/publication/269690552_An_Efficient_Cost_Model_for_Data_Storage_with_Horizontal_Layout_in_the_Cloud

[2] Donovan, Jim. "Does it Pay to Move from On-Premises to Public Cloud Storage?" Wasabi, 10 September 2020, https://wasabi.com/blog/on-premises-vs-cloud-storage/

[3] Duffin, Erin. "U.S. – Projected Inflation Rate 2008-2024." Statista, 7 May 2020, https://www.statista.com/statistics/244983/projected-inflation-rate-in-the-united-states/

[4] Florescu, D. and Kossmann, D. (2009). Rethinking Cost and Performance of Database Systems. Retrieved from https://dl.acm.org/doi/10.1145/1558334.1558339

[5] Islam, S. and Abedin, Z. (2013). Impacts of Data Mining on Relational Database Management System Centric Business Environments. Retrieved from https://www.researchgate.net/publication/276936082_Impacts_of_Data_Mining_on_Relational_Database_Management_System_Centric_Business_Environments

[6] Lang, W., and Patel, J. (2009). Towards Eco-friendly Database Management Systems. Retrieved from https://www.semanticscholar.org/paper/Towards-Eco-friendly-Database-Management-Systems-Lang-Patel/48aa5ae8f93680699e1a4d0d6d8815dcd16a64bd

[7] "Pricing." Google Cloud, 06 December 2020, https://cloud.google.com/bigquery/pricing

[8] Sentarl, I., Erdursun, A., and Caman, D. Development of a Database Management System Design Involving Quality Related Costs. Lund University Campus Helsingborg.

[9] Svensson, P., Podehl, M., Stephenson, G., and Cawson, M. (2006). Database Management Systems for Statistical and Scientific Applications: Are Commercially Available DBMS Good Enough? Retrieved from https://www.researchgate.net/publication/225112547_Database_management_systems_for_statistical_and_scientific_applications_Are_commercially_available_DBMS_good_enough

[10] Alam, M., and Shakil, K. (2015). Cloud Database Management System Architecture. Retrieved from: https://www.researchgate.net/publication/270791476_Cloud_Database_Management_System_Architecture

# 9. Appendices

| | |
|---|---|
| Storage hardware | $500,000 |
| Annual Maintenance | $500,000 |
| Ancillary hardware/software | $50,000 |
| Ancillary maintenance | $50,000 |
| Power and cooling | $120,000 |
| Headcount | $750,000 |
| **Total** | $1,970,000 |

**Figure 9.1:** The estimation of 1 PB of On-Premises Storage (Five Year Cost) in 2018 [2]

| Operation | Pricing | Details |
|---|---|---|
| Active Storage (per GB per month) | $0.023 | The first 10 GB is free each month |
| Long-term storage (per GB per month) | $0.016 | The first 10 GB is free each month |
| Streaming Inserts (per 200MB) | $0.01 | You are charged for rows that are successfully inserted. Individual rows are calculated using a 1 KB minimum size |
| Queries (on-demand) (per TB) | $5.00 | The first 1 TB per month is free. |

**Table 9.2**: Prices of cloud storage of Google Cloud by December 2020 (Los Angeles) [7]

| Operation | Details |
|---|---|
| Loading data | You are not charged for loading data from Cloud Storage or from local files into BigQuery. However, you are charged for storing data in Cloud Storage |
| Copying data | You are not charged for copying a table, but you do incur charges for storing the new table and the table you copied. |
| Exporting data | When you export data from BigQuery to Cloud Storage, you are not charged for the export operation, but you do incur charges for storing the data in Cloud Storage |
| Deleting datasets | You are not charged for deleting a dataset. |
| Deleting tables, views, partitions, and functions | You are not charged for deleting a table, deleting a view, deleting individual table partitions, or deleting a user-defined function. |
| Metadata operations | You are not charged for list, get, patch, update and delete calls. Examples include (but are not limited to): listing datasets, updating a dataset's access control list, updating a table's description, or listing user-defined functions in a dataset. |

**Table 9.3**: Some free operations of Google Cloud BigQuery by December 2020 [7]