Optimized Data Extraction Model

Project report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

In

Computer Science and Engineering

By:

Yash Bhardwaj

191397

Under the supervision of

Dr. Jagpreet Sidhu

Department of Computer Science & Engineering And

Information Technology

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT, SOLAN, HIMACHAL PRADESH – 173234

CERTIFICATE

This is to certify that the work which is being presented in the internship report titled **"Optimized Data Extraction Model"** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat, is an authentic record of work carried out by **Yash Bhardwaj** during the said period; February 2023 – till date, ensuring proper care towards the rules and regulations as specified by the Non-Disclosure Agreement signed between Yash Bhardwaj and Testbook, Navi Mumbai dated 13 february, 2023.

Yash Bhardwaj

191397

Jaypee University of Information Technology Waknaghat, Solan, H.P.

The above statement made is correct to the best of our knowledge.

Dr Jagpreet Sidhu (Assistant Professsor, SG)
Jaypee University Of Information Technology, Solan

ACKNOWLEDGEMENT

This is a matter of pleasure for me to acknowledge my deep sense of gratitude to my college, Jaypee University of Information Technology for giving me an opportunity to explore my abilities via this internship program. I would like to express my sincere gratitude to our Training and Placement officer, Mr. Pankaj Kumar and our faculty Coordinator, Dr. Nafis U Khan for this opportunity. I also wish to express my gratitude to my internship supervisors, for their valuable guidance and advice towards my internship.

I would like to record my sincere appreciation and gratitude towards all the officials, coaches, trainers, mentors and employees of Testbook., without whose kind assistance, my internship program would not have been proceeding in a swift direction. The facts and other vital information provided by them have contributed towards making this report as comprehensive as possible. I am indeed thankful to them.

Last but not the least, I would like to express my sincere thanks to all my family members, friends and well-wishers for their immense support and best wishes throughout the internship duration and the preparation of this report and I wish they would continue to contribute towards my well-being.

I believe that this report will be a valuable asset not only for academic institutions, but will also be useful for all those who are interested to learn about internship experiences in an Ed-Tech firm.

Yash Bhardwaj

191397

Jaypee University of Information Technology,

Waknaghat, Solan, H.P

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled "Optimized Data Extraction Model" in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering/Information Technology submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from July 2022 to May 2023 under the supervision of Dr Jagpreet Sidhu, Assistant Professor (SG). The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Yash Bhardwaj, 191397

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)

Supervisor Name: Dr Jagpreet Sidhu

Designation: Assistant Professor (SG)

Department name: CSE & IT

Plagiarism Certificate

As provided by LRC of JUI

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: .....

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: _____ **Department:** _____ **Enrolment No** _____

Contact No. _____ **E-mail.** _____

Name of the Supervisor: _____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): _____

_____

_____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at .....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature **Librarian**

.....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

Table Of Contents:

LIST OF FIGURES

ABSTRACT

An effective and precise framework for extracting data from many sources is the optimized data extraction model. It is a smart and creative method of data extraction that makes use of state-of-the-art algorithms and methods to locate and extract pertinent data swiftly and accurately. The concept can be significantly customized to satisfy the unique requirements of particular users or organizations. It is made to be extremely scalable and adaptable, and it is capable of processing massive volumes of data. The optimized data extraction model is a crucial tool for companies and organizations that must quickly, precisely, and efficiently extract data from numerous sources.

CHAPTER: 1 INTRODUCTION

1.1 Introduction

IIT graduates launched Testbook.com in January 2014 to help students pass government and public sector exams. Ashutosh Kumar, Praveen Agrawal, Arpit Oswal, Narendra Agrawal, and Manoj Munna are alumni of IIT Bombay and make up the core team. An educational innovation company with its headquarters in Navi Mumbai is called Testbook. It enables students to get ready for entrance examinations for the government and banking, such as Bank Clerk, SSC, Bank PO, Bank SO, Defence examinations, Railway RRB, Insurance, and so on.

I wish to study for government examinations, but "Why is government exam preparation so expensive? ", "Why is no defined path available easily?" before deciding to become an aspirant, everyone asked. The employees of the company and Testbook.com have been working to make things simple for everyone. A Mumbai-based business called Testbook educates people getting ready for government exams. Regardless of whether you use a home computer or a mobile device, the company offers a platform that is user-friendly for everyone. Even a few cities have offline operations centers for the platform.

The main component of academics that gives students the practical experience they've been wanting for the last three years of college life is internships. This opportunity was provided to me by Testbook.com, who gave me a Data Analyst Internship.

The main objective of this internship is to prepare us for the workplace. Tasks are given to us; initially, the majority of them involved fixing issues. The work complexity rises in accordance with the intern's performance chart, which also allows the interns to track their development. The corporate culture consistently gave interns a sense of belonging. Daily work includes regular scrum meetings, which are followed by team meetings. I was given a data analyst assignment, which required me to think about the product and user experience in addition to an analyst logic so that every visitor to the website would feel comfortable using the platform.

The most popular and best current affairs application is daily current affairs GK quiz. For Hindi and English tests for the Bank, SSC, Railways, Civil Services, and MBA, this method is preferred in India. This useful software motivates you to get ready in Hindi and English. All of your daily notes and tests will be provided in both languages, just like on test day! This programme encourages you to focus your efforts on only the most important current affairs updates by providing you with a side-by-side summary of the day's most important events and updates in less than 100 words overall.

Nearly 75% of applicants for state jobs come from Tier II, III, and IV cities, so Testbook's founders Narendra Agrawal and Ashutosh Kumar purposefully created a platform that enables applicants to prepare for and then pass fictitious exams for state employment selection tests. In light of this, the company introduced the "Testbook Pass" gift card in March 2016 to enable students to pay with real money. It functioned much like a resurrection voucher that could be bought with cash and used on the website to offer their examinations.

The foundational problem was the next one. Nearly 80% of the researchers preparing for state exams are unable to afford a PC, so the majority go to digital bistros. At that time, the other donors created PC laboratories that were Testbook-marked and provided contestants with a space where they could practice their examinations under actual test conditions. The company moved 250 understudies to the crucial Patna center in the first three months alone.

1.1 Testbook Logo

Testbook has received a series of investments from different financial experts. The business received a $250,000 seed investment from Shankar Narayanan, Utsav Somani, LetsVenture, and Ah! Adventures in October 2014. After that, in March 2016, Testbook received an undisclosed amount of Series A investment from S. Chand Group, India's most established and largest distribution and training management company. After that, Testbook received an undisclosed sum from Matrix Partners in March 2017.

Products:

TestBook Edu is a product-based organization that offers a few key items to its consumers throughout India to assist them in preparing for government exams utilizing the greatest technology currently on the market.

A dedicated team of seasoned educators and content writers at Testbook works nonstop to produce excellent study guides and practice tests for various exams. In order to guarantee that students have access to the most recent and pertinent study materials, the organization has also partnered with a number of top publishers and content providers.

Testbook has created a mobile app that enables users to access its services while on the road in addition to its online classes and study materials. Students who have tried the software give it rave ratings and it is available for both Android and iOS devices.

Testbook is a fantastic illustration of how technology can be utilized to raise the standard and accessibility of education. The business has quickly emerged as a dominant force in the Indian ed-tech market and is well-positioned for future expansion.

A variety of items are available from Testbook to aid students in getting ready for challenging exams. The following are some of the main items that Testbook sells:

1. **Testbook Super Coaching** is a premium coaching service that gives students access to individualized instruction and guidance from knowledgeable teachers. To assist students in achieving their academic objectives, this coaching service offers individualized study schedules, doubt-clearing sessions, and one-on-one coaching.

2. **Testbook Pass**: For a full year, students who subscribe to this programme have unrestricted access to all Testbook courses and practice exams. Students can study for several competitive exams with Testbook Pass without needing to buy additional courses or practice exams.

3. **Testbook Pass Pro** is a premium subscription service that offers students extra advantages over and above those offered by Testbook Pass. Testbook Pass Pro users gain access to individualized mentoring from knowledgeable educators, priority support, and exclusive study materials in addition to unlimited access to all courses and mock exams.

4. **Testbook Skill Academy** is a collection of online courses created to aid students in learning new skills that are practical in the job. These courses include a wide range of subjects, such as programming, data analytics, and digital marketing.

5. **Daily current affairs** updates are also available from Testbook via its website and mobile app. These current affairs updates are intended to assist students stay current with current affairs, which is a crucial component of many competitive exams. They cover the most recent news and events from across the world.

In general, the services offered by Testbook are created to give students a comprehensive educational experience that goes beyond standard exam preparation. These services are intended to aid students in developing new abilities and knowledge that they may use in the job and in keeping up with current events, which are an important part of many competitive exams.

The above products of Testbook includes features such as:

1. Online Courses: For a variety of competitive tests, including SSC, Banking, Railways, Defence, and Teaching exams, Testbook offers extensive online courses created by subject matter experts. Practice questions, mock exams, and video lectures are all included in these courses.

2. Mock Exams: Testbook provides a huge collection of practice exams for a variety of competitive exams. These practice exams are intended to assist students better understand their strengths and shortcomings by simulating the experience of taking an exam.

3. Study Resources: Testbook offers top-notch study resources online as well as in the form of PDFs and ebooks. These study tools are intended to give students a thorough understanding of the exam syllabus and aid in efficient preparation.

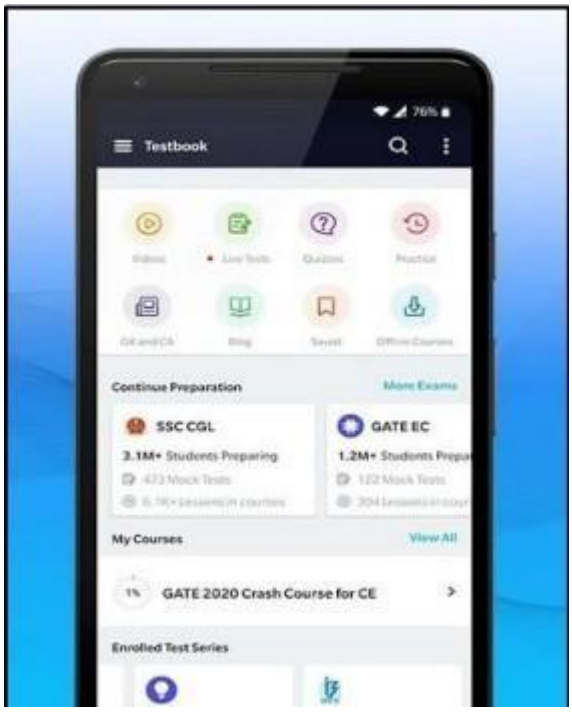
4. Live Classes: Testbook provides live classes so that students may speak with knowledgeable instructors and have their questions answered. These live classes give students a classroom-like experience because they are held in real-time.

5. Test Series: Testbook offers test series, which feature a number of practice questions and mock exams, for a number of competitive exams. These test series are made to aid students in increasing their response times and precision when answering exam questions.

Testbook platform available in the market:

1.2 Website:

1.3 App:



1.4 Current Affairs App:

1.2 Problem Statement

Basically as a data analyst intern my work involves examination of data to find significant customer insights and potential users for the information. We also advise the company's management and other stakeholders of this information.

Therefore, the goal of my project is to efficiently build a model to extract the required user data as the conditions provided, from the company database in order to facilitate hiring of new employees in accordance with the demands of the company's human resource department.

1.3 Objectives

However, the architecture created to carry out the task can be used by other people who want to automate the data for further processing while utilizing the resources. The full problem statement is specific for the optimisation of the platform and the resource used. The company's primary goal is to improve the system's usability and efficiency. The secret to Testbook's recent success and growth is that it approaches change from the very beginning, at the most fundamental level.

The necessary model must meet the following requirements:

1. Minimum period
2. Constant Data Access
3. Captivating Insights
4. An improved user interface
5. To integrate the dashboard with advanced customer relationship tools.

1.4 Methodology

This model will revolve around a number of different approaches:

1.Redash Query Methodology

2. Python Script Method

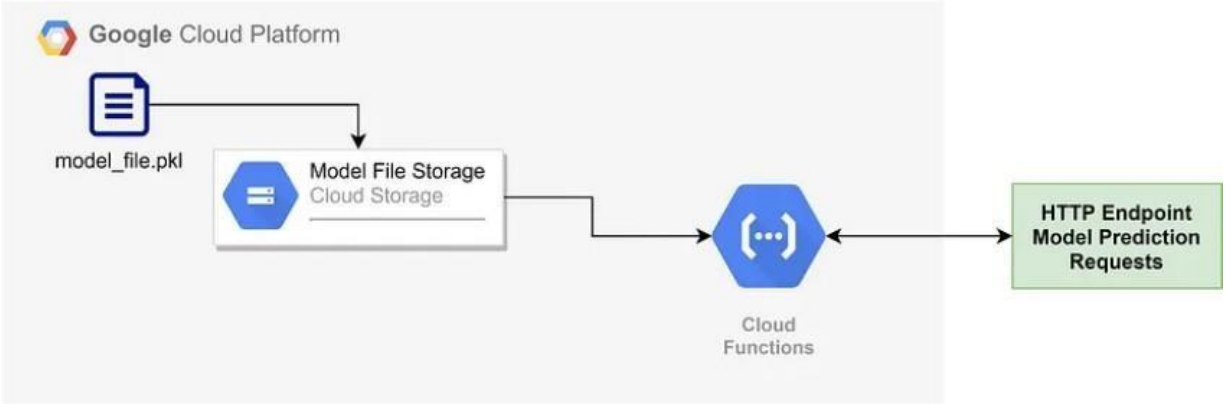
3. The Google Cloud Service

CHAPTER 2: LITERARY SURVEY

The study on the cloud function journal by Vijay Kumar explains a serverless execution environment for constructing and linking cloud services is called Cloud Functions. You can create straightforward, one-purpose functions with Cloud Functions that are linked to events released by your cloud infrastructure and services. When an event being watched fires, your function is called. In a completely managed environment, your code runs. Neither server management nor infrastructure provisioning are required. On the Google Cloud Platform, cloud functions can be created using the JavaScript, Python 3, Go, or Java runtimes. It's simple to test locally and port your function to any standard Node.js (Node.js 10 or 12), Python 3 (Python 3.7 or 3.8), Go (Go 1.11 or 1.13), or Java (Java 11) environment.

You can create code to connect and augment cloud services using Cloud Functions, which offers a connective layer of logic. A log change, a file upload to cloud storage, or an incoming message on a Pub/Sub topic can all be heard and handled. With the addition of Cloud Functions, you may address an expanding range of use cases with arbitrary programming logic. The majority of Google Cloud services, including Cloud Vision and many others, effortlessly authenticate with Cloud Functions since they have access to the Google Service Account credential.

There are no servers to install, maintain, patch, or update with Cloud Functions. Features are fault resilient, highly available, and automatically scaled. Building serverless backends, analyzing real-time data, and creating intelligent apps are all made possible by cloud capabilities. Events that occur in your cloud environment are referred to as cloud events. These could include things like updates to database data, the addition of files to a storage system, or the creation of a new virtual machine instance. Whether you choose to react to certain events or not is up to you. With a trigger, you can produce a response to an event. Declaring your interest in a specific event or series of events is known as a trigger. You can record occurrences and take appropriate action by tying a function to a trigger.



Serverless Machine Learning Model HTTP Endpoint using Google Cloud Functions

2.1 GCF Flow

Based on Google's Cloud Functions, this FAAS gives an overview of cloud functions. Environment variables, configuration types, and trigger types can all be defined in functions before they are executed on a serverless platform.

A. Simple Steps

Select the Create button.
Give your job a name.
Choose HTTP in the Trigger field.
Select Allow unauthenticated invocations in the Authentication field.
After making your changes, click Save and then Next.
Choose the Inline editor option in the Source code field. You will employ the editor's default function for this exercise.
Choose the desired Node.js runtime using the Runtime selection.
Click Deploy at the bottom of the page.
Click Test the function on the testing page to begin.

B. Method

Establish environment variables for relevant data, such as API details, database connections, date-and-time verification, etc.

Include the proper dependencies in the package.json

Create code for data access:

a. Provide the API connection credentials.

b. Use filters to gather data.

b. Correctly format the output.

d. Add the data to the database.

e. Disconnect from the connections/pool

4. Verify the database's data. Make the time to accomplish this.

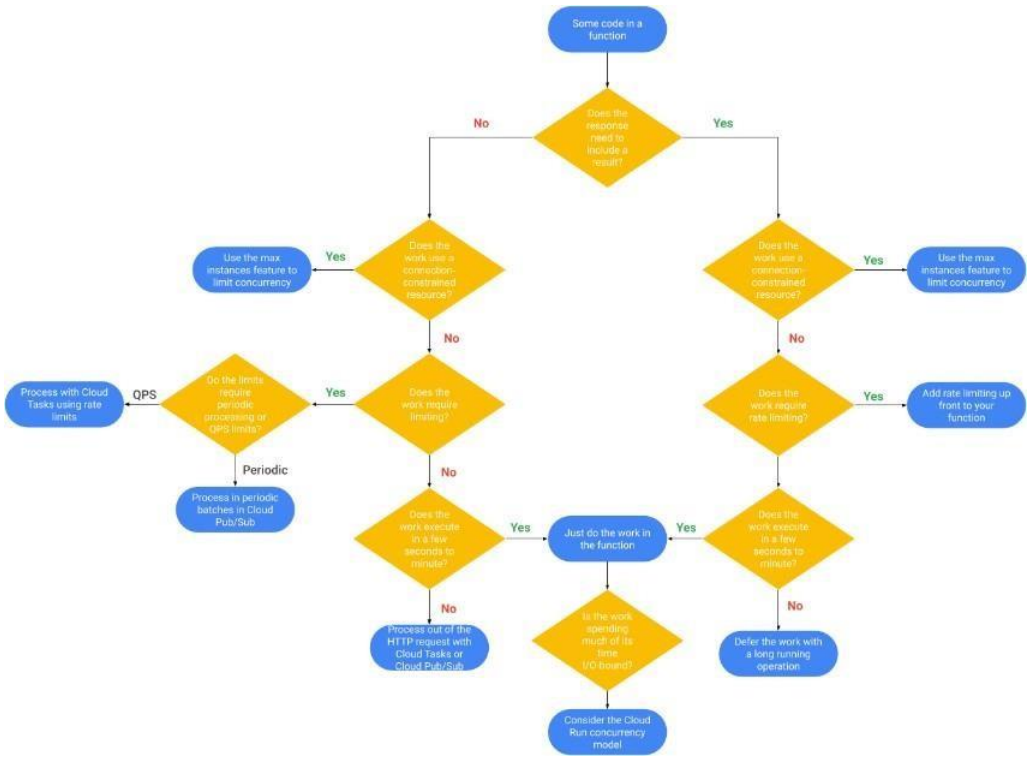
C. Final Verdict

You can run your code using scalable pay-as-you-go functions as a service (FaaS) with no server management.

There are no servers to install, maintain, or update.

Integrated monitoring, logging, and debugging capability Automatic scaling based on load

built-in security based on the idea of least privilege at the role and per function level.



2.2 Flow Chart Of Google Cloud Function

The study conducted by Khushboo Vachiyatwala and Rupal Snehkunj focuses on using Python's Pandas library. This comprehensive library offers numerous integrated support for data analysis. It is helpful for grouping searches and designing tabular data graphically. This module serves as the structural framework for statistical data processing in Python using various Pandas API. The research is done using a structured data set file that can access formats like xls, csv, pdf, and many others. The project uses a randomly generated personnel database for various functions and Python's pandas package for data visualization.

To work with data structures effectively, the Python panda is utilized. It supports the NumPy and Matplotlib libraries. The Matplotlib library is used for graphic performance, while the NumPy library is exclusively used for numerical array performance. The Pandas library offers a better framework for statistical computing and data analysis. Additionally, the Pandas library supports SQL tools for data manipulation, including the merging of

data using different joins (inner, right, and left). Panel data analytics, which tries to offer comparable capabilities and has implemented several features including automatic data alignment and hierarchical indexing, is where the name of the Pandas library originated.

The Pandas library is a contemporary, high-level, object-oriented programming toolkit that includes a wide range of add-on packages.

Numpy, a built-in library in Pandas, is used for numerical data.

Numpy uses the array data type for operations like sorting, indexing, and reshaping. Because array data types support homogenous data, Numpy does. Additionally, Pandas supports the Matplotlib library and lets users save files in a variety of formats, including excel, csv, and json.

Three types of data structures are primarily supported by Pandas: (1) Series (2) Data Frame (3) Panel. One-dimensional array is a component of the series data structure. It is compatible with homogeneous data types. Two-dimensional arrays are included in Data Frame data structures, which also accept heterogeneous types of data and can be used to scale and modify data.

There is a three-dimensional array on the panel.

There are various methods involved using Pandas discussed in this paper:

1. File access in pandas
2. Graphics using panda’s library
3. API with pandas
4. Merge with pandas

Conclusion:

This study concentrated on the Python Pandas library's numerous features. This library serves as the structural base for statistical computing and data analysis. The Pandas library provides data structures and methods for working with time series and mathematical tables. The study used manually compiled structured data sets that were plotted, examined, and visualized in Python using several pandas tools.

The paper of “Python Using database and SQL” by Shweta J. Patil focuses on The database is a collection of structured data that is simple to use, maintain, and update. The data are categorized based on how they are organized. In that they map from keys to values, most databases are structured similarly to dictionaries. Similar to a dictionary, database software is made to insert and access vast volumes of data exceedingly quick.The performance of database software is maintained by creating indexes as new data is added to the database so that the computer can access a specific entry rapidly.

Oracle, MySQL, Microsoft SQL Server, PostgreSQL, and SQLite are just a few examples of the many diverse database systems that are used for a wide range of reasons.

You must first create a connection between your Python file and the database. Your database can then be added to, searched, deleted from, or updated. Additionally, you can take data out of the database, alter it, and then put it back in again. SQL statements are used to carry out database operations[1]. In this essay, we first go over the resources that are available and the fundamentals of database modules that are known to be utilized with Python.

Each review is accompanied by a sample Python programme that demonstrates how to use the module in question. A description of how to connect Python to a database is given in the second section. A brief review of the fundamental SQL statements is given in the third section. Python is used to carry out the primary database operations in the fourth segment.

Python has an easy-to-use API that supports working with databases. Python comes with modules for SQLite and Berkeley DB, among others. There are third-party modules available for MySQL, PostgreSQL, FirebirdSQL, and other databases.

Before use, the latter must be downloaded and installed. For instance, the debian package "python mysqldb" can be used to install the package MySQLdb.a some of the supported databases:

- GadFly
- MySQL
- PostgreSQL
- SQLite
- Oracle

Within a database is a table. For MySQL, this is particularly true.

A database must initially be constructed or at the very least be present in order to build a table. Therefore, a connection to the database is required in order to retrieve data from a table. Utilizing the connect() method accomplishes this. In other words, connect serves as the phpMyAdmin's constructor. The specifications are as follows:

The host is the identifier for the computer operating the MySQL server. An IP address or a name could be used. Localhost is used as the default value if no value is given.

The user id, which needs to be verified, is user. In other words, this is the legitimate ID for utilizing Server services. The current effective user is the default value. Usually, it is either "nobody" or "root."

Password -- A user is authenticated by the MySQL server (or any server, for that matter) using a user id and password combination. There are no passwords set as the default. This implies that the string for this argument is null.

Once a connection has been made with the server, the database called db must be used. The connection is useless, though, if the database that will be utilized is not chosen. This option has no default setting.

Conclusions

While working on the project, I made an effort to evaluate each database server in search of the best one. After careful consideration, MySQL Server is selected since it fulfils several of the 14 necessary qualities for Python implementation.

One of the most well-known advanced programming languages is Python, which owes a lot to both its own inherent expressiveness and the variety of support modules that help extend its advantages. As a result, Python is a perfect fit for creating a reliable connection between the programme and the database.

CHAPTER 3: SYSTEM DEVELOPMENT

Using different analytical approaches we have:

There are three approaches for this project model that I used in the course of finishing the project:

1. Redash Query Methodology
2. Python Script Method
3. The Google Cloud Service

Hardware/Software Requirements For this Project Completion:

Redash:

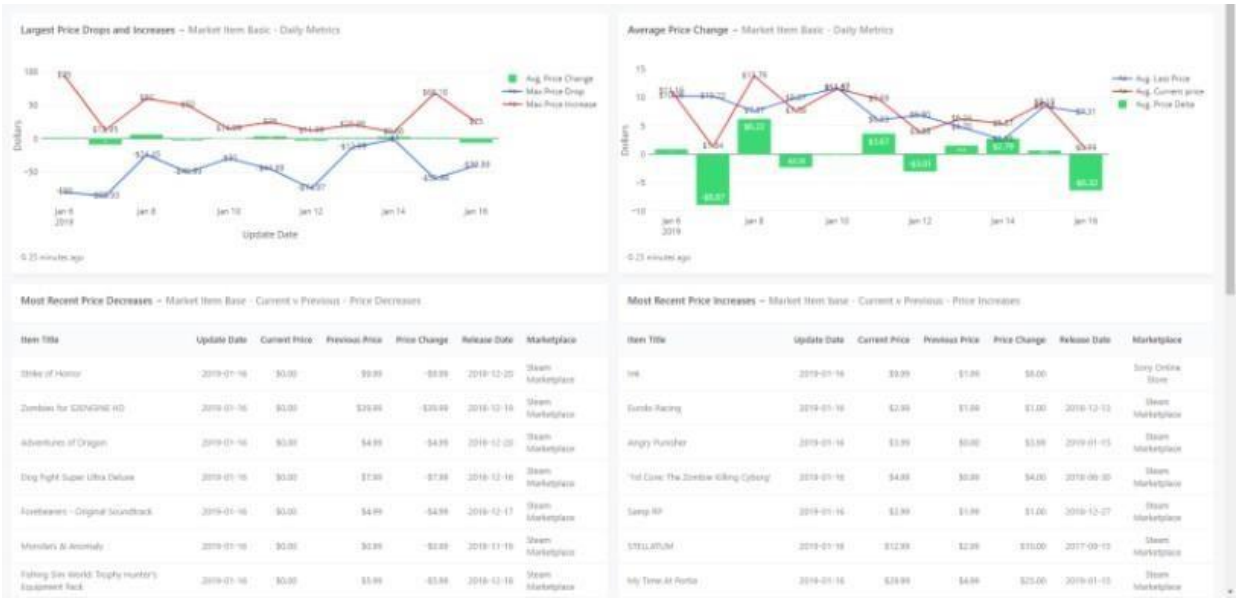
Redash is a collaborative and open-source platform for data visualization that enables users to connect with different data sources, create and share dashboards, and make data-driven decisions. Its objective is to provide users with an easy way to access, query, and visualize data in a collaborative environment.

Redash users can access various data sources such as relational and NoSQL databases, cloud storage services, and API services. They can create interactive dashboards, reports, and visualizations using a user-friendly web-based interface, with support for different visualization types such as bar charts, line charts, and scatter plots.

In addition, Redash includes a powerful query editor with support for SQL and other query languages to create complex queries for building dashboards and visualizations. It offers features that enable easy collaboration on data analysis, including sharing dashboards and reports with other users, commenting and annotating on visualizations, and setting up alerts based on specific data criteria.

Redash is open-source, meaning it can be customized and extended to suit specific user needs, and it has an active community of users and contributors that provides support and new feature integrations.

Overall, Redash is a robust and flexible data visualization and collaboration platform suitable for organizations of any size that require data-driven decision-making and enhanced data analysis capabilities.



3.1 View Of Redash

MongoDB:

MongoDB is an open-source NoSQL database that stores data in flexible, JSON-like documents. Its design enables it to handle large volumes of data with high scalability, availability, and performance. MongoDB also allows users to store unstructured data and offers dynamic schema capabilities, meaning that data can be added, modified, and deleted without having to restructure the database.

The database provides a flexible and robust query language that supports a wide range of queries, including ad-hoc queries, range queries, and full-text searches. MongoDB also comes with a built-in aggregation framework that allows users to analyze and summarize large volumes of data in real-time.

MongoDB is particularly suitable for modern web and mobile applications due to its features, which make it easy to develop and deploy applications quickly. It supports multiple programming languages and platforms and comes with a comprehensive set of drivers and APIs, making it easy to integrate with other tools and services.

Moreover, MongoDB offers enterprise-grade security features, such as role-based access control, encrypted storage, and network isolation. These features make it a reliable and secure option for storing sensitive data.

In conclusion, MongoDB is a flexible, scalable, and high-performing NoSQL database that can handle large volumes of data. It is an excellent choice for organizations of all sizes that require a fast and flexible database solution, thanks to its powerful query language, dynamic schema, and real-time aggregation capabilities.

MongoDB Syntax:

```
db.demo.group({
    "key": {
        "person": true
    },
    "initial": {
        "sumscore": 0,
        "sumforaverageaveragescore": 0,
        "countforaverageaveragescore": 0,
        "countstar": 0
    },
    "reduce": function(obj, prev) {
        prev.sumscore = prev.sumscore + obj.score - 0;
        prev.sumforaverageaveragescore += obj.score;
        prev.countforaverageaveragescore++;
        prev.minimumvaluescore = isNaN(prev.minimumvaluescore) ? obj.score : Math.min(pr
        prev.maximumvaluescore = isNaN(prev.maximumvaluescore) ? obj.score : Math.max(pr
        if (true != null) if (true instanceof Array) prev.countstar += true.length;
        else prev.countstar++;
    },
    "finalize": function(prev) {
        prev.averagescore = prev.sumforaverageaveragescore / prev.countforaverageaverage
        delete prev.sumforaverageaveragescore;
        delete prev.countforaverageaveragescore;
    },
    "cond": {
        "score": {
            "$gt": 0
        },
        "person": {
            "$in": ["bob", "jake"]
        }
    }
});
```

3.2 Example Of MongoDB Query

SQL:

SQL, which stands for Structured Query Language, is a widely-used programming language for managing and manipulating relational databases. This language is used to

create, modify, and query databases in various organizations. SQL is designed to work with relational databases, which are databases that organize data into tables with defined relationships between them.

With SQL, users can carry out several operations on the data stored in a database. Such operations include creating new tables, inserting data into tables, updating existing data, and querying data to retrieve specific information. SQL supports a wide range of data types, including text, numeric, date and time, and Boolean values, among others.

SQL has many features that make it a powerful language for working with data, including support for transaction management, data integrity, and security. This programming language is also highly standardized, which means that code written in SQL can be easily ported between different database systems.

Overall, SQL is a powerful and flexible language used for managing and manipulating relational databases. It is widely used in various industries and plays a vital role in data management and analysis.

Python:

As an interpreted, high-level, and general-purpose programming language, Python is designed to emphasize code readability and ease of use. It is a popular language for a wide range of applications, including web development, data analysis, machine learning, artificial intelligence, and more.

One of the key strengths of Python is its simplicity and ease of use. Its syntax is designed to be easy to read and understand, which makes it a popular choice for beginners and experts alike. Python also has a large and active community of developers who contribute to its development and support a wide range of libraries and frameworks.

Python has a wide range of built-in data types and structures, including lists, tuples, dictionaries, and sets, which makes it easy to work with data. It also supports object-oriented programming and functional programming paradigms, giving developers flexibility in how they structure and organize their code.

Python is known for its vast library of modules and packages, which can be easily imported and used in projects. This makes it easy to add functionality to a project without

having to write all the code from scratch. Some of the most popular Python libraries include NumPy, Pandas, Matplotlib, and Scikit-learn, which are widely used in data analysis and machine learning applications.

Overall, Python is a powerful and versatile programming language that is widely used in various industries and applications. Its simplicity, ease of use, and vast library of modules and packages make it a popular choice for developers of all levels of experience.

Pandas:

Pandas is a popular open-source data manipulation and analysis library for the Python programming language. It provides easy-to-use data structures and data analysis tools for handling structured data. Pandas offers two main classes for data manipulation: DataFrame and Series.

A DataFrame is a two-dimensional table that stores data in rows and columns, similar to a spreadsheet. It is a powerful tool for data manipulation, as it allows users to filter, group, and aggregate data easily. A Series, on the other hand, is a one-dimensional labeled array that can hold data of any type. It can be thought of as a column in a DataFrame.

Pandas offers a wide range of data manipulation and analysis functions, such as merging and joining datasets, reshaping and pivoting data, and handling missing values. It also supports data visualization, making it easy to create plots and charts to explore and communicate data.

Pandas is widely used in data science and analytics, as it allows users to work with large datasets quickly and efficiently. It is also compatible with many other data analysis tools and libraries in the Python ecosystem, such as NumPy, SciPy, and Scikit-learn.

Overall, Pandas is a powerful and flexible data manipulation and analysis library for Python. Its easy-to-use data structures and data analysis tools make it a popular choice for data scientists, analysts, and developers working with structured data.

NumPy:

NumPy is a popular Python library used for numerical computing and data analysis. It provides a powerful array and matrix processing functionality that allows users to perform mathematical operations on large sets of data with high efficiency and

performance. NumPy provides a range of mathematical functions, including linear algebra, Fourier transform, and random number generation, making it a versatile library for scientific computing.

One of the key features of NumPy is its powerful multi-dimensional array object, called ‘n’D array, which provides efficient storage and manipulation of large arrays. The ‘n’ D array object provides a number of methods for performing mathematical operations on arrays, such as element-wise operations, matrix multiplication, and linear algebra functions.

NumPy also provides tools for integrating with other popular Python libraries, such as SciPy, Pandas, and Matplotlib, making it a valuable tool for scientific computing and data analysis.

Overall, NumPy is a versatile library for numerical computing and data analysis in Python. Its powerful array processing capabilities, efficient storage and manipulation of large arrays, and integration with other scientific Python libraries make it a popular choice for data scientists and researchers.

PyMongo:

Pymongo is a Python library designed to interact with MongoDB databases through the Python programming language. It offers developers an interface to perform various operations such as inserting, updating, deleting, and querying documents. Pymongo supports all of MongoDB's features, including its query language and advanced aggregation framework. Additionally, it provides features such as indexing and replication, making it a reliable choice for creating large-scale applications.

Pymongo is known for its simplicity and ease of use. Its API is clear and concise, making it easy for developers to get started with MongoDB even without prior experience with the database. Pymongo can be seamlessly integrated with other Python libraries, such as NumPy and Pandas, making it ideal for processing and analyzing data.

Advanced features such as sharding and transactions are also supported by Pymongo, enabling developers to build highly scalable and fault-tolerant applications. It includes robust error handling and detailed logging for easy debugging and issue diagnosis.

Overall, Pymongo is a powerful and flexible library that offers developers the ability to create scalable and reliable applications with MongoDB. Its ease of use, advanced features, and integration with Python make it a popular choice for working with MongoDB databases.

Anaconda:

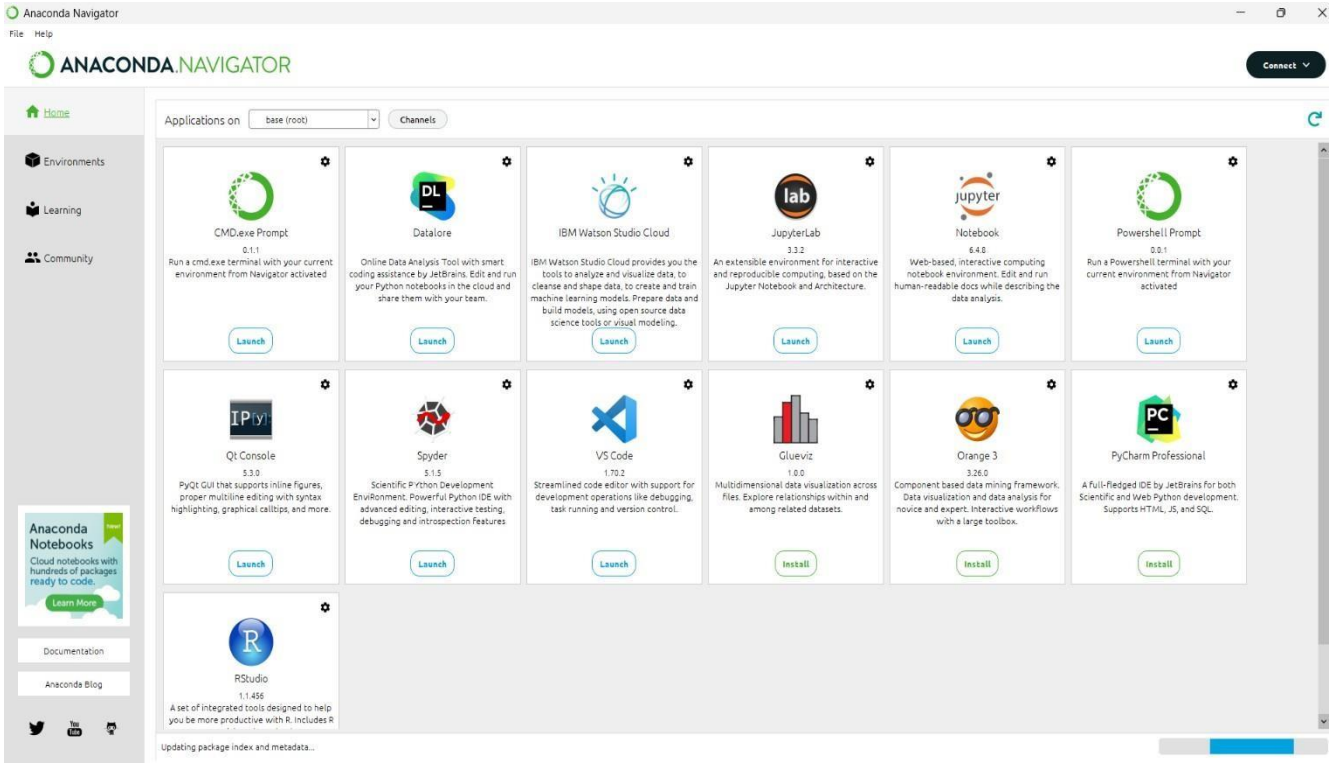
Anaconda is an open-source software distribution that is used for data science, machine learning, and scientific computing. It provides a platform that streamlines the process of installing, managing, and updating Python and R packages and their dependencies, as well as the popular data science libraries such as Pandas, NumPy, and Matplotlib.

Anaconda also includes its own package manager, Conda, which is a cross-platform package manager that enables users to easily install, update, and remove packages and their dependencies. Conda also allows users to create and manage virtual environments, which enables them to isolate their development environments and maintain consistent dependencies across different projects.

In addition, Anaconda includes a range of tools and applications that are used for data science and machine learning, such as Jupyter Notebooks, Spyder, and RStudio. These tools provide a range of capabilities, such as data visualization, data exploration, statistical analysis, and machine learning.

Another advantage of Anaconda is its cross-platform compatibility. It is available for Windows, macOS, and Linux operating systems, and it provides a consistent development environment across all platforms.

Overall, Anaconda is a powerful and flexible platform that provides a streamlined way to manage Python and R packages and their dependencies, as well as a range of tools and applications that are used for data science and machine learning. Its ease of use, cross-platform compatibility, and rich set of features make it a popular choice for data scientists and developers.



3.3 Anaconda Interface

Jupyter Notebook:

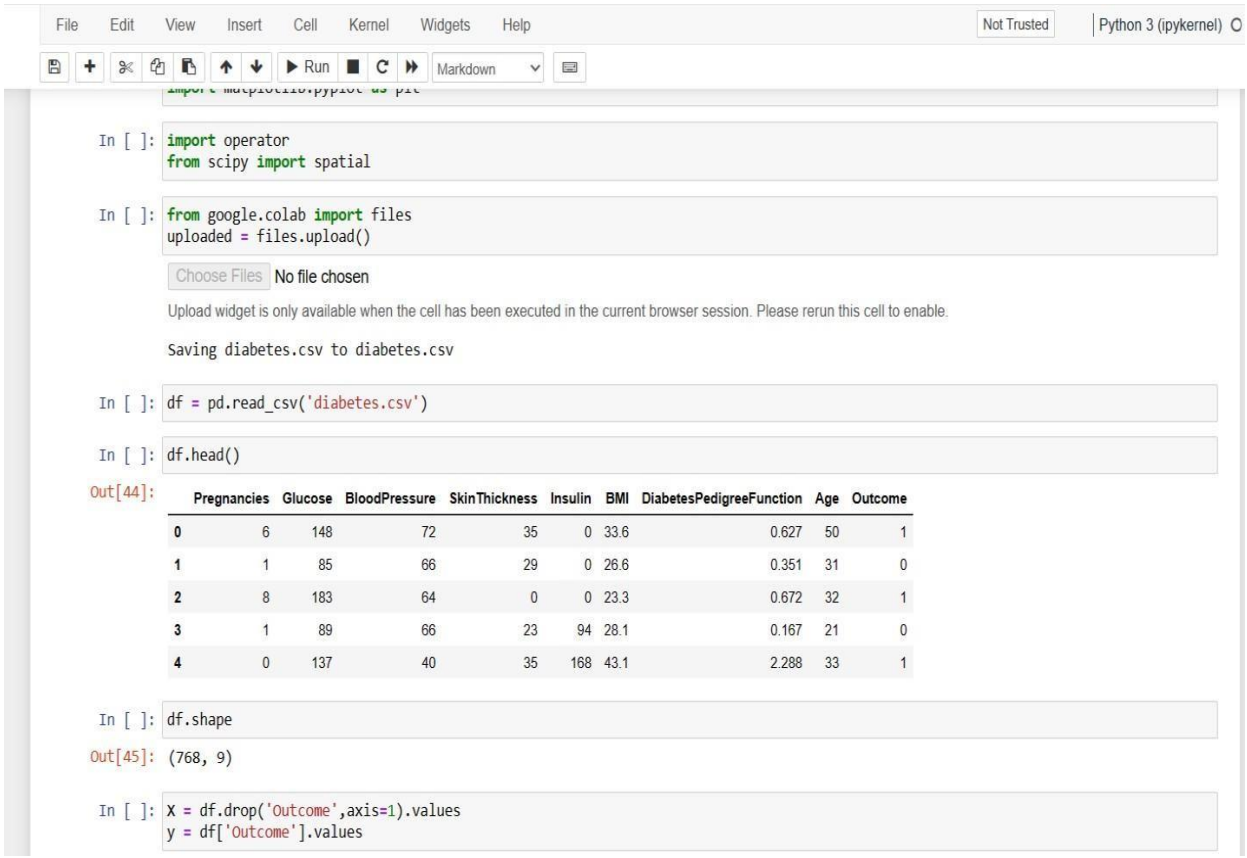
Jupyter Notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations, and narrative text. It is a popular tool among data scientists and researchers for interactive data exploration, analysis, and visualization.

The Jupyter Notebook is based on the IPython kernel and supports a wide range of programming languages, including Python, R, and Julia. Users can write and execute code in a cell-by-cell fashion, which makes it easy to test and iterate on code. Additionally, Jupyter Notebook allows users to mix code, text, and visualizations in a single document, which enhances the clarity and reproducibility of their work.

Jupyter Notebook also provides a rich set of features for data exploration and visualization, including support for interactive widgets, 3D visualizations, and rich media. The platform also supports a variety of output formats, including HTML, PDF, and LaTeX, which makes it easy to share and publish notebooks.

One of the key advantages of Jupyter Notebook is its open-source nature, which allows users to customize and extend the platform to meet their specific needs. It also has a large and active community of contributors and users, who provide support and contribute to the development of new features and integrations.

Overall, Jupyter Notebook is a powerful and flexible tool for interactive data exploration, analysis, and visualization that is widely used in the data science community. Its support for multiple programming languages, rich set of features, and open-source nature make it an ideal choice for data scientists and researchers who require a flexible and customizable platform for their work.



3.4 Jupyter Notebook Interface

MongoDB Connection Establishment With Python Sheet:

Python can establish a connection with MongoDB using the PyMongo library, which allows developers to interact with MongoDB databases from Python. PyMongo enables

developers to perform basic database operations, such as inserting, updating, deleting, and querying documents, using the Python programming language.

To connect Python to MongoDB using PyMongo, developers must first install PyMongo and then import it into their Python script. Once imported, they can create a connection to the MongoDB server by specifying the host and port number.

Developers can then create a reference to a specific database and collection within the database using PyMongo's API. From there, they can perform various database operations, such as inserting new documents, updating existing documents, deleting documents, and querying the database to retrieve specific data.

Overall, PyMongo provides a flexible and powerful way to connect to MongoDB databases from Python and is widely used by developers for building applications that use MongoDB for storing and managing data.

```
#Change Client location according to your machine setup
import pymongo
from pymongo import MongoClient
client_maindb  = MongoClient("")
db_main =  client_maindb.tb_dev
import numpy as np
from datetime import datetime
import datetime as dt
import pandas as pd
from bson.objectid import ObjectId
client_test  = MongoClient("")
db_test =  client_test.tb_dev
client_engage  = MongoClient("")
db_engage =  client_engage.tb_dev
client_combine  = MongoClient("")
db_combine =  client_combine.tb_dev
```

3.5 Python Connection Establishment With MongoDb

Google Cloud Function (GCF):

Google Cloud Functions is a serverless computing platform that allows developers to run code in response to various events without having to manage any underlying infrastructure. With Cloud Functions, developers can write code in a variety of programming languages, including Python, Node.js, and Go, and deploy that code to the cloud with a simple command.

Cloud Functions are designed to be highly scalable and can be triggered by a wide range of events, including HTTP requests, changes to Cloud Storage buckets, and events from other Google Cloud services such as Pub/Sub and Firestore. When an event is triggered, the corresponding Cloud Function automatically runs, executes the developer's code, and then shuts down when the task is complete.

One of the key advantages of using Cloud Functions is that developers only pay for the exact amount of computing time used, rather than having to pay for a fixed amount of server time regardless of usage. This makes Cloud Functions a cost-effective option for running small, intermittent tasks and allows developers to focus on writing code rather than managing infrastructure.

Cloud Functions also integrates seamlessly with other Google Cloud services, such as Cloud Pub/Sub and Firebase, making it easy to build complex applications that rely on multiple services working together. With its ease of use, scalability, and cost-effectiveness, Cloud Functions is a popular choice for developers looking to build serverless applications on the Google Cloud platform.

BigQuery:

BigQuery is a Google Cloud Platform cloud-based data warehouse that allows users to analyze and query big datasets in real-time using SQL-like queries. Many organizations use it for business intelligence, data analysis, and machine learning. It is built to handle petabyte-scale data.

BigQuery's capacity to quickly process massive amounts of data is one of its key advantages. Users do not have to worry about infrastructure setup or maintenance because it is entirely managed. BigQuery is also extremely scalable, enabling customers to start with a modest dataset and quickly grow up to handle larger datasets as required.

BigQuery integrates with several other Google Cloud services, including Google Cloud Storage and Google Cloud Dataflow, and supports a wide range of data formats. Additionally, it offers a wide range of sophisticated features for data analysis and visualization, including data connectors for well-liked BI tools like Tableau and Looker and machine learning models.

BigQuery is a robust and adaptable data warehousing system that may assist businesses in swiftly and easily processing and analyzing enormous amounts of data.

Why is Google Cloud Function Preferred Over AWS ?

In comparison to AWS, Google Cloud Functions has a number of benefits:

1. Pricing: Compared to AWS, Google Cloud Functions has a more reasonable pricing structure. While AWS Lambda charges you for the entire time your function runs, regardless of the amount of resources used, Google Cloud Functions only charges you for the exact amount of resources you use.

2. Better Google service integration: BigQuery, Cloud Storage, and Cloud Pub/Sub are just a few of the Google Cloud services that Google Cloud Functions interfaces with without any hiccups. As a result, creating and deploying serverless applications that use these services is made simpler.

3. Quicker startup times: Functions can be executed more quickly because to Google Cloud Functions' quicker startup times compared to AWS Lambda. This is particularly crucial for applications that demand quick responses.

4. Greater language support: Compared to AWS Lambda, Google Cloud Functions offers more language support. Google Cloud Functions also supports Java,.NET, and Ruby in addition to Node.js, Python, and Go.

5. superior tooling: Compared to AWS Lambda, Google Cloud Functions offers superior developer experience and tooling. For managing functions, the Google Cloud Console offers a simpler user interface, and the Cloud Functions Emulator enables developers to test functions locally before deploying them.

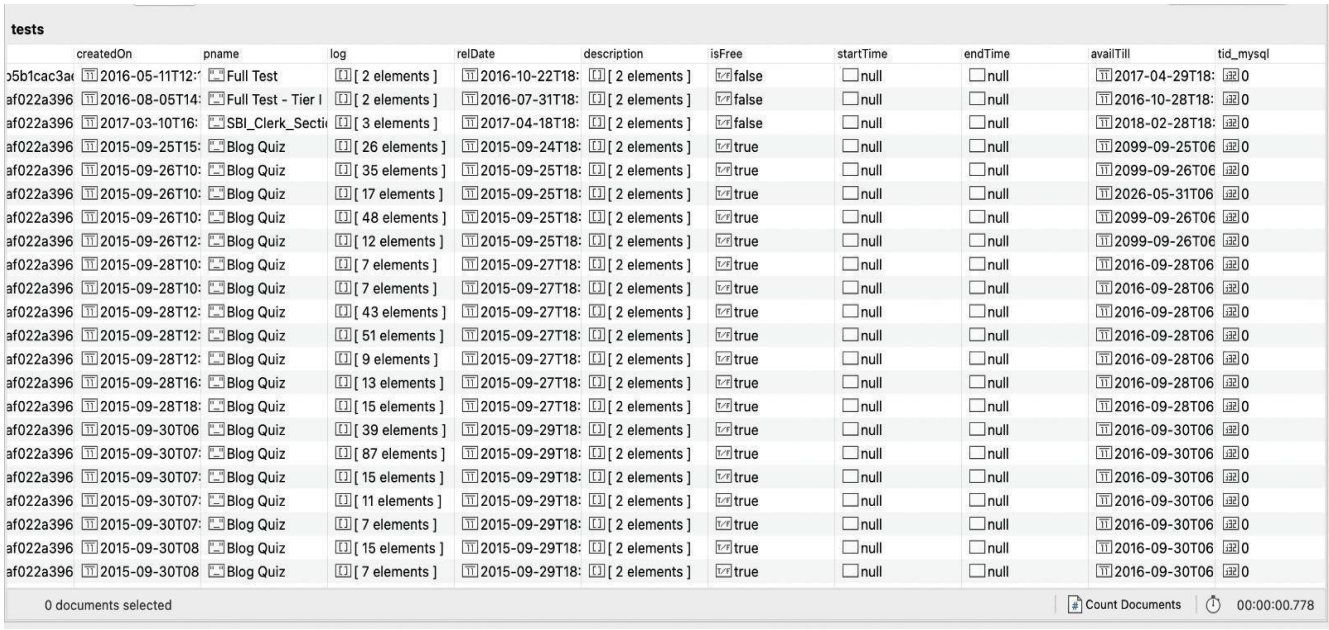
Overall, compared to AWS Lambda, Google Cloud Functions provides a serverless computing solution that is more affordable, integrated, and developer-friendly.

CHAPTER 4: EXPERIMENTS AND RESULTS ANALYSIS

Database Used:

tests:

This database (DB) holds all test information available in the testbook database, with a unique identifier known as a Tid.



tests

createdOn	pname	log	relDate	description	isFree	startTime	endTime	availTill	tid_mysql
5b1cac3a	2016-05-11T12:	Full Test	[2 elements]	2016-10-22T18:	[2 elements]	false	null	null	2017-04-29T18:	0
af022a396	2016-08-05T14:	Full Test - Tier I	[2 elements]	2016-07-31T18:	[2 elements]	false	null	null	2016-10-28T18:	0
af022a396	2017-03-10T16:	SBI_Clerk_Secti	[3 elements]	2017-04-18T18:	[2 elements]	false	null	null	2018-02-28T18:	0
af022a396	2015-09-25T15:	Blog Quiz	[26 elements]	2015-09-24T18:	[2 elements]	true	null	null	2099-09-25T06	0
af022a396	2015-09-26T10:	Blog Quiz	[35 elements]	2015-09-25T18:	[2 elements]	true	null	null	2099-09-26T06	0
af022a396	2015-09-26T10:	Blog Quiz	[17 elements]	2015-09-25T18:	[2 elements]	true	null	null	2026-05-31T06	0
af022a396	2015-09-26T10:	Blog Quiz	[48 elements]	2015-09-25T18:	[2 elements]	true	null	null	2099-09-26T06	0
af022a396	2015-09-26T12:	Blog Quiz	[12 elements]	2015-09-25T18:	[2 elements]	true	null	null	2099-09-26T06	0
af022a396	2015-09-28T10:	Blog Quiz	[7 elements]	2015-09-27T18:	[2 elements]	true	null	null	2016-09-28T06	0
af022a396	2015-09-28T10:	Blog Quiz	[7 elements]	2015-09-27T18:	[2 elements]	true	null	null	2016-09-28T06	0
af022a396	2015-09-28T12:	Blog Quiz	[43 elements]	2015-09-27T18:	[2 elements]	true	null	null	2016-09-28T06	0
af022a396	2015-09-28T12:	Blog Quiz	[51 elements]	2015-09-27T18:	[2 elements]	true	null	null	2016-09-28T06	0
af022a396	2015-09-28T12:	Blog Quiz	[9 elements]	2015-09-27T18:	[2 elements]	true	null	null	2016-09-28T06	0
af022a396	2015-09-28T16:	Blog Quiz	[13 elements]	2015-09-27T18:	[2 elements]	true	null	null	2016-09-28T06	0
af022a396	2015-09-28T18:	Blog Quiz	[15 elements]	2015-09-27T18:	[2 elements]	true	null	null	2016-09-28T06	0
af022a396	2015-09-30T06	Blog Quiz	[39 elements]	2015-09-29T18:	[2 elements]	true	null	null	2016-09-30T06	0
af022a396	2015-09-30T07:	Blog Quiz	[87 elements]	2015-09-29T18:	[2 elements]	true	null	null	2016-09-30T06	0
af022a396	2015-09-30T07:	Blog Quiz	[15 elements]	2015-09-29T18:	[2 elements]	true	null	null	2016-09-30T06	0
af022a396	2015-09-30T07:	Blog Quiz	[11 elements]	2015-09-29T18:	[2 elements]	true	null	null	2016-09-30T06	0
af022a396	2015-09-30T07:	Blog Quiz	[7 elements]	2015-09-29T18:	[2 elements]	true	null	null	2016-09-30T06	0
af022a396	2015-09-30T08	Blog Quiz	[15 elements]	2015-09-29T18:	[2 elements]	true	null	null	2016-09-30T06	0
af022a396	2015-09-30T08	Blog Quiz	[7 elements]	2015-09-29T18:	[2 elements]	true	null	null	2016-09-30T06	0

0 documents selected | Count Documents | 00:00:00.778

4.1 **tests DB**

test_summary:

All of the user's test-taking information, including the amount of time it took, the results, and so forth, is included in the test summary.

test_summary

a_end	sectionalAnalysis	e_end	resp	exam	client	lang	isIgnored	isLive	pdfUrl	normal
2021-03-27T10:	[1 elements]	2021-03-27T10:	[20 elements]	Rajasthan SSB	android	en	false	false		
2021-07-03T11:	[3 elements]	2021-07-03T11:	[91 elements]	Bank Clerk	android	en	false	false		
2021-07-07T11:	[3 elements]	2021-07-07T11:	[100 elements]	Bank Clerk	android	en	false	false		
2021-07-08T13:	[3 elements]	2021-07-08T13:	[92 elements]	Bank Clerk	android	hn	false	false		
2021-07-09T03:	[3 elements]	2021-07-09T03:	[94 elements]	Bank Clerk	android	hn	false	false		
2021-07-11T05:	[3 elements]	2021-07-11T05:	[100 elements]	Bank Clerk	android	hn	false	false		
2021-07-21T12:	[2 elements]	2021-07-21T12:	[80 elements]	Bank PO	android	en	false	false		
2021-07-22T10:	[2 elements]	2021-07-22T10:	[67 elements]	Bank PO	android	en	false	false		
2021-07-23T15:	[2 elements]	2021-07-23T15:	[60 elements]	Bank PO	android	en	false	false		
2021-07-25T08:	[1 elements]	2021-07-25T08:	[24 elements]	DFCCIL	android	hn	false	false		
2021-08-06T15:	[2 elements]	2021-08-06T15:	[80 elements]	Bank Clerk	android	en	false	false		
2021-08-07T10:	[2 elements]	2021-08-07T10:	[80 elements]	Bank Clerk	android	en	false	false		
2021-08-16T04:	[1 elements]	2021-08-16T04:	[10 elements]	Bank PO	android	en	false	false		
2021-08-16T05:	[1 elements]	2021-08-16T05:	[30 elements]	Bank Clerk	android	hn	false	false		
2021-08-17T05:	[1 elements]	2021-08-17T05:	[10 elements]	Bank PO	android	hn	false	false		
2021-08-17T13:	[1 elements]	2021-08-17T13:	[10 elements]	Bank PO	android	en	false	false		
2021-08-17T13:	[1 elements]	2021-08-17T13:	[10 elements]	Bank PO	android	en	false	false		
2021-08-17T13:	[1 elements]	2021-08-17T13:	[10 elements]	Bank PO	android	hn	false	false		
2021-08-17T13:	[1 elements]	2021-08-17T13:	[10 elements]	Bank PO	android	hn	false	false		
2021-08-17T13:	[1 elements]	2021-08-17T13:	[10 elements]	Bank PO	android	hn	false	false		
2021-08-17T13:	[1 elements]	2021-08-17T13:	[10 elements]	Bank PO	android	hn	false	false		
2021-08-17T14:	[1 elements]	2021-08-17T14:	[10 elements]	Bank PO	android	hn	false	false		

0 documents selected | Count Documents | 00:00:00.358

4.2 test_summary DB

Table | + Add Visualization

Search userEmail and tid...

createdOn	userEmail	issue	client	testStage	comment	language	tid	testName
2021-01-18 15:27	nikhilkumar95@gmail.com	Wrong Question	android	freeze	proper condition is not given in this question	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-18 15:32	roshan3019@gmail.com	Wrong Question	web	freeze	No. of people speaking Japanese only would be Zero.	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-18 16:09	richasharma1023@gmail.com	Wrong Question	android	freeze	data looks incorrect.. total should be more than 75	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-18 17:18	pratikmid.2015@gmail.com	Wrong Question	web	freeze	ans will be 0. here option is missing	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-18 18:26	gunjan1805@gmail.com	Wrong Question	android	freeze	I think this question is not right my answer is 0 so please check this question	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-18 18:26	gunjan1805@gmail.com	Formatting Issue	android	freeze		en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-19 06:55	abhijadhavcsn@gmail.com	Wrong Question	android	freeze	no Japanese person remain to speak only Japanese	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-19 08:36	shubham232528@gmail.com	Wrong Question	web	freeze	the number of people is more than 50	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper
2021-01-19 09:43	rahulgurjar@gmail.com	Formatting Issue	android	freeze	language change nhi ho rha	en	60055a77a1bf62b3299c231c	RRB NTPC 18 January 2021 Shift 1 Memory Based Paper

< 1 2 >

4.3 Output Of Redash Method Of Extraction:

Preference Of MongoDB Over SQL in Redash:

In several circumstances, MongoDB may be preferred to SQL databases for a number of reasons:

1. Flexibility: Because it's a document-oriented database, MongoDB enables dynamic and flexible schema design. When data models are not clearly defined or are anticipated to change frequently, this is helpful.

2. Scalability: MongoDB can handle high data and traffic levels by adding more servers to a cluster because it is designed to grow horizontally. It's a fantastic option for applications that need great scalability because of this.

3. Performance: For some types of queries, particularly those involving unstructured or semi-structured data, MongoDB can offer high performance. In order to increase performance, it also offers sharding, which enables the dissemination of data across numerous servers.

4. Productivity of developers: MongoDB's flexible query language and schema can make it simpler for developers to deal with data and make quick iterations on application features. Additionally, for developers accustomed to working with JavaScript or other web technologies, MongoDB's JSON-based data model may be more familiar and simpler to use.

While traditional transactional applications still rely on SQL databases, MongoDB is a fantastic choice for contemporary applications that demand adaptable data models and high scalability. The decision between SQL and MongoDB will ultimately come down to the particular requirements of the application and the preferences of the development team.

Drawbacks:

Redash is a popular business intelligence and visualization tool that can help organizations make sense of their data. While it is a powerful and versatile platform, there are some potential drawbacks to consider when working with large amounts of data.

1. One potential drawback is that Redash can be slow when working with large datasets. This is because it relies on a database to store and manage the data, and as the dataset grows in size, queries can take longer to run and load times can become slow. This can be particularly problematic when dealing with complex queries or aggregations, as these operations can be resource-intensive and slow down the system even further.

2. Another potential issue is that Redash may not be as flexible as other tools when it comes to handling large amounts of data. For example, it may not be able to handle certain types of data sources or may not be able to handle large data sets with complex schemas. This can limit the types of analyses and visualizations that can be performed, and may require additional data processing or transformation before the data can be used in Redash.

3. Finally, Redash may not be as scalable as other tools when it comes to handling large amounts of data. This is because it is designed to be run on a single server or cluster, which can limit the amount of data that can be processed and analyzed. As data volumes increase, organizations may need to consider alternative tools or technologies that can handle larger volumes of data more efficiently.

Overall, while Redash is a powerful and flexible tool for working with data, it may not be the best choice for all use cases, particularly those involving very large data sets or complex queries. Organizations should carefully consider their needs and requirements before choosing a business intelligence or visualization tool, and should evaluate a range of options to find the best fit for their specific use case.

Approach 2: Using Python Script Method:

We have two databases: "tests" and "test_summary". The "tests" database contains information about the tests conducted, including the test ID (Tid), student ID (Sid), and

their corresponding marks. On the other hand, the "test_summary" database contains information about the test attempts made by the students, including the test ID (Tid), student ID (Sid), and other details.

To extract the maximum marks, Tids, and Sids for the details of user exam attempts from the "test_summary" database using the "tests" database, we can use a query that joins the two databases on the Tid and Sid columns.

"tests" that contain information about various exams taken by users. We also have another database named "test_summary" that contains a summary of each user's test results, including their total marks, test ID (Tid), and student ID (Sid).

To extract the maximum marks, Tids, and Sids for each user's exam attempts, we can join the "tests" and "test_summary" databases on the Tid and Sid fields, and then use an aggregate function to get the maximum marks.

Once we have this data, we can extract category-wise data from the "test_summary" database to get a more detailed breakdown of each user's test results. Finally, we can extract student details from another database (let's call it "students") using the Sids we obtained earlier. However, we cannot present the actual student details due to data confidentiality concerns.

4.4 Code Using test_summary DB:

```
print ("Time:",dt.datetime.now())
t=db_test.test_summary.aggregate([
   {
            "$match": {
               "_id": {
                    "$gte":ObjectId.from_datetime(start_date),
                    "$lte":ObjectId.from_datetime(end_date)
                },
                "tid":{"$in": tids },
                "status": "finished"
            }
        },
        {"$project":{"_id":0,"sid":"$sid","tid":"$tid","marks":"$marks"}}
         ],allowDiskUse = True)
print("Got Cursor")
temp=pd.DataFrame(list(t))
print(len(temp))
print("Got data from mongo")
sids=pd.concat([temp],axis=0,ignore_index=True)
del temp
print("Time:",dt.datetime.now())
```

4.5 Code Using tests DB:

```
print ("Time:",dt.datetime.now())
p=db_test.tests.aggregate([
    {
    "$match":{
        "stage":"freeze",
        "metaData.type":{
            "$in":["test","liveTest"]
        }
        }
    },
    {
        "$unwind":"$targetSuperGroup"
    },
    {
    "$project":{
        "category":{ "$arrayElemAt": ["$targetSuperGroup.title.value" ,0] },
        "maxM":{"$sum":"$sections.maxM"}
    }
    },
    {"$match":{"maxM":{"$gt":0}}}
])
print("Got Cursor")
t1=pd.DataFrame(list(p))
category=pd.concat([t1],axis=0,ignore_index=True)
len(category)
```

	sid	tid	marks
0	607d7155fb2915314d5e40e1	63ad7c05604565c43a94f4c5	30.00
1	5a76deecf8f7c20f84faafd4	6332fd04dff3a91b9c6cb675	112.00
2	59708a712be36008dc5d5280	63ce3551b05bee5db483d006	77.00
3	629a596c23b02b53d42439dd	6396f5d1b31836c610e26399	13.25
4	62f8a30d509578788de1d441	63cf89e3a109a78780ec3be2	57.50
...	...	...	...
6257	5fabdfd02aa59fff12558d0f	63a078ff4b6017fd9ddf156a	18.50
6258	63968531195376f310252f8a	63650ea1f51c11218320922a	2.00
6259	62b46f7cc0e53cf6087410a2	63d8bfd87919a82533bfe9ab	99.50
6260	5d7376a89cbd530e25db0615	6391c308b169e450773c6755	27.50
6261	558dbb5f2a396512bc15ce39	6329bcb89aecf6929473c754	21.00

6262 rows × 3 columns

	sid	tid	marks	_id	category	maxM	Result
0	607d7155fb2915314d5e40e1	63ad7c05604565c43a94f4c5	30.0	63ad7c05604565c43a94f4c5	SSC Exams	50.0	60.000000
1	5f62374f423e175c2e86b378	63ad7c05604565c43a94f4c5	30.0	63ad7c05604565c43a94f4c5	SSC Exams	50.0	60.000000
2	60119582d3405f9cfe1c1c43	63ad7c05604565c43a94f4c5	32.5	63ad7c05604565c43a94f4c5	SSC Exams	50.0	65.000000
3	61c34e39a5623d84edfb0e35	63ad7c05604565c43a94f4c5	23.0	63ad7c05604565c43a94f4c5	SSC Exams	50.0	46.000000
4	5a76deecf8f7c20f84faafd4	6332fd04dff3a91b9c6cb675	112.0	6332fd04dff3a91b9c6cb675	SSC Exams	225.0	49.777778
...	...	...	...	...	...	...	...
6257	5efb562fbd7c6a0d0df387b9	6391e40ab7b34d1e7f558fd7	15.0	6391e40ab7b34d1e7f558fd7	SSC Exams	20.0	75.000000
6258	5df86bfcca762b0cfb92e126	63650ed4222cfcc4a18adaed	6.0	63650ed4222cfcc4a18adaed	SSC Exams	30.0	20.000000
6259	62c294f941456a0a1f9e8b69	6329cbe9b32e6f4a683ddedd	2.0	6329cbe9b32e6f4a683ddedd	SSC Exams	30.0	6.666667
6260	62ee949ea4cb47b9747f87ec	63ad7c29868595743da986dd	-0.5	63ad7c29868595743da986dd	SSC Exams	50.0	-1.000000
6261	5d7376a89cbd530e25db0615	6391c308b169e450773c6755	27.5	6391c308b169e450773c6755	SSC Exams	50.0	55.000000

6262 rows × 7 columns

4.6 Script Results

Drawbacks:

1. Python is a popular programming language used for working with data, but it can have some drawbacks when dealing with large datasets. One issue is memory usage, as Python needs to load the entire dataset into memory to perform operations on it. This can cause performance issues and even crashes when working with datasets that are too large for the available memory.
2. Another drawback of using Python with large datasets is that it can be slower than other languages, such as C++ or Java, especially when performing computationally intensive tasks. Python is an interpreted language, meaning that it is not compiled and executed directly by the computer's CPU, which can slow down the execution of code.
3. Additionally, Python is not optimized for parallel processing, which can limit its ability to perform tasks efficiently on multicore or distributed systems. While there are tools and libraries available to enable parallel processing in Python, such as multiprocessing and Dask, they can add complexity to the code and require additional resources.
4. Finally, Python's dynamic typing and lack of compile-time type checking can make it more difficult to identify errors in code when working with large datasets. This can lead to unexpected results and longer debugging times, especially when working with complex data structures.

Despite these drawbacks, Python remains a popular language for working with data due to its ease of use, flexibility, and extensive ecosystem of libraries and tools. By using best practices for memory management, optimizing code for performance, and leveraging parallel processing tools, developers can mitigate some of the issues when working with large datasets in Python.

3. Cloud Function Approach:

GCF is a serverless computing platform that allows developers to write and deploy code without having to worry about infrastructure management. Cloud Functions is built on top of Google Cloud Platform (GCP) and is designed to execute a single function in response to a trigger.

To implement code using Google Cloud Functions, developers can write their code in a supported language, such as Python, Node.js, Go, or Java, and deploy it to Cloud

Functions. The function can then be triggered by various event sources, such as HTTP requests, Cloud Storage events, Pub/Sub messages, and more.

Cloud Functions provides a flexible and scalable way to implement code, as it automatically scales based on the demand for the function and can handle high volumes of traffic. It also allows developers to focus on writing code and not on infrastructure management, reducing development time and costs.

Another advantage of using Cloud Functions is its integration with other Google Cloud services, such as Cloud Storage, Pub/Sub, and Firestore. This allows developers to easily create and integrate cloud-native applications using a wide range of Google Cloud services.

Overall, Google Cloud Functions provides a powerful and flexible platform for implementing code without worrying about infrastructure management. Its scalability, ease of use, and integration with other Google Cloud services make it a popular choice for building cloud-native applications.

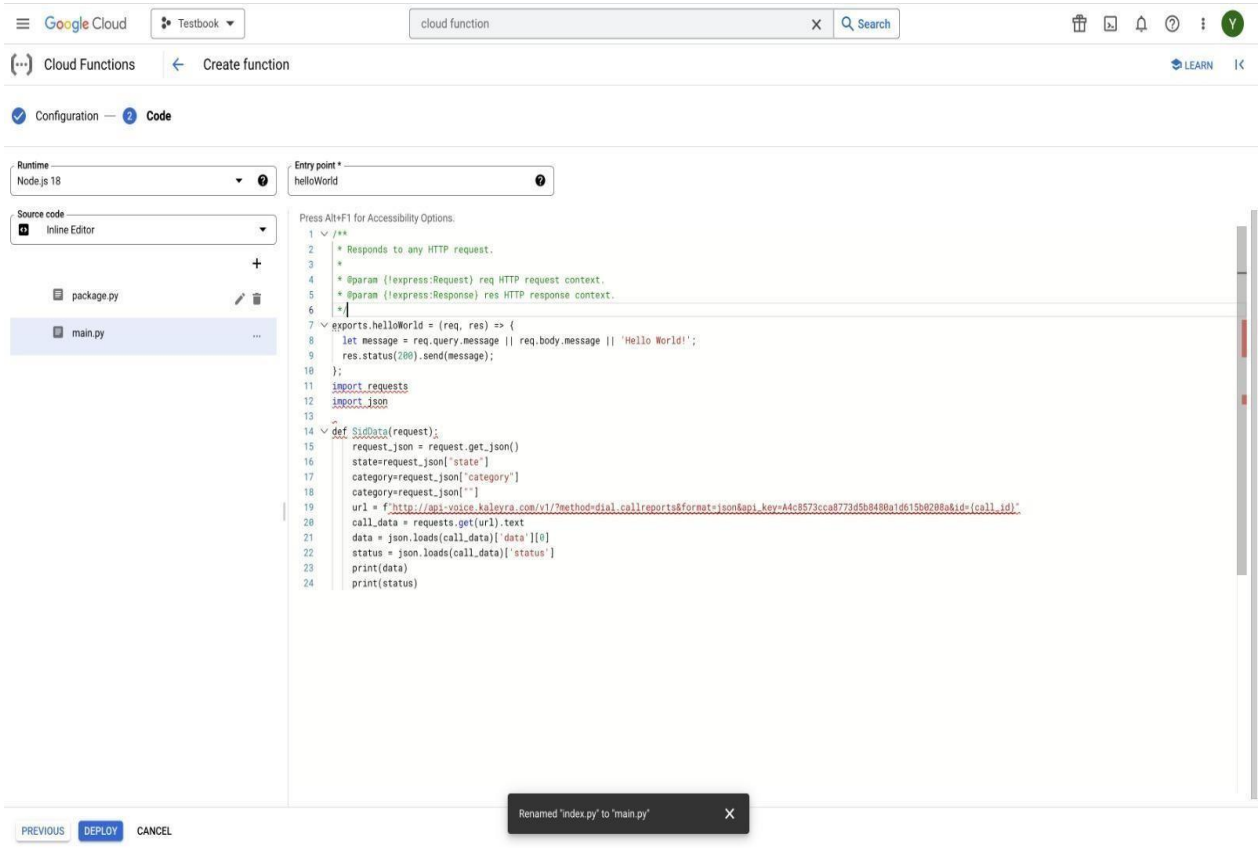
APIs are commonly used in cloud functions as they allow functions to interact with other systems and services. This is typically done through REST APIs, which enable functions to send and receive HTTP requests and responses.

APIs provide a range of benefits for cloud functions, including the ability to access external data sources and services, automate workflows, and integrate with other applications. By leveraging APIs, cloud functions can easily interact with other services and systems, without having to build and maintain complex integrations from scratch.

In addition, many cloud services provide their own APIs that can be used within cloud functions. For example, Google Cloud Platform provides APIs for a wide range of services, including storage, machine learning, and analytics. These APIs can be used to integrate cloud functions with other Google Cloud services, making it easier to build complex applications and workflows.

Overall, APIs play a crucial role in the functionality and usefulness of cloud functions, allowing developers to build powerful and efficient applications that can integrate with a wide range of services and systems.

4.7 Google Cloud Function Console



Cloud Functions offer several advantages over running Python scripts or Redash queries:

1. Scalability: Cloud Functions can handle a large number of requests concurrently, making them ideal for applications with unpredictable or high workloads. In contrast, running Python scripts or Redash queries on a single server may result in slower performance and potential downtime.
2. Cost-effectiveness: Cloud Functions are charged only for the time they run and the resources they consume, making them a cost-effective solution for small to medium-sized applications. Python scripts and Redash queries may require dedicated server resources, which can be more expensive to maintain.

3. Flexibility: Cloud Functions are designed to be flexible and can be triggered by various events, such as HTTP requests, changes to data in a database, or scheduled time intervals. This allows developers to automate processes and build complex workflows easily.
4. Integration: Cloud Functions can easily integrate with other Google Cloud services, such as Cloud Storage, BigQuery, and Pub/Sub, providing a complete cloud solution for data processing and analytics. In contrast, Python scripts and Redash queries may require additional configuration to work with these services.

Overall, Cloud Functions offer a cost-effective, scalable, flexible, and integrated solution for running code in the cloud.

CHAPTER 5: CONCLUSIONS

Conclusion:

In conclusion, effective data analysis and decision-making depend on an optimized data extracting methodology. Developers may create and put into use data extraction workflows that are scalable, dependable, and secure by employing the right tools and methods. In order to do this, the appropriate database management system must be chosen, such as MongoDB or SQL, and data processing activities must be handled using cloud computing services like Google Cloud Functions. A further benefit of integrating API services is that it can speed up data extraction and lessen the demand for manual labor. Overall, an improved data extracting methodology can help organizations achieve a competitive edge in their particular industries and significantly improve data-driven decision-making.

Data extraction procedures can be streamlined and made more effective by utilizing contemporary technologies and best practices, which results in a faster and more accurate examination of the data. This can assist businesses in making wise decisions, enhancing client experiences, and streamlining business operations. The particular requirements of each organization must be taken into account when selecting the data extraction tools and methods, though. In today's data-driven world, an optimized data extracting methodology can help organizations stay ahead of the competition and provide significant benefits.

Future Aspect:

Optimized data extraction models appear to have a bright future as new tools and methods are made available as technology develops. These potential future benefits of optimized data extraction models are listed below:

1. Machine learning integration: With the volume of data being produced, automated data processing is becoming more and more necessary. Organizations can increase data accuracy and decrease the need for manual intervention by incorporating machine learning algorithms into data extraction methods.

2. Real-time data extraction: By giving businesses current information about their operations, real-time data extraction enables organizations to take speedier decisions. We may anticipate seeing more sophisticated tools for real-time data extraction as technology advances, including the application of AI and machine learning.
3. Increased use of cloud services: Organizations are already utilizing the cloud for data extraction, and we may anticipate seeing even more of them in the future. Using serverless computing tools like Google Cloud Functions, which can save costs and increase scalability, is one example of this.
4. Enhanced data security: Organizations will continue to place a high priority on data security as a result of the growing volume of data being retrieved and processed. We may anticipate the implementation of increasingly sophisticated security measures in the future, including the usage of blockchain technology and cutting-edge encryption techniques.

   The use of technology by businesses to obtain insights into their operations and enhance decision-making bodes well for the future of optimized data extraction techniques. Organizations can make sure they're maximizing the value of their data and staying ahead of the competition by remaining up to date with the most recent tools and practices.

Applications:

The optimized data extraction model is applicable in a wide range of fields and use situations. Here are a few instances:

1. E-commerce: Online merchants can gather client information, such as browsing habits, past purchases, and preferences, by using the optimized data extraction methodology. This information can be utilized to tailor marketing efforts, enhance pricing options, and increase client loyalty.

2. Healthcare: To gather and analyze patient data, including medical history, treatments, and outcomes, healthcare professionals can make use of the optimized data extraction model. In order to enhance patient care, find trends and patterns, and advance research and development, this data can be exploited.

3. Finance: Financial organizations can gather and analyze financial data, such as market trends, investment performance, and client transactions, using the optimized data extraction methodology. Investment choices, risk management, and customer engagement can all be improved with the use of this data.

4. Manufacturing: Manufacturing businesses can gather and analyze production data, such as equipment performance, maintenance history, and quality control, using the optimized data extraction methodology. Utilizing this information can enhance product quality, decrease downtime, and optimize production procedures.

   Overall, the optimized data extraction approach has the ability to completely change how businesses function and make choices, resulting in increased productivity, better customer experiences, and increased market competitiveness.

REFERENCES

[1] Vijay Kumar, Talwinder Kaur “Cloud Functions and Serverless Computing”, publisher: IJRASET, vol. 10, doi : 6.05.2022

[2] Rupal Snehkunj and Khushboo Vachiyatwala “Data Analysis Using Pandas Library of Python”, vol. 3, doi : 3.03.2022

[3] Shweta J. Patil “Python – Using Database and SQL”, publisher: IJSR, vol. 1, doi : 6.06.2019

[4] https://redash.io/

[5]https://www.mongodb.com/community/forums/t/process-of-storing-images-in-mongodb/15093

[6] https://www.python.org/

YASH BHARDWAJ 191397

ORIGINALITY REPORT

11%	5%	3%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijraset.com Internet Source	2%
2	Submitted to Florida Institute of Technology Student Paper	1%
3	Submitted to Glasgow Caledonian University Student Paper	1%
4	Submitted to Tower Hamlets College Student Paper	1%
5	Submitted to Cerritos College Student Paper	1%
6	Submitted to Liverpool John Moores University Student Paper	1%
7	Submitted to University of Greenwich Student Paper	<1%
8	cloud.google.com Internet Source	<1%
9	Submitted to Barnet and Southgate College Student Paper	<1%