mA Capstone Project Report on

Project Title

Submitted by

Aditi Jain, One-Data Intern

Aditya Kejariwal, One-Data Intern

Koundinya Katamoni Goud, One-Data Intern

SkyDive Global Campus Academy

April 2024

DEDICATION

I dedicate this project to my beloved family, whose unwavering support and encouragement have been my pillars of strength. Your love, belief in me, and sacrifices made this journey possible, and I

am forever grateful. To my parents, for instilling in me the values of perseverance and determination, and to my siblings, for their unwavering faith in my abilities.

This project is a testament to the love and resilience of my family. I carry your lessons and blessings with me, knowing that your unwavering presence is the driving force behind my success.

ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to all those who contributed to the successful completion of this project.

First and foremost, I want to express my sincere thanks to [Any of the mentors], for his/her invaluable guidance, mentorship, and patience. Their expertise and insights greatly enriched this project.

I am deeply appreciative of my colleagues and fellow team members whose tireless collaboration brought diverse skills and perspectives to the project. Your dedication and teamwork were instrumental in achieving our project objectives.

I acknowledge the support of Genpact for providing the necessary resources and facilities that made this project possible.

Special thanks go to my family for their unwavering encouragement and understanding during the project's demanding phases.

Finally, I want to express my appreciation to all the individuals and resources, online databases, and research materials that I accessed throughout this project.

This project has been a significant learning experience, and I am thankful to everyone who played a part in it.

ABSTRACT

This project addresses the critical need for streamlined invoice processing in the retail domain, aiming to mitigate the inefficiencies and errors inherent in manual accounts payable processes. By leveraging innovative technologies such as Optical Character Recognition (OCR), Robotic Process Automation (RPA), and cloud computing, our solution automates the capture, validation, and reconciliation of invoices from diverse sources. The implementation of Google Cloud Platform (GCP) services such as Cloud Function, Pub/Sub, BigQuery, and Looker Studio enables seamless integration and real-time data synchronization, enhancing operational efficiency and accuracy. Through a rigorous research methodology, including data collection, processing, and analysis, our project demonstrates the effectiveness of automated invoice processing in reducing processing times and error rates while improving financial transparency and decision-making. The findings underscore the potential value of automated invoice processing systems in optimizing accounts payable operations for retailers, highlighting their relevance and significance in the evolving landscape of retail accounting.

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Project Introduction

Our project's goal is to use an Automated Invoice Processing System to transform retail accounting. This innovative technology, which includes Python and Google Cloud Platform services, automates the processes of invoice capture, validation, reconciliation, and approval. As the retail industry continues to expand and evolve, efficient and accurate accounts payable operations become increasingly critical. Mismanaged invoices can result in financial discrepancies, overlooked payments, strained vendor relationships, and operational inefficiencies.

1.2 Background and Related Works

Within the retail industry, accounts payable management poses difficulties because manual methods are labor-intensive and prone to errors, delaying payment reconciliation. Accounts payable management involves handling and processing payments owed by a company to its suppliers or vendors for goods and services received. With a surge in invoices from varied sources like email and digital portals, there's a pressing need for automation.

Emerging technologies like Optical Character Recognition (OCR) and robotic process automation (RPA) offer potential in streamlining accounts payable processes. Moreover, cloud computing advancements enable seamless integration, allowing real-time data synchronization. The Automated Invoice Processing System aims to address these challenges by automating invoice capture, validation, reconciliation, and approval workflows. Additionally, it provides reporting and analytics capabilities for optimizing processes and monitoring performance.

The rationale behind the development of the Automated Invoice Processing System lies in addressing these challenges and leveraging emerging technologies to revolutionize retail accounting practices. By automating invoice capture, validation, reconciliation, and approval workflows, the system aims to enhance efficiency, accuracy, and transparency in accounts payable operations for retailers. Moreover, by providing reporting and analytics capabilities, the system enables retailers to gain valuable insights into their invoice processing metrics, facilitating process optimization and performance monitoring.

1.3 Key Terminology and Concepts

This section provides essential definitions and explanations of key terms and concepts relevant to understanding the Automated Invoice Processing System, including technologies, cloud computing services, and analytical tools utilized in the project.

- **Invoices: Invoices are documents issued by suppliers or vendors to request payment for goods or services provided to a company.**

- **Purchase Orders: Purchase orders are documents issued by a buyer to a supplier, outlining the details of goods or services to be purchased, including quantities, prices, and delivery dates.**

- **Accounts Payable Processes: Accounts payable processes refer to the series of steps involved in managing and processing payments owed by a company to its suppliers or vendors for goods and services received.**

- **Optical Character Recognition (OCR): OCR technology converts scanned or photographed text into machine-readable data, facilitating automated extraction of invoice information.**

- **Robotic Process Automation (RPA): RPA automates repetitive tasks involved in invoice processing, reducing manual effort and improving efficiency.**

- **ETL (Extract, Transform, Load): Process of extracting data from various sources, transforming it into a format suitable for analysis, and loading it into a target database or data warehouse.**

- **Cloud Computing: Cloud computing provides scalable and flexible computing resources for automated invoice processing systems, enabling cost-effective and reliable solutions.**

- **Google Cloud Platform (GCP): GCP offers a suite of cloud computing services utilized in the project, including Cloud Function, Pub/Sub, BigQuery, and Looker Studio.**

- **Cloud Function: Cloud Function is a serverless compute service that executes event-driven code, enabling seamless integration and automation within the GCP ecosystem.**

- **Pub/Sub (Publish/Subscribe): Pub/Sub is a messaging service that facilitates communication between independent applications, enabling real-time data exchange and event-driven workflows.**

- **BigQuery: BigQuery is a fully managed, serverless data warehouse solution on GCP, utilized for storing and analyzing large datasets efficiently in the project.**

- **Looker Studio: Looker Studio is a business intelligence and data visualization platform, leveraged for generating reports, dashboards, and analytics to gain insights into invoice processing metrics.**

1.4 Outline of the Report

Chapter 1: Introduction

- **Introduces the project, its context, and objectives.**

- **Reviews relevant background information and related works.**

- **Defines key terminology and concepts.**

- **Outlines the structure and content of the report.**

Chapter 2: Project Overview and Objectives

- **Provides a summary of background research.**

- **Identifies the problem statement and gaps in current practices.**

- **Clarifies the aim and specific objectives of the project.**

- **Discusses the significance and relevance of the project.**

- **Previews the structure and organization of the report.**

Chapter 3: Project Methodology

- **Describes the research methodology employed in the project.**

- **Explains the process of data collection and processing.**

- **Details the tools and technologies utilized in the project.**

- **Discusses the experimental design or data sources used for analysis.**

Chapter 4: Results and Discussions

- **Presents the findings and outcomes of the project.**
- **Analyzes and interprets the results in relation to the project objectives.**
- **Discusses how the findings align with the initial problem statement.**
- **Compares the results with existing literature and research.**

Chapter 5: Conclusion and Future Recommendations

- **Summarizes the key findings and achievements of the project.**
- **Evaluates the extent to which project objectives were met.**
- **Offers recommendations for future research and improvements.**
- **Concludes with final thoughts and implications for practice.**

CHAPTER 2: PROJECT OVERVIEW AND OBJECTIVES

2.1 Summary of Background Works

In the realm of retail accounting, manual accounts payable processes have long been plagued by inefficiencies and errors, stemming from the labor-intensive nature of invoice handling. Retailers face challenges in managing a high volume of invoices from diverse sources while ensuring accuracy and compliance with regulatory standards.

A review of existing literature reveals a growing body of research and industry reports focused on automated invoice processing as a solution to these challenges. Studies highlight the potential of technologies such as Optical Character Recognition (OCR) and Robotic Process Automation (RPA) in streamlining invoice capture, validation, and reconciliation processes. These technologies offer opportunities for retailers to automate repetitive tasks, reduce processing times, and improve data accuracy.

Existing solutions in the market offer a range of features and functionalities for automated invoice processing, including cloud-based platforms, ERP integrations, and advanced analytics capabilities. However, despite these advancements, there remain gaps in addressing the specific needs of

retailers, particularly in terms of scalability, customization, and seamless integration with existing systems.

This background research underscores the significance of our project, which aims to develop an Automated Invoice Processing System tailored specifically for retailers. By leveraging innovative technologies and integrating with Google Cloud Platform (GCP) services such as Cloud Function, Pub/Sub, BigQuery, and Looker Studio, our solution seeks to address the shortcomings of existing approaches and deliver tangible benefits in terms of efficiency, accuracy, and compliance.

Overall, the insights gleaned from the background works section inform the objectives and methodology of our project, laying the groundwork for the development of a comprehensive solution to the challenges of manual invoice processing in the retail sector.

2.2 Problem Statement and Identified Gaps

Retail operations involve a plethora of processes, including the management of accounts payable, that typically require human effort and are susceptible to errors. The processing and balancing of invoices can be a prolonged and cumbersome task, resulting in delays, inefficiencies, and inaccuracies in financial records. These challenges underscore the necessity for a solution that can automate invoice processing for retail businesses.

1.1.1 Project Context

In today's dynamic retail landscape, the need for efficient and accurate invoice processing is paramount. Manual handling of invoices leads to errors, delays, and increased costs. With the rise of e-commerce and diverse supplier networks, retailers face mounting pressure to optimize their accounts payable processes. Our Automated Invoice Processing System leverages cutting-edge technologies to address these challenges and propel retailers towards streamlined and agile financial operations.

1.1.2 Objectives of the Study

The specific aims of the project are outlined as follows:

1.1.2.1 Primary Objectives

- **Automated Invoice Capture: Implement streamlined processes to efficiently capture invoices from various channels, enhancing data acquisition agility.**
- **Data Extraction and Validation: Utilize OCR technology to extract and validate invoice data, ensuring compliance with business rules and criteria.**

• **Invoice Reconciliation: Establish robust links between invoices, purchase orders, and goods received for accurate reconciliation and financial transparency.**

1.1.2.2 Secondary Objectives

• **Exception Handling Workflows: Develop efficient workflows and escalation mechanisms to promptly address exceptions and discrepancies in invoice processing.**

• **Approval Workflow Management: Define and implement approval workflows and routing rules to streamline invoice review and approval processes.**

• **Reporting and Analytics: Provide comprehensive reporting and analytics capabilities to monitor invoice processing metrics and optimize operational performance.**

2.3 Aim and Objectives

The primary aim of this project is to revolutionize retail accounting practices by developing an Automated Invoice Processing System tailored specifically for retailers. This system aims to address the inherent challenges and inefficiencies associated with manual accounts payable processes, ultimately streamlining operations, reducing manual effort, and enhancing accuracy in invoice handling and reconciliation. The objectives are,

1. Automated Invoice Capture: Implement automated mechanisms to capture invoices from diverse sources, including email, paper documents, and electronic invoice portals. By automating the capture process, the system aims to streamline the intake of invoices, reducing the manual effort required for data entry and processing.

2. Data Extraction and Validation: Utilize Optical Character Recognition (OCR) technology to extract relevant data fields from invoices and validate them against predefined business rules and validation criteria. This objective aims to enhance accuracy in invoice processing by automating the extraction and validation of critical invoice information.

3. Invoice Matching: Develop algorithms to match invoices with purchase orders (POs) and goods receipts to ensure accuracy in invoice-to-PO and invoice-to-receipt reconciliation. This objective seeks to minimize discrepancies and errors in invoice processing by automating the matching process and ensuring alignment with procurement and inventory data.

4. Exception Handling: Design workflows and escalation mechanisms to handle exceptions, discrepancies, and invoice disputes efficiently. By automating exception handling processes, the system aims to expedite resolution and minimize disruptions in the accounts payable workflow.

5. Approval Workflows: Define approval workflows and routing rules to route invoices for review and approval based on predefined approval hierarchies and thresholds. This objective aims to streamline the invoice approval process, ensuring timely review and adherence to organizational policies and compliance requirements.

6. Reporting and Analytics: Provide reporting and analytics capabilities to track invoice processing metrics, such as invoice cycle times, error rates, and vendor performance, for process optimization

and performance monitoring. This objective aims to empower stakeholders with actionable insights for continuous improvement and informed decision-making.

2.4 Significance and Relevance

Manual invoice handling in the retail sector presents numerous challenges, including lengthy processing times, increased error rates, and difficulties in maintaining compliance with regulatory standards. These inefficiencies not only hinder operational efficiency but also pose risks to financial transparency and decision-making processes within retail organizations. However, the implementation of an Automated Invoice Processing System has the potential to address these challenges comprehensively. By automating invoice capture, validation, reconciliation, and approval workflows, retailers can significantly reduce processing times, minimize errors, and ensure adherence to regulatory requirements. This streamlined approach enhances operational efficiency, reduces costs associated with manual labor, and improves overall financial management practices.

The adoption of automated invoice processing offers a range of benefits for retailers. Beyond cost savings achieved through reduced manual effort and error correction, retailers can also benefit from time efficiency gains by accelerating the invoice processing cycle. Real-time data analytics capabilities enable improved decision-making, while timely and accurate payments enhance vendor relationships and supplier trust. In line with broader industry trends towards digital transformation, automation, and data-driven decision-making, automated invoice processing represents a critical component of modern retail operations. By embracing automation technologies, retailers can stay ahead of the curve and adapt to changing market dynamics, gaining a competitive edge in the process.

Moreover, the implications of automated invoice processing extend beyond the retail sector. The integration of automation technologies in financial management practices sets a precedent for increased efficiency, accuracy, and transparency across industries. Insights gained from automated invoice processing can inform strategic decision-making and drive business growth in an increasingly competitive and data-driven business environment. As organizations across various sectors recognize the value of automation in optimizing operations and enhancing decision-making processes, automated invoice processing stands as a testament to the transformative power of technology in modern business practices.

2.5 Report Structure

This report is structured into five chapters, each providing valuable insights into different aspects of the project:

- **Chapter 1: Introduction**

Provides a foundational understanding of the project, including its aim, objectives, relevance within the retail domain, and an overview of the report's structure. It sets the stage for the subsequent

chapters by establishing the context and importance of the project.

- **Chapter 2: Project Overview and Objectives**

Offers a summary of background works, defines the problem statement and identified gaps, outlines the aim and objectives of the project, and explores its significance and relevance. It provides a comprehensive overview of the project's scope and objectives.

- **Chapter 3: Project Methodology**

Details the research methodology, data collection and processing techniques, tools and technologies utilized, and the experimental design or data sources employed. It provides insights into the methodology employed to carry out the project.

- **Chapter 4: Results and Discussions**

Presents the project results, interprets findings, assesses alignment with the problem statement, and compares results with existing literature. It offers a detailed analysis of the project outcomes and their implications.

- **Chapter 5: Conclusion and Future Recommendations**

Summarizes key findings, evaluates the achievement of objectives, suggests future research directions and improvements, and concludes with final thoughts and implications. It synthesizes the findings of the project and provides recommendations for future work.

CHAPTER 3: PROJECT METHODOLOGY

3.1 Research Methodology

The primary objective of this project was derived from the problem statement. Our problem statement can be stated as:

Developing an Automated Invoice Processing System for Retailers to streamline accounts payable processes, reduce manual effort, and enhance accuracy in invoice handling and reconciliation.

The problem statement was first identified by understanding the domain of retail and the different mechanisms behind invoice processing. A literature review was conducted in order to understand the current scenario and techniques to implement several components pertaining to invoice handling. Based on the understanding of how these current systems were implemented, we

identified the shortcomings or the gaps that need to be addressed, in order to result in a more efficient and complete workflow.

Upon identifying the functionalities that needed to be added, we also looked into the methods and platforms to implement the same. Several tools and technologies were explored, keeping in mind their costs, feasibility, ease of access and scalability.

In order to understand the kind of data needed for this project, first the problem statement and our proposed functionalities were analyzed. Based on this, it was concluded that invoice data containing purchase order information, payment amounts and methods, agencies and vendors and other such information was relevant.

3.2 Data Collection and Processing

For this project, data pertaining to invoices and corresponding purchase orders was required, preferably in the form of images. To demonstrate the full functionality of the project, 2 datasets were utilized: the FATURA dataset and the PASS dataset.

An optical character recognition feature is essential to implement. For the same, a dataset that consists of invoice in the form of images is required. The FATURA dataset consists of 10000 jpg images of invoices, each having dimensions 595 x 841 pixels. The invoices contain information such as company name, addresses, telephone numbers, item information and invoice number.

Since all the data is in the form of images, and not text that can be utilized further, the very first step was to implement OCR on the dataset, to get the information contained in each invoice image in a structured, text format.

Prior to extracting the text from images, the images must be processed to extract the text accurately.

The following were the steps to implement OCR:

i. Performing OTSU thresholding to convert a greyscale image into a black and white (binary image)

ii. Create a rectangular structuring element. This element will help perform dilation to detect text boxes from the rest of the image.

iii. Implement dilation on the image. Dilation is a morphological image processing operation which expands the size of an object in foreground. Since our image is binary, each pixel is either black (1, the foreground), or white (0, the background). The structuring element created in the previous step moves across the image and adds pixels to the edges of the objects, making them larger.

iv. Implement contouring: Contours are the outlines of objects in an image that are represented as a list of points. This is done to detect those regions which have text to achieve a more accurate extraction.

v. Implement Tesseract Extraction on each individual text box (contoured region).

Essentially, for OCR, the colored image is first made into a binary image, followed by dilation to make the text more prominent, after which contouring is applied to detect text boxes. Finally, the text is extracted from each text box.

Now, this extracted text is also stored in a structured format in a CSV file. For this, a list of columns with names consisting of the different common fields in an invoice was initialized:

[due date, date, po number, ITEMS, quantity price, quantity, price, invoice id, invoice number, invoice #, invoice, address, ship to, ship_to, sub_total, balance, balance due, balance_due, buyer, bill to, email, tel, gstin, bank name, branch name, bank account number, bank swift code, subtotal, discount, tax, total, site"]

For every segment of text extracted, Regular Expressions were formulated. The extracted text was then put into the column corresponding to the regular expression it matches, and the text data was structured accordingly.

This data was then processed further and cleaned. Data validation was performed to ensure numeric values in columns such as payments, date formats were made uniform and an ETL pipeline was implemented to automate the process to clean any new incoming data.

Since the FATURA dataset did not come with a corresponding Purchase Orders dataset, in order to implement further processing and features in our project, we have also incorporated the PASS datasets which consists of a dataset of Payments, and another dataset of Purchase Orders, on which all the functionalities apart from OCR are demonstrated.

PASS Dataset: Payments from PASS.

This dataset contains payment information that the District of Columbia Government made to vendors. It is made available by the Open Data DC under the City of Washington, DC. This dataset includes information pertaining to contract numbers, invoice payment dates and methods, suppliers and agencies involved, etc. Along with this, the corresponding Purchase Orders dataset was utilized.

This dataset contained information regarding purchase orders, commodities, suppliers and dates. Both these datasets were available in a CSV format via government websites.

Preprocessing: In order to further utilize this data, both the datasets were first validated in terms of format and all the fields with dates were converted to a datetime format. All floating and numeric values were found to be formatted and the dataset was mostly clean. Null values were detected, but not dropped for the demonstration of the exception handling functionality ahead.

3.3 Tools and Technologies

The main of this project is to incorporate multiple functionalities into one and to automate these modules such that anyone may utilize them without any hassle. For the same, a platform was required which could integrate all the proposed services.

The main platform on which this project was implemented was Google Cloud Platform. The following is the segment-wise explanation of the tools that were used:

Data Collection: For the FATURA dataset, once the images were collected, a script was developed in Python using OpenCV, Pandas and PyTesseract to perform OCR and accordingly accumulate data into a structured CSV file. OpenCV is a Python library that is used for image processing operations. The structuring, dilation and text-box element aspects of the preprocessing were implemented via this library. PyTesseract is a python library for extracting text from images and was utilized to get the text within each individual text box identified. Lastly, Pandas is a library allowing for convenient data analysis and manipulation, which assisted the structuring and exporting of the data as CSV.

Similarly, in order to clean and analyse the PASS datasets, Python was used along with Pandas.

Implementing the functionalities proposed in this project required the usage of Google Cloud Platform, and several aspects of cloud computing. For storing, importing and exporting data, Buckets were used as a warehouse.

To implement the ETL pipelining, Purchase Order matching as well as exception handling, Google Cloud functions written in Python were utilized which were triggered every time a file is added or updated in the base location. Post ETL, cleaned files were exported to the final bucket.

For the exception handling aspect, the SendGrid API was incorporated into the function in order to add the functionality of alerting emails. Google Cloud Workflows was utilized to build the approval mechanism step by step. Lastly, looker studio was utilized for developing an analysis report.

3.4 Experimental Design or Data Sources

The proposed experimental design of our project can be broken into the following components:

- **OCR**
- **ETL Diagram**
- **Exception handling**
- **Workflow**
- **Report**

OCR:

As described above, the OCR methodology was implemented with the purpose of ensuring that even that are not electronically generated in nature were able to be processed and vendors can migrate to the platform seamlessly via images of their invoices.

The OCR design was further broken into two constituents: identifying regions of text (text boxes) and extracting the text from each of those regions. A common issue with current OCR mechanisms is that due to the lack of context, text will often be extracted across the image. This means that if an image has two columns of text side by text, OCR readers will tend to merge the column sentences as they read from left to right across an image.

Hence, in order to retain the context and increase the accuracy of the meaning of the text extracted, a region-segmenting functionality was necessary.

ETL:

Figure 1:Detailed ETL

Figure 2: ETL Overview

Exception Handling:

Exception handling is a critical aspect of ensuring the reliability and accuracy of the Automated Invoice Processing System. After the ETL process, the system identifies any anomalies or errors in the processed invoices. This is done through a Cloud Function, whose base trigger is any change in the dataset. Upon detection, the system consolidates the row numbers with corresponding exceptions into a dictionary, where each row number serves as a key and the associated exceptions as values.

Here is a list of exceptions that are being handled:

- **Invalid Payments Contract Number: Rows with missing or invalid contract numbers in the payments section.**
- **Invalid Purchase Order Contract Number: Rows with missing or invalid contract numbers in the purchase orders section.**
- **Mismatched Contract Numbers: Rows where the contract numbers in the payments and purchase orders sections do not match.**
- **Mismatched Invoice and Payment Dates: Rows where the invoice date is later than the payment date.**
- **Invalid Invoice Number: Rows with missing or invalid invoice numbers.**
- **Mismatched Fiscal Years: Rows where the payment date falls outside the fiscal year range.**
- **Mismatched Created and Ordered Dates: Rows where the created date is later than the ordered date in the purchase orders section.**

Subsequently, this dictionary is converted into a pandas DataFrame, which is then transformed into an HTML table format. This formatted table containing the exceptions is then sent via email to the designated recipient for further review and resolution.

This exception handling process ensures that any discrepancies or issues in the invoicing data are promptly identified and communicated, allowing for timely intervention and resolution by the appropriate personnel. By implementing such a robust exception handling mechanism, the Automated Invoice Processing System enhances reliability, streamlines operations, and ensures the integrity of accounts payable processes for retailers.

Approval Workflow:

Ideally, any invoice generated requires an approval from an entity. The basic idea of this component is to implement a request mechanism for incoming invoices and alert the person in charge of approval. Based on the values present in the dataset, four tiers were hypothesized:

o 1st tier: Invoices with amount less than 10,000 USD – Managers

o 2nd tier: Invoices with amount greater than 10,000 USD but under 50,000 USD – Senior Managers

o 3rd tier: Invoices with amount greater than 50,000 USD but under 100,000 USD – Directors

o 4th tier: Invoices with amount greater than 100,000 USD – Vice Presidents

Accordingly, the workflow will get triggered every time a new invoice is added, then identify the position that should be approving the workflow, and finally send the alerting mail.

Figure 3: Approval Workflow Diagram

Figure 5: Sub-Workflow (Sending Mail)

The workflow consists of the following:

- Every time a new query is added to the bigQuery table of invoices, a cloud function is triggered.

- This cloud function publishes a message to a pub/sub topic pertaining to triggering the workflow.

- The workflow is triggered upon pulling the message published.

- The workflow first extracts the query that was added most recently – i.e. the query that triggered the function.

- The workflow then implements a Switch Case to determine the appropriate entity that needs to approve the invoice.

- Based on this information, the workflow formulates a mail and sends it to the identified position to request their approval.

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 Presentation of Project Results

a) OCR

Base image:

Figure 6:Base image

Binary Image:

Figure 7:Binary Image

Image after contouring (text boxes identified):

Figure 8:Image after contouring

Snippet of the final extracted data in CSV Format:

Figure 9:final extracted data

b) ETL Pipeline

Base bucket (Input):

Currently empty

Figure 10: Empty Bucket

Upon Transfer of data:

Figure 11: Bucket with Data

****

ETL Pipeline Cloud Function is triggered:

Figure 12:ETL Pipeline Cloud Function Logs

Cleaned File is output in destination bucket:

Figure 13: destination bucket with cleaned file

Cleaned data snippet:

Figure 14:Cleaned data snippet

c) Exception Handling

The empty bucket, with no data in it.

Figure 15:empty bucket

Bucket with data input:

Figure 16:Bucket with data input

Triggering of cloud function for exception handling mechanism:

Figure 17:Logs of cloud function for exception handling

Final Output: The email sent from the cloud platform, highlighting each exception and error present in the invoices.

Figure 18:email sent from the cloud platform

d) Approval Workflow

Inserting a new record

Figure 19: Inserting a new record

into the invoices and PO table:

Triggers a workflow:

Figure 20: Trigger a workflow

Workflow sends a mail to the appropriate entity requesting approval:

Figure 21: Workflow sends a mail

e) Report Generation

Figure 22: Looker Studio Report

4.2 Interpretation of Findings

Google Cloud was observed to be a comprehensive platform to implement an entire start-to-end system for invoice processing. All modules were successfully implemented without any high-end costs and proved to be a sustainable, scalable model.

From the final analysis generated on the data that had been cleaned with exceptions handled, several key observations were found. A trend was seen in terms of Purchase orders placed every November, with the average amount during November reaching 900 while remaining considerably less and more stable throughout the rest of year. A critical reason for this can be the end of the fiscal year occurring in September in the United States, with the new FY starting in October, this is also reason why we see a significant increase in the invoices volume processed in September.

Another key finding was that the total amount paid to vendors by agencies tended to reach a yearly high during October and November. The highest amount paid was observed in January 2023, with November 2022 and November 2023 reaching almost similar values.

Analyzing the payment types, it was found that nearly half the transactions were made as direct deposits, with first class mail coming in next and electronic and cheque payments being preferred the least. An invoice volume spike was observed in the month of September – meaning that the highest number of invoices generated was in September, nearly one month before the ending of the fiscal year.

The agency with the highest payment amount was Department of General Services (DGS) at about 4,000,000,000, followed by the District Department of Transportation at nearly half the amount.

4.3 Alignment with Problem Statement

Problem Statement: Developing an Automated Invoice Processing System for Retailers to streamline accounts payable processes, reduce manual effort, and enhance accuracy in invoice handling and reconciliation.

The problem statement of this project was to build a platform for automated invoice approval and to incorporate several crucial aspects regarding invoice processing. Keeping these factors in mind, all the components have been successfully delivered and implemented, and are shown to be running within extremely short time limits – less than 3 seconds in most scenarios.

The functionality of OCR allows users from small scale and street-side businesses that do not tend to generate electronic bills to use this system as well. Manual effort is greatly reduced by implementing functionalities online that can be availed at any point by simply entering or scanning new data.

4.4 Comparison with Literature

CHAPTER 5: CONCLUSION AND FUTURE RECOMMENDATIONS

5.1 Summary of Key Findings

After implementing Google Cloud Platform for our invoice processing system, a detailed analysis was conducted to uncover key insights. The findings highlight efficiency, seasonal trends, payment preferences, and agency transactions. These insights offer valuable understanding for optimizing processes. Key findings include:

- **Efficient Google Cloud Implementation: Cost-effective and scalable, Google Cloud facilitated end-to-end invoice processing without high expenses.**
- **Seasonal Purchase Order Trends: Notable increases in purchase orders during November, coinciding with fiscal year-end transitions.**
- **Seasonal Payment Variations: Peaks in vendor payments during October and November, with January 2023 recording the highest payments.**
- **Payment Preferences: Direct deposits were favored, followed by first-class mail, while electronic and cheque payments were less common.**
- **September Invoice Surge: A significant spike in invoices generated in September, preceding fiscal year-end deadlines.**
- **Top-Paying Agencies: The Department of General Services (DGS) led in payments, followed by the District Department of Transportation.**
- **These findings provide actionable insights into invoice processing dynamics, aiding future process optimization.**

5.2 Achievement of Objectives

The achievement of objectives within the Automated Invoice Processing System project reflects the successful implementation of key functionalities aimed at revolutionizing retail accounting practices, leveraging a range of technologies:

- **Automated Invoice Capture: The system utilizes cloud-based email processing services and document scanning technology to automate the capture of invoices from diverse sources, including email, paper documents, and electronic portals.**
- **Data Extraction and Validation: Leveraging Optical Character Recognition (OCR) technology like tesseract, the system accurately extracts relevant data fields from invoices. Additionally, custom validation algorithms ensure the accuracy of extracted data against predefined business rules and validation criteria.**

• Invoice Matching: Advanced algorithms, written in python, match invoices with purchase orders (POs) and goods receipts. The system utilizes data synchronization mechanisms to ensure accuracy in invoice-to-PO and invoice-to-receipt reconciliation.

• Exception Handling: Robust workflows and escalation mechanisms, made in Google Cloud Function, handle exceptions, discrepancies, and invoice disputes efficiently. Custom business rules engine and decision support systems enable automated resolution of common exceptions.

• Approval Workflows: Automated approval workflows, built in the Google Cloud Workflows, route invoices for review and approval based on predefined hierarchies and thresholds. Role-based access controls. Implemented using Google Pub/Sub, ensure compliance with organizational policies and regulatory requirements.

• Reporting and Analytics: The system provides reporting and analytics capabilities using Google BigQuery for data storage and analysis. Dashboards enable stakeholders to monitor invoice processing metrics and gain insights for process optimization.

The successful achievement of these objectives underscores the system's effectiveness in enhancing efficiency, accuracy, and transparency in accounts payable operations for retailers. By leveraging OCR technology and automated workflows, the system streamlines invoice processing, minimizes errors, and optimizes financial processes effectively.

5.3 Future Research and Improvements

Several avenues for future improvements and enhancements can be explored to further enhance the effectiveness and efficiency of the Automated Invoice Processing System that we have made:

1. Enhanced OCR Capabilities: Invest in advanced OCR technology to improve accuracy in text extraction and interpretation, especially for complex invoice formats or handwritten documents. Integration with natural language processing (NLP) models can enable better understanding of unstructured data.

2. Intelligent Exception Handling: Develop machine learning algorithms to automatically identify and categorize exceptions, reducing the need for manual intervention. Implement predictive analytics to anticipate potential issues and proactively resolve them before they impact the workflow.

3. Vendor Self-Service Portal: Develop a self-service portal for vendors to submit invoices, track payment status, and resolve billing discrepancies autonomously. Implement real-time notifications and alerts to keep vendors informed about invoice processing status and payment timelines.

4. Mobile Accessibility and Chatbot Support: Enhance system accessibility by developing mobile applications for invoice submission, approval, and tracking. Implement chatbot support for invoice-related inquiries and issue resolution, improving user experience and reducing response times.

5.4 Final Thoughts and Implications

As we conclude our journey through the development and implementation of the Automated Invoice Processing System, several insights and consequences surface, revealing this job's larger relevance.

Firstly, the successful deployment of this system underscores the transformative potential of technology in revolutionizing traditional accounting practices, particularly within the retail sector. By automating manual invoice handling processes, streamlining data extraction and validation, and implementing intelligent workflows, retailers can achieve significant improvements in efficiency, accuracy, and compliance.

Furthermore, the ramifications extend beyond operational enhancements to encompass broader strategic considerations for retailers. The integration of cloud-based solutions, machine learning algorithms, and advanced analytics not only streamlines internal processes but also empowers retailers to delve deeper into their financial operations, vendor relationships, and overall business performance.

Moreover, this project's significance transcends retail accounting, extending to the broader domains of digital transformation and innovation. Through the embrace of emerging technologies and the adoption of agile methodologies, organizations can unlock new avenues for growth, competitiveness, and sustainability in an increasingly dynamic and disruptive business landscape.

In conclusion, the development and implementation of the Automated Invoice Processing System represent a significant milestone in the journey towards digital transformation and innovation in retail accounting. By harnessing the power of technology to automate routine tasks, streamline processes, and enhance decision-making, retailers can position themselves for success in an ever-evolving digital economy. As we look towards the future, it is essential to continue embracing innovation, collaboration, and continuous improvement to drive meaningful change and create lasting value for retailers and stakeholders alike.

REFERENCES

- **Purchase Orders from PASS | Open Data DC**
- **Payments from PASS | Open Data DC**
- **FATURA Dataset**

- **Looker Studio Report**