-

Industry Project

Report

On

Weekly Task as Data Analyst

Developed By:-

Het Parikh (20162101021)

Guided By:-

Prof. Aniket Patel

Mr. Javed Nilgar

Submitted to

Faculty of Engineering and Technology

Institute of Computer Technology

Ganpat University

Year - 2024

CERTIFICATE

This is to certify that the Industry Project work entitled “Weekly Task as Data Analyst”by Het Parikh(20162101021) of Ganpat University, towards the partial fulfillment of requirements of the degree of Bachelor of Technology – Computer Science and Engineering, carried out by them in the CSE(CBA/BDA/CS) Department at ICT The results/findings contained in this Project have not been submitted in part or full to any other University / Institute for award of any other Degree/Diploma.

Name & Signature of Internal Guide

Name & Signature of Head

Place: ICT - GUNI

Date: 04-05-2024

ACKNOWLEDGEMENT

An Industry Internship project is a golden opportunity for learning and self-development. I consider myself very lucky and honored to have so many wonderful people lead me through in completion of this project. First and foremost, I would like to thank Rohit Patel, Principal, ICT and Prof. Dharmesh Darji, Head, ICT who allowed us to undertake this project. My thanks to Prof. Aniket Patel & Mr. Javed Nilgar for their guidance in project work, who despite being extraordinarily busy, took time out to hear, guide, and keep us on the correct path. We do not know where we would have been without his help. The CSE department monitored our progress and arranged all facilities to make life easier. We choose this moment to acknowledge their contribution gratefully.

Het Parikh [20162101021]

Travel Designer
Group

Date — 30th Apr 2024

TO WHOM IT MAY CONCERN

This is to certify that Mr. Het Parikh, student of Ganpat University, has successfully completed an internship in the field of Data Analyst from 17th Jan 2024 till 30th Apr 24 under the guidance of Mr. Tarak Thakor designated as a Chief Technology Officer.

During the period of his internship program with us, he had been exposed to different process and found diligent, hardworking and inquisitive.

We wish him every success in him life and career.

Best of luck.

For and on behalf of,

Travel Designer group

COO

Travel Designer India Pvt. Ltd.

Corporate Office: B Wing - 1402, Mondeal Heights, Near Wide Angle Cinema, S. G. Highway, Ahmedabad - 380015, Gujarat, India
Tel: +91-79-6617 6000 I Fax: +91-79-6617 6060 I All India Helpline: 1860 233 3322 I Email: info@traveldesigner.in
www.traveldesignergroup.com

CIN NO: U63040GJ2004PTC44927

ABSTRACT

As a data analyst, I have been fortunate to experience a deeply gratifying journey thus far. Within the scope of my role, I am entrusted with the vital responsibility of furnishing the sales team with indispensable analytics on a weekly basis. This pivotal task entails the intricate development of dashboards meticulously tailored to steer discerning business decisions effectively. Leveraging sophisticated tools such as Jupyter Notebook, I adeptly extract, refine, and meticulously cleanse data to ensure the utmost precision in visualization. Furthermore, my exploration of Pentaho, an ETL tool, has afforded me the opportunity to seamlessly automate intricate data cleansing processes, thereby optimizing operational efficacy to the fullest extent. These concerted efforts underscore my unwavering dedication to not only harnessing but also pioneering advanced analytics methodologies, culminating in the refinement of decision-making paradigms and the propelling of organizational success to greater heights.

INDEX

CHAPTER: 1 OVERVIEW

Overview

1.1 Scope of work

The scope of work for the role of a data analyst at Travel Designer Group, a renowned company in the travel industry, encompasses several key responsibilities. Primarily, the focus is on providing insightful dashboards tailored to queries from senior management. These dashboards serve as crucial decision-making tools for the sales team, enabling them to make informed business calls based on comprehensive data analysis. Additionally, a significant aspect of the role involves meticulous data cleaning to ensure accuracy and reliability in the extracted datasets. The data analyst will be tasked with refining and preparing the extracted data to facilitate a clear understanding for all stakeholders. Furthermore, the scope may extend to exploring and implementing advanced data analytics techniques to enhance the quality and depth of insights provided to the sales team and other relevant departments. Overall, the data analyst will play a pivotal role in driving data-driven decision-making processes and contributing to the overarching success of Travel Designer Group.

CHAPTER: 2 WEEK-1 PROGRESS

CHAPTER 2 WEEK-1 PROGRESS

2.1 Team Meetings

Active participation in team meetings aimed at introducing and educating colleagues on the utilization of new technologies within the organization. This involves understanding the specific roles and responsibilities of team members, comprehending the functionality and significance of the rezlive.com product, and gaining insights into the operational procedures of the company.

2.2 Installation and Configuring ELK Stack

- Installation and configuration of the ELK stack, a comprehensive process that demands meticulous attention to detail and intricate setup steps. This task typically requires a duration of approximately two days to ensure the seamless integration and functioning of Elasticsearch, Logstash, and Kibana, the core components of the ELK stack.
- Detailed explanation of the ELK stack components:
    - Elasticsearch: Providing an in-depth overview of Elasticsearch, focusing on its role as a distributed, RESTful search and analytics engine designed for horizontal scalability, reliability, and real-time search capabilities.
    - Logstash: Delving into the functionalities of Logstash, emphasizing its role as a data processing pipeline that ingests, transforms, and enriches diverse data sources before indexing them into Elasticsearch for analysis and visualization.
    - Kibana: Exploring the features and capabilities of Kibana, highlighting its role as a powerful data visualization and exploration tool that enables users to interact with data stored in Elasticsearch through dynamic dashboards, visualizations, and analytical tools.

2.3 Acquaintance to ELK Stack

Familiarization with the ELK Stack through hands-on practice and experimentation with sample datasets. This phase involves actively engaging with Elasticsearch, Logstash, and Kibana to gain

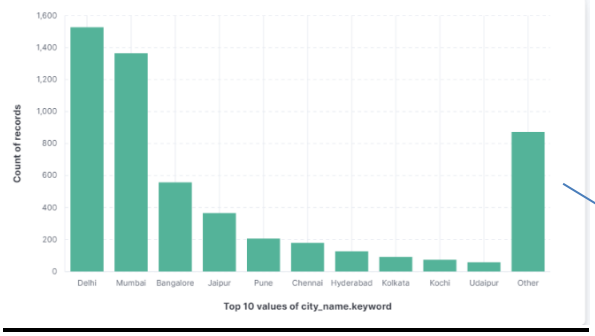
practical experience in data ingestion, transformation, indexing, and visualization processes. By working with sample data, the goal is to develop proficiency in navigating and utilizing the ELK stack effectively for data analysis and visualization purposes.

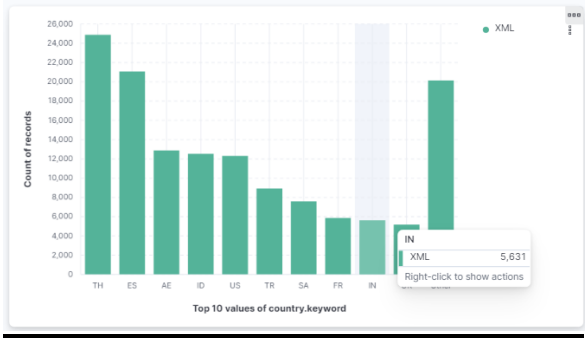
CHAPTER: 3 WEEK-2 PROGRESS

CHAPTER 3 WEEK-2 PROGRESS

3.1 Dashboard Creation

The development and implementation of dashboards play a pivotal role in catering to the specific requirements of two distinct client categories: XML clients, who utilize the hotel booking API, and POS clients, who directly book through rezlive.com. These meticulously designed dashboards serve as invaluable monitoring tools, allowing for the comprehensive tracking and analysis of essential metrics and data points. A recent illustrative example underscores the paramount importance of these dashboards: they facilitated the timely identification of a top client originating from Saudi Arabia, who had conspicuously abstained from engaging in any hotel searches for the preceding four days. This insightful observation, made possible through the utilization of the searches dashboard, enabled the sales team to swiftly respond and capitalize on potential business opportunities. Thus, the strategic utilization of dashboards not only enhances operational efficiency but also empowers proactive decision-making and client management practices within the organization.





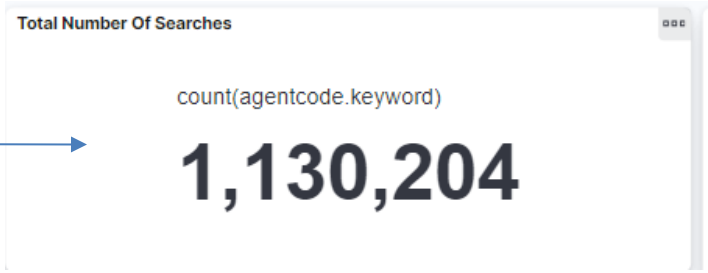
These XML client searches represent global activity, with a focus on the top 10 countries as parameters. Notably, Thailand stands out as the leading country in terms of search frequency. Additionally, the provided image illustrates the capability to delve deeper into specific countries; for instance, clicking on India reveals a separate dashboard showcasing the top cities with the highest XML client

Screenshots:

agentcode.keyword	Count of records	Last value of timestamp_ist	agentcode.keyword2
Agent-1	1,05,820	Feb 8, 2024 @ 16:52:51.000	1
Agent-2	12,109	Feb 8, 2024 @ 16:52:48.000	1
Agent-3	6,635	Feb 8, 2024 @ 16:52:41.000	1
Agent-4	4,837	Feb 8, 2024 @ 16:51:36.000	1
Agent-5	2,554	Feb 8, 2024 @ 16:52:47.000	1
Agent-6	2,039	Feb 8, 2024 @ 16:45:58.000	1
Agent-7	1,310	Feb 8, 2024 @ 16:43:47.000	1
Agent-8	967	Feb 8, 2024 @ 16:51:18.000	1
Agent-9	490	Feb 8, 2024 @ 16:52:02.000	1
Agent-10	457	Feb 8, 2024 @ 16:52:37.000	1
Agent-11	428	Feb 8, 2024 @ 16:50:26.000	1
Agent-12	419	Feb 8, 2024 @ 16:45:10.000	1
Agent-13	320	Feb 8, 2024 @ 16:52:28.000	1
Agent-14	290	Feb 8, 2024 @ 16:36:35.000	1
Agent-15	252	Feb 8, 2024 @ 16:27:22.000	1
Agent-16	224	Feb 8, 2024 @ 16:51:02.000	1
Agent-17	181	Feb 8, 2024 @ 16:52:37.000	1
Agent-18	167	Feb 8, 2024 @ 16:50:11.000	1
Agent-19	144	Feb 8, 2024 @ 16:52:11.000	1
Agent-20	127	Feb 8, 2024 @ 16:47:46.000	1
Agent-21	104	Feb 8, 2024 @ 16:52:39.000	1
Agent-22	62	Feb 8, 2024 @ 16:33:00.000	1
Agent-23	60	Feb 8, 2024 @ 16:33:18.000	1
Agent-24	55	Feb 8, 2024 @ 16:42:45.000	1
Agent-25	48	Feb 8, 2024 @ 16:42:58.000	1
Agent-26	42	Feb 8, 2024 @ 16:32:25.000	1
Agent-27	34	Feb 8, 2024 @ 16:42:34.000	1
Agent-28	24	Feb 8, 2024 @ 16:34:07.000	1
Agent-29	24	Feb 8, 2024 @ 16:32:18.000	1
Agent-30	12	Feb 8, 2024 @ 16:27:46.000	1
Agent-31	9	Feb 8, 2024 @ 16:14:29.000	1
Agent-32	4	Feb 8, 2024 @ 16:46:25.000	1

This dashboard provides insights into the cumulative count of searches conducted by clients within a given day, along with the most recent date and timestamp of their last search activity. This dashboard provides insights into the cumulative count of searches conducted by clients within a given day, along with the most recent date and timestamp of their last search activity.

Total number of searches by XML and POS clients combined.



3.2 Data Extraction

The process involves the extraction of data from the File Transfer Protocol (FTP) server followed by meticulous cleaning utilizing Excel. This multi-step procedure begins with accessing the FTP server, where relevant data is located, and transferring it to a local storage system for further analysis. Upon retrieval, the data undergoes thorough cleaning procedures within Excel, involving various

techniques such as removing duplicates, formatting inconsistencies, and addressing missing or erroneous entries. Additionally, data validation processes may be employed to ensure accuracy and completeness. Each step of the extraction and cleaning process requires careful attention to detail and adherence to best practices to maintain data integrity and reliability. This comprehensive approach guarantees that the extracted data is refined and prepared for subsequent analysis, contributing to informed decision-making and strategic insights within the organization.

3.3 Hitachi Pentaho

Engaging in an immersive learning experience with Pentaho, specifically focusing on Extract, Transform, Load (ETL) processes, through practical application with sample datasets. This hands-on approach facilitates a comprehensive understanding of data extraction, transformation, and loading methodologies, empowering proficiency in leveraging Pentaho for efficient data management and analysis tasks.

agentcode.keyword	checkin.keyword	city_name.keyword	Count of records
Agent-1	04-03-2024	Paris	165
Agent-2	04-03-2024	Saint-Denis	17
Agent-3	04-03-2024	Clichy	14
Agent-4	04-03-2024	Gennevilliers	13
Agent-5	04-03-2024	Courbevoie	12
Agent-6	04-03-2024	Levallois-Perret	12
Agent-7	04-03-2024	Roissy-en-France	12
Agent-8	04-03-2024	Boulogne-Billancourt	10
Agent-9	04-03-2024	Le Kremlin-Bicetre	10
Agent-10	04-03-2024	Neuilly-sur-Seine	10
Agent-11	04-03-2024	Chevilly-Larue	8
Agent-12	04-03-2024	Creteil	8
Agent-13	04-03-2024	Issy-Les-Moulineaux	8
Agent-14	04-03-2024	Montrouge	8
Agent-15	04-03-2024	Puteaux	8
Agent-16	04-03-2024	Colombes	7
Agent-17	04-03-2024	Malakoff	7
Agent-18	04-03-2024	Massy	7
Agent-19	04-03-2024	Rungis	7

The table presents data on agents' most recent hotel bookings, detailing the cities they booked in and the corresponding search frequencies for hotels in those cities. This information offers valuable insights into booking patterns and agent activity, guiding strategic decisions and enhancing customer satisfaction. Additionally, it enables monitoring of agent performance, facilitating data-driven adjustments to optimize business outcomes.

CHAPTER: 4 WEEK-3 PROGRESS

CHAPTER 4 WEEK-3 PROGRESS

4.1 Data Cleaning

Data cleaning plays a crucial role in ensuring the accuracy and reliability of data used for analysis. Despite the widespread use of Excel for this purpose, its manual nature makes it time-consuming, especially for large datasets. Recognizing this inefficiency, I was tasked by senior management to develop a Python script to automate the data cleaning process, aiming to enhance efficiency and save valuable time. This initiative has progressed meticulously, with significant strides towards completing the assigned task. Through diligent efforts and unwavering dedication, the development of the Python script is nearing completion, poised to revolutionize our data cleaning workflow and streamline analytical processes within the organization.

As part of this endeavor, a sample dataset was provided for me to clean before accessing the mail file. The provided dataset includes:

CustomerID
Name
Email
Phone Number
Address

The dataset exhibits common issues such as missing values, inconsistent formatting, and duplicate entries. By leveraging Python's libraries such as pandas and numpy, I've devised a systematic approach to address these issues programmatically. The script identifies and rectifies missing values, standardizes formatting across fields, and eliminates duplicate records. Additionally, it performs data validation checks to ensure accuracy and consistency.

CHAPTER: 5 WEEK-4 PROGRESS

CHAPTER 5 WEEK-4 PROGRESS

5.1 Data Cleaning – II

In the fourth week of my tenure, I was entrusted with a significant task involving the maintenance and updating of an extensive Excel sheet containing information on hotels collaborating with our company. This sheet encompasses a staggering 1.2 million hotel entries and requires regular updates every 15 days. To streamline operations, two distinct Excel files are generated from this master sheet—one tailored for internal office use and the other customized for agent utilization.

For the agent-specific Excel file, certain columns need to be removed while others are added, aligning with the specific requirements and workflows of our agent network. This customization ensures that agents have access to pertinent information while maintaining data integrity and confidentiality.

Over the past week, I dedicated myself to developing a Python script capable of efficiently cleaning and formatting the data contained within the master Excel sheet, specifically focusing on the version intended for office use. This script employs advanced techniques to address common data anomalies such as missing values, inconsistent formatting, and duplicate entries.

Furthermore, the script facilitates seamless integration of updated information into the office-specific Excel file, ensuring that stakeholders have access to the most accurate and up-to-date data for decision-making and strategic planning.

By automating this data cleaning and formatting process, we not only optimize efficiency but also enhance the reliability and usability of the information available to our teams. This proactive approach aligns with our commitment to leveraging technology to streamline operations and drive sustainable growth.

Screenshots

```python
import csv
def process_csv(input_file, output_file):
    with open(input_file, 'r', encoding='utf-8') as f_in, \
         open(output_file, 'w', newline='', encoding='utf-8') as f_out:
        reader = csv.DictReader(f_in)
        fieldnames = reader.fieldnames
        replacements = {
            'address': ('|', ','),
            'email': ('|', ''),
            'phone': ('|', ','),
            'address': ('â€“', '-'),
        }
        writer = csv.DictWriter(f_out, fieldnames=fieldnames)
        writer.writeheader()
        for row in reader:
            for column, value in row.items():
                row[column] = value.replace('&amp;', '&')
                row[column] = row[column].replace('&Amp;', '&')
                if column in replacements:
                    old_value, new_value = replacements[column]
                    row[column] = row[column].replace(old_value, new_value)
                row[column] = row[column].replace('â€“', '-')
            writer.writerow(row)
input_file = r'C:\Users\Admin\Desktop\Python\Automation/merged_export.csv'
output_file = 'merged_export2.csv'
process_csv(input_file, output_file)
```

In the address column, phone, and email, the "|" character was present, which needed replacement since none of these columns originally contained "|", as verified from the portal. I replaced "|" with "," for phone and address, and with " " for the email column.

```python
import pandas as pd
import numpy as np
import os
import io
```
[8] Python

```python
file_path = r'C:\Hotel Master/12_2.csv'
```
[13] Python

```python
df = pd.read_csv(file_path, sep='|')
```
[14] Python

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_4100\985970564.py:1: DtypeWarning: Columns (12,13) have mixed types. Specify dtype option on import or set low_memory=
  df = pd.read_csv(file_path, sep='|')
```

```python
df.to_csv("export2.csv")
```
[15] Python

CHAPTER: 6 WEEK-5 PROGRESS

CHAPTER 6 WEEK-5 PROGRESS

6.1 Data Cleaning – III

In the preceding week, my responsibilities encompassed the handling of a master sheet provided to me, housing critical data regarding our collaboration with numerous hotels. This master sheet, consisting of approximately 17 columns, served as the foundation for creating two distinct sub-files: one tailored for internal office utilization and the other meticulously crafted for the convenience of our Agents and Suppliers.

While constructing the sub-file intended for Agent/Supplier use, careful attention was paid to customizing the data to suit their operational requirements. This entailed the removal of redundant columns that held no relevance to their workflows, thereby streamlining their access to pertinent information. Simultaneously, additional columns such as country code were incorporated to enhance the comprehensiveness of the dataset, ensuring that Agents and Suppliers have access to all requisite details for efficient decision-making.

A paramount consideration during this process was the avoidance of UTF-8 characters within the dataset, as their presence could potentially disrupt the seamless integration of data into the agents' databases. Characters such as '/n' pose a particular risk, as they have the potential to break lines and disrupt the integrity of entire rows within the Excel file. Thus, meticulous measures were implemented to meticulously scrub the dataset of any such characters, safeguarding the integrity and usability of the data for our Agents and Suppliers.

```
import pandas as pd
df = pd.read_csv(r'c:\Hotel Master\merged_export_modified2.csv')
df['Hotel_Name'] = df['Hotel_Name'].str.replace('/', '-')
columns_to_drop = ['GIATA_ID', 'Hotel_Old_Name', 'EMail', 'Phone', 'Create_Date', 'Last_Update_Date', 'status','Unnamed: 0']
df.drop(columns=columns_to_drop, inplace=True)
new_column_names = {
    'RezLive_Hotel_Code': 'HotelCode',
    'Hotel_Name': 'Name',
    'Address': 'HotelAddress',
    'City_Code': 'CityId',
    'Zip_Code': 'HotelPostalCode',
    'Country': 'CountryId'
}
df.rename(columns=new_column_names, inplace=True)
df.head()
df.to_csv('Hotel_Master_27-02-2024_no_code.csv', index=False)
```

Several columns were removed, and some were renamed to align with the format suitable for Agents' use.

CHAPTER: 7 WEEK-6 PROGRESS

CHAPTER 7 WEEK-6 PROGRESS

7.1 Data Cleaning – IV

In the sixth week of my tenure, I embarked on a mission to revolutionize our data management process by implementing a comprehensive automation solution. At the heart of this endeavor lay the task of extracting pertinent data from a master file and transforming it into two distinct formats—one tailored for internal office use and the other finely tuned to meet the specific needs of our agents.

To achieve this ambitious goal, I meticulously crafted a sophisticated Python script capable of seamlessly integrating with our FTP server infrastructure. The script's first order of business was to retrieve the master file from the FTP server, laying the groundwork for subsequent processing steps.

With the master file in hand, the script swung into action, orchestrating a symphony of data manipulation techniques to ensure that the extracted data met the unique requirements of both our internal office operations and our agent network. For the office-specific file, the script meticulously formatted the data, arranging columns, and standardizing entries to optimize usability and facilitate streamlined analysis for our internal teams.

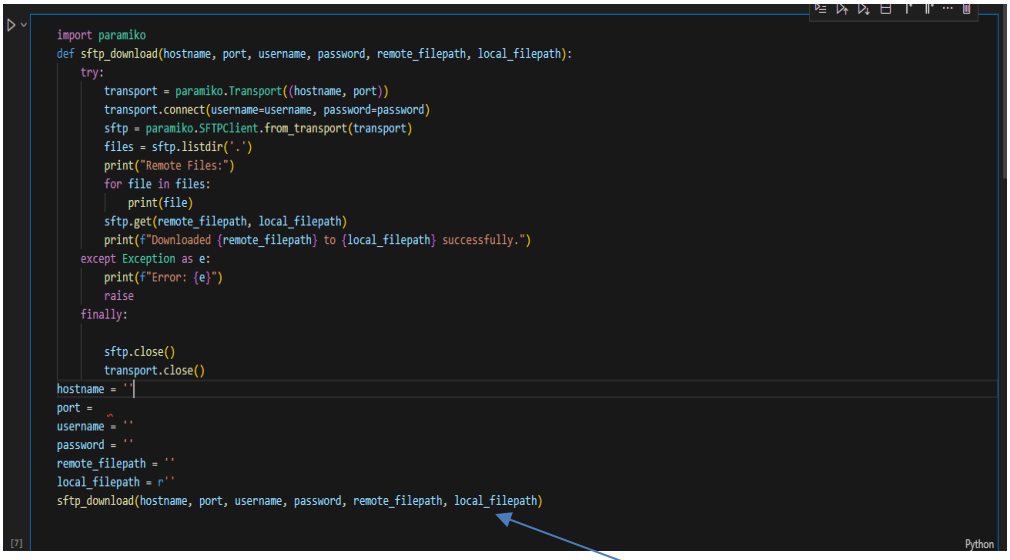
Once the office-focused formatting was complete, the script pivoted its attention to the agent-specific version of the file. Here, the challenge lay in tailoring the data to suit the specific workflows and preferences of our agents. This involved not only removing extraneous columns but also incorporating additional fields, such as country codes, to enhance the comprehensiveness and utility of the dataset for our agent partners.

But the script's work didn't stop there. Recognizing the importance of data integrity and security, particularly in transit, the script implemented stringent cleaning procedures to ensure that the final files were free of any UTF-8 characters that could potentially disrupt data integrity upon transmission. Characters such as '/n' were systematically eradicated to safeguard the seamless integration of the data into our agents' databases.

With the data thoroughly formatted and cleansed, the script seamlessly transitioned to the final phase of the automation process: packaging the files into a compressed zip archive. This compression not only reduced file size but also streamlined the uploading process, ensuring swift and efficient transmission back to the FTP server for distribution to our agents.

In essence, this comprehensive automation solution represents a quantum leap forward in our data management capabilities. By harnessing the power of Python scripting and leveraging our existing infrastructure, we've not only optimized efficiency but also elevated the reliability, consistency, and security of our data handling processes. This strategic investment in automation reaffirms our commitment to excellence and positions us for continued success in an increasingly data-driven landscape.

Screenshots:



```python
import paramiko
def sftp_download(hostname, port, username, password, remote_filepath, local_filepath):
    try:
        transport = paramiko.Transport((hostname, port))
        transport.connect(username=username, password=password)
        sftp = paramiko.SFTPClient.from_transport(transport)
        files = sftp.listdir('.')
        print("Remote Files:")
        for file in files:
            print(file)
        sftp.get(remote_filepath, local_filepath)
        print(f"Downloaded {remote_filepath} to {local_filepath} successfully.")
    except Exception as e:
        print(f"Error: {e}")
        raise
    finally:

        sftp.close()
        transport.close()
hostname = ''
port =
username = ''
password = ''
remote_filepath = ''
local_filepath = r''
sftp_download(hostname, port, username, password, remote_filepath, local_filepath)
```

Code to connect FTP server

```python
import pandas as pd
df_hotels = pd.read_csv(r'c:\Hotel Master\Hotel_Master_27-02-2024_no_code.csv')
df_countries = pd.read_csv(r'c:\Hotel Master\CountryCode.csv')

df_merged = pd.merge(df_hotels, df_countries, on='CountryId', how='left')
df_merged.loc[df_merged['CountryId'] == 'Namibia', 'CountryCode'] = 'NA'
columns = list(df_merged.columns)
country_code_index = columns.index('CountryCode')
columns.pop(country_code_index)

country_id_index = columns.index('CountryId')
columns.insert(country_id_index + 1, 'CountryCode')

df_merged = df_merged[columns]

output_file = r'c:\Hotel Master\hotels.csv'
df_merged.to_csv(output_file, index=False)

print(df_merged.head())
```

This code snippet is a segment of the process for transforming the Excel file to suit the agents' requirements, which involves appending country codes to match the format of the agents' file.

CHAPTER: 8 WEEK-7 PROGRESS

CHAPTER 8 WEEK-7 PROGRESS

8.1 Threat Detection Dashboard

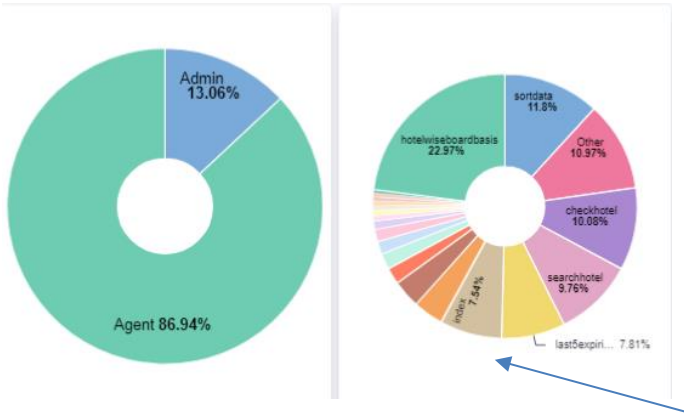
During the seventh week, our company faced a severe cyber attack targeting our database, prompting urgent measures to fortify our defenses against future breaches. As part of the comprehensive response effort, I was tasked with a critical assignment: developing a dashboard to provide higher authorities with real-time insights into user activities within our systems.

This dashboard served as a vital tool for monitoring and analyzing user behavior, enabling swift detection of any suspicious or unauthorized activities. It provided granular details such as user actions, timestamps, associated usernames, and IP addresses, empowering decision-makers to identify potential security threats and take proactive measures to mitigate risks.

The development of this dashboard involved a meticulous process of data collection, aggregation, and visualization. Leveraging cutting-edge technologies and best practices in cybersecurity, I engineered a robust solution capable of providing comprehensive visibility into user interactions across our systems.

By centralizing this information in a user-friendly dashboard interface, I aimed to enhance transparency and accountability within our organization's cybersecurity framework. This proactive approach not only strengthens our defenses against future attacks but also fosters a culture of vigilance and responsiveness to emerging threats.

As we continue to navigate the ever-evolving landscape of cybersecurity, the implementation of this dashboard stands as a testament to our commitment to safeguarding sensitive data and preserving the trust of our stakeholders. Through ongoing monitoring and analysis, we remain vigilant in our efforts to uphold the highest standards of security and resilience in the face of evolving cyber threats.

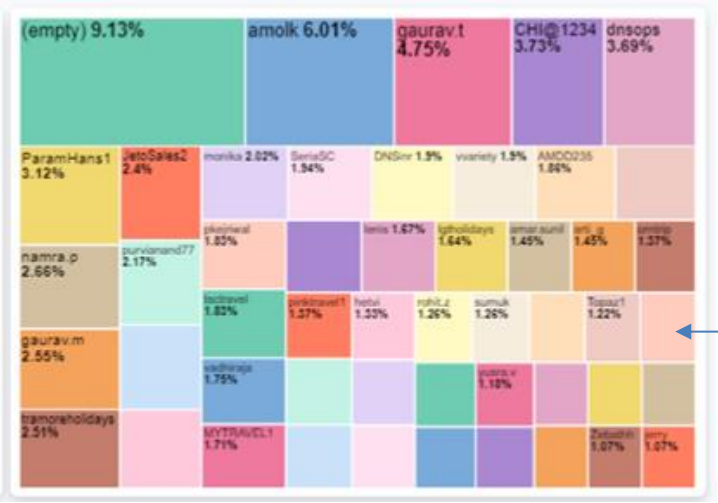


Here are two pie charts illustrating user types, distinguishing between Admin and Agent, and actions performed by users.

IP_Address.keyword	timestamp_ist per 30 seconds	Count of records
115.242.149.___	12:34:00	2
115.242.149.___	12:34:30	11
115.242.149.___	12:35:00	11
115.242.149.___	12:35:30	22
115.242.149.___	12:36:00	19
115.242.149.___	12:36:30	18
115.242.149.___	12:37:00	28
115.242.149.___	12:37:30	34
115.242.149.___	12:38:00	19
115.242.149.___	12:38:30	21
115.242.149.___	12:39:00	12
115.242.149.___	12:39:30	16
115.242.149.___	12:40:00	22
115.242.149.___	12:40:30	17
115.242.149.___	12:41:00	20
115.242.149.___	12:41:30	30
115.242.149.___	12:42:00	22
115.242.149.___	12:42:30	22
115.242.149.___	12:43:00	28
115.242.149.___	12:43:30	25

IP_Address.keyword	Top 3 values of usern	timestamp_	Count of records
119.160.128.__	A	12:55:00	13
119.160.128.__	A	12:55:30	1
119.160.128.__	A	12:56:00	null
119.160.128.__	A	12:56:30	null
119.160.128.__	A	12:57:00	null
119.160.128.__	A	12:57:30	null
119.160.128.__	A	12:58:00	1
119.160.128.__	A	12:58:30	null
119.160.128.__	A	12:59:00	null
119.160.128.__	A	12:59:30	1
119.160.128.__	A	13:00:00	5
119.160.128.	A	13:00:30	null
119.160.128.__	A	13:01:00	null
119.160.128.__	A	13:01:30	null
119.160.128.	A	13:02:00	null
119.160.128.__	A	13:02:30	null
119.160.128.__	A	13:03:00	null
119.160.128.__	A	13:03:30	null
119.160.128.__	A	13:04:00	null
119.160.128.__	A	13:04:30	null
119.160.128.__	A	13:05:00	null
119.160.128.__	A	13:05:30	2
119.160.128.__	A	13:06:00	4
119.160.128.__	A	13:06:30	1
119.160.128.__	A	13:07:00	28
119.160.128.__	A	13:07:30	49

The left image showcases a table or lens created within the dashboard, displaying Office IP addresses in India and Dubai. On the right side, the image illustrates IP addresses outside the organization, aiding in tracking external IPs effectively.



This dashboard displays the usernames of individuals with the highest number of actions performed.

CHAPTER: 9 WEEK-8 PROGRESS

CHAPTER 9 WEEK-8 PROGRESS

9.1 Machine Learning model.

During the eighth week, I spearheaded an initiative to develop a machine learning model using linear regression to forecast future booking trends for agents in the travel industry. Drawing from a dataset spanning the previous month, encompassing agent codes, daily booking metrics, and total monthly booking figures, our aim was to provide predictive insights into upcoming booking patterns. However, as we delved deeper into the intricacies of the travel sector, we came to realize the inherent challenges in accurately predicting booking trends. The industry's dynamic nature, coupled with various external factors such as seasonal changes, economic fluctuations, and unforeseen events, renders traditional forecasting methods ineffective.

This realization prompted a strategic pivot in our approach. Rather than relying solely on predictive models, we recognized the importance of agility and adaptability in responding to the rapidly evolving landscape of travel demand. While machine learning techniques offer valuable insights, we acknowledged that the travel industry is uniquely susceptible to sudden shifts in consumer behavior. Factors such as sudden changes in travel advisories, unexpected weather events, or global crises can swiftly alter booking patterns, making static predictions obsolete.

In light of this, our focus shifted towards leveraging real-time analytics and adopting a more dynamic decision-making framework. By closely monitoring market trends, analyzing consumer behavior patterns, and remaining responsive to emerging developments, we position ourselves to make informed decisions in a timely manner. This shift towards agile data-driven decision-making underscores our commitment to innovation and our proactive stance in navigating the complexities of the travel industry.

CHAPTER: 10 WEEK-9
PROGRESS

CHAPTER 10 WEEK-9 PROGRESS

10.1 Time Analysis Dashboard

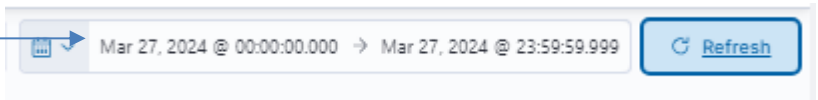
During the ninth week, I was assigned the task of addressing a concerning drop in searches and bookings by seven of our top-performing agents. Following a discussion between senior management and these agents to understand the underlying reasons for the decline, I was tasked with creating a dashboard to closely monitor and analyze their search activity.

This dashboard needed to provide a comprehensive overview of recent search volumes, allowing for comparison over time to identify any notable fluctuations. Additionally, it required a feature to drill down into the data to examine search activity on an hourly basis, enabling us to pinpoint specific trends and patterns.

Each day, I was responsible for generating reports detailing any fluctuations in search activity and providing insights into potential factors driving these changes. This proactive approach aimed to keep the agents informed and empowered to address any issues impacting their search performance promptly.

Overall, this task underscored the importance of proactive monitoring and communication in responding to shifts in agent activity, ensuring that we can swiftly identify and address any challenges that may arise.

This dashboard showcases the total number of searches conducted by agents, along with the corresponding country and nationality. The total searches are presented for each hour.



CHAPTER: 11 WEEK-10
PROGRESS

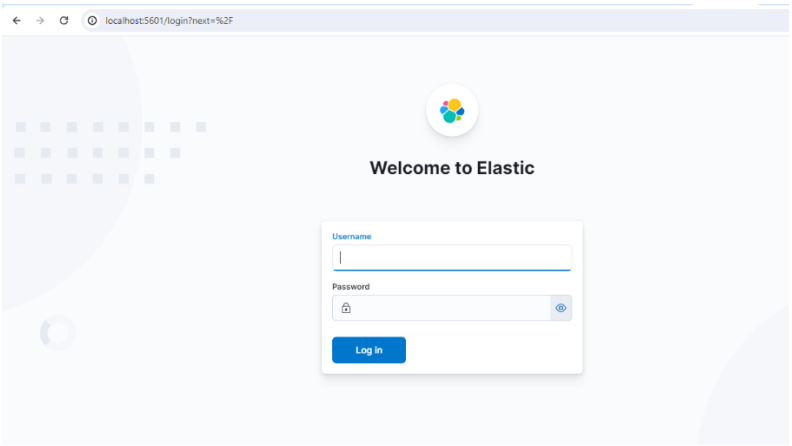
CHAPTER 11 WEEK-10 PROGRESS

11.1 Configuring ElK stack Locally

In the tenth week, we embarked on a comprehensive initiative aimed at emulating the office environment within a localized context. Recognizing the importance of meticulously replicating each aspect of our operational setup, we adopted a systematic approach to achieve this goal. The first pivotal step in this endeavor involved the installation and configuration of the ELK (Elasticsearch, Logstash, Kibana) stack within our local infrastructure.

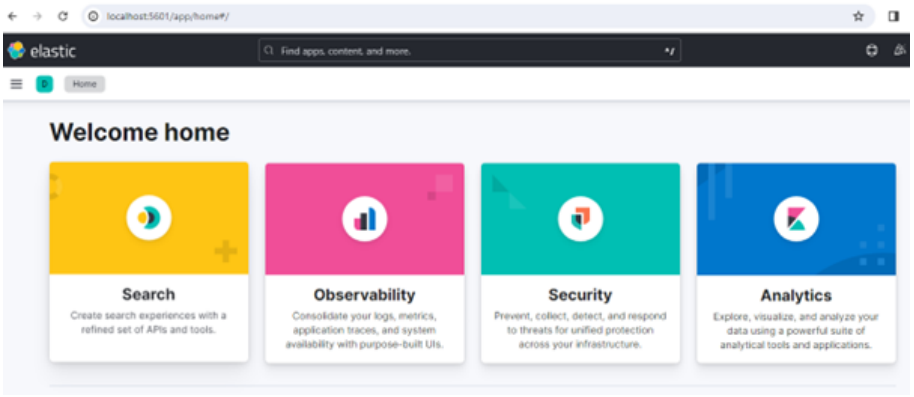
This process necessitated careful attention to detail, as the successful deployment and setup of the ELK stack were crucial for facilitating seamless data processing and analysis akin to our office environment. Leveraging our expertise and resources, we meticulously configured each component of the ELK stack to ensure optimal performance and functionality.

Once the installation and configuration were complete, the next imperative was to validate the functionality of the ELK stack. To accomplish this, we utilized pre-existing data provided by ELK, allowing us to simulate real-world scenarios and gauge the system's responsiveness and effectiveness. By thoroughly testing the ELK stack against this dataset, we aimed to verify its capability to handle and analyze data in a manner consistent with our office environment.



Use Following Commands in cmd to run ELK locally:

- Elasticsearch.bat
- Kibana.bat
- Logstash -f logstash.conf



CHAPTER: 12 WEEK-11
PROGRESS

CHAPTER 12 WEEK-11 PROGRESS

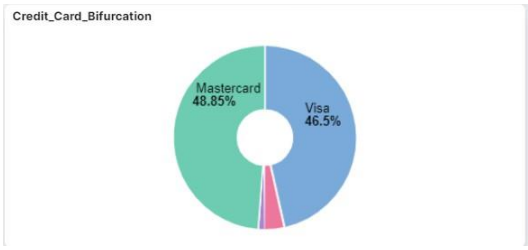
12.1 Risk analysis Dashboard

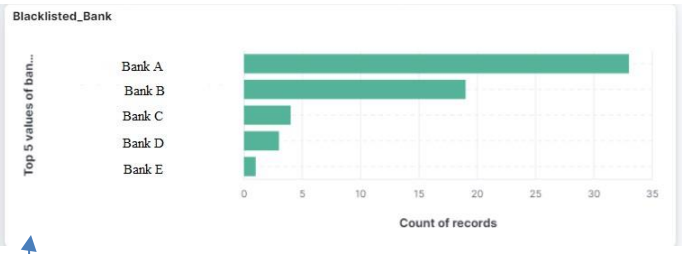
In the 11th week, I was assigned the task of developing a dashboard to monitor the ratio of risky transactions to non-risky transactions. This involved compiling a list of blacklisted banks and implementing the dashboard functionality. Before deploying the system to the live server, we opted to test it locally to ensure smooth operation. After several iterations and adjustments to field mappings, we successfully crafted a dashboard capable of identifying the volume of risky transactions.

```
PUT /card_details
{
  "settings": {
    "number_of_shards": 1,
    "number_of_replicas": 1
  },
  "mappings": {
    "properties": {
      "ConsultantName": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "agency_name": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
```

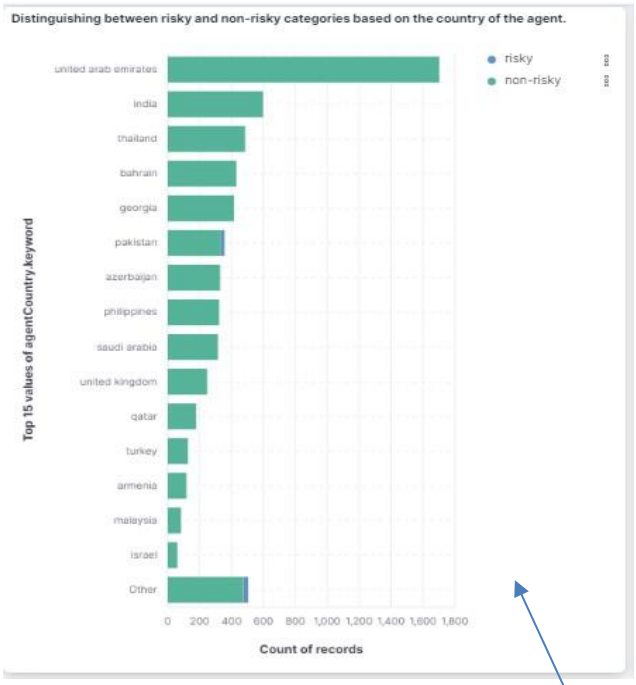
This excerpt pertains to the data type mapping of a specific field, indicating its nature within the system.

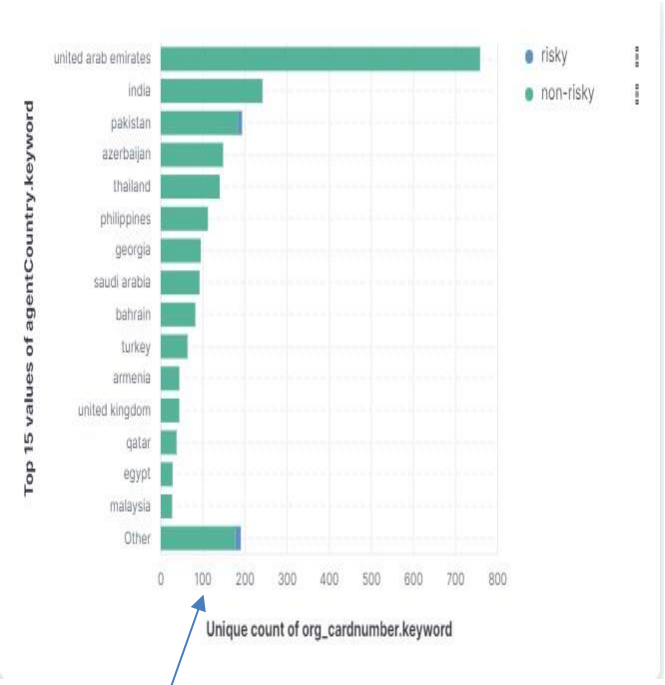
The provided images showcase the overall transaction count, a breakdown of risky versus non-risky transactions depicted through a pie chart, as well as individual counts for both risky and non-risky transactions.





Above picture shows the bifurcation of type of card used using pie chart and Total number of transaction done under blacklisted banks.





Three sets of bar charts were produced in a uniform style to distinguish between risky and non-risky transactions across different parameters: country of the agent, agent code, and consultant name, with transaction count as the measure. Additionally, an identical set of three graphs was replicated, with the only distinction being the metric used, which was the unique number of cards used.

CHAPTER: 13 WEEK-12
PROGRESS

CHAPTER 13 WEEK-12 PROGRESS

13.1 Risk analysis Dashboard on Live server

We replicated the dashboard and conducted the final adjustments directly on the live server. While transitioning to the live environment, we encountered several issues that initially disrupted the accuracy of the data presentation. Through diligent troubleshooting and resolution efforts, we addressed these challenges and ensured that the dashboard functioned as intended. With the adjustments in place, we proceeded to make the dashboard accessible for internal use, providing our team with a reliable tool for monitoring and analysis.

This query within the development tool of the ELK stack facilitates the complete removal of data. It serves as a fail-safe mechanism to eradicate any data that may not have been retrieved accurately, ensuring a clean slate for subsequent operations.

CHAPTER: 14 WEEK-13
PROGRESS

CHAPTER 14 WEEK-13 PROGRESS

14.1 Data Cleaning for Cities.csv

During Week 13, I undertook the responsibility of cleansing and transforming the "cites.csv" file. This file boasted approximately 220 columns, necessitating a meticulous approach to streamline its structure according to the specified format. As part of this transformation, I identified and retained only four pertinent columns, discarding the remaining ones to streamline the dataset. Additionally, the prescribed format stipulated the use of a pipeline separator, which I meticulously integrated into the data. Leveraging Python, I successfully executed these tasks, ensuring the file adhered to the designated format and met the project requirements.

```python
import pandas as pd


file_path = 'C:/Hotel Master/Cities.csv'
df = pd.read_csv(file_path,dtype={'country_code': str},na_values=[], keep_default_na=False)
columns_to_drop = ['is_preferredcity', 'destination_id', 'country_id', 'state_code', 'new_state_name', 'status', 'expedia_code', 'hotels

df.drop(columns=columns_to_drop, inplace=True)



combined_header = '|'.join(df.columns.tolist())
df[combined_header] = df.apply(lambda row: '|'.join(row.astype(str)), axis=1)


df_final = df.iloc[:, 4:]


output_nodified_path = 'C:/Users/Admin/Desktop/Python/Automation/cities.csv'
df_final.to_csv(output_nodified_path, index=False)
```

```python
import zipfile
import os

def csv_to_zip(csv_file, zip_file):
    with zipfile.ZipFile(zip_file, 'w') as zf:
        zf.write(csv_file, os.path.basename(csv_file))

csv_file = 'C:/Users/Admin/Desktop/Python/Automation/cities\.csv'
zip_file = 'C:/Users/Admin/Desktop/Python/Automation/cities.zip'

csv_to_zip(csv_file, zip_file)
```

Following the successful cleansing and transformation of the "cites.csv" file, I proceeded to compress it into a ZIP archive as per the project requirements. Subsequently, I uploaded the ZIP file to the designated location as instructed.

CHAPTER: 15 WEEK-14
PROGRESS

CHAPTER 15 WEEK-14 PROGRESS

15.1 Updating Risk Analysis Dashboard

In the fourteenth week, an imperative task emerged: the refinement of our Risk Analysis Dashboard following an assessment from a senior. This evaluation brought to light the necessity for augmenting the dashboard with additional analytical perspectives, one of which centered on the total transaction amount. However, our endeavors encountered a formidable challenge: the data type assigned to the amount field was text. This posed a significant hindrance as it prevented us from accurately computing the sum of transaction amounts per country—a vital metric for our risk assessment. To overcome this obstacle and ensure the integrity of our analysis, we made the strategic decision to initiate a data cleansing process. This involved clearing the existing dataset and subsequently fetching fresh data. Additionally, we implemented a crucial modification by redefining the mapping of the amount field, transitioning it from a text to a double data type. This adjustment not only facilitated seamless summation operations but also enhanced the overall robustness and accuracy of our Risk Analysis Dashboard, aligning it more closely with the evolving needs and expectations of our senior.

CHAPTER: 16 WEEK-15
PROGRESS

CHAPTER 16 WEEK-15 PROGRESS

16.1 Documentation

During the final week of my internship, I was tasked with compiling comprehensive documentation detailing all the projects and tasks I completed throughout my tenure at the company. This documentation aimed to serve as a valuable resource for anyone seeking to understand the work I undertook during my internship and facilitate knowledge transfer for future reference. Additionally, I was assigned the responsibility of orienting a new incoming intern by providing an overview of the tasks I accomplished thus far, enabling them to seamlessly continue the projects and initiatives I had been working on.

CHAPTER: 17 CONCLUSION AND DISCUSSION

CHAPTER 17 CONCLUSION AND DISCUSSION

17.1 Conclusion

In conclusion, the organization's journey towards enhancing data analysis capabilities has been transformative, marked by the mastery of tools like the ELK stack and Excel for insightful dashboard creation and data cleaning, respectively. The pursuit of automation through Python scripting promises to revolutionize workflows, optimizing efficiency and time savings. Additionally, the integration of Pentaho for ETL operations enriches the analytical toolkit, emphasizing the commitment to leveraging technology for operational excellence. The Python scripts developed play a pivotal role in this journey, automating tasks and freeing up time for more strategic analysis, thus contributing significantly to the organization's data analysis capabilities. This commitment to innovation underscores the organization's dedication to driving operational excellence, poised for further growth and learning as it continuously refines analytical processes to meet evolving challenges and achieve greater success.

Plagiarism Report :