INTERNSHIP PROJECT PHASE- RESEARCH

PROJECT 6 – DEVELOPMENT AND TESTING OF YELP DATA ANALYSIS PROJECT ON IBM CLOUD

1 Introduction

The digital age has seen a significant surge in user-generated content, especially in the form of online reviews and check-ins. Yelp, one of the leading review platforms, has accumulated vast amounts of data about local businesses, users, and their interactions. This project focuses on analyzing Yelp's business and check-in datasets to extract meaningful insights about user behavior, business distribution, and engagement patterns. The implementation transitions this analysis to IBM Cloud, leveraging its powerful data science and machine learning tools.

2 Project Objectives and Dataset Overview

Objectives:

- Perform exploratory data analysis (EDA) on Yelp's business and check-in data.
- Identify key patterns in business categories, locations, and customer check-ins.
- Visualize and summarize data trends.

3 Dataset Overview:

1.yelp_academic_dataset_business.json: Contains details about businesses such as name, location, star rating, review count, and categories.

2.yelp_academic_dataset_checkin.json: Records customer check-in times for the businesses.

The datasets are semi-structured in JSON format and are read using the pandas library in Python.

4 System Architecture and IBM Cloud Environment

The system architecture consists of the following layers:

Data Storage Layer: IBM Cloud Object Storage to store raw Yelp datasets.

Processing Layer: IBM Watson Studio and IBM Cloud Pak for Data used for data analysis.

Visualization Layer: IBM Cognos Dashboard Embedded or Matplotlib/Seaborn within Jupyter Notebooks.

IBM Cloud Components:

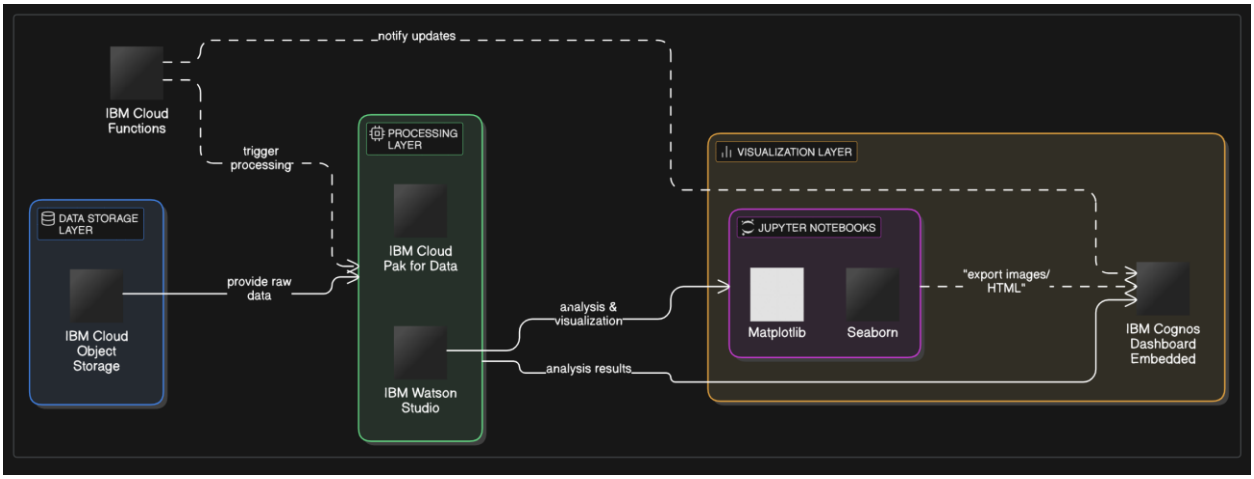
IBM Watson Studio: Integrated environment for data science, allows use of Jupyter Notebooks.

IBM Cloud Object Storage: Used to store and retrieve the dataset.

IBM Cloud Functions (Optional): For automating tasks.

IBM Cognos Dashboard Embedded: For visualizing analysis results.

CLOUD ARCHITECTURE



5 Implementation Steps on IBM Cloud

Step 1: Setup IBM Cloud Account

Create or login to an IBM Cloud account.

Provision IBM Cloud Object Storage instance.

Provision IBM Watson Studio instance.

Step 2: Upload Dataset

Upload yelp_academic_dataset_business.json and yelp_academic_dataset_checkin.json to Object Storage.

Step 3: Create a Watson Studio Project

Launch Watson Studio.

Create a new project.

Associate Object Storage with the project.

Step 4: Data Import in Jupyter Notebook

Create a new Jupyter Notebook in Watson Studio.

Insert a code cell to read the datasets using pandas.read_json().

Example:

```
import pandas as pd

business_df = pd.read_json("path/to/yelp_academic_dataset_business.json", lines=True)

checkin_df = pd.read_json("path/to/yelp_academic_dataset_checkin.json", lines=True)
```

Step 5: Exploratory Data Analysis (EDA)

Use functions like .info(), .describe(), .value_counts() to summarize data.

Check for missing values.

Perform groupby analysis on city, stars, categories.

Example visualizations:

```
import matplotlib.pyplot as plt

import seaborn as sns

sns.countplot(x='stars', data=business_df)

plt.title("Distribution of Business Ratings")
```

Step 6: Analyze Check-In Data

Extract date and time features.

Aggregate check-in counts by day/time.

Merge with business data for enriched analysis.

Step 7: Visualize Results

Use Seaborn/Matplotlib or integrate IBM Cognos Dashboard to build interactive dashboards.

6 IBM Cloud Services and Tools Used

Description:

IBM Cloud Object Storage

To upload and store large Yelp datasets

IBM Watson Studio

Development environment for data analysis and modeling

Jupyter Notebooks

Interactive Python coding for analysis

IBM Cognos Dashboard

Optional: Interactive dashboards for non-technical users

IBM Cloud CLI & Functions

For automation and deployment (optional)

7 Implementation Steps on IBM Cloud

Step 1: Setup IBM Cloud Account

Create or login to an IBM Cloud account.

Provision IBM Cloud Object Storage instance.

Provision IBM Watson Studio instance.

Step 2: Upload Dataset

Upload yelp_academic_dataset_business.json and yelp_academic_dataset_checkin.json to Object Storage.

Step 3: Create a Watson Studio Project

Launch Watson Studio.

Create a new project.

Associate Object Storage with the project.

Step 4: Data Import in Jupyter Notebook

Create a new Jupyter Notebook in Watson Studio.

Insert a code cell to read the datasets using pandas.read_json().

Example:

```
import pandas as pd

business_df = pd.read_json("path/to/yelp_academic_dataset_business.json", lines=True)

checkin_df = pd.read_json("path/to/yelp_academic_dataset_checkin.json", lines=True)
```

Step 5: Exploratory Data Analysis (EDA)

Use functions like .info(), .describe(), .value_counts() to summarize data.

Check for missing values.

Perform groupby analysis on city, stars, categories.

Example visualizations:

```
import matplotlib.pyplot as plt

import seaborn as sns

sns.countplot(x='stars', data=business_df)

plt.title("Distribution of Business Ratings")
```

Step 6: Analyze Check-In Data

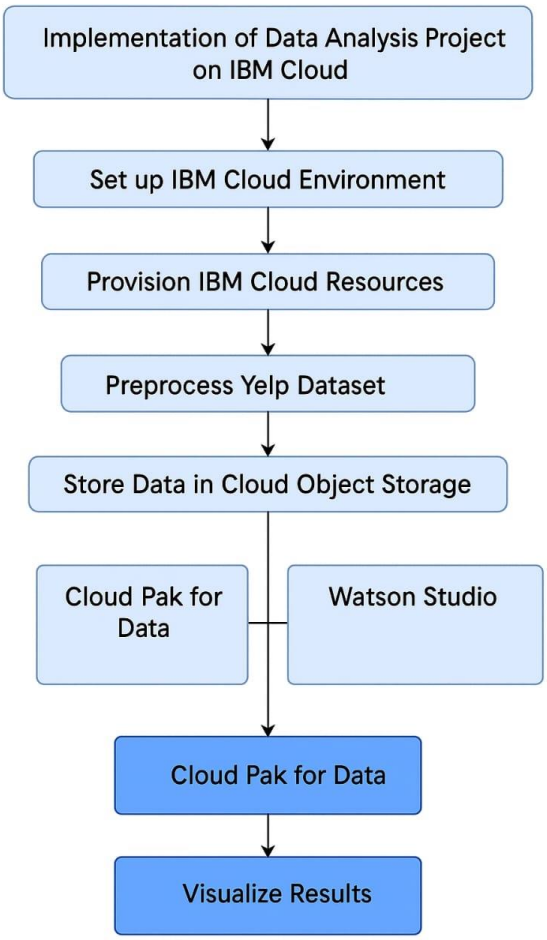
Extract date and time features.

Aggregate check-in counts by day/time.

Merge with business data for enriched analysis.

Step 7: Visualize Results

Use Seaborn/Matplotlib or integrate IBM Cognos Dashboard to build interactive dashboards.



IMPLEMENTATION ON IBM CLOUD

8 IBM Cloud Services and Tools Used

Description:

IBM Cloud Object Storage

To upload and store large Yelp datasets

IBM Watson Studio

Development environment for data analysis and modeling

Jupyter Notebooks

Interactive Python coding for analysis

IBM Cognos Dashboard

Optional: Interactive dashboards for non-technical users

IBM Cloud CLI & Functions

For automation and deployment (optional)

9 Data Analysis and Results Overview

From the initial EDA, the following insights were obtained:

- Businesses are concentrated in certain cities, with a skewed distribution of reviews.
- Categories such as Restaurants, Shopping, and Nightlife dominate.
- Check-ins show peak activity during weekends and evenings.
- There is a correlation between star rating and average check-ins.

10 Conclusion and Future Scope

This project demonstrates how large-scale review and check-in data can be effectively analyzed in a cloud environment like IBM Cloud. Using Watson Studio, Object Storage, and visualization tools, the Yelp dataset was explored for actionable insights. In the future, the project can be extended to:

- Perform sentiment analysis using NLP on user reviews (additional dataset).
- Build predictive models for customer footfall.
- Deploy dashboards using IBM Cloud Functions for real-time analysis.

- The cloud implementation makes the solution scalable, portable, and more accessible to collaborative teams.

REFERENCES:

[1] IBM Cloud Documentation

[2] Yelp Open Dataset

[3] Python pandas/seaborn/matplotlib documentation