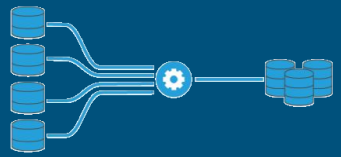
Generation

# Scalable-ETL-development for The café using AWS cloud



## Team 2 Group Project

De-Lon 6

a

# Contents

- Meet the Team
- Problem statement
- Planning
- Technologies used
- Architecture
- Schema
- Extract / Clean
- Transform
- Load
- CI/CD architecture
- CI/CD using codeBuild
- Successful implementation
- CI/CD GitHub Integration - AWS Lambda
- Analysis & Trends (Grafana)
- Monitoring Infrastructure
- Insights / Maximum Viable Product (MVPs)
- What?  Where?  When?
- Wrap up

a

# Problem Statement

Client: Pop-up café

- Who want us to help them log and track orders
- To track transactions across all outlets for new and returning customers
- Identify the latest trends and make business decisions, to maximise revenue and profits
- Current software has limitations and is time consuming to gather reports from all branches

Our Approach

- Building a fully scalable ETL (Extract, Transform, Load) pipeline to handle large volumes of transaction data from multiple stores
- Use Grafana for data visualization by querying on the transformed data, enabling the client to identify company-wide trends and insights

# Technologies

**Python:**

Developing our application

**RDBMS:**

Postgresql

**GitHub:**

Source control

**AWS Services:**

ETL from S3 to Redshift Data warehouse

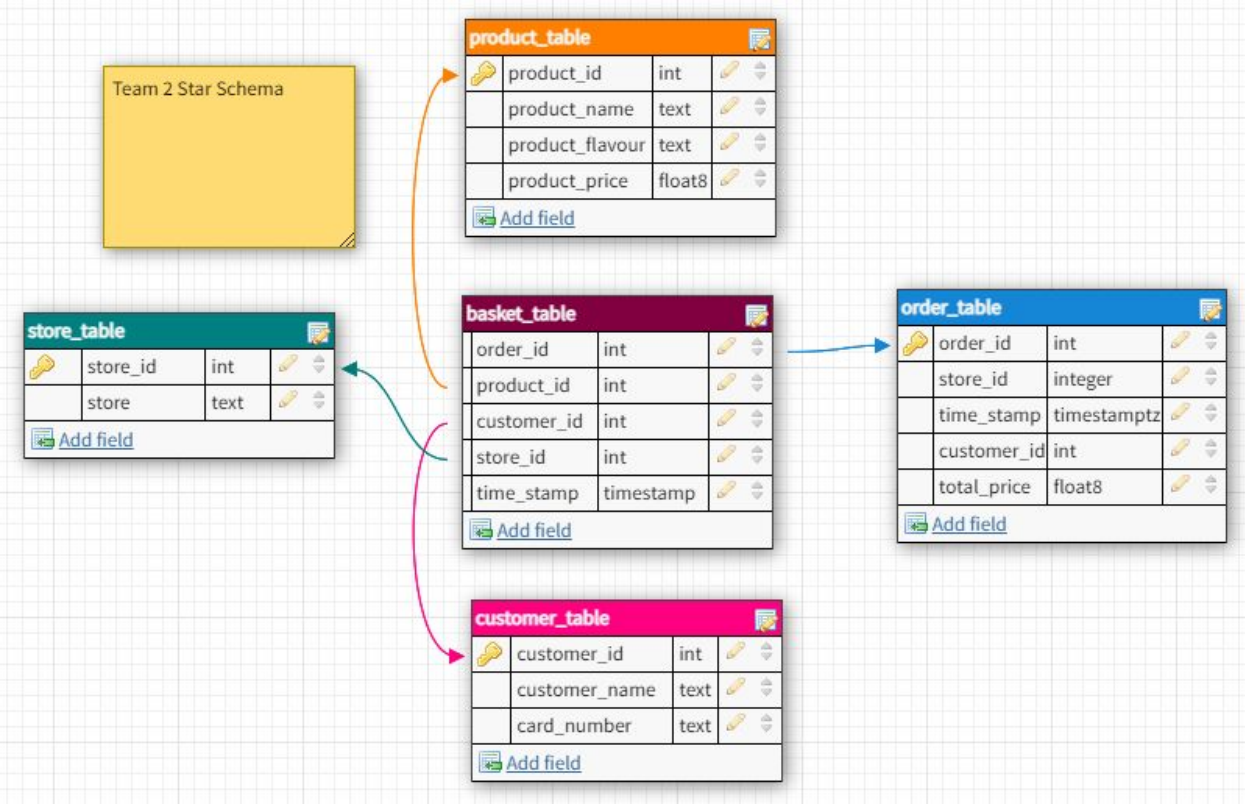
**Grafana:**

Application monitoring & Visualisation

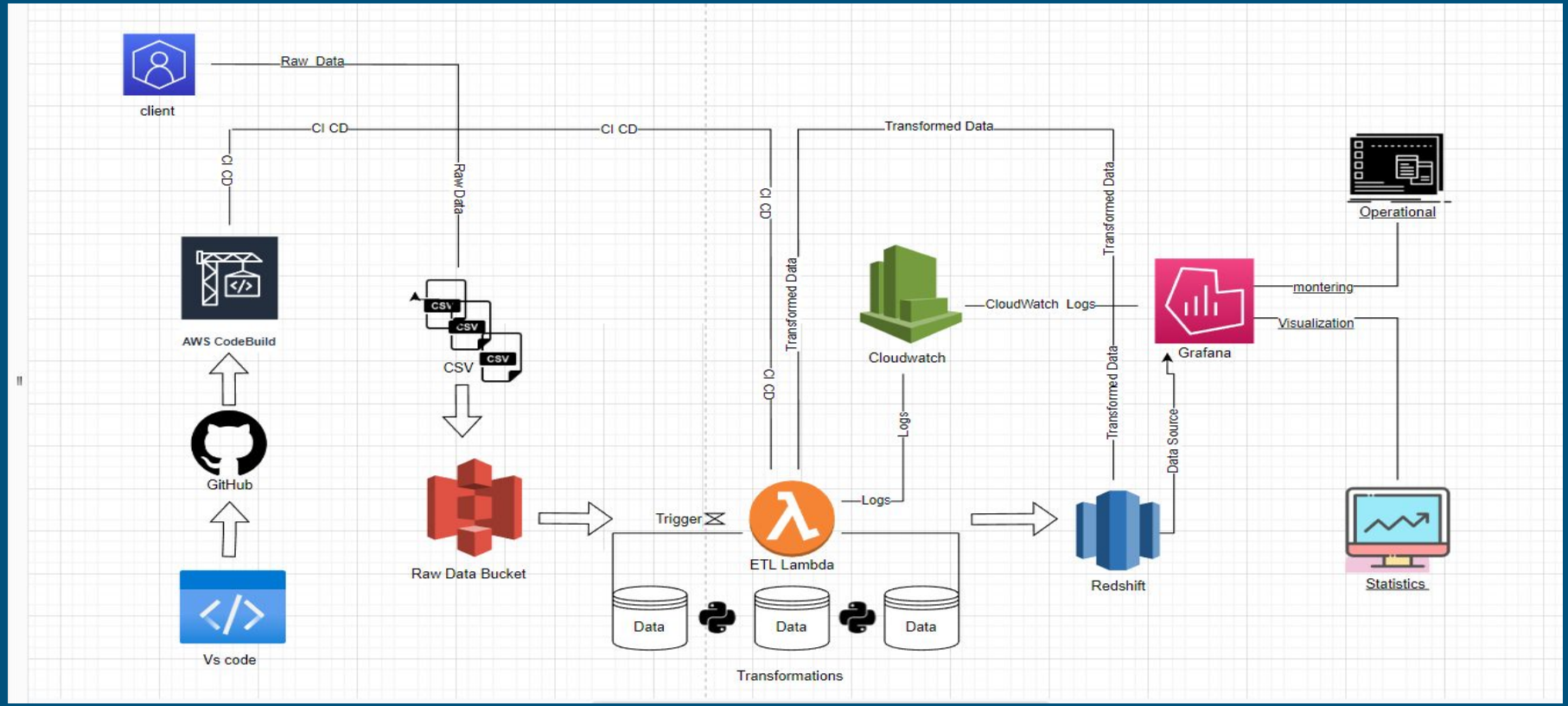
**Trello:**

Agile project management principles - SCRUM sprints

# Star schema


Team 2 Star Schema
product_table
product_id int
product_name text
product_flavour text
product_price float8
Add field
store_table
store_id int
store text
Add field
basket_table
order_id int
product_id int
customer_id int
store_id int
time_stamp timestamp
Add field
order_table
order_id int
store_id integer
time_stamp timestamptz
customer_id int
total_price float8
Add field
customer_table
customer_id int
customer_name text
card_number text
Add field

# ETL Pipeline Architecture




a

# Extract / Clean

Raw data extracted from CSV file & read as dataframe (Data Ingestion)



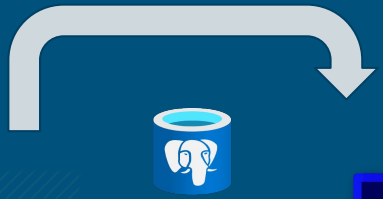

```python
import pandas as pd

file = 'chesterfield_10-06-2022_09-00-00.csv'
fn = ['timestamp', 'store', 'customer_name',
                    'basket_items', 'total_price', 'cash_or_card', 'card_number']
df = pd.read_csv(file, names =fn )

print(df)
```

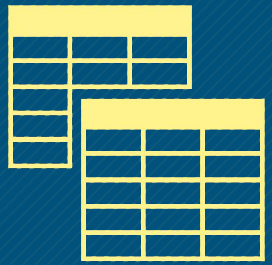
```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  JUPYTER

Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\ahmed\Desktop\testing> & C:/Python310/python.exe c:/Users/ahmed/Desktop/testing/app.py
           timestamp        store   customer_name                                        basket_items  total_price cash_or_card   card_number
0    10/06/2022 09:00  Chesterfield     Allen Ford  Regular Flavoured iced latte - Hazelnut - 2.75...         10.20         CASH           NaN
1    10/06/2022 09:02  Chesterfield    Nicole Miles  Regular Flavoured iced latte - Caramel - 2.75,...         12.00         CARD  8.435947e+15
2    10/06/2022 09:04  Chesterfield    Arlen Calvert  Regular Latte - 2.15, Regular Flavoured iced l...          4.90         CARD  9.116675e+15
3    10/06/2022 09:06  Chesterfield  Delaine Crosby      Regular Flavoured iced latte - Hazelnut - 2.75          2.75         CARD  6.555526e+15
4    10/06/2022 09:08  Chesterfield  Daniel Pettrey         Large Flavoured iced latte - Vanilla - 3.25          3.25         CASH           NaN
..                ...           ...             ...                                                ...           ...          ...           ...
263  10/06/2022 16:45  Chesterfield    Traci Abeles  Regular Flavoured iced latte - Hazelnut - 2.75...         10.50         CASH           NaN
264  10/06/2022 16:49  Chesterfield    Dante Jackson            Large Flavoured latte - Hazelnut - 2.85          2.85         CASH           NaN
265  10/06/2022 16:52  Chesterfield    Hector Rosel  Large Flavoured latte - Hazelnut - 2.85, Large...         11.10         CASH           NaN
266  10/06/2022 16:56  Chesterfield    Mary Starner                          Large Flat white - 2.45          2.45         CASH           NaN
267  10/06/2022 16:59  Chesterfield    June Paulson  Large Flavoured latte - Hazelnut - 2.85, Regul...          8.15         CASH           NaN

[268 rows x 7 columns]
PS C:\Users\ahmed\Desktop\testing>
```

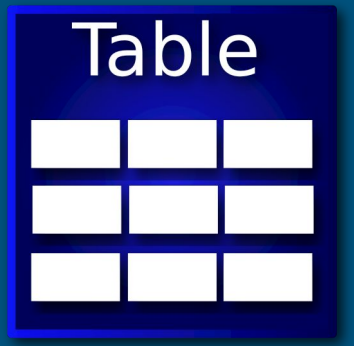
a

# Transform

Dataframe transformed into enriched dataframe with desired schema and respective values






```
import pandas as pd

file = 'chesterfield_10-06-2022_09-00-00.csv'
fn = ['timestamp', 'store', 'customer_name',
      'basket_items', 'total_price', 'cash_or_card', 'card_number']
df = pd.read_csv(file, names =fn )

print(df)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  JUPYTER

Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\ahmed\Desktop\testing> & C:/Python310/python.exe c:/Users/ahmed/Desktop/testing/app.py
            timestamp        store  customer_name                                        basket_items  total_price cash_or_card   card_number
0    10/06/2022 09:00  Chesterfield     Allen Ford  Regular Flavoured Iced Latte - Hazelnut - 2.75,...        10.20         CASH           NaN
1    10/06/2022 09:02  Chesterfield    Nicole Miles  Regular Flavoured Iced Latte - Caramel - 2.75,...        12.00         CARD  8.435947e+15
2    10/06/2022 09:04  Chesterfield   Arlen Calvert  Regular Latte - 2.15, Regular Flavoured Iced L...         4.90         CARD  9.116675e+15
3    10/06/2022 09:06  Chesterfield  Delaine Crosby     Regular Flavoured Iced Latte - Hazelnut - 2.75         2.75         CARD  6.555526e+15
4    10/06/2022 09:08  Chesterfield  Daniel Pettrey          Large Flavoured Iced Latte - Vanilla - 3.25         3.25         CASH           NaN
..               ...           ...             ...                                                ...          ...          ...           ...
263  10/06/2022 16:45  Chesterfield    Traci Abeles  Regular Flavoured Iced Latte - Hazelnut - 2.75...        10.50         CASH           NaN
264  10/06/2022 16:49  Chesterfield   Dante Jackson                Large Flavoured Latte - Hazelnut - 2.85         2.85         CASH           NaN
265  10/06/2022 16:52  Chesterfield   Hector Rosel  Large Flavoured Latte - Hazelnut - 2.85, Large...        11.10         CASH           NaN
266  10/06/2022 16:56  Chesterfield    Mary Starmer                          Large Flat White - 2.45         2.45         CASH           NaN
267  10/06/2022 16:59  Chesterfield    June Paulson  Large Flavoured Latte - Hazelnut - 2.85, Regul...         8.15         CASH           NaN

[268 rows x 7 columns]
PS C:\Users\ahmed\Desktop\testing>
```

# Load

Loading transformed data tables into Redshift using database connection credentials

```
import pandas as pd

file = 'chesterfield_10-06-2022_09-00-00.csv'
fn = ['timestamp', 'store', 'customer_name',
                    'basket_items', 'total_price', 'cash_or_card', 'card_number']
df = pd.read_csv(file, names =fn )

print(df)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  JUPYTER

Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\ahmed\Desktop\testing> & C:/Python310/python.exe c:/Users/ahmed/Desktop/testing/app.py
           timestamp         store   customer_name                                      basket_items  total_price cash_or_card   card_number
0    10/06/2022 09:00  Chesterfield      Allen Ford  Regular Flavoured iced latte - Hazelnut - 2.75...        10.20         CASH           NaN
1    10/06/2022 09:02  Chesterfield    Nicole Miles  Regular Flavoured iced latte - Caramel - 2.75,...        12.00         CARD  8.435947e+15
2    10/06/2022 09:04  Chesterfield   Arlen Calvert  Regular Latte - 2.15, Regular Flavoured iced l...         4.30         CARD  9.116675e+15
3    10/06/2022 09:06  Chesterfield  Delaine Crosby     Regular Flavoured iced latte - Hazelnut - 2.75         2.75         CARD  6.555526e+15
4    10/06/2022 09:08  Chesterfield  Daniel Pettrey        Large Flavoured iced latte - Vanilla - 3.25         3.25         CASH           NaN
..               ...           ...             ...                                                ...          ...          ...           ...
263  10/06/2022 16:45  Chesterfield    Traci Abeles  Regular Flavoured iced latte - Hazelnut - 2.75...        10.50         CASH           NaN
264  10/06/2022 16:49  Chesterfield   Dante Jackson           Large Flavoured latte - Hazelnut - 2.85         2.85         CASH           NaN
265  10/06/2022 16:52  Chesterfield    Hector Rosel  Large Flavoured latte - Hazelnut - 2.85, Large...        11.10         CASH           NaN
266  10/06/2022 16:56  Chesterfield    Mary Starner                        Large Flat white - 2.45         2.45         CASH           NaN
267  10/06/2022 16:59  Chesterfield    June Paulson  Large Flavoured latte - Hazelnut - 2.85, Regul...         8.15         CASH           NaN

[268 rows x 7 columns]
PS C:\Users\ahmed\Desktop\testing>
```

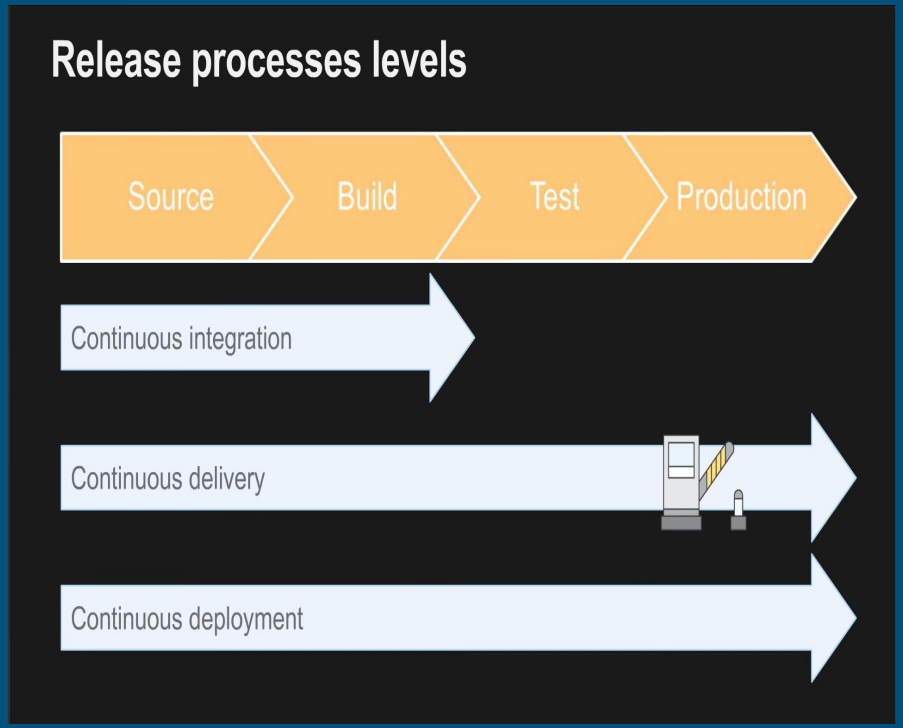
# CI/CD Using CodeBuild

Why CI/CD?

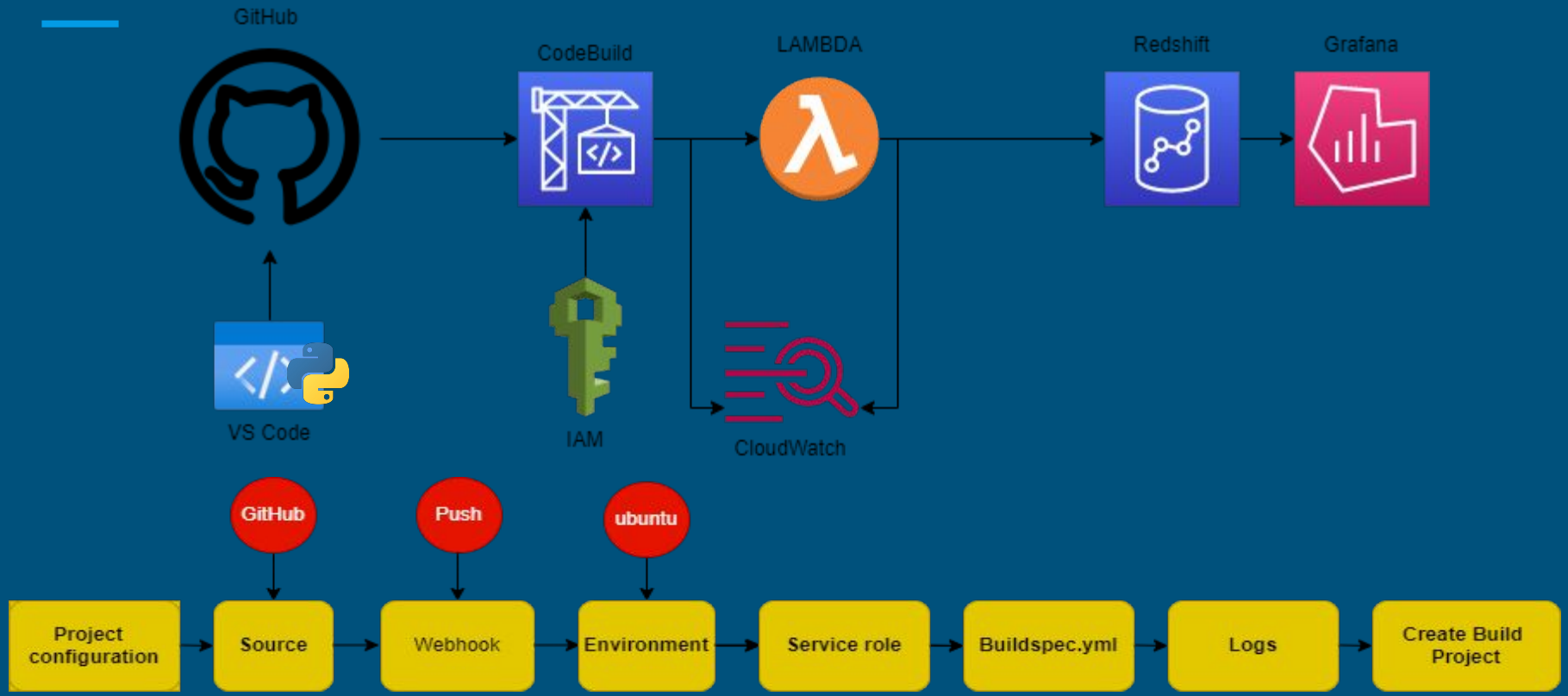
- To deliver a new version of software - series of steps required
- CI/CD automate these steps to improve software delivery throughout the software development life cycle

Why CodeBUILD?

- Fully managed continuous integration service
- Readily deployable software packages
- Continuous scaling and concurrent processing of multiple builds
- Quick start by using prepackaged build environments
- Charged by the minute for the compute



# CI/CD Architecture



Generation

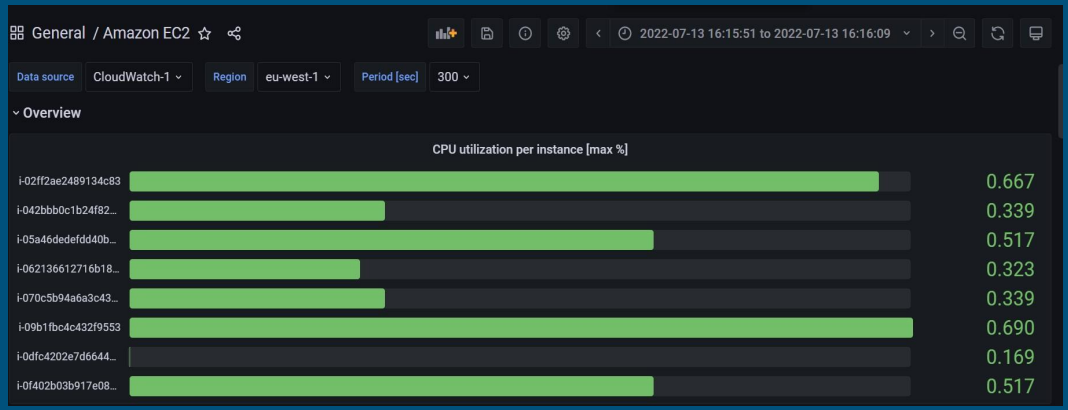
# Analysis & Trends

Grafana configured with Redshift for database access & Cloudwatch for Lambda, EC2 monitoring

# Monitoring Infrastructure

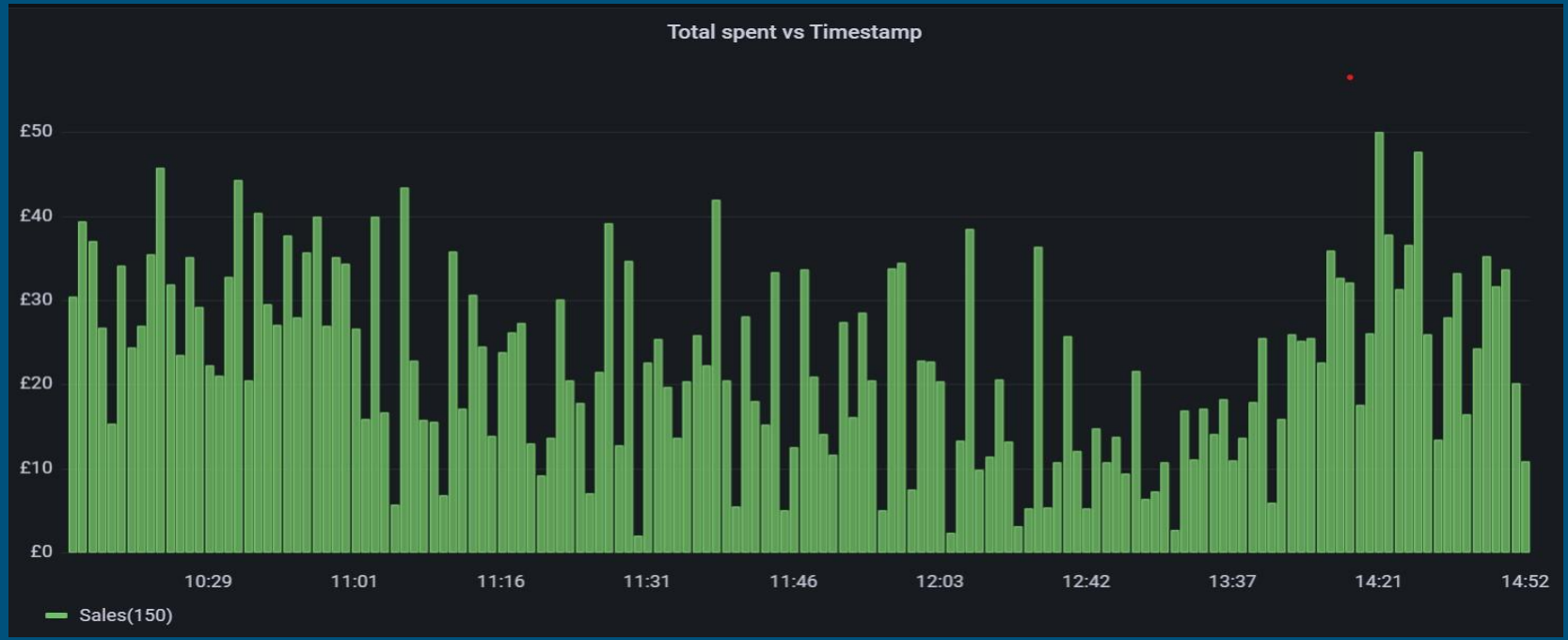

Lambda Durations

EC2 Metrics

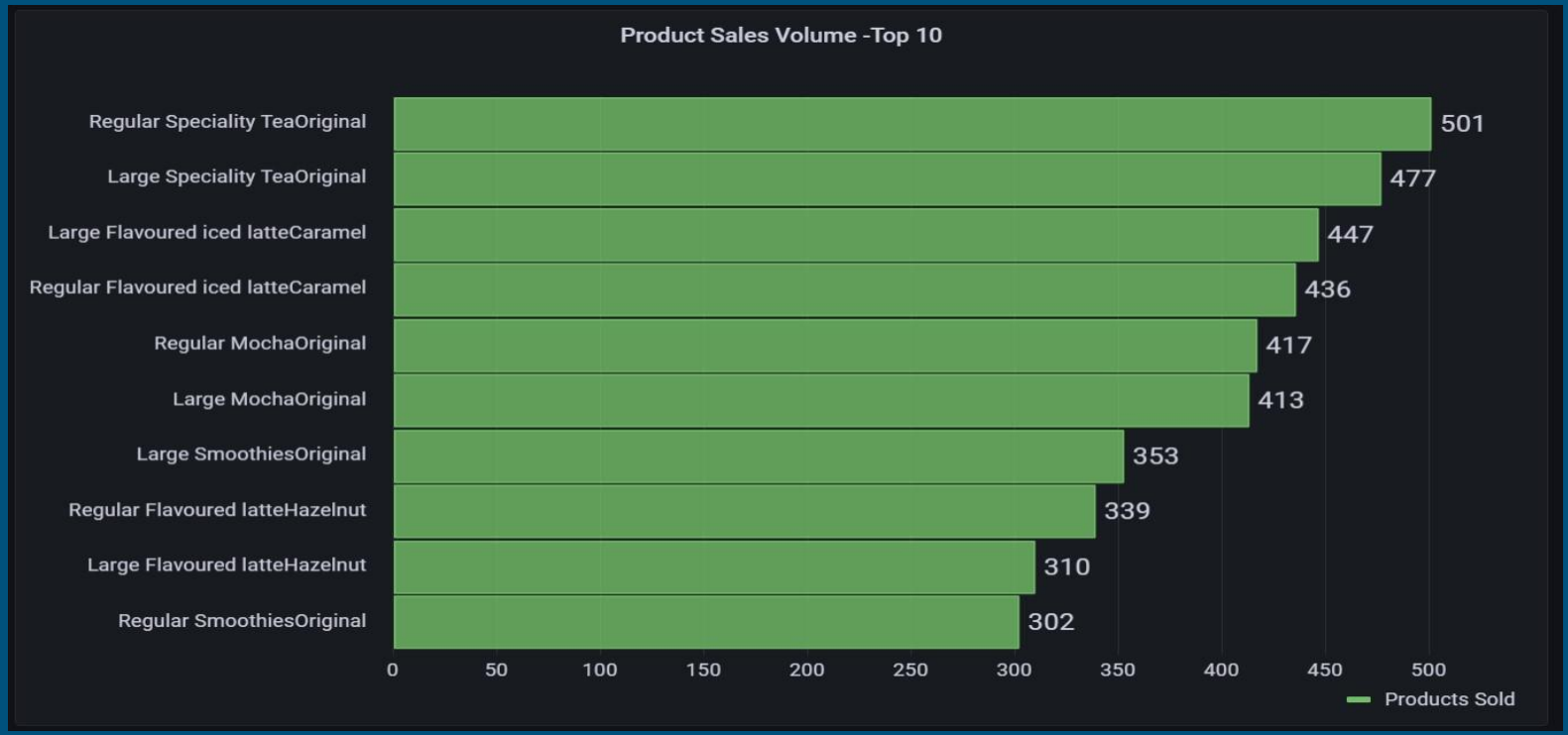


# Insights from multiple variables

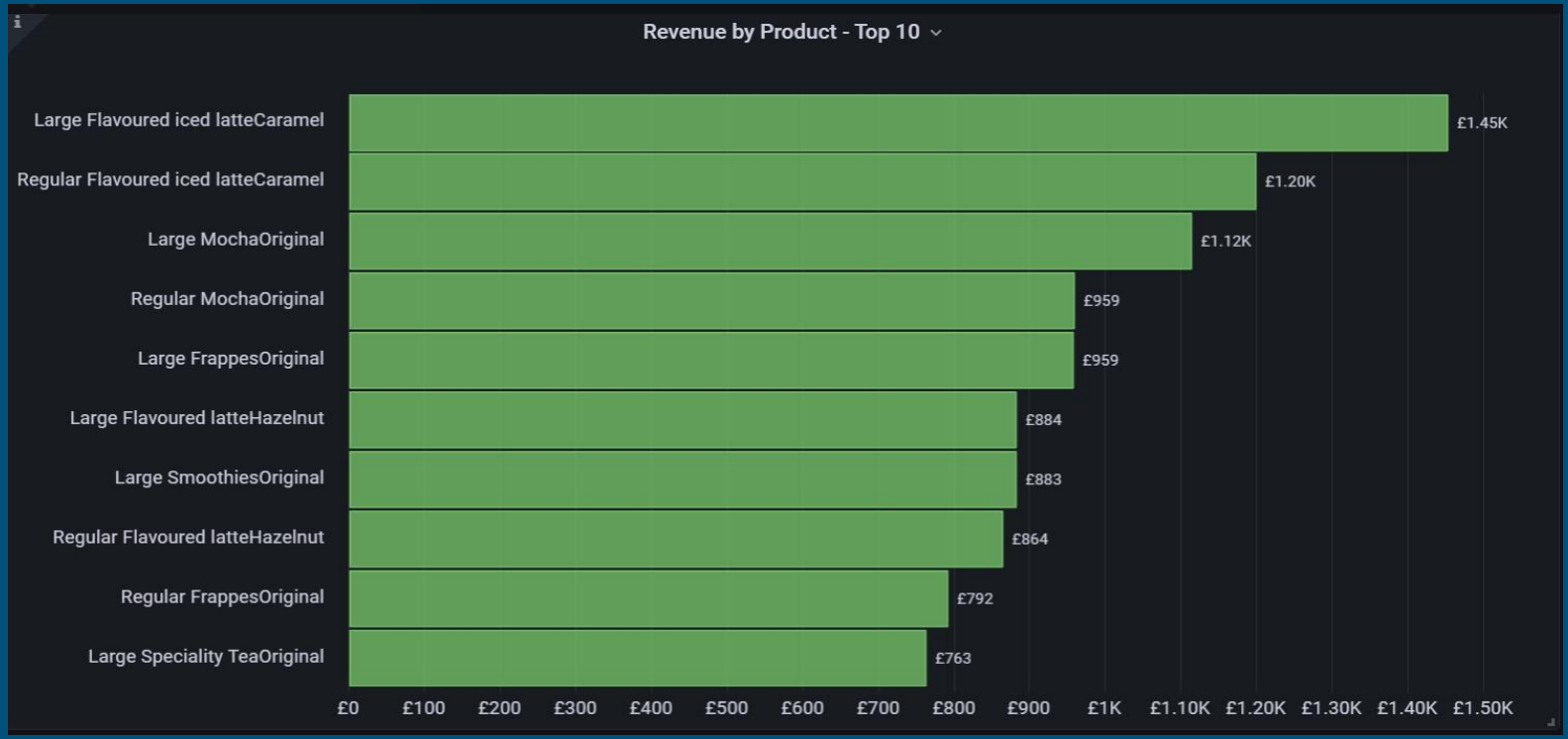

# Sales trend over a period of 5 hrs



# Highest Selling Product by Volume



# Maximum Viable Product (MVP)



# What? Where? When?

**What** - what items were selling well

**Where** - best performing store

**When** - peak business hours



# Thanks...for listening

*Applause to Generation team and Infinity Works from Team 2*

## Thanks to .....

- Jakub and Rachel

  For all your support and motivation

## Special thanks to .....

- Bala and Darren

  For being there for us & helping us through out

- Entire Generation team

  Working at backend to make this programme a smooth experience