Building a Data Pipeline to Support Analyzing Clickstream Data with AWS

Ho Areykal

American University of Phnom Penh

ITM 380: Cloud Computing

November 30, 2023

Contents

1. Introduction

AnyCompany Café, a global business that offers desserts and coffee through its website. The café owner wants to use data to improve the website and the business. This project aims to create a data analytics pipeline to collect, analyze the data of every click that a website visitor makes, and make an analytics dashboard to show the customer behavior patterns. The team hopes to use the data to optimize the website and the marketing strategies. The data might also help the owner decide where to open new cafés.

1.1 System Issue

The current system of AnyCompany Café has several challenges, such as:

- Lack of understanding regarding user behavior: The company does not have sufficient data to understand how potential customers interact with their website. This includes which pages they visit, how long they stay on each page, and what items they click on. This information can help the company tailor their website and marketing efforts to better suit their customers' preferences.
- Inability to establish targeted advertising: Without the ability to analyze the clickstream data, the company can't identify trends and patterns in user behavior. The new solution aims to enable more targeted advertising efforts, potentially leading to increased sales.
- Inefficient location planning: The company does not have the data to help decide where to open additional locations. By understanding where their online customers are located, they can make more informed decisions about where there is demand for their cafés.

1.2 Objective

The objectives of this project are:

- Deploy a data analytics pipeline on AWS that supports the analysis of website clickstream data.
- Transform clickstream data before it arrives in the visualization layer.
- Use AWS services to analyze clickstream data.
- Design a dashboard reporting mechanism for clickstream data analysis.
- Adjust the data analytics pipeline.

1.3 Project Scope

The scope of this project involves several key components:

- Data Collection: This involves setting up a system to collect and store clickstream data from the café's website. This data will include information about every click a user makes while browsing the site.
- Data Processing: The collected data will need to be processed and cleaned to ensure it's ready for analysis. This might involve removing irrelevant data, dealing with missing or inconsistent data, and transforming the data into a suitable format for analysis.
- Data Integration: If the company has other relevant data sources (like sales data or customer demographic data), these might need to be integrated with the clickstream data. This will allow for more comprehensive analysis.
- Data Analysis: The team will analyze the processed data to gain insights into user behavior on the website. This might involve statistical analysis, machine learning, or other data analysis techniques.
- Insight Application: The insights gained from the data analysis will be used to guide the company's decisions. This might involve focusing advertising efforts, deciding where to

open additional locations, or making changes to the website to improve the user experience.

- Dashboard Creation: An analytics dashboard will be created to allow the café owner to easily observe customer behavior. This will involve designing and implementing a user-friendly interface that presents the data in a clear and understandable way.
- Maintenance and Updates: Once the system is set up, it will need to be maintained and updated regularly to ensure it continues to provide accurate and relevant insights. This might involve updating the data collection and processing methods, adding new features to the dashboard, or retraining machine learning models as new data is collected.

2 Similar Systems

There are several systems that implement similar solutions for analyzing clickstream data, including:

- Datastream Clickstream Browser Data Feed
- Machintel Clickstream Data
- Swash Web Browsing Clickstream Data

3 Method of Implementing

For a project like this, which involves multiple stages and requires collaboration between different roles (data engineers, data analysts, web developers), an Agile development methodology would be the best fit because of its:

- Iterative Development: Agile methodology promotes iterative and incremental development. This means the project is broken down into small parts (called “sprints”), which are planned, developed, tested, and reviewed in cycles. This allows for continuous improvement and flexibility in responding to changes or new requirements.

- Collaboration and Communication: Agile emphasizes regular communication and close collaboration between team members and stakeholders. This can help ensure everyone has a clear understanding of the project goals and progress.
- Feedback and Adaptation: Agile encourages frequent feedback and allows for quick adaptation to changes. This is particularly useful for a data analytics project like this, where insights from early stages of data analysis might inform changes or additions to the data collection or processing methods.

3.1 Project Lifecycle

The project lifecycle for creating a data analytics pipeline and dashboard for AnyCompany Café's clickstream data can be broken down into the following stages:

1. Requirement Gathering: Understand the business needs and objectives of AnyCompany Café. Identify the key performance indicators (KPIs) that the dashboard should track.

2. Data Collection: Set up the infrastructure to collect clickstream data from the website. This could involve implementing tracking scripts on the website or using a third-party service.

3. Data Processing and Integration: Clean and preprocess the collected data. Integrate it with other data sources like sales data, customer demographic data.

4. Data Analysis: Analyze the integrated data to gain insights into user behavior, advertising effectiveness, and potential locations for new cafés.

5. Dashboard Design and Development: Design a user-friendly dashboard that presents the analysis results in an easily digestible format. Develop the dashboard using a suitable data visualization tool.

6. Testing: Test the data pipeline and dashboard to ensure they work as expected. This includes checking data accuracy, dashboard functionality, and performance.

7. Deployment: Deploy the data pipeline and dashboard. This could involve setting up a server to host the dashboard and scheduling regular data updates.

8. Maintenance and Updates: Regularly maintain and update the system to ensure it continues to meet the company's needs. This could involve updating the data processing scripts, adding new features to the dashboard.

9. Review and Iteration: Regularly review the system's performance and make necessary adjustments. This could involve refining the data analysis algorithms, redesigning the dashboard.

4. List of Cloud Tools Used

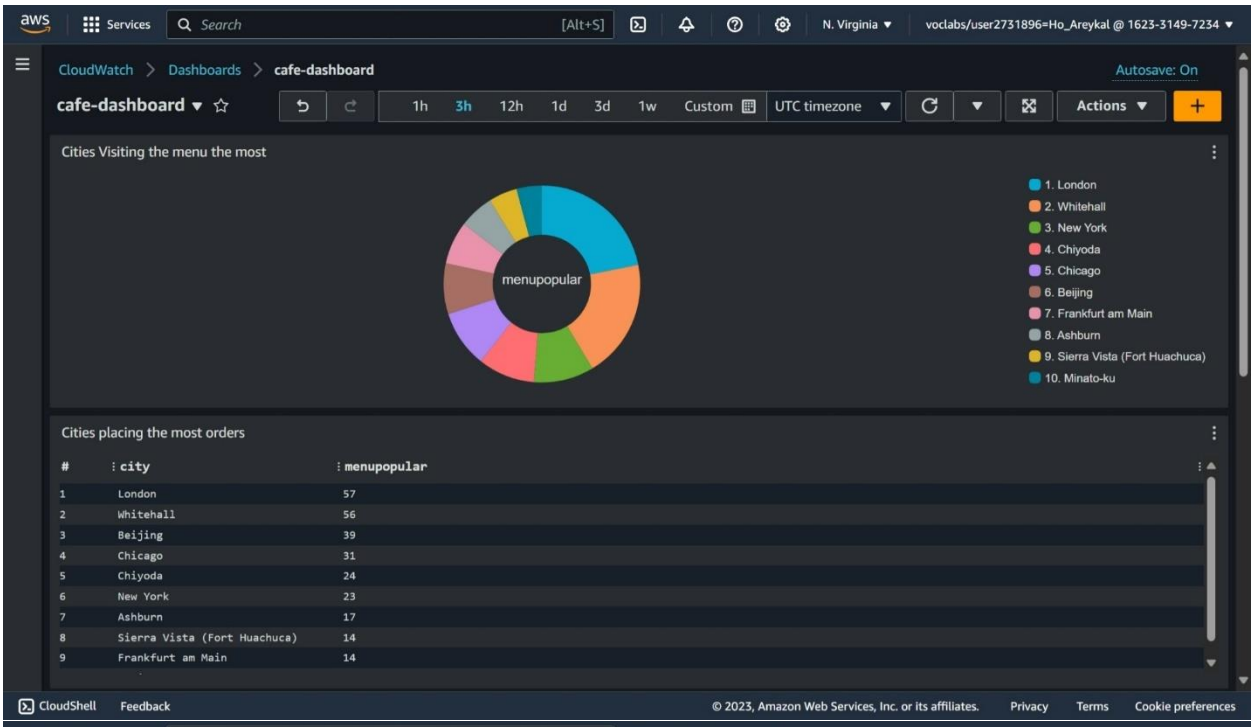
4.1 Hardware

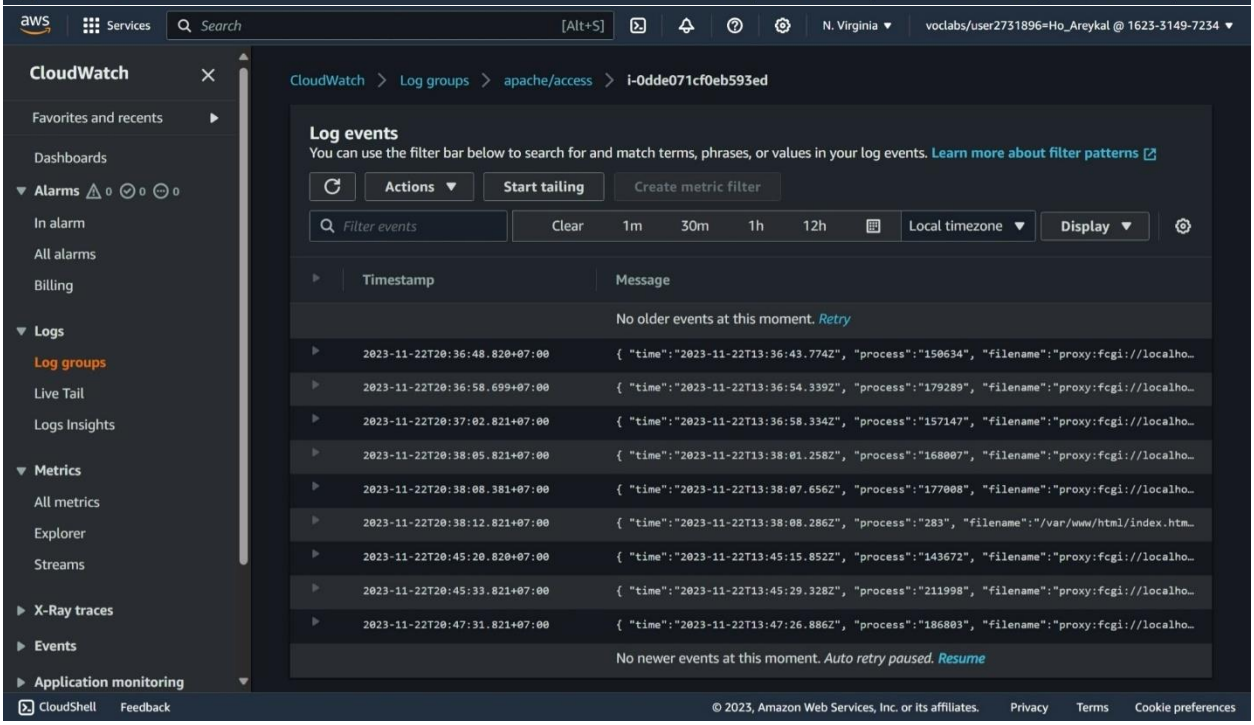
- Amazon EC2
- Amazon S3

4.2 Software

- Amazon Cloud9
- Amazon CloudWatch
- Amazon VPC

5. Project Design, Functions, and Outcomes





6. Limitations

Possible limitations of this system include:

- Data Quality: The quality of insights is directly dependent on the quality of the data collected. If there are issues with the data collection process, such as missing data or inaccurate data, it could lead to misleading insights.
- Privacy Concerns: Collecting and analyzing user data can raise privacy concerns. It's important to ensure that all data collection and analysis activities comply with relevant privacy laws and regulations.
- Complexity of User Behavior: User behavior on a website can be complex and multifaceted. Clickstream data might not capture the full context of user behavior. For example, a user might browse the menu extensively but not make a purchase due to factors outside the website's control.
- Resource Intensive: Depending on the volume of data, the process can be resource-intensive, requiring substantial computational power and storage capacity.
- Location Decisions: While the data can provide insights into online user behavior, it might not directly translate into physical café location success. Other factors such as local market conditions, rent prices, and accessibility also play a crucial role.

7. Future Work

Some are areas of the project could be improved in the future, including:

- Real-Time Analytics: Improve the system to provide real-time insights, which can help the company react quickly to changes in customer behavior.
- Improved Privacy Measures: Implement advanced privacy measures to protect user data, such as differential privacy or federated learning.

- Integration with Business Operations: Integrate the insights from the data analytics pipeline directly into business operations, such as inventory management or targeted marketing campaigns.

8. Conclusion

In conclusion, this project plays a crucial role in helping foster the learning experience of using various AWS services to achieve certain goals. Although the services are limited to the lab's environment, they are enough for new users to dive in and dig deeper into the ecosystem. The main challenges when completing this project are the unfamiliarity with Linux commands and the directories within the environment.