Transform Unstructured Log E-Commerce Apps to Structured Data

Damar Arba Pramuditya

Department of Computer Sciences and Electronics, Faculty of Mathematics and Natural Sciences, University Gadjah Mada

Abstract

Access logs for electronic ecommerce sites are a useful source of data for understanding user behavior . The purpose of this study was to transform unstructured terminal history to structural data and analysis data log, by inspecting and analyzing from one week eyar access logs (10,364,865) We used private sources from system apps. The dataset was downloaded from Harvard's Dataverse and contains logs from an Iranian ecommerce site (zanbil.ir).

Keywords: Log application, e-commerce log, analysis log, unstructured data

Introduction

While our applications have always grown because of market demand, we intend to give the best solution and support to our customers at minimum time. Many of our issues are solved from machine to machine connection, which means they must use either linux, windows or even macOS.

Manual log analysis depends on the proficiency of the person running the analysis. If they have a deep understanding of the system, they may gain some momentum reviewing logs manually. However, this has serious limitations. It puts the team at the mercy of one person. As long as that person is unreachable, or unable to resolve the issue, the entire operation is put at risk.

Goals

Personalize the terminal history to reflect behavior user. Comply with security policies, regulations & audits and automate task

Analysis

Types of structured data analysis

- Pattern Detection and Recognition

refers to filtering incoming messages based on a pattern book. Detecting patterns is an integral part of log analysis as it helps spot anomalies.

- Log Normalization

is the function of converting log elements such as IP addresses or timestamps, to a common format.

- Classification and Tagging

is the process of tagging messages with keywords and categorizing them into classes. This enables

you to filter and customize the way you visualize data.

Security Consideration

Do you have especially sensitive data via HTTP like credit card number, that may require HTTPS?

In log system we don't intend to log specific data like password or credentials, or credit card numbers.

Financial Consideration

Since this will add the number of characters we store in the log, we probably need to check the impact of cost log. The cost is related to the number of characters we write.

Data Tracking, Data Analytics & Marketing Integration Consideration

Does this feature require additional monitoring & tracking? How about mobile tracking, web tracking? It's important that tracking requirements be identified from Day 1.

Log data should not be track, however we may need some kind of analytics dashboard to integrate with our logging system.

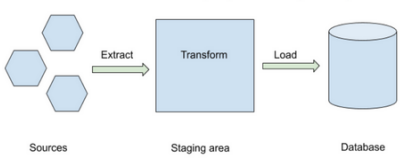
Operation Consideration

Does this feature introduce new behavior that needs to be communicated to customers? new features / changes used by the operation team?

This improvement should be announce to engineer team whom interact with log per daily basis.

Design

System Design



Extract challenges

The data extraction part of the ETL process poses several challenges. A lot of the problems arise from the architectural design of the extraction system:

Data volume.

The volume of data extraction affects system design. The solutions for low-volume data do not scale well as data quantity increases. In this study will use more than 2000 rows of data, with size 3,3 gigs.

Source limits

We used private sources from system apps. The dataset was downloaded from Harvard's Dataverse and contains logs from an Iranian ecommerce site (zanbil.ir)

Implementation

In this study we focused on how the input data being processed and finally came as output. The input data is unstructured and the expected output is structured data. Given the Design system above using ETL in this implementation phase divided into 2 main parts : backend and frontend. Before that will be describe each technology stack use in this implementation.

Stack Use

Mysql

The database served is MySQL. MySQL is an open-source relational database management system. Its name is a combination of "My", the name of co-founder Michael Widenius's daughter My, and "SQL", the abbreviation for Structured Query Language. The database is play role as Load in ETL

Jupyter Notebook

Jupyter Notebook (formerly IPython Notebook) is a web-based interactive computational environment for creating notebook documents. Jupyter Notebook is built using several open-source libraries, including IPython, ZeroMQ, Tornado, jQuery, Bootstrap, and

MathJax. A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension. The Jupyter Notebook role is as Extract and Transform in ETL

Metabase

Metabase is the easy, open-source way to help everyone in your company work with data like an analyst. Its like a database client Interface but with extra intelligence that can visualize the dataset.

Results

Given dataset will be extract by common log format approach. This approach assumes the common log format and/or the combined one, which are two of the most commonly used. Eventually other formats can be incorporated. We start with the below regular express taken from Regular Expressions Cookbook.

For load into database MySQL using library SQLAlchemy. SQLAlchemy is the Python SQL toolkit and Object Relational Mapper that gives application developers the full power and flexibility of SQL. It provides a full suite of well known enterprise-level persistence patterns, designed for efficient and high-performing database access, adapted into a simple and Pythonic domain language.

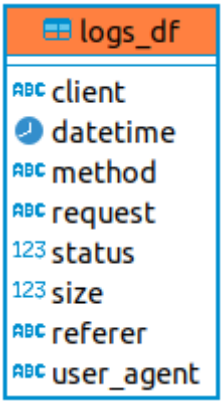


Figure 1: ER Table data logs.

Visualize Data

As the final process transform data from unstructured data to structured data, In this study will provide frontend tools to serve the data, we choose open source tools Metabase

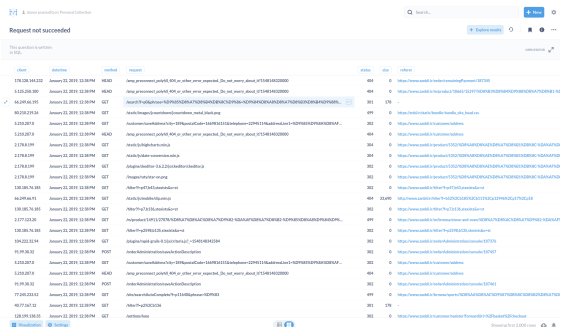


Figure 2: Metabase stream from mysql database

Conclusion

Terminal has always been the close friend to developers. We may overlook the terminal as only to work on a daily basis, but it turns out we can personalize person behavior based on dump history terminal log.

With ETL approach this study successfully transform log application e-commerce data from unstructured to structured data. Then we utilize open source tools to visualize the data.

Future Works

This initial study leads to very various integration and improvement, this study able to process raw data into structured data.

So for the future works, field machine learning can absorb the data and make classification or prediction.

References

- Functional Specification Document | The Complete Guide. (n.d.). Retrieved October 4, 2022, from https://www.xenonstack.com/blog/functional-specification-document
- Structured vs Unstructured Data: 5 Key Differences | Integrate.io. (n.d.). Retrieved October 4, 2022, from https://www.integrate.io/blog/structured-vs-unstructured-data-key-differences/
- What Is Log Analysis Tutorial: Logging Use Cases & Benefits—Sematext. (n.d.). Retrieved October 4, 2022, from https://sematext.com/blog/log-analysis/
- What is Log Analysis? | Sumo Logic. (n.d.). Retrieved October 4, 2022, from https://www.sumologic.com/glossary/log-analysis/
- Asquith, G. B., Krygowski, D., & Gibson, C. R. (2004). Basic well log analysis (Vol. 16). Tulsa: American Association of Petroleum Geologists.
- Peters, T. A. (1993). The history and development of transaction log analysis. Library hi tech.