JACOBS
UNIVERSITY

Advance Project 1
by
Konrad Burchardt
20-06-2019

Search Engine Data Integration Dashboard

Professor: Dr. Aldabert Wilhelm
Spring Term 2019
MRD005-340001

Executive Summary

Task and Goal

The task for this project is to create an ETL pipeline using Google search organic traffic and load the data in a user friendly dashboard. ETL process performs data cleaning during extraction process and load significant data into data warehouse. Our main goal for this project is by having this pipeline we will be able to have the data in one place and make it easier for a businesses stakeholders to access this data, analyze it and discover different business insights.

Data Background

We will be working with Monthly Google search organic traffic from the domain https://www.tuves.cl. This data will be acquire by using the Google search API. We will use 9 different data sets of search traffic and they are the following:

- **Mobile, Desktop and Tablet - Day to day Organic data:** These are multiples CSV files that are organized by date and includes Keywords, Clicks, Impressions, CTR, Ranking, Device and Date.
- **Mobile, Desktop and Tablet - Monthly Top 10 Keywords:** One data set that includes Keyword, Clicks, Impressions, Avg CTR and Avg Ranking for the specific device and time frame.
- **Mobile, Desktop and Tablet - Monthly Top 10 URLs:** One data set that includes URL, Clicks, Impressions, Avg CTR and Avg Ranking for the specific device and time frame.

Approach and Methods used

This project is composed by 5 steps.

1. **Extraction**: Extract the Google search data using the Google search API. Our data will be saved in CSV files.
2. **Transform**: Transform the day to day data into 3 monthly data sets. A mobile data set, a desktop and a tablet. The Keywords and URLs are ready to go.
3. **Load** Load the data to the data warehouse. We will be using a MySQL database.
4. **Load** Build a REST API that will parse all of our MySQL tables into JSON.
5. **Analyze** Create a dashboard where we will load our data and analyze it using tables and visualizations.

Results

Our objective was to create a easy-to-use interface that displays Organic Search data in a simple way. Our outcome was positive, we were able to create a user friendly dashboard that allowed the SEO Managers of the business, find important insights and use it as an easy monthly reporting tool across the company.

The next steps for this project would be to take this report to the next level. First thing will be to automate each of the process, then we will include new KPI's and include machine learning to find which pages need attention and optimization.

Contents

Chapter 1

Introduction

This report is focused on the creation of an ETL pipeline and a user friendly SEO dashboard. ETL pipelines refers to a set of processes extracting data from one system, transforming it, and loading into some database or data-warehouse. A SEO data dashboard is a information tool that includes visualization and display important KPI's Related to search engine traffic. These metrics and key data points are used to monitor the health of the SEO effort in a business and to make important business decisions. In this report we used Google search data(Organic traffic) from the domain https://www.tuves.cl. Tuves HD is an international provider of satellite television in high definition that transmits hundreds of digital signals , with headquarters in Santiago de Chile and customers throughout South and Central America[1]. Organic traffic refers to the visitors that come to our website as a result of unpaid search results[2].

In this project we used different methods, tools and steps to create our pipeline as well as our dashboard. The steps and technologies used are the following *(All of these steps will be explained in depth in the Data Prepossessing section* :

1. **Extraction**: Extract the Google search data using the Google search API.We will use a Python Script that will make an API call to Google servers and export the desired data into multiples CSV files.

2. **Transform**: Transform the day to day data into 3 monthly data sets. Each of the CSV files is equivalent to o day of data. Our focus is to create 3 data sets with 30 days of data from these CSV files. For each day we will create a line in our main CSV file. The outcome will be 3 data sets, Mobile, Desktop and Tablet, with the 30 of day Organic data. Here we use a python script.

3. **Load** Load the data to the data warehouse. We will create an empty database in MySQL and every time we load one of our CSV files it will generate a table with the CSV columns.

4. **Load** Build a REST API that will parse all of our MySQL tables into JSON.We will send our data to our app in JSON format using NodeJS and ExpressJS. The API will have 9 different urls that include all of the desired that that we need to make our dashboard app.

5. **Analyze** Create a dashboard using ReactJS and ChartJS. We will load our data from our RESTapi. After loading the data we will generate user friendly charts and tables. We will give a simple analysis of what can we find using this dashboard.

Chapter 2

Background of Data

In this chapter we will explain what SEO is and how does it work, What is Organic traffic, Explain how our data structure and where does it come from.

2.1 What is SEO?

SEO stands for Search Engine Optimization and is the process of getting free organic traffic from search engines. SEO is different than paid search, in SEO you can't pay to be in the number one position of the search result page. There are multiple ways you can optimize your website for a target keyword to rank in top positions. Keyword have different search volumes and the better you rank for a high volume keyword the more clicks.

2.2 How does SEO Work?

Major search engines use web crawlers to discover publicly available web pages. These crawlers look at web pages and follow links on these web pages. They go link to link gathering information and taking notes of key signals, from keywords to website freshness . All the information that the crawler find is added to the search index, where they contain billions of web pages and is over 100,000,000 gigabytes in size[3]. This index is fed to a search algorithm that checks all the information and signals of the web page and creates an entry that is assigned to a keyword.

2.3 Organic Traffic

The term organic traffic is used to refer to the traffic that go to a website as a result of unpaid search results. As mentioned before, Organic traffic is the opposite of paid traffic. Traffic that is considered organic, are visitors that finds a website after using a search engine like Yahoo, Bing or Google.

2.4 Our Data

The Data we are using is all organic traffic from May 2019(30 days) from the website https://www.tuves.cl. Our data is divided into 9 different data sets. Overall Day to day traffic, mobile, desktop and tablet top keywords and mobile, desktop and tablet top urls.

2.4.1 Day to day Traffic

Day to day traffic is all traffic that we had in the month of May in 2019. This is composed by multiple CSV files. Each CSV file is a day of the month of May and are named with the date for example **20190501.csv**

Each of the CSV files contain the following rows:

Keywords	Clicks	Impressions	CTR	Position	Device	Date

Each line correspond to a Keyword that brought traffic to our site that day. For example the first 3 lines of **20190501.csv** would look like this:

Keywords	Clicks	Impressions	CTR	Position	Device	Date
tuves hd	80	113	70.8%	1	Mobile	1/05/19
tuves	52	114	45%	1.2	desktop	1/05/19
tu ves	43	94	45.7%	1.2	desktop	1/05/19

The columns have the following meanings:

- **Keywords**: Keyword that the website was ranking for.
- **Clicks**: How many times a user clicked through to the site. How this is counted depends on the search result type.
- **Impressions**:How many times a user saw a link to the site in search results. This is calculated differently for images and other search result types, depending on whether or not the result was scrolled into view.
- **CTR**: Is the percentage of impressions that resulted in a click.
- **Position**: Is the average position of the site in search results, based on its highest position whenever it appeared in a search.
- **Device**: Device the user searched for that keywords.
- **Date**: Date the search was made.

2.5 Google Search Console API

The following symbols characters are reserved by LATEX because they introduce a command and have a special meaning.

Chapter 3

Data Prepossessing

3.1 Extraction: Google API Requests

3.1.1 Day to day Data

3.1.2 Top 10 Keywords

3.1.3 Top 10 Urls

3.2 Transform: Data Cleaning

3.2.1 Day to Day Data Cleaning

3.3 Load: Data to data warehouse(MySQL)

3.3.1 Creating Database

3.3.2 Desktop data

3.3.3 Mobile data

3.3.4 Tablet data

3.3.5 Mobile, Desktop and Tablet Top 10 URLs

3.3.6 Mobile, Desktop and Tablet Top 10 Keywords

3.4 Load: RestAPI(NodeJS)

3.4.1 API Index

3.4.2 Desktop API

3.4.3 Mobile API

3.4.4 Tablet API

3.4.5 Mobile, Desktop and Tablet Top 10 URLs API

3.4.6 Mobile, Desktop and Tablet Top 10 Keywords API

3.5 Transform: SEO Dashboard(ReactJS)

aaaaa

3.5.1 Dashboard structure

aa

3.5.2 Dashboard Visualization

3.5.3 Components

aa

3.5.4 Total KPI's

3.5.5 Charts

3.5.6 Tables

The following symbols characters are reserved by LATEX because they introduce a command and have a special meaning.

Chapter 4

Data Exploration

4.1 Mobile

4.2 Desktop

4.3 Tablet

The following symbols characters are reserved by LATEX because they introduce a command and have a special meaning.

Chapter 5

Results and Conclusions

The following symbols characters are reserved by LATEX because they introduce a command and have a special meaning.

Chapter 6

References

[1]“TV Satelital HD Prepago.” Tuves HD, www.tuves.cl/.

[2]“Search Engine Users.” Pew Research Center: Internet, Science and Tech, Pew Research Center: Internet, Science and Tech, 8 Apr. 2014, www.pewinternet.org/2005/01/23/search-engine-users/.

[3] ”How Search Works - Indexing, Google”, www.google.com/search/howsearchworks/crawling-indexing/.

Chapter 7

Appendix

7.1 Extraction: Google API Requests

7.1.1 Top 10 Keywords

7.1.2 Top 10 Urls

7.1.3 Day to day Data

7.2 Transform: Data Cleaning

7.2.1 Day to Day Data Cleaning

7.3 Load:Data to data warehouse(MySQL)

7.3.1 Desktop data

7.3.2 Mobile data

7.3.3 Tablet data

7.3.4 Mobile, Desktop and Tablet Top 10 URLs

7.3.5 Mobile, Desktop and Tablet Top 10 Keywords

7.4 Load: RestAPI

7.4.1 Desktop API

7.4.2 Mobile API

7.4.3 Tablet API

7.4.4 Mobile, Desktop and Tablet Top 10 API

7.4.5 Mobile, Desktop and Tablet Top 10 API

7.5 Transform: SEO Dashboard

7.5.1 ReactJS Dashboard structure

7.5.2 ReactJS components

7.5.3 Charts Api Calls

7.5.4 Tables API Calls