Hit Data Revenue

Agenda

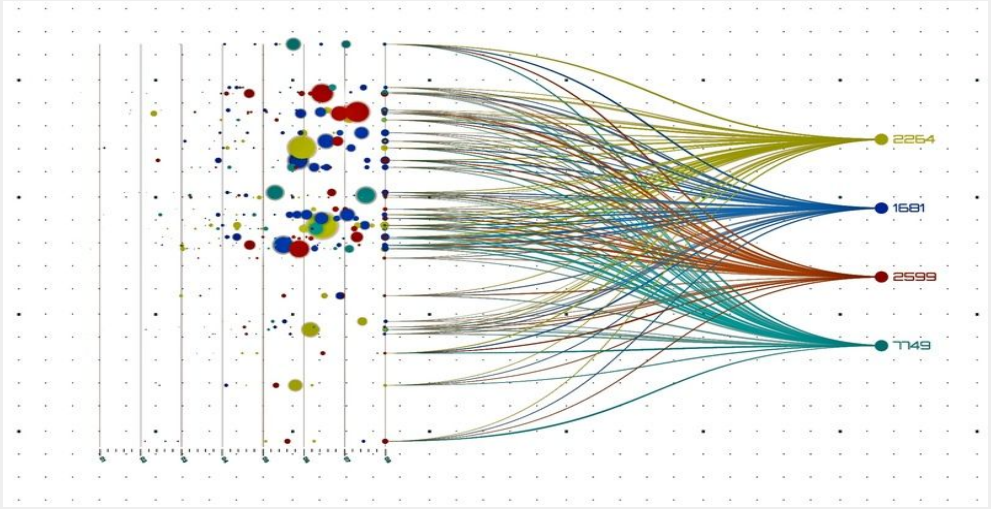
- Business Problem Statement
- Introduction
- Data Observations and Considerations
- Data Processing Approach
- AWS Deployment and Automation
- Results And Discussion
- Further Improvements
- Application Demo

Business Problem Statement

Having a lot advertising sources in the trend any business would like to know which sources are contributing to their revenue to justify the spending/resources they are allocating and adjust accordingly.

Introduction

- Stream of continuous activities performed by a user from their web searches to any final activity.
- Gives a good narrative to users online journey by linking the actions taken by users within a single session.



How it can benefit business

- Identity Customer Trends
- Different Pathways lead to Same destination
- Preventing abandonments
- Navigation optimization

Data Observations and Considerations

ip	event	geo_city	pagename	page_url	product_list	referrer
112.33.98.231		Salt Lake City	Home	http://www.esshopzilla.com		http://search.yahoo.com/search?p=cd+pla
23.8.61.21	2	Rochester	Zune - 32 GB	http://www.esshopzilla.com/product/?pid=asfe13	Electronics;Zune - 328GB;1;;	http://www.bing.com/search?q=Zune&go=
23.8.61.21	12	Rochester	Shopping Cart	http://www.esshopzilla.com/cart/		http://www.esshopzilla.com/product/?pid
23.8.61.21	11	Rochester	Order Checkout Detail	https://www.esshopzilla.com/checkout/		http://www.esshopzilla.com/cart/
23.8.61.21		Rochester	Order Confirmation	https://www.esshopzilla.com/checkout/?a=confirm		https://www.esshopzilla.com/checkout/
23.8.61.21	1	Rochester	Order Complete	https://www.esshopzilla.com/checkout/?a=comple	Electronics;Zune - 32GB;1;250;	https://www.esshopzilla.com/checkout/?a
44.12.96.2		Duncan	Hot Buys	http://www.esshopzilla.com/hotbuys/		http://www.google.com/search?hl=en&cli
44.12.96.2	2	Duncan	Ipod - Nano - 8 GB	http://www.esshopzilla.com/product/?pid=as3221	Electronics;Ipod - Nano - 8GB;1;;	http://www.esshopzilla.com/hotbuys/
44.12.96.2	12	Duncan	Shopping Cart	http://www.esshopzilla.com/cart/		http://www.esshopzilla.com/product/?pid
44.12.96.2	11	Duncan	Order Checkout Detail	https://www.esshopzilla.com/checkout/		http://www.esshopzilla.com/cart/
44.12.96.2		Duncan	Order Confirmation	https://www.esshopzilla.com/checkout/?a=confirm		https://www.esshopzilla.com/checkout/
44.12.96.2	1	Duncan	Order Complete	https://www.esshopzilla.com/checkout/?a=comple	Electronics;Ipod - Nano - 8GB;1;190;	https://www.esshopzilla.com/checkout/?a
67.98.123.1		Salem	Home	http://www.esshopzilla.com		http://www.google.com/search?hl=en&cli
67.98.123.1		Salem	Search Results	http://www.esshopzilla.com/search/?k=Ipod		http://www.esshopzilla.com
67.98.123.1	2	Salem	Ipod - Nano - 8 GB	http://www.esshopzilla.com/product/?pid=as3221	Electronics;Ipod - Nano - 8GB;1;;	http://www.esshopzilla.com/search/?k=Ip
67.98.123.1		Salem	Search Results	http://www.esshopzilla.com/search/?k=Ipod		http://www.esshopzilla.com/product/?pid
67.98.123.1	2	Salem	Ipod - Touch - 32 GB	http://www.esshopzilla.com/product/?pid=as2323	Electronics;Ipod - Touch - 32GB;1;;	http://www.esshopzilla.com/search/?k=Ip
67.98.123.1	12	Salem	Shopping Cart	http://www.esshopzilla.com/cart/		http://www.esshopzilla.com/product/?pid
67.98.123.1	11	Salem	Order Checkout Detail	https://www.esshopzilla.com/checkout/		http://www.esshopzilla.com/cart/
67.98.123.1		Salem	Order Confirmation	https://www.esshopzilla.com/checkout/?a=confirm		https://www.esshopzilla.com/checkout/
67.98.123.1	1	Salem	Order Complete	https://www.esshopzilla.com/checkout/?a=comple	Electronics;Ipod - Touch - 32GB;1;290	https://www.esshopzilla.com/checkout/?a

Data Observations and Considerations

- Hit Data is grouped based on IP address.
- Sequence of activities are tracked based on the event ids provided.
- Data stream can be tied together with page_url (merchant) and referrer columns (Search Engine)
- Only completed orders with price availability are considered for this analysis.
- To consider any final activity towards revenue, search engine keyword should match with the purchased product.

UDF Functions

- Required to retrieve Search Engine and Keyword

```
def url_search_engine_fun(url):
    if not url:
        return ''
    return urlparse(url).netloc

def url_search_keyword_fun(url):
    parsed=urlparse(url)
    if parsed.query:
        dic=parse_qs(parsed.query)
        if parsed.netloc =='search.yahoo.com':
            return dic['p'][0].upper()
        else :
            return dic['q'][0].upper()
    else:
        return ''

url_search_engine_udf = udf(lambda m: url_search_engine_fun(m))
url_search_keyword_udf = udf(lambda m: url_search_keyword_fun(m))
```

Data Processing Approach

- Data is grouped into 2 buckets to identify the search engine referrals and associated completed orders

Search Engine Results

SearchIPaddress	pagename	Search_Engine_Domain	Search_Keyword
67.98.123.1	Home	www.google.com	IPOD
23.8.61.21	Zune - 32 GB	www.bing.com	ZUNE
112.33.98.231	Home	search.yahoo.com	CD PLAYER
44.12.96.2	Hot Buys	www.google.com	IPOD
123.88.100.21	Zune - 32 GB	www.bing.com	ZUNE

Completed Orders

SoldIPaddress	pagename	merchent_site	sold_product	price
23.8.61.21	Order Complete	www.esshopzilla.com	Zune - 32GB	250
44.12.96.2	Order Complete	www.esshopzilla.com	Ipod - Nano - 8GB	190
67.98.123.1	Order Complete	www.esshopzilla.com	Ipod - Touch - 32GB	290
123.88.100.21	Order Complete	www.esshopzilla.com	IPAD	650

Data Processing Approach

- Results received by joining 2 Data Frames

SearchIPaddress	Search_Engine_Domain	Search_Keyword	SoldIPaddress	pagename	merchent_site	sold_product	price
23.8.61.21	www.bing.com	ZUNE	23.8.61.21	Order Complete	www.esshopzilla.com	Zune - 32GB	250
44.12.96.2	www.google.com	IPOD	44.12.96.2	Order Complete	www.esshopzilla.com	Ipod - Nano - 8GB	190
67.98.123.1	www.google.com	IPOD	67.98.123.1	Order Complete	www.esshopzilla.com	Ipod - Touch - 32GB	290
123.88.100.21	www.bing.com	ZUNE	123.88.100.21	Order Complete	www.esshopzilla.com	IPAD	650

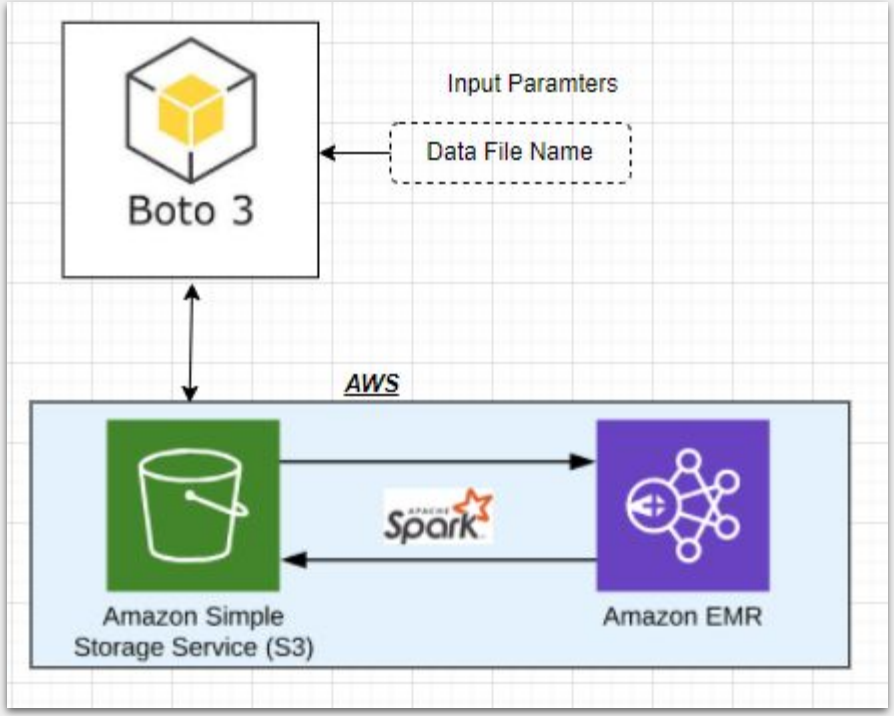
Data Processing Approach

- Aggregating previous data frame on Search engine and search keyword.

Search_Engine_Domain	Search_Keyword	Revenue
www.google.com	IPOD	480
www.bing.com	ZUNE	250

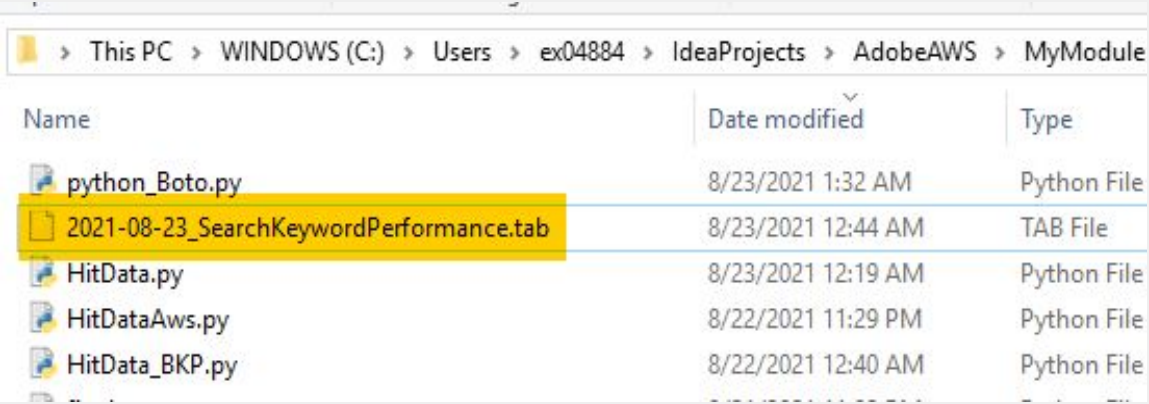
AWS EMR Deployment and Boto3 Automations

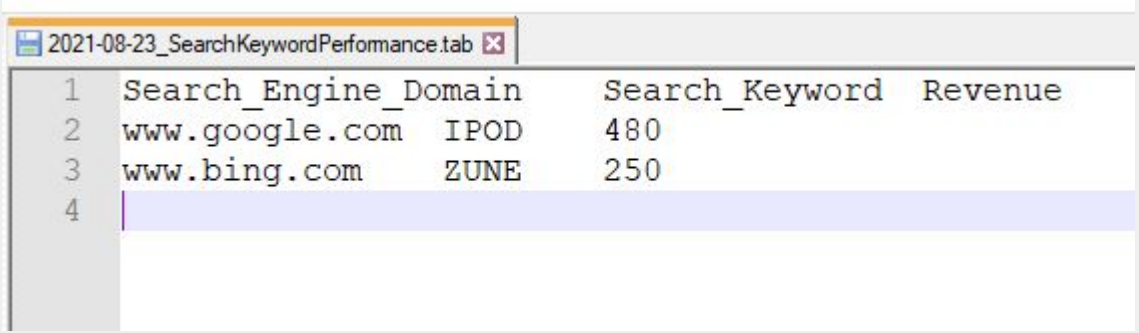
- Executed pyspark code using AWS EMR cluster



Results

File Locations and Contents





Further Improvements

- Based on business decisions
  - What if multiple products are bought from Same IP address, are we going to give user credit for 2 items or is it supposed to be for 1
  - What if user left items in the cart for a week and then made a purchase (what should be the timeline)
  - We can refine this process more based on the “event” column

Further Improvements..

- Also instead of running boto script from local, a simple application can be written to accept data file uploaded by user and execute AWS automation from there.

Application Demo

Questions?