Big Data Ingestion Using Hadoop : 1038-A

Faculty Advisor and Sponsor:
Dr. Rohit Aggarwal

David Eccles School of Business
THE UNIVERSITY OF UTAH

Team Members and Roles

Aditya Kannan

Software Engineer and Project Manager

Ekta Jaiswal

Data Engineer

Vikal Gupta

Software Developer and Reporting Analyst

Agenda

- Data Scenario today
- Why this project and Objective
- Business Questions to be answered
- Project Infrastructure
- Project Workflow
- Learnings and Conclusion

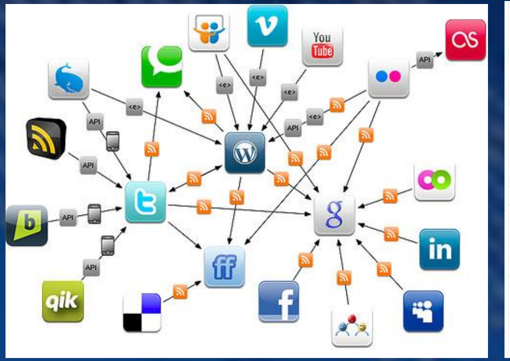
Data Scenario today

- According to IBM, 80% of data generated today is unstructured

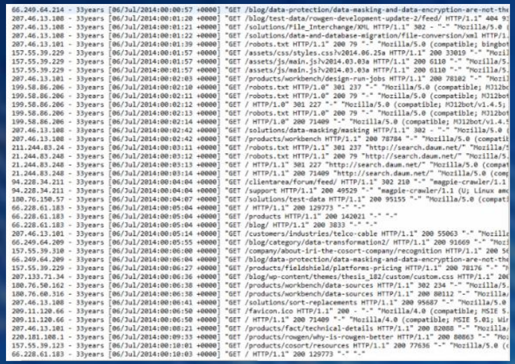
Outsourcing Data

Video Streaming Data

Social Media Data



Logs Data



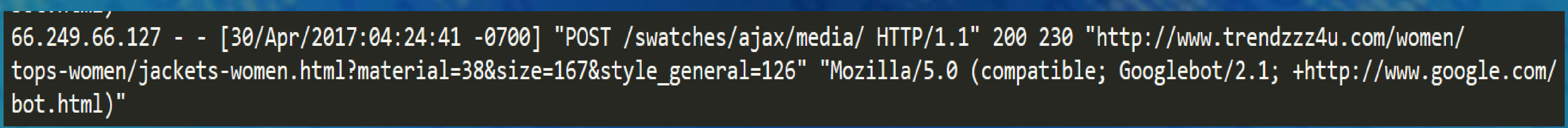
```
66.249.64.214 - 33years [06/Jul/2014:00:00:57 +0000] "GET /blog/data-protection/data-masking-and-data-encryption-are-not-the
207.46.13.108 - 33years [06/Jul/2014:00:01:20 +0000] "GET /blog/test-data/rowgen-development-update-2/feed/ HTTP/1.1" 404 93
207.46.13.108 - 33years [06/Jul/2014:00:01:21 +0000] "GET /solutions/File_Interchange/XML HTTP/1.1" 302 - "-" "Mozilla/5.0 (
207.46.13.108 - 33years [06/Jul/2014:00:01:22 +0000] "GET /solutions/data-and-database-migration/file-conversion/xml HTTP/1.
207.46.13.101 - 33years [06/Jul/2014:00:01:39 +0000] "GET /robots.txt HTTP/1.1" 200 79 "-" "Mozilla/5.0 (compatible; bingbot
157.55.39.229 - 33years [06/Jul/2014:00:01:57 +0000] "GET /assets/css/styles.css?v2014.06.25a HTTP/1.1" 200 33019 "-" "Mozil
157.55.39.229 - 33years [06/Jul/2014:00:01:57 +0000] "GET /assets/js/main.js?v2014.03.03a HTTP/1.1" 200 6110 "-" "Mozilla/5.
157.55.39.229 - 33years [06/Jul/2014:00:01:57 +0000] "GET /assets/js/main.js?v2014.03.03a HTTP/1.1" 200 6110 "-" "Mozilla/5.
207.46.13.101 - 33years [06/Jul/2014:00:02:03 +0000] "GET /products/workbench/design-run-jobs HTTP/1.1" 200 78102 "-" "Mozil
199.58.86.206 - 33years [06/Jul/2014:00:02:10 +0000] "GET /robots.txt HTTP/1.0" 301 237 "-" "Mozilla/5.0 (compatible; MJ12bo
199.58.86.206 - 33years [06/Jul/2014:00:02:11 +0000] "GET /robots.txt HTTP/1.0" 200 79 "-" "Mozilla/5.0 (compatible; MJ12bot
199.58.86.206 - 33years [06/Jul/2014:00:02:12 +0000] "GET / HTTP/1.0" 301 227 "-" "Mozilla/5.0 (compatible; MJ12bot/v1.4.5;
199.58.86.206 - 33years [06/Jul/2014:00:02:13 +0000] "GET /robots.txt HTTP/1.0" 200 79 "-" "Mozilla/5.0 (compatible; MJ12bot
199.58.86.206 - 33years [06/Jul/2014:00:02:14 +0000] "GET / HTTP/1.0" 200 71409 "-" "Mozilla/5.0 (compatible; MJ12bot/v1.4.5
207.46.13.108 - 33years [06/Jul/2014:00:02:42 +0000] "GET /solutions/data-masking/masking HTTP/1.1" 302 - "-" "Mozilla/5.0 (
207.46.13.108 - 33years [06/Jul/2014:00:02:42 +0000] "GET /products/workbench HTTP/1.1" 200 78784 "-" "Mozilla/5.0 (compatib
211.244.83.24 - 33years [06/Jul/2014:00:03:11 +0000] "GET /robots.txt HTTP/1.1" 301 237 "http://search.daum.net/" "Mozilla/5
21.244.83.248 - 33years [06/Jul/2014:00:03:12 +0000] "GET /robots.txt HTTP/1.1" 200 79 "http://search.daum.net/" "Mozilla/5.
21.244.83.248 - 33years [06/Jul/2014:00:03:13 +0000] "GET / HTTP/1.1" 301 227 "http://search.daum.net/" "Mozilla/5.0 (compat
21.244.83.248 - 33years [06/Jul/2014:00:03:14 +0000] "GET / HTTP/1.1" 200 71409 "http://search.daum.net/" "Mozilla/5.0 (comp
94.228.34.211 - 33years [06/Jul/2014:00:04:04 +0000] "GET /clientarea/forum/feed/ HTTP/1.1" 302 210 "-" "magpie-crawler/1.1
94.228.34.211 - 33years [06/Jul/2014:00:04:04 +0000] "GET /support HTTP/1.1" 200 49529 "-" "magpie-crawler/1.1 (U; Linux amd
180.76.150.57 - 33years [06/Jul/2014:00:04:07 +0000] "GET /solutions/test-data HTTP/1.1" 200 95155 "-" "Mozilla/5.0 (compati
66.228.61.183 - 33years [06/Jul/2014:00:05:04 +0000] "GET / HTTP/1.1" 200 129773 "-" "-"
66.228.61.183 - 33years [06/Jul/2014:00:05:04 +0000] "GET /products HTTP/1.1" 200 142021 "-" "-"
66.228.61.183 - 33years [06/Jul/2014:00:05:04 +0000] "GET /blog/ HTTP/1.1" 200 3833 "-" "-"
207.46.13.101 - 33years [06/Jul/2014:00:05:14 +0000] "GET /customers/industries/telco-cable HTTP/1.1" 200 55063 "-" "Mozilla
66.249.64.209 - 33years [06/Jul/2014:00:05:55 +0000] "GET /blog/category/data-transformation2/ HTTP/1.1" 200 91669 "-" "Mozi
157.55.39.310 - 33years [06/Jul/2014:00:06:00 +0000] "GET /company/about-iri-the-cosort-company/recognition HTTP/1.1" 200 56
66.249.64.209 - 33years [06/Jul/2014:00:06:04 +0000] "GET /blog/data-protection/data-masking-and-data-encryption-are-not-the
157.55.39.229 - 33years [06/Jul/2014:00:06:27 +0000] "GET /products/fieldshield/platforms-pricing HTTP/1.1" 200 78176 "-" "M
207.133.71.34 - 33years [06/Jul/2014:00:06:36 +0000] "GET /blog/wp-content/themes/thesis_182/custom/custom.css HTTP/1.1" 200
180.76.50.162 - 33years [06/Jul/2014:00:06:38 +0000] "GET /products/workbench/data-sources HTTP/1.1" 302 234 "-" "Mozilla/5.
180.76.60.316 - 33years [06/Jul/2014:00:06:38 +0000] "GET /products/workbench/data-sources HTTP/1.1" 200 80112 "-" "Mozilla/
207.46.13.108 - 33years [06/Jul/2014:00:06:41 +0000] "GET /solutions/sort-replacements HTTP/1.1" 200 95687 "-" "Mozilla/5.0
209.11.120.66 - 33years [06/Jul/2014:00:06:50 +0000] "GET /favicon.ico HTTP/1.1" 200 - "-" "Mozilla/4.0 (compatible; MSIE 5.
209.11.120.66 - 33years [06/Jul/2014:00:06:50 +0000] "GET / HTTP/1.1" 200 71409 "-" "Mozilla/4.0 (compatible; MSIE 5.01; Wir
207.46.13.101 - 33years [06/Jul/2014:00:08:21 +0000] "GET /products/fact/technical-details HTTP/1.1" 200 82088 "-" "Mozilla/
220.181.108.1 - 33years [06/Jul/2014:00:09:33 +0000] "GET /products/rowgen/why-is-rowgen-better HTTP/1.1" 200 80863 "-" "Moz
157.55.39.123 - 33years [06/Jul/2014:00:10:01 +0000] "GET /products/cosort/resources HTTP/1.1" 200 77636 "-" "Mozilla/5.0 (c
66.228.61.183 - 33years [06/Jul/2014:00:10:03 +0000] "GET / HTTP/1.1" 200 129773 "-" "-"
```

- Need to process unstructured data to structured data

Why this Project ?

- To learn and implement big data technologies which are used to process log data
- Clickstream Log Data as the data source
- What is Clickstream data?

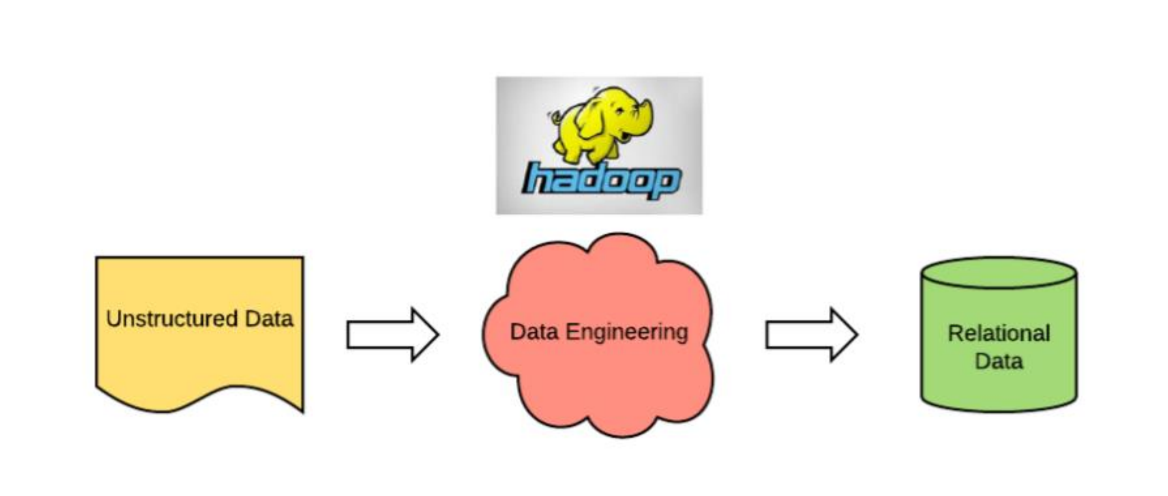
  Clickstream Data is user navigation data on any website



```
66.249.66.127 - - [30/Apr/2017:04:24:41 -0700] "POST /swatches/ajax/media/ HTTP/1.1" 200 230 "http://www.trendzzz4u.com/women/
tops-women/jackets-women.html?material=38&size=167&style_general=126" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/
bot.html)"
```

David Eccles School of Business THE UNIVERSITY OF UTAH

- Big data world comprises of many technologies
- Focus of this project is to learn and implement Apache Hadoop ecosystem



- Hadoop is primarily used for Data Engineering by many companies notably Amazon, eBay, Walmart

Business Questions to be answered

- Most popular browsing time
- Most popular product category
- Weekly Distribution of Clicks per page
- Customer Conversion Rate

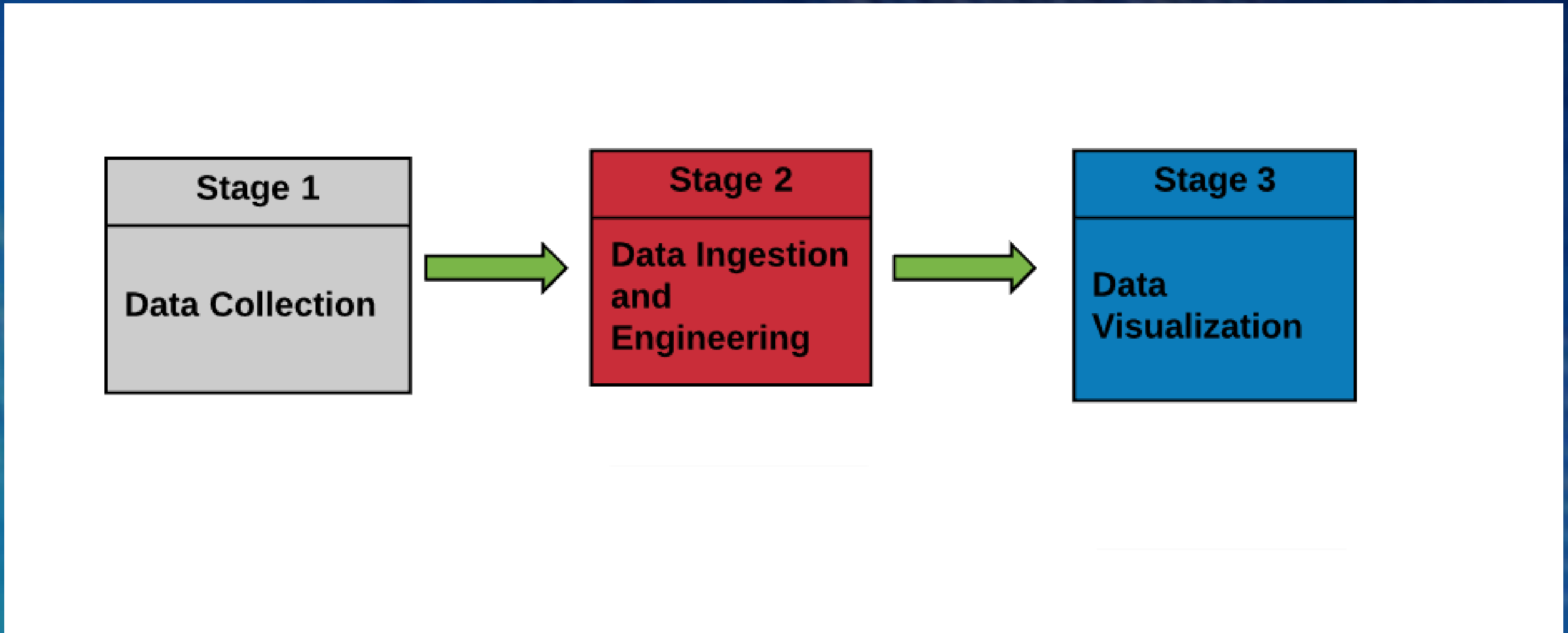
Project Infrastructure

Major components of infrastructure

- eCommerce Website Setup: Magento eCommerce platform, goDaddy cPanel hosting
- Apache Hadoop Setup: Multi Node Cloudera Hadoop cluster on aws EC2
- Data Collection Server: Divolte.js setup on aws EC2 to track custom events from website

Project Workflow

- Primarily three stages
- Performed several iterations of the below workflow



Stage 1

Data Collection

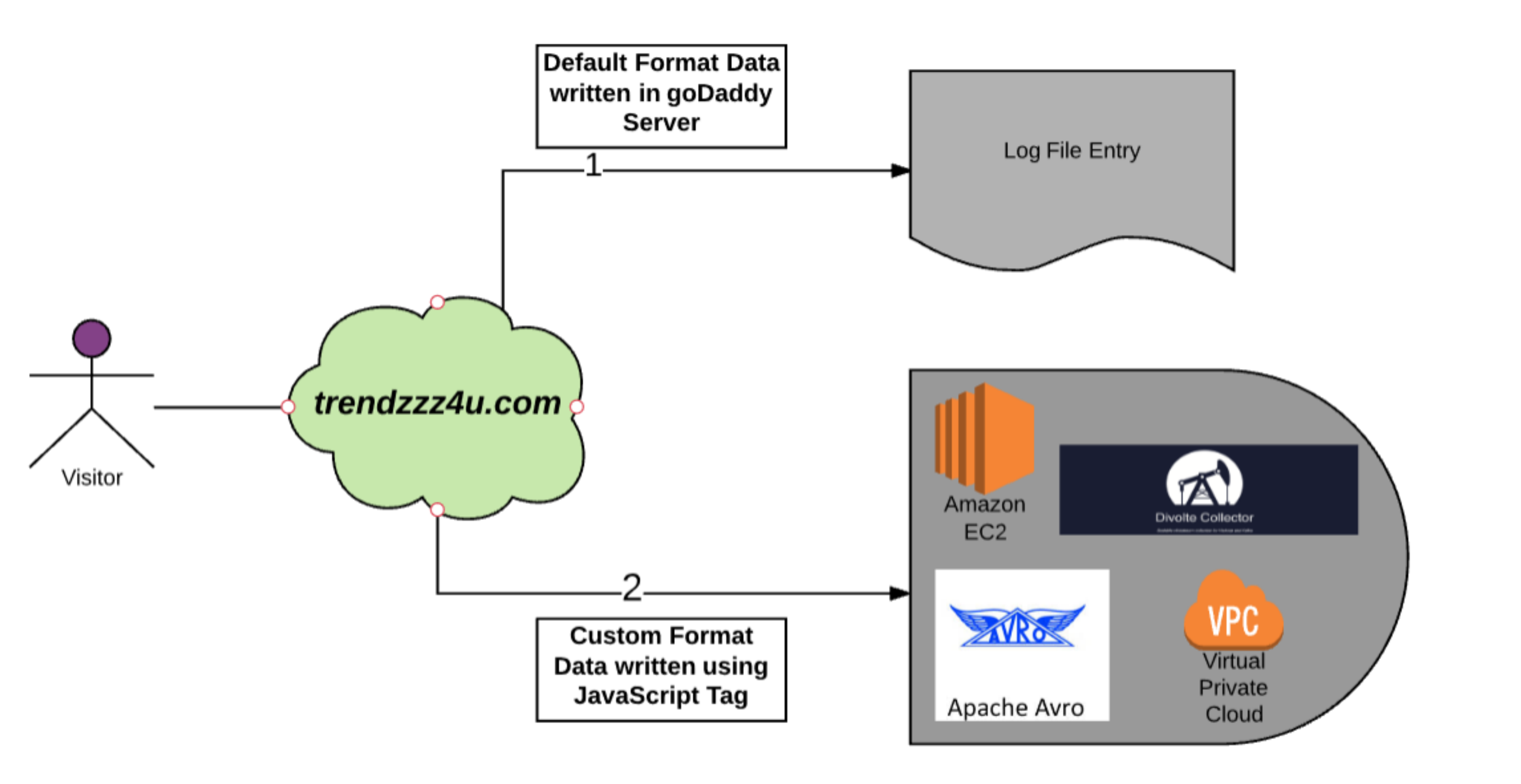
Stage 2

Data Ingestion and Engineering

Stage 3

Data Visualization

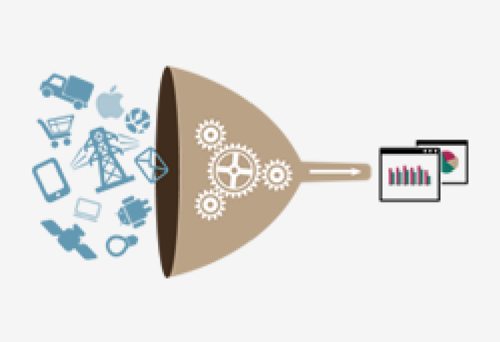
Stage 1 : Data Collection



Two Ways to track Website Visitor

Stage 2: Data Ingestion and Engineering

- The process of accessing and importing data for immediate usage or storage in a database is called as Data Ingestion



- Build Data Pipeline
- Transfer files from local Filesystem to HDFS (Hadoop File System)

E.g. Apache Sqoop is a popular tool used in big data ecosystem to transfer bulk data

jargons

- Data engineering is a process of converting unstructured data to meaningful relational data using set of sophisticated tools or procedures
  - ✓ ELT Process (Extract Load Transform)
  - ✓ Focus on data transformation using Map Reduce

E.g. Components used in this project:

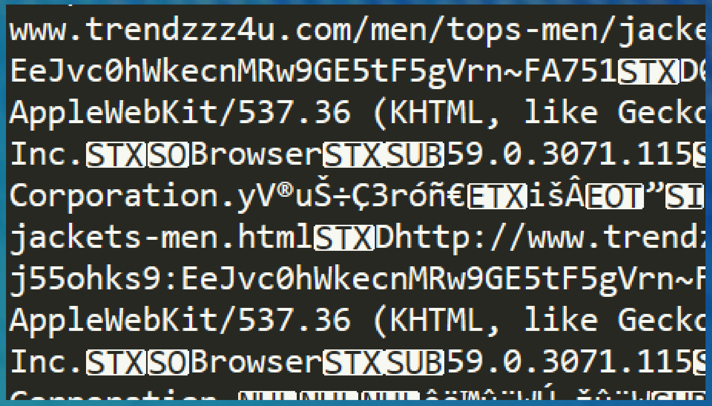
- Apache Hive
- Apache Pig
- Python

Apache Hive

- It provides a SQL like interface to query data stored in various databases and file system

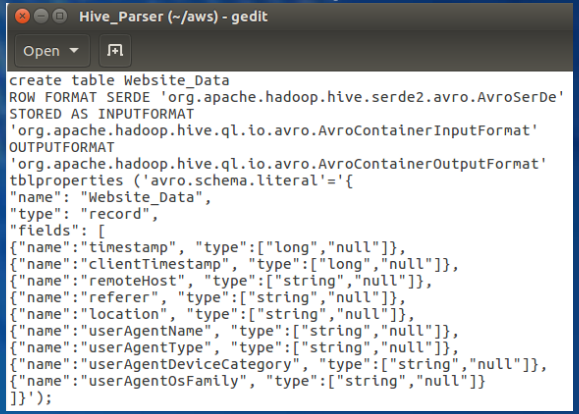
Application in our project: Avro File (Data Source)

Before



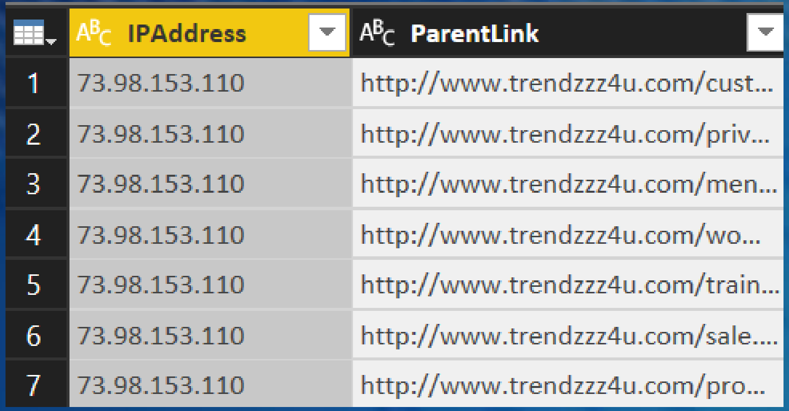
```
www.trendzzz4u.com/men/tops-men/jacke
EeJvc0hWkecnMRw9GE5tF5gVrn~FA751STXD
AppleWebKit/537.36 (KHTML, like Geck
Inc.STXSOBrowserSTXSUB59.0.3071.115
Corporation.yV®uŠ÷Ç3róñ€ETXišÂEOT”SI
jackets-men.htmlSTXDhttp://www.trend
j55ohks9:EeJvc0hWkecnMRw9GE5tF5gVrn~
AppleWebKit/537.36 (KHTML, like Geck
Inc.STXSOBrowserSTXSUB59.0.3071.115
```

Run Script



```
create table Website_Data
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.avro.AvroSerDe'
STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat'
OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat'
tblproperties ('avro.schema.literal'='{
"name": "Website_Data",
"type": "record",
"fields": [
{"name":"timestamp", "type":["long","null"]},
{"name":"clientTimestamp", "type":["long","null"]},
{"name":"remoteHost", "type":["string","null"]},
{"name":"referer", "type":["string","null"]},
{"name":"location", "type":["string","null"]},
{"name":"userAgentName", "type":["string","null"]},
{"name":"userAgentType", "type":["string","null"]},
{"name":"userAgentDeviceCategory", "type":["string","null"]},
{"name":"userAgentOsFamily", "type":["string","null"]}
]}');
```

After



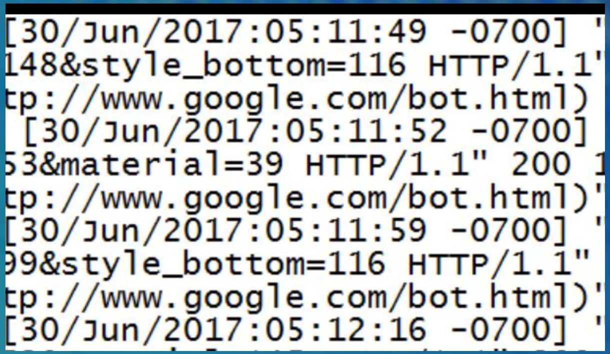
	IPAddress	ParentLink
1	73.98.153.110	http://www.trendzzz4u.com/cust...
2	73.98.153.110	http://www.trendzzz4u.com/priv...
3	73.98.153.110	http://www.trendzzz4u.com/men...
4	73.98.153.110	http://www.trendzzz4u.com/wo...
5	73.98.153.110	http://www.trendzzz4u.com/train...
6	73.98.153.110	http://www.trendzzz4u.com/sale....
7	73.98.153.110	http://www.trendzzz4u.com/pro...

Apache Pig and Python

- It is a high level platform for creating programs on Hadoop. The language used is called as Pig Latin

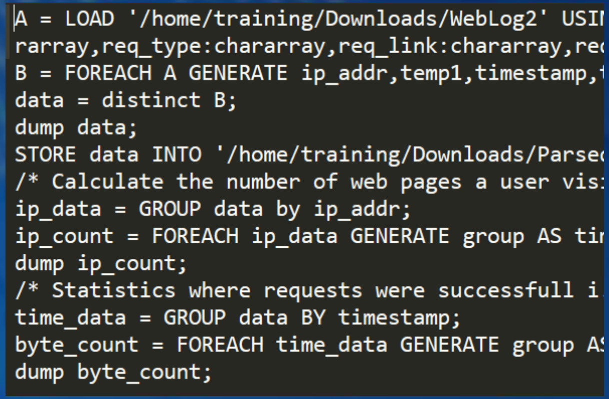
Application in our project: Web Server Log File (Data Source)

Before



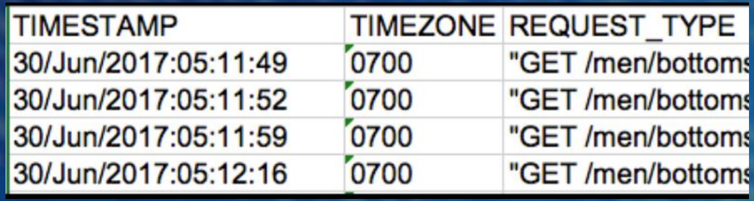
```
[30/Jun/2017:05:11:49 -0700] '
148&style_bottom=116 HTTP/1.1'
tp://www.google.com/bot.html)
 [30/Jun/2017:05:11:52 -0700]
53&material=39 HTTP/1.1" 200 1
tp://www.google.com/bot.html)'
[30/Jun/2017:05:11:59 -0700] '
99&style_bottom=116 HTTP/1.1"
tp://www.google.com/bot.html)'
[30/Jun/2017:05:12:16 -0700] '
```

Run Script



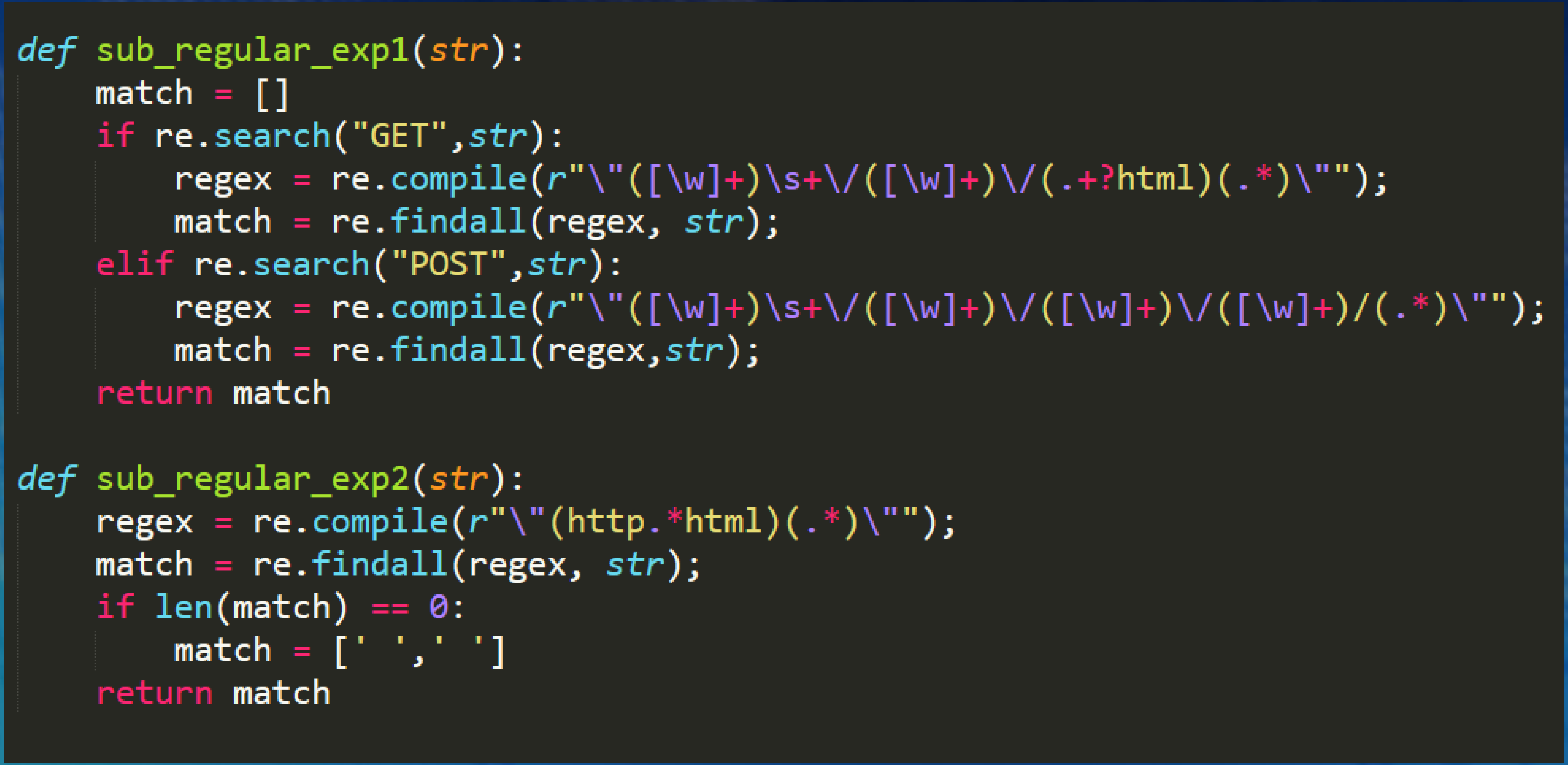
```
A = LOAD '/home/training/Downloads/WebLog2' USI
rarray,req_type:chararray,req_link:chararray,re
B = FOREACH A GENERATE ip_addr,temp1,timestamp,
data = distinct B;
dump data;
STORE data INTO '/home/training/Downloads/Parse
/* Calculate the number of web pages a user vis
ip_data = GROUP data by ip_addr;
ip_count = FOREACH ip_data GENERATE group AS ti
dump ip_count;
/* Statistics where requests were successfull i
time_data = GROUP data BY timestamp;
byte_count = FOREACH time_data GENERATE group A
dump byte_count;
```

After



TIMESTAMP	TIMEZONE	REQUEST_TYPE
30/Jun/2017:05:11:49	0700	"GET /men/bottoms
30/Jun/2017:05:11:52	0700	"GET /men/bottoms
30/Jun/2017:05:11:59	0700	"GET /men/bottoms
30/Jun/2017:05:12:16	0700	"GET /men/bottoms

Sample of Python Script



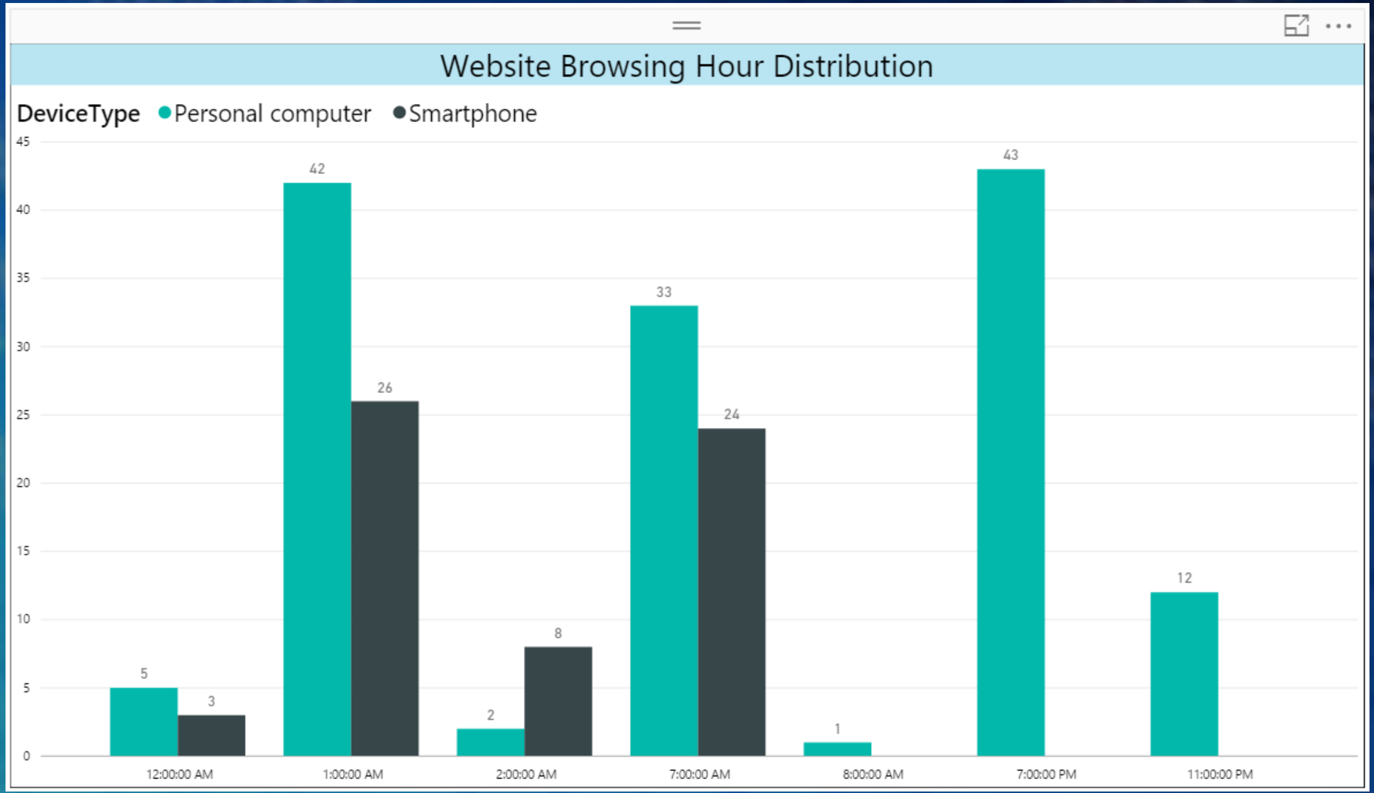
```
def sub_regular_exp1(str):
    match = []
    if re.search("GET",str):
        regex = re.compile(r"\"([\w]+)\s+\/([\w]+)\/(.+?html)(.*)\"");
        match = re.findall(regex, str);
    elif re.search("POST",str):
        regex = re.compile(r"\"([\w]+)\s+\/([\w]+)\/([\w]+)\/([\w]+)/(.*)\"");
        match = re.findall(regex,str);
    return match

def sub_regular_exp2(str):
    regex = re.compile(r"\"(http.*html)(.*)\"");
    match = re.findall(regex, str);
    if len(match) == 0:
        match = [' ',' ']
    return match
```

Stage 3: Data Visualization

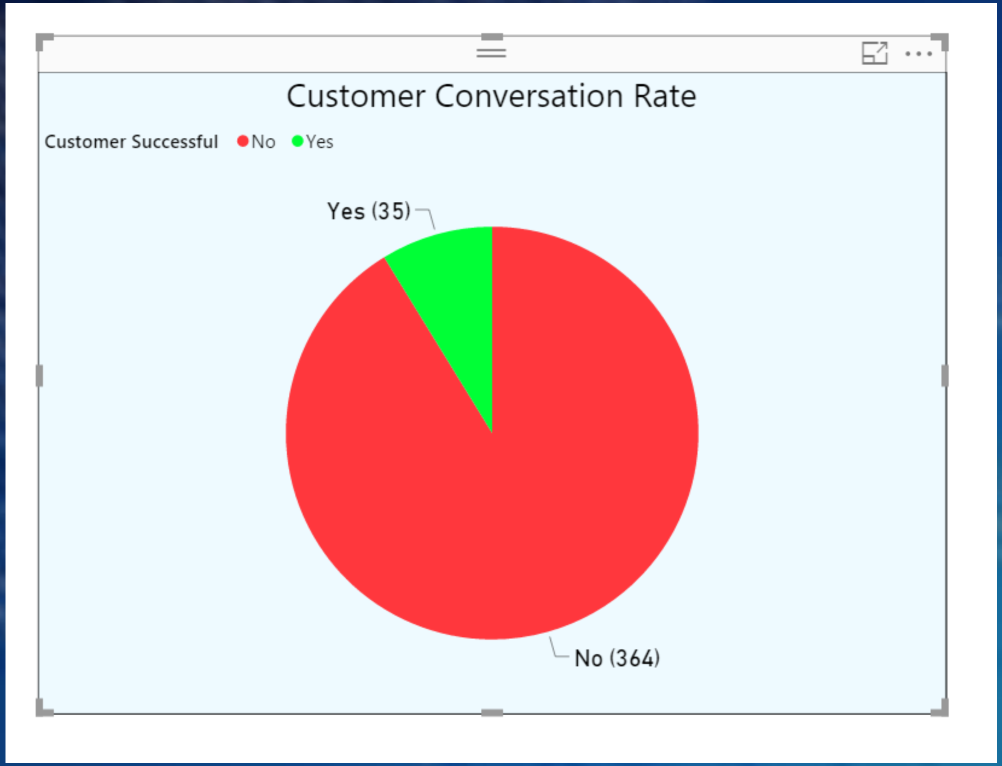
- Final stage of the process
- Structured Data exported from Hadoop to csv format
- Use of Business Intelligence tools such as Tableau and Power BI

Sample Reports

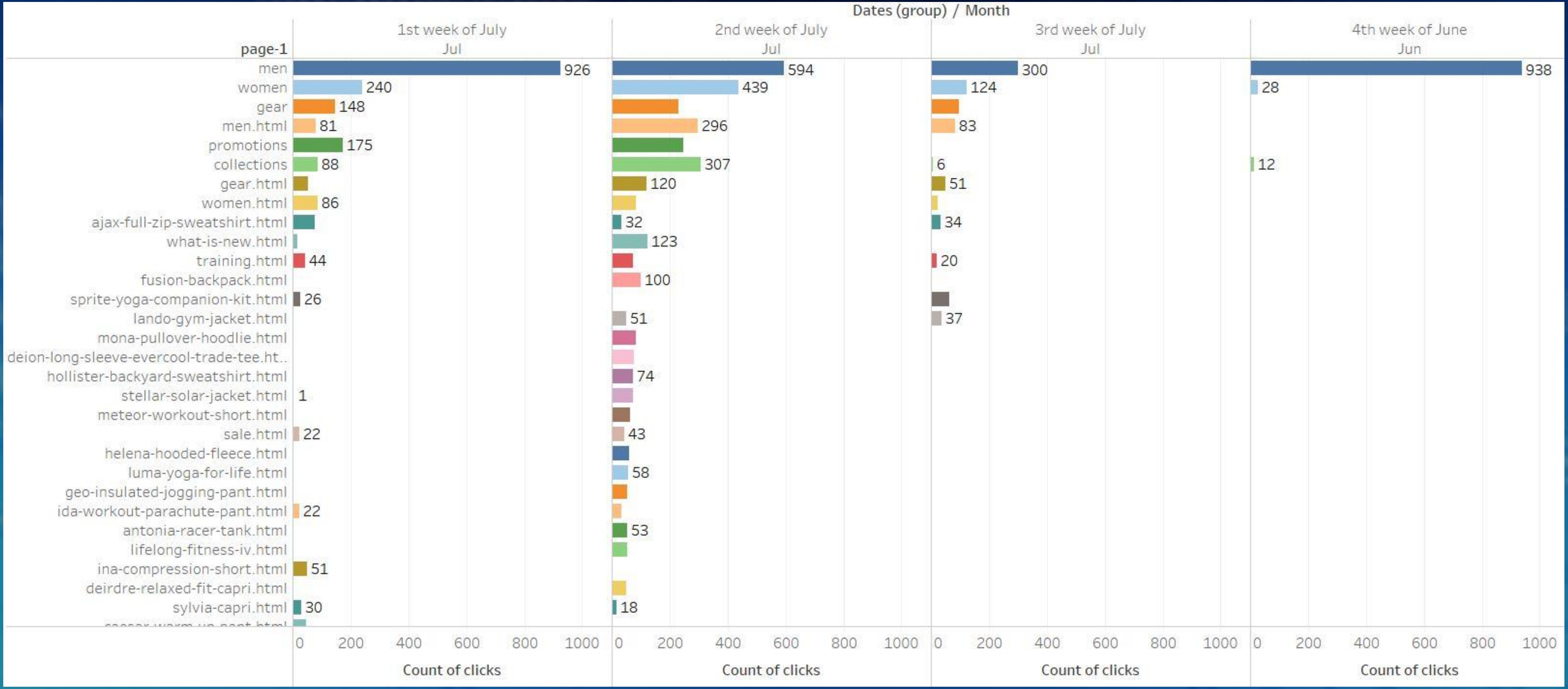
Website Browsing Hour Distribution



Customer Conversion Rate

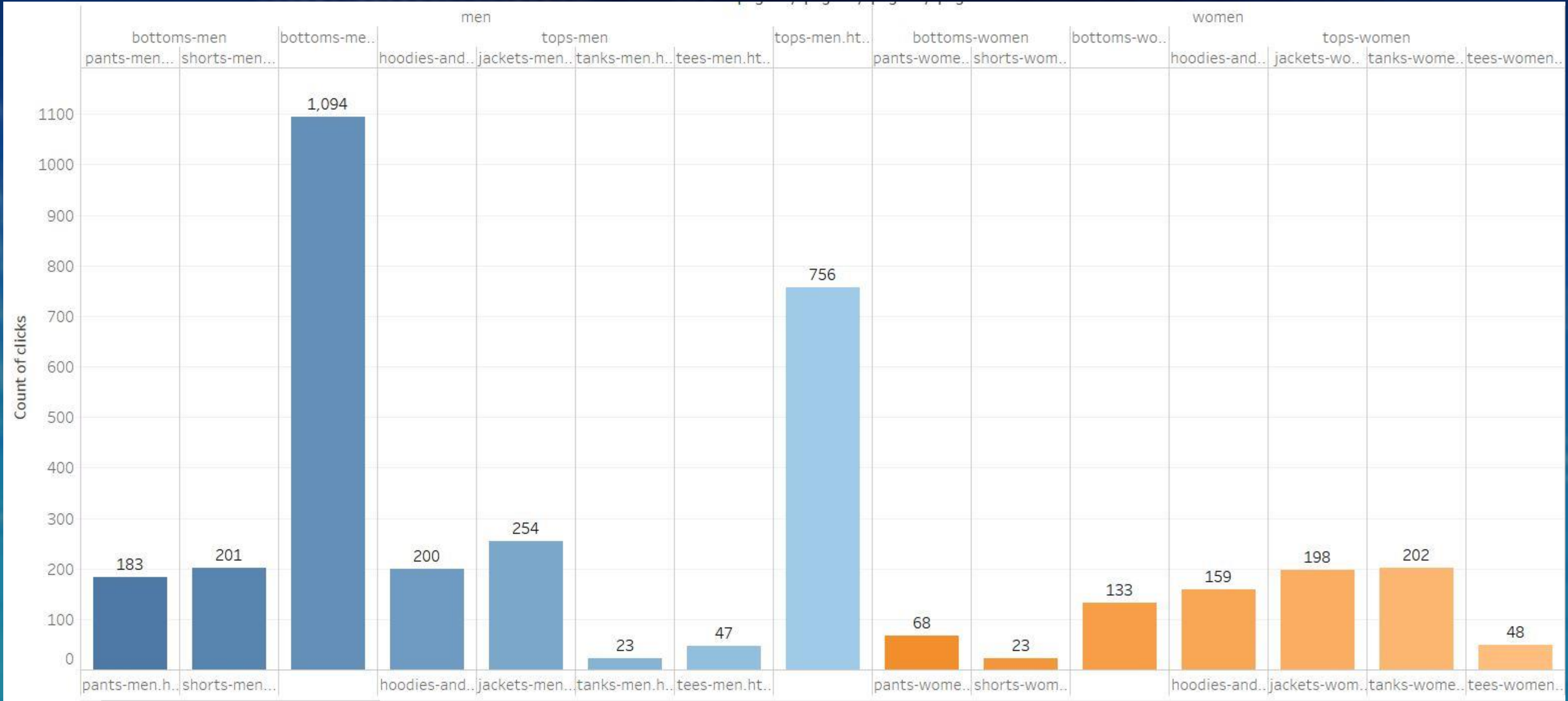


Weekly Distribution of Clicks per page



David Eccles School of Business
THE UNIVERSITY OF UTAH

Most Popular Product Category

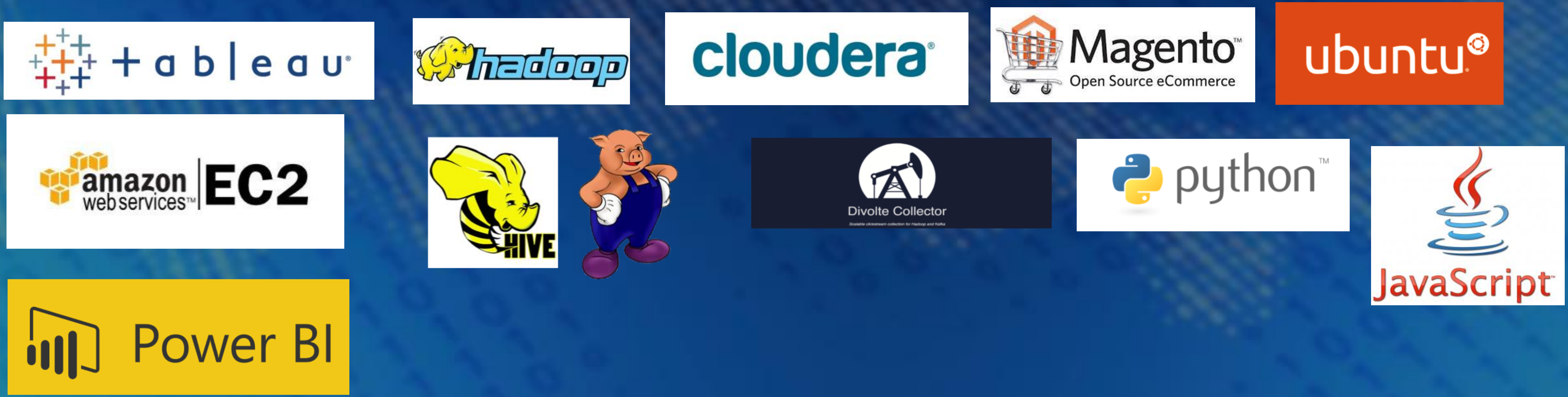


Challenges Faced

- eCommerce Website Setup
- Multi-node cluster creation on Amazon Ec2
- Cloudera Hadoop Installation on cluster
- Parsing links to get detailed information about Product Category and sub-category

Learnings and Conclusion

- Implementation of an end to end project
- Technology Stack worked on:



Special Thanks

- Thanks to our entire Capstone Faculty team and Sponsor for timely guidance
- Bi-weekly progress reports helped us to get a reality check of the project
- Great learning experience

Any Questions?

David Eccles School of Business
THE UNIVERSITY OF UTAH

Thank You!!!

David Eccles School of Business
THE UNIVERSITY OF UTAH