2019-2020

BLM4120 Big Data Processing and Analytics

MERVE ÜLKÜ ÖZKARA
15011028

ORÇUN ÇELİK
15011053

İçindekiler

1.General information about our project

We have raw data from a large multi-category online store.

We want to do several statistical operations with this data. Our operations are:

- Users behaviour of event types (view,purchase, etc.)
- The average prices of brands
- The brand that most viewed (Top 5)
- Time period that users have purchased products most

2.Goal

We will be using the “eCommerce behavior data from multi-category store” as a dataset. This file contains behavior data for October 2019 from a large multi-category online store. Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users. File structure is given below.

File structure

event_time

Time when event happened at (in UTC).

event_type

Events can be:

- view - a user viewed a product
- cart - a user added a product to shopping cart
- remove*from*cart - a user removed a product from shopping cart
- purchase - a user purchased a product

Typical funnel: view ⇒ cart ⇒ purchase.

product_id

ID of a product

category_id

Product's category ID

category_code

Product's category taxonomy (code name) if it was possible to make it. Usually present for meaningful categories and skipped for different kinds of accessories.

brand

Downcased string of brand name. Can be missed.

price

Float price of a product. Present.

user_id

Permanent user ID.

user_session

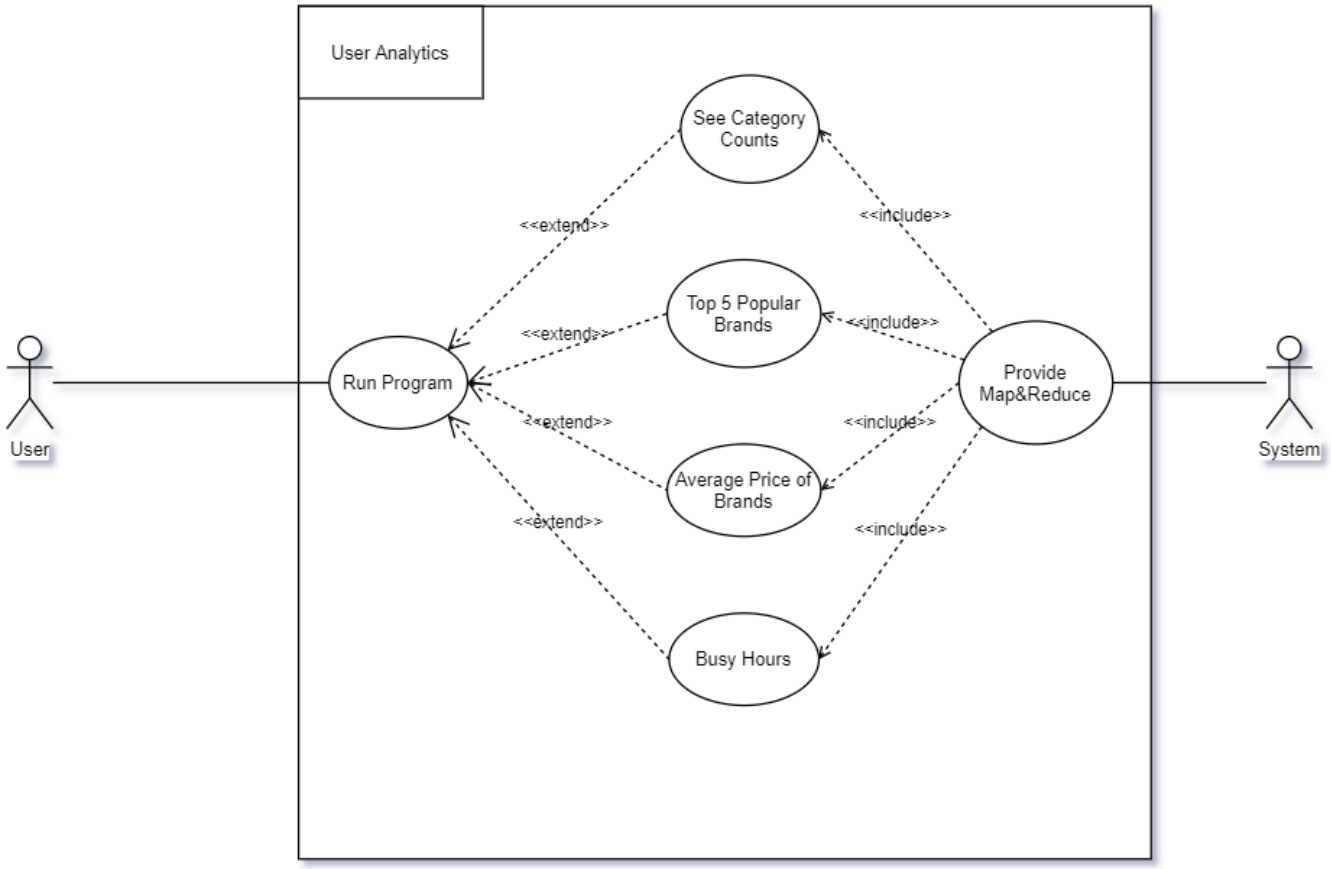
Temporary user's session ID. Same for each user's session. Is changed every time user come back to online store from a long pause.

3. Use Case

3.1.Use Case Scenerio

Name of Use Case Scenario:	**eCommerce Behavior Statistics**
Actor:	**User**
Definition:	**This use case scenario text consists eCommerce behavior's statistics**
Precondition:	**Program must be started**
Last Condition:	**The chosen statistical function must be shown to the user**
Main Scenario:	**1.System shows statistical methods to choose** **2.User chooses the statistical function** **3.System reads data from the file** **4.System shows results to the user**

3.2.Use Case Diagram



4.Graphicle User Interface

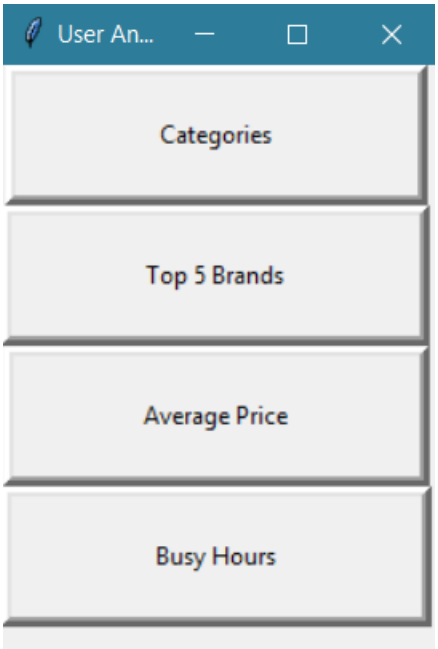
We've designed a simple GUI for several functions.

Categories: Finds user behaviour of site. (numbers of purchase,view etc.)

Top 5 Brands: Finds top 5 viewed brands.

Average Price: Finds average price for each brand.

Busy Hours: Finds busy times in website.



5.Technical challenges

We have faced many challenges along with this project. The biggest challenge was that the performance of our system was not adequate to start the Hadoop(Hortonworks Virtual Image) on our computer. After that, we've decided to do our project on python with MRJob and we used AWS Elastic Map Reduce Service for multi-node Hadoop operations.

6.Performance Evaluation

Amazon Storage Services

We've uploaded our data into Amazon S3 Storage Services. We could not upload big amounts of data due to upload speed.

Amazon S3 > mydata1501102815011053

mydata1501102815011053

Overview | Properties | Permissions | Management | Access points

Type a prefix and press Enter to search. Press ESC to clear.

Upload | Create folder | Download | Actions

US West (Oregon)

Viewing 1 to 4

Name	Last modified	Size	Storage class
26mb.csv	May 27, 2020 2:37:18 PM GMT+0300	25.7 MB	Standard
5mb.csv	May 27, 2020 2:29:11 PM GMT+0300	5.1 MB	Standard
midfile.csv	May 27, 2020 9:23:24 AM GMT+0300	123.4 MB	Standard
smallfile.csv	May 27, 2020 3:44:27 AM GMT+0300	1.3 MB	Standard

Viewing 1 to 4

6.1.In this section we use MRCategories function on AWS

-EMR Results of Small Data Set : 1.3MB with 1 node (Result = 4 Min taken)

Cluster:

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: --

Master public DNS: ec2-54-188-90-108.us-west-2.compute.amazonaws.com SSH

History service: --

Tags: __mrjob_label = MRCategorieswithmain, __mrjob_owner = ORCUN, __mrjob_version = 0.7.1 View All

Summary

ID: j-1EGWIOQGCBPR5
Creation date: 2020-05-27 13:44 (UTC+3)
End date: 2020-05-27 13:49 (UTC+3)
Elapsed time: 4 minutes
After last step completes: Cluster auto-terminates
Termination protection: Off

Configuration details

Release label: emr-5.27.0
Hadoop distribution: Amazon 2.8.5
Applications:
Log URI: s3://mrjob-b1bc9cdadbc79696/tmp/logs/
EMRFS consistent view:

Network and hardware

Availability zone: us-west-2b
Subnet ID: --
Master: Terminated 1 m5.xlarge
Core: --
Task: --

Command prompt when we communicate with AWS EMR

```
(base) C:\Users\ORCUN>python MRCategorieswithmain.py -r emr --num-core-instances 2 s3://mydata1501102815011053/smallfile.csv > smallwith2nodes.txt
No configs found; falling back on auto-configuration
No configs specified for emr runner
Using s3://mrjob-b1bc9cdadbc79696/tmp/ as our temp dir on S3
Creating temp directory C:\Users\ORCUN\AppData\Local\Temp\MRCategorieswithmain.ORCUN.20200527.111014.594565
writing master bootstrap script to C:\Users\ORCUN\AppData\Local\Temp\MRCategorieswithmain.ORCUN.20200527.111014.594565\b.sh
uploading working dir files to s3://mrjob-b1bc9cdadbc79696/tmp/MRCategorieswithmain.ORCUN.20200527.111014.594565/files/wd...
Copying other local files to s3://mrjob-b1bc9cdadbc79696/tmp/MRCategorieswithmain.ORCUN.20200527.111014.594565/files/
```

```
Using s3://mrjob-b1bc9cdadbc79696/tmp/ as our temp dir on S3
Creating temp directory C:\Users\ORCUN\AppData\Local\Temp\MRCategorieswithmain.ORCUN.20200527.111014.594565
writing master bootstrap script to C:\Users\ORCUN\AppData\Local\Temp\MRCategorieswithmain.ORCUN.20200527.111014.594565\b.sh
uploading working dir files to s3://mrjob-b1bc9cdadbc79696/tmp/MRCategorieswithmain.ORCUN.20200527.111014.594565/files/wd...
Copying other local files to s3://mrjob-b1bc9cdadbc79696/tmp/MRCategorieswithmain.ORCUN.20200527.111014.594565/files/
Created new cluster j-U89K7V10ZV8J
Added EMR tags to cluster j-U89K7V10ZV8J: __mrjob_label=MRCategorieswithmain, __mrjob_owner=ORCUN, __mrjob_version=0.7.1
Waiting for Step 1 of 1 (s-3A2KH1N0TSBK6) to complete...
  PENDING (cluster is STARTING)
  PENDING (cluster is STARTING)
  PENDING (cluster is STARTING)
  PENDING (cluster is STARTING: Configuring cluster software)
  PENDING (cluster is BOOTSTRAPPING: Running bootstrap actions)
  PENDING (cluster is BOOTSTRAPPING: Running bootstrap actions)
  PENDING (cluster is BOOTSTRAPPING: Running bootstrap actions)
  master node is ec2-34-219-161-213.us-west-2.compute.amazonaws.com
  RUNNING for 0:00:21
  COMPLETED
Attempting to fetch counters from logs...
Waiting for cluster (j-U89K7V10ZV8J) to terminate...
  TERMINATING
  TERMINATING
  TERMINATING
  TERMINATED
```

-Small Data Set: 1.3MB with 2 nodes (Result = 6 Min taken)

Cluster: MRCategorieswithmain.ORCUN.20200527.111014.594565 Terminated Steps completed

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: --

Master public DNS: ec2-34-219-161-213.us-west-2.compute.amazonaws.com SSH

History service: --

Tags: __mrjob_label = MRCategorieswithmain, __mrjob_owner = ORCUN, __mrjob_version = 0.7.1 View All

Summary

ID: j-U89K7V10ZV8J
Creation date: 2020-05-27 14:10 (UTC+3)
End date: 2020-05-27 14:17 (UTC+3)
Elapsed time: 6 minutes
After last step completes: Cluster auto-terminates
Termination protection: Off

Configuration details

Release label: emr-5.27.0
Hadoop distribution: Amazon 2.8.5
Applications: --
Log URI: s3://mrjob-b1bc9cdadbc79696/tmp/logs/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-west-2b
Subnet ID: --
Master: Terminated 1 m5.xlarge
Core: Terminated 2 m5.xlarge
Task: --

-EMR Results of Small Data Set : 5 MB with 2 Nodes (Result = 6 Min taken)

Cluster: MRCategorieswithmain.ORCUN.20200527.113028.085539 Terminated Steps completed

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: --

Master public DNS: ec2-54-212-27-254.us-west-2.compute.amazonaws.com SSH

History service: --

Tags: __mrjob_label = MRCategorieswithmain, __mrjob_owner = ORCUN, __mrjob_version = 0.7.1 View All

Summary

ID: j-H4IYM8SHJCPY
Creation date: 2020-05-27 14:30 (UTC+3)
End date: 2020-05-27 14:37 (UTC+3)
Elapsed time: 6 minutes
After last step completes: Cluster auto-terminates
Termination protection: Off

Configuration details

Release label: emr-5.27.0
Hadoop distribution: Amazon 2.8.5
Applications: --
Log URI: s3://mrjob-b1bc9cdadbc79696 /tmp/logs/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-west-2b
Subnet ID: --
Master: Terminated 1 m5.xlarge
Core: Terminated 2 m5.xlarge
Task: --

-EMR Results of Mid-Range Data Set : 128 MB with 4 Nodes (Result = 5 Min taken)

Cluster: MRCategorieswithmain.ORCUN.20200527.114223.865043 Terminated Steps completed

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: --

Master public DNS: ec2-34-219-80-34.us-west-2.compute.amazonaws.com SSH

History service: --

Tags: __mrjob_label = MRCategorieswithmain, __mrjob_owner = ORCUN, __mrjob_version = 0.7.1 View All

Summary

ID: j-1Z4RORANXYXL5
Creation date: 2020-05-27 14:44 (UTC+3)
End date: 2020-05-27 14:49 (UTC+3)
Elapsed time: 5 minutes
After last step completes: Cluster auto-terminates
Termination protection: Off

Configuration details

Release label: emr-5.27.0
Hadoop distribution: Amazon 2.8.5
Applications: --
Log URI: s3://mrjob-b1bc9cdadbc79696 /tmp/logs/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-west-2b
Subnet ID: --
Master: Terminated 1 m5.xlarge
Core: Terminated 4 m5.xlarge
Task: --

6.2.In this section we use MRAvgPrice function on AWS

-EMR Result of 128 MB File with 4 Nodes (Result 7 Minitues taken)

Cluster: MRAvgPricemain.ORCUN.20200527.115841.216173 Terminated Steps completed

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: --

Master public DNS: ec2-35-167-65-28.us-west-2.compute.amazonaws.com SSH

History service: --

Tags: __mrjob_label = MRAvgPricemain, __mrjob_owner = ORCUN, __mrjob_version = 0.7.1 View All

Summary

ID: j-1SXTS994WUC6N
Creation date: 2020-05-27 14:59 (UTC+3)
End date: 2020-05-27 15:06 (UTC+3)
Elapsed time: 7 minutes
After last step completes: Cluster auto-terminates
Termination protection: Off

Configuration details

Release label: emr-5.27.0
Hadoop distribution: Amazon 2.8.5
Applications: --
Log URI: s3://mrjob-b1bc9cdadbc79696/tmp/logs/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-west-2b
Subnet ID: --
Master: Terminated 1 m5.xlarge
Core: Terminated 4 m5.xlarge
Task: --

-EMR Results of 128 MB File with 6 Nodes (Result 6 Minitues taken)

Cluster: MRAvgPricemain.ORCUN.20200527.132306.072760 Terminated Steps completed

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: --

Master public DNS: ec2-54-202-89-235.us-west-2.compute.amazonaws.com SSH

History service: --

Tags: __mrjob_label = MRAvgPricemain, __mrjob_owner = ORCUN, __mrjob_version = 0.7.1 View All

Summary

ID: j-NPAKQX6YS360
Creation date: 2020-05-27 16:23 (UTC+3)
End date: 2020-05-27 16:29 (UTC+3)
Elapsed time: 6 minutes
After last step completes: Cluster auto-terminates
Termination protection: Off

Configuration details

Release label: emr-5.27.0
Hadoop distribution: Amazon 2.8.5
Applications: --
Log URI: s3://mrjob-b1bc9cdadbc79696/tmp/logs/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-west-2b
Subnet ID: --
Master: Terminated 1 m5.xlarge
Core: Terminated 6 m5.xlarge
Task: --

-128 MB File 8 Nodes (Result Error)

Cluster:

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: --

Master public DNS: --

History service: --

Tags: __mrjob_label = MRAvgPricemain, __mrjob_owner = ORCUN, __mrjob_version = 0.7.1 View All

Summary

ID: j-1JA1K16WYTK1X
Creation date: 2020-05-27 16:16 (UTC+3)
End date: 2020-05-27 16:20 (UTC+3)
Elapsed time: 4 minutes
After last step completes: Cluster auto-terminates
Termination protection: Off

Configuration details

Release label: emr-5.27.0
Hadoop distribution: Amazon 2.8.5
Applications:
Log URI: s3://mrjob-b1bc9cdadbc79696/tmp/logs/
EMRFS consistent view:

Network and hardware

Availability zone: us-west-2b
Subnet ID: --
Master: Terminated 1 m5.xlarge
Core: Terminated 8 m5.xlarge
Task: --

7.Conclusion

In smaller data sets:

Map Reduce time does not change when using a large number of clusters.

In contrary total time spend on AWS increases due to software installation on different clusters.

In larger data sets:

Using different clusters to run Map-Reduce saves time.