

Big Data Analytics Project

By:

SAAD AHMED KHAN – 24440

HAIDER SAEED KHAN – 24123

TALAL KHAN - 25253

Step By Step Procedure

Setting up Hadoop Cluster

Command used: docker-compose up -d

```
PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\HBASE\docker-hbase> docker-compose -f "docker-compose-standalone.yml" up -d
time="2024-12-31T23:09:59+05:00" level=warning msg="C:\\Users\\Saad Ahmed Khan\\Projects\\BIG DATA\\BigDataProject\\BDA_try11\\HBASE\\docker-hbase\\docker-compose-standalone.yml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid potential confusion"
[+] Running 11/11
 ✓ Network bda_network      Created                                0.0s
 ✓ Container nodemanager    Started                               2.5s
 ✓ Container namenode       Started                               2.5s
 ✓ Container spark-master-bda Started                               2.6s
 ✓ Container hbase          Started                               2.5s
 ✓ Container historyserver  Started                               2.2s
 ✓ Container resourceanage  Started                               2.2s
 ✓ Container datanode       Started                               2.6s
 ✓ Container zookeeperbda   Started                               2.6s
 ✓ Container kafka          Started                               2.9s
 ✓ Container spark-worker-bda1 Started                               2.7s
PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\HBASE\docker-hbase>
```

Ingesting Data into HDFS

Running the upload_to_namenode.sh script (in the scripts folder)

```
PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\scripts> bash upload_to_namenode.sh
Successsfully copied 844MB to namenode:/data/temp_data1.csv
File successfully copied and renamed to data.csv in /data inside namenode container.
PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\scripts>
```

Setting up and running kafka producer.py file to generate data

```
PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\scripts> python ./producer.py
start
Message delivered to data4 [0]
Message delivered to data4 [0]
Message delivered to data4 [0]
Message delivered to data4 [0]
Message delivered to data4 [0]
Message delivered to data4 [0]
□
```

Setting up and running kafka consumer.py to ingest data inside namenode

```

PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\scripts> python consumer.py
Consumer started
Listening to topic: data4
Received message: {'OrderID': '4668779', 'CustomerID': '68406', 'ProductID': '9', 'Quantity': '7', 'OrderDate': '2022-08-30 12:00:00', 'ShippingAddress': '329 Elliott Crossroad Suite 082 Ericaview, AR 59111', 'ShippingDate': '2020-01-29 12:00:00', 'Name': 'Jason Gomez', 'Age': '36', 'Country': 'France', 'RegistrationDate': '2018-08-11 12:00:00', 'ProductName': 'Smartphone', 'Category': 'Home Appliances', 'Price': '606.62', 'TotalAmount': ''}
Received message: {'OrderID': '4668780', 'CustomerID': '31603', 'ProductID': '14', 'Quantity': '9', 'OrderDate': '2020-11-25 12:00:00', 'ShippingAddress': '92365 Olivia Port Lake Amanda, GU 20908', 'ShippingDate': '2020-06-08 12:00:00', 'Name': 'Rebecca Rasmussen', 'Age': '48', 'Country': 'India', 'RegistrationDate': '2017-02-02 12:00:00', 'ProductName': 'Printer', 'Category': 'Electronics', 'Price': '737.01', 'TotalAmount': ''}
Received message: {'OrderID': '4668781', 'CustomerID': '3551', 'ProductID': '8', 'Quantity': '8', 'OrderDate': '2023-10-19 12:00:00', 'ShippingAddress': '90363 Reynolds Harbor Suite 407 Deborahnton, DE 43533', 'ShippingDate': '2020-03-29 12:00:00', 'Name': 'Jason Pittman', 'Age': '80', 'Country': 'India', 'RegistrationDate': '2017-09-20 12:00:00', 'ProductName': 'Desk Lamp', 'Category': 'Toys', 'Price': '703.35', 'TotalAmount': ''}
Received message: {'OrderID': '4668782', 'CustomerID': '85697', 'ProductID': '16', 'Quantity': '8', 'OrderDate': '2023-03-17 12:00:00', 'ShippingAddress': '0879 James Place Apt. 960 North Deborah, NJ 01378', 'ShippingDate': '2022-08-01 12:00:00', 'Name': 'Matthew Carson', 'Age': '36', 'Country': 'Pakistan', 'RegistrationDate': '2021-07-13 12:00:00', 'ProductName': 'Backpack', 'Category': 'Electronics', 'Price': '599.01', 'TotalAmount': ''}
Received message: {'OrderID': '4668783', 'CustomerID': '41728', 'ProductID': '18', 'Quantity': '5', 'OrderDate': '2023-07-12 12:00:00', 'ShippingAddress': '086 Kerri View Barnettville, PA 01333', 'ShippingDate': '2023-12-08 12:00:00', 'Name': 'Stephen Johnson', 'Age': '72', 'Country': 'India', 'RegistrationDate': '2019-07-12 12:00:00', 'ProductName': 'Backpack', 'Category': 'Electronics', 'Price': '420.1', 'TotalAmount': '2100.5'}
Received message: {'OrderID': '4668784', 'CustomerID': '24390', 'ProductID': '9', 'Quantity': '4', 'OrderDate': '2020-02-21 12:00:00', 'ShippingAddress': 'Unit 5987 Box 2353 DPO AE 55766', 'ShippingDate': '2020-01-26 12:00:00', 'Name': 'Jessica Morales', 'Age': '73', 'Country': 'UK', 'RegistrationDate': '2019-10-20 12:00:00', 'ProductName': 'Smartphone', 'Category': 'Home Appliances', 'Price': '606.62', 'TotalAmount': '242

```

Inserting Data into HDFS using the upload_to_hdfs.sh script

```

PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\scripts> bash upload_to_hdfs.sh
24/12/31 18:18:02 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /data/data.csv
File uploaded to HDFS: /data/data.csv
PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\scripts>

```

Finally verifying that the data has been moved to hdfs:

```

C:\Users\Saad Ahmed Khan>docker exec -it namenode /bin/bash
root@8d4d8b9dea16:/# hdfs dfs -ls /data
Found 1 items
-rw-r--r-- 3 root supergroup 843532177 2024-12-31 18:18 /data/data.csv
root@8d4d8b9dea16:/#

```

Printing data using cat:

```

r,20,USA,2021-01-07 12:00:00,Washing Machine,Home Appliances,159.92,
51248,19709,6,8,2022-10-13 12:00:00,Unit 7300 Box 9573 DPO AA 37231,2021-10-02 12:00:00,Matthew Ferguson,20,USA,2016-04-
25 12:00:00,GConsole,Electronics,169.32,1354.56
51249,62858,20,6,2023-10-26 12:00:00,"552 Peter Points Apt. 087 Lisamouth, MD 44212",2020-02-08 12:00:00,Jennifer Gutier
rez,31,Canada,2014-03-30 12:00:00,Smartphone,Electronics,420.68,
51250,473,19,4,2022-12-06 12:00:00,"0933 Brooks Plains Gibsonburgh, NV 72823",2020-02-22 12:00:00,John Ortiz,40,USA,2021
-11-26 12:00:00,Desk Lamp,Electronics,864.39,
51251,34791,3,4,2021-03-23 12:00:00,"0588 April Heights Hoffmanfurt, LA 11092",2021-01-03 12:00:00,Lisa Branch,55,Austra
lia,2014-08-19 12:00:00,Backpack,Electronics,147.35,
51252,55898,10,4,2022-07-04 12:00:00,"9985 Whitney Expressway Singhhaven, CT 47997",2021-03-18 12:00:00,Michael Jones,21
,Australia,2016-09-07 12:00:00,Washing Machine,Home Appliances,159.92,
51253,7677,17,4,2020-01-05 12:00:00,"80021 Rojas Fields Lake Ronald, WA 80245",2022-08-09 12:00:00,Michael Phillips,68,A
ustralia,2016-04-15 12:00:00,Desk Lamp,Electronics,364.13,
51254,60458,20,4,2022-06-18 12:00:00,"1128 Erin Avenue Port Danielton, IN 81718",2023-10-20 12:00:00,Bryan Martinez,59,U
SA,2022-03-04 12:00:00,Smartphone,Electronics,420.68,
51255,74111,19,4,2020-01-16 12:00:00,"69903 Moore Circle Apt. 827 Lake Joseton, WA 11635",2020-12-24 12:00:00,Brittany W
right,63,France,2017-08-19 12:00:00,Desk Lamp,Electronics,864.39,
51256,40751,6,4,2022-11-01 12:00:00,USNV Bowman FPO AE 18650,2022-09-24 12:00:00,Kendra Snyder,73,USA,2016-04-02 12:00:0
0,GConsole,Electronics,169.32,
51257,6044,14,1,2020-10-05 12:00:00,"192 Ashley Trace Port Francisco, RI 71179",2023-05-12 12:00:00,Todd Arnold,41,UK,20
22-09-10 12:00:00,Printer,Electronics,737.01,737.01
51258,21875,7,3,2023-03-13 12:00:00,"8535 Schwartz Roads Suite 640 Griffinside, OK 54102^C
51324,74346,12,4,2023-01-23 12:00:00,"42411 Jensen Mall Suite 577 Wilsonton, MO 50346",2023-07-17 12:00:00,Scott Thomas,
52,UK,2014-04-16 12:00:00,GConsole,Clothing,38.31,153.24
51325,67089,16,2,2023-12-19 12:00:00,"1885 Garza Field Apt. 536 Susanmouth, NM 36444",2020-04-26 12:00:00,Jeremy Mccoy,4
9,India,2023-12-09 12:00:00,Backpack,Electronics,599.01,1198.02
51326,23473,15,9,2021-09-20 12:00:00,"7011 Chad Summit Garrisontown, TX 73393",2022-06-15 12:00:00,Jessica Jones,42,Fran
ce,2016-05-17 12:00:00,Mouse,Electronics,589.75,
51327,41803,14,7,2023-04-12 12:00:00,cat: Filesystem closed
root@8d4d8b9dea16:/#

```

EDA on ingested data using python:

All the EDA steps can be found in the attached EDA notebook. A brief summary is provided below.

1) Displaying column info;

```
df = pd.read_csv(StringIO(file_content))

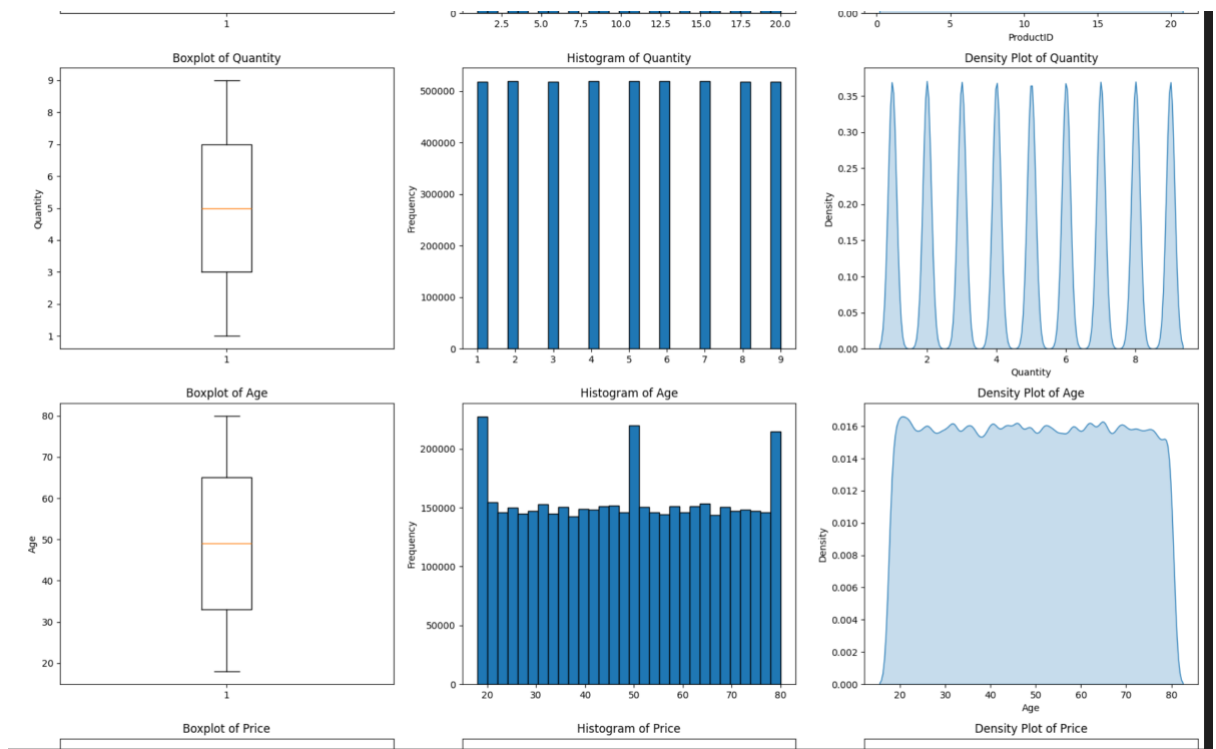
# Show the DataFrame
print(df.head())
```

	OrderID	CustomerID	ProductID	Quantity	OrderDate	ShippingAddress	ShippingDate
0	1	92256	15	1	2023-06-16 12:00:00	485 Jackson Falls Apt. 624 Lake Michelleland, ...	2021-05-23 12:00:00
1	2	7170	2	7	2020-11-09 12:00:00	074 Brown Estate Suite 175 Robinbury, MO 37341	2022-10-01 12:00:00
2	3	57706	20	6	2021-11-14 12:00:00	0362 Adam Vista Patriciaborough, UT 39097	2023-09-27 12:00:00
3	4	78850	16	4	2020-08-13 12:00:00	329 Isabel Ranch Grimesshire, HI 37219	2022-03-05 12:00:00
4	5	37213	7	2	2021-11-28 12:00:00	711 Skinner Street Suite 169 East Annette, SC ...	2020-06-23 12:00:00

	Name	Age	Country	RegistrationDate	ProductName
0	Kyle Hunt	20	Canada	2022-07-01 12:00:00	Mouse
1	Taylor Black	22	Pakist	2021-12-21 12:00:00	Headphones
2	Mario Bartlett	73	UK	2014-01-19 12:00:00	Smartphone
3	Danielle Williams	57	Pakist	2018-01-27 12:00:00	Backpack
4	Jerry Hall	27	Germany	2014-10-23 12:00:00	Printer

	Category	Price	TotalAmount
0	Electronics	589.75	589.75
1	Electronics	803.02	NaN
2	Electronics	420.68	2524.08
3	Electronics	599.01	NaN
4	Books	251.30	NaN

2) Box plot, histograms and frequency density charts for all numerical columns:



3) Frequency tables for categorical columns

```
start
Frequency Table for OrderDate:
OrderDate
2020-12-28 12:00:00    3406
2022-10-29 12:00:00    3391
2021-03-18 12:00:00    3369
2022-08-01 12:00:00    3364
2023-05-30 12:00:00    3355
...
2023-08-16 12:00:00    3036
2022-04-13 12:00:00    3030
2021-06-15 12:00:00    3024
2023-11-01 12:00:00    3012
2021-11-28 12:00:00    3010
Name: count, Length: 1461, dtype: int64

Frequency Table for ShippingAddress:
ShippingAddress
USNS King FPO AE 57227    2
USNV Smith FPO AA 93867    2
USCGC Wagner FPO AA 61054    2
USNV Johnson FPO AP 44491    2
USS Torres FPO AE 29227    2
..
PSC 4349, Box 2735 APO AE 79661    1
...
08300 Benjamin Shoals East Annettetfurt, HI 73623    1
864 Garcia Ramp Gonzalezborough, NH 74311    1
Name: count, Length: 4668759, dtype: int64
```

4) Finding numerical Column outliers

```

outliers = df[(df[column] < lower_bound) | (df[column] > upper_bound)][column]

if not outliers.empty:
    print(f"{column} Outliers:")
    for outlier in outliers:
        print(f"- {outlier}")
    print()

numerical_outliers(data)

[7]

... start
TotalAmount Outliers:
- 7779.51
- 7227.18
- 7227.18
- 7227.18
- 7072.2
- 7072.2
- 7227.18
- 7072.2
- 7779.51
- 7227.18
- 7779.51
- 7779.51
- 7227.18
- 7779.51
- 7779.51
- 7779.51
- 7072.2
- 7072.2
- 7072.2
- 7779.51
- 7779.51
- 7779.51
...

```

5) Missing Value analysis

```

print('start')
def check_missing_data(df):
    missing_summary = df.isnull().sum()
    missing_percentage = (missing_summary / len(df)) * 100

    print("Missing Values Summary:")
    print(pd.DataFrame({
        'Missing Count': missing_summary,
        'Percentage': missing_percentage
    }))

    return missing_summary, missing_percentage

missing_summary, missing_percentage = check_missing_data(data)
print(missing_summary)

```

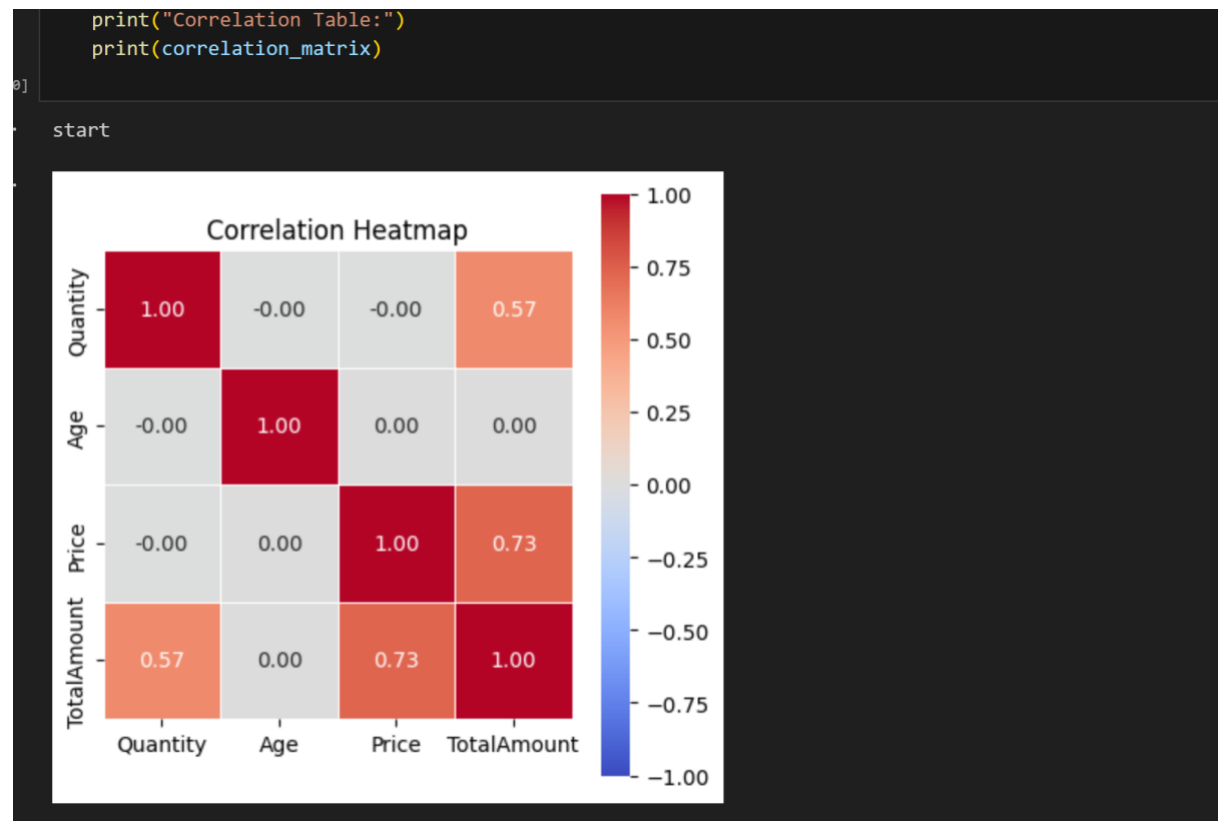
```

start
Missing Values Summary:

```

	Missing Count	Percentage
OrderID	0	0.000000
CustomerID	0	0.000000
ProductID	0	0.000000
Quantity	0	0.000000
OrderDate	0	0.000000
ShippingAddress	0	0.000000
ShippingDate	0	0.000000
Name	0	0.000000
Age	0	0.000000
Country	0	0.000000
RegistrationDate	0	0.000000
ProductName	0	0.000000
Category	0	0.000000
Price	0	0.000000
TotalAmount	3268360	70.004269
OrderID	0	
CustomerID	0	

6) Correlation Analysis:



Applying Transformation on ingested data

Using the run_spark_job.sh file

```
PS C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_try11\scripts> bash run_spark_job.sh
24/12/31 18:26:20 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /transformation
Successfully copied 4.1kB to spark-master-bda:/spark/spark_analysis.py
24/12/31 18:26:23 INFO SparkContext: Running Spark version 3.3.0
24/12/31 18:26:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/12/31 18:26:23 INFO ResourceUtils: =====
24/12/31 18:26:23 INFO ResourceUtils: No custom resources configured for spark.driver.
24/12/31 18:26:23 INFO ResourceUtils: =====
24/12/31 18:26:23 INFO SparkContext: Submitted application: Transform and Combine Partitioned Files
24/12/31 18:26:23 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -
> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.
0)
24/12/31 18:26:23 INFO ResourceProfile: Limiting resource is cpu
24/12/31 18:26:23 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/12/31 18:26:23 INFO SecurityManager: Changing view acls to: root
24/12/31 18:26:23 INFO SecurityManager: Changing modify acls to: root
24/12/31 18:26:23 INFO SecurityManager: Changing view acls groups to:
24/12/31 18:26:23 INFO SecurityManager: Changing modify acls groups to:
24/12/31 18:26:23 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); groups with view
permissions: Set(); users with modify permissions: Set(root); groups with modify permissions: Set()
24/12/31 18:26:23 INFO Utils: Successfully started service 'sparkDriver' on port 45249.
24/12/31 18:26:23 INFO SparkEnv: Registering MapOutputTracker
24/12/31 18:26:23 INFO SparkEnv: Registering BlockManagerMaster
24/12/31 18:26:24 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/12/31 18:26:24 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/12/31 18:26:24 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/12/31 18:26:24 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-c68cc44e-c480-4e0d-b3d2-0f7ceed61ac4
24/12/31 18:26:24 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
```



```

nments Map()
24/12/31 18:26:34 INFO Executor: Running task 0.0 in stage 2.0 (TID 17)
24/12/31 18:26:34 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
24/12/31 18:26:34 INFO FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
24/12/31 18:26:34 INFO SQLHadoopMapReduceCommitProtocol: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
24/12/31 18:26:34 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 0-52982905, partition values: [empty row]
24/12/31 18:26:34 INFO CodeGenerator: Code generated in 10.536971 ms
24/12/31 18:26:38 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 52982905-105965810, partition values: [empty row]
24/12/31 18:26:42 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 105965810-158948715, partition values: [empty row]
24/12/31 18:26:45 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 158948715-211931620, partition values: [empty row]
24/12/31 18:26:48 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 211931620-264914525, partition values: [empty row]
24/12/31 18:26:52 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 264914525-317897430, partition values: [empty row]
24/12/31 18:26:55 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 317897430-370880335, partition values: [empty row]
24/12/31 18:26:59 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 370880335-423863240, partition values: [empty row]
24/12/31 18:27:02 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 423863240-476846145, partition values: [empty row]
24/12/31 18:27:05 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 476846145-529829050, partition values: [empty row]
24/12/31 18:27:09 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 529829050-582811955, partition values: [empty row]
24/12/31 18:27:12 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 582811955-635794860, partition values: [empty row]
24/12/31 18:27:16 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 635794860-688777765, partition values: [empty row]
24/12/31 18:27:19 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 688777765-741760670, partition values: [empty row]
24/12/31 18:27:22 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 741760670-794743575, partition values: [empty row]
24/12/31 18:27:25 INFO FileScanRDD: Reading File path: hdfs://namenode:9000/data/data.csv, range: 794743575-843532177, partition values: [empty row]
24/12/31 18:27:30 INFO FileOutputCommitter: Saved output of task 'attempt_local1491283818_0001_m_000006_0' to hdfs://namenode:9000/data/output/attempt_local1491283818_0001_m_000006_0

```

Verifying that spark job has done its job (saved the transomed data into transformation directory in hdfs)

```

C:\Users\Saad Ahmed Khan>docker exec -it namenode /bin/bash
root@8d4d8b9dea16:/# hdfs dfs -ls /
Found 8 items
drwxrwxrwt   - root root                0 2024-12-30 17:37 /app-logs
drwxr-xr-x   - root supergroup          0 2024-12-31 18:18 /data
-rw-r--r--   3 root supergroup 843411626 2024-12-30 16:08 /data.csv
drwxr-xr-x   - root supergroup          0 2024-12-31 18:10 /hbase
drwxr-xr-x   - root supergroup          0 2024-12-30 17:52 /output
drwxr-xr-x   - root supergroup          0 2024-12-30 14:19 /rmstate
drwx-----  - root supergroup          0 2024-12-30 17:36 /tmp
drwxr-xr-x   - root supergroup          0 2024-12-31 18:27 /transformation
root@8d4d8b9dea16:/# hdfs dfs -ls /transformation
Found 2 items
-rw-r--r--   3 root supergroup                0 2024-12-31 18:27 /transformation/_SUCCESS
-rw-r--r--   3 root supergroup 860892470 2024-12-31 18:27 /transformation/part-00000-95176dd5-8e08-4a55-af2a-eda6706a0db0-c000.csv
root@8d4d8b9dea16:/# |

```

Ingesting data from hdfs to hbase

Running hdfs_to_hbase script

```

2024-12-31 18:32:28,128 INFO [LocalJobRunner Map Task Executor #0] client.ConnectionManager$HConnectionImplementation: Closing zookeeper sessionId=0x1941de92f650015
2024-12-31 18:32:28,133 INFO [LocalJobRunner Map Task Executor #0] zookeeper.ZooKeeper: Session: 0x1941de92f650015 closed
2024-12-31 18:32:28,133 INFO [LocalJobRunner Map Task Executor #0-EventThread] zookeeper.ClientCnxn: EventThread shut down
2024-12-31 18:32:28,980 INFO [main] mapreduce.Job: map 100% reduce 0%
2024-12-31 18:32:34,037 INFO [communication thread] mapred.LocalJobRunner: map > map
2024-12-31 18:32:34,954 INFO [main] mapreduce.Job: map 95% reduce 0%
2024-12-31 18:32:37,038 INFO [communication thread] mapred.LocalJobRunner: map > map
2024-12-31 18:32:37,305 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner: map > map
2024-12-31 18:32:37,337 INFO [LocalJobRunner Map Task Executor #0] client.ConnectionManager$HConnectionImplementation: Closing zookeeper sessionId=0x1941de92f650014
2024-12-31 18:32:37,339 INFO [LocalJobRunner Map Task Executor #0] zookeeper.ZooKeeper: Session: 0x1941de92f650014 closed
2024-12-31 18:32:37,339 INFO [LocalJobRunner Map Task Executor #0-EventThread] zookeeper.ClientCnxn: EventThread shut down
2024-12-31 18:32:37,349 INFO [LocalJobRunner Map Task Executor #0] mapred.Task: Task:attempt_local1491283818_0001_m_000006_0 is done. And is in the process of committing
2024-12-31 18:32:37,350 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner: map
2024-12-31 18:32:37,350 INFO [LocalJobRunner Map Task Executor #0] mapred.Task: Task 'attempt_local1491283818_0001_m_000006_0' done.
2024-12-31 18:32:37,350 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner: Finishing task: attempt_local1491283818_0001_m_000006_0
2024-12-31 18:32:37,351 INFO [Thread-34] mapred.LocalJobRunner: map task executor complete.
2024-12-31 18:32:37,960 INFO [main] mapreduce.Job: map 100% reduce 0%
2024-12-31 18:32:37,960 INFO [main] mapreduce.Job: Job job_local1491283818_0001 completed successfully
2024-12-31 18:32:38,036 INFO [main] mapreduce.Job: Counters: 24

```

```

File System Counters
  FILE: Number of bytes read=179718186
  FILE: Number of bytes written=183172633
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=3679575350
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=42
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=0
Map-Reduce Framework
  Map input records=4668792
  Map output records=4668792
  Input split bytes=777
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=8096
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=879886336
ImportTsv
  Bad Lines=0
File Input Format Counters
  Bytes Read=860917046
File Output Format Counters

```

Verifying that data has been inserted inside hbase table

```

hbase(main):004:0> scan 'ec'
ROW COLUMN+CELL
1 column=customer:Age, timestamp=1735669811231, value=20
^C 1 column=customer:Country, timestamp=1735669811231, value=Canada

1 column=customer:CustomerID, timestamp=1735669811231, value=92256
1 column=customer:Name, timestamp=1735669811231, value=Kyle Hunt
1 column=customer:RegistrationDate, timestamp=1735669811231, value=2022-07-01 12:00:00
1 column=order:OrderDate, timestamp=1735669811231, value=2023-06-16 12:00:00
1 column=order:Quantity, timestamp=1735669811231, value=1
1 column=order:TotalAmount, timestamp=1735669811231, value=589.75
1 column=product:Category, timestamp=1735669811231, value=Electronics
1 column=product:Price, timestamp=1735669811231, value=589.75
1 column=product:ProductID, timestamp=1735669811231, value=15
1 column=product:ProductName, timestamp=1735669811231, value=Mouse
1 column=shipment:ShippingAddress, timestamp=1735669811231, value=485 Jackson Falls Apt. 6
24 Lake Michelleland LA 31239
1 column=shipment:ShippingDate, timestamp=1735669811231, value=2021-05-23 12:00:00

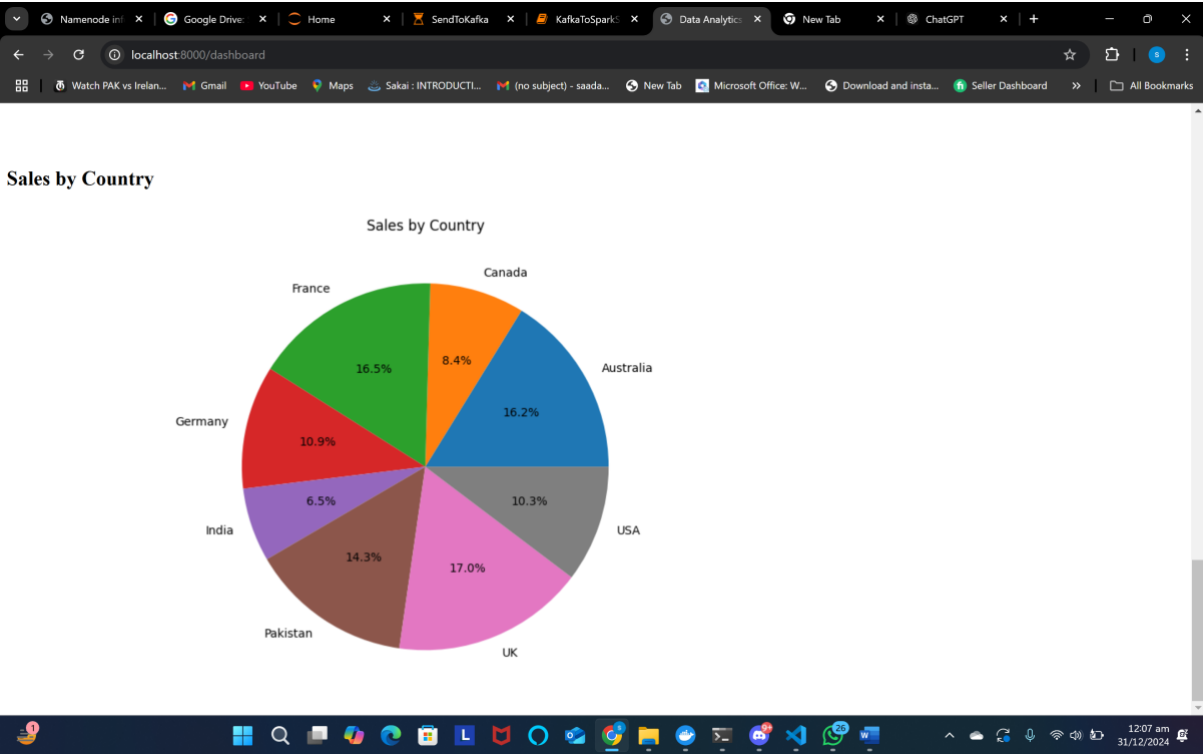
1000742 column=product:Price, timestamp=1735669811231, value=38.31
1000742 column=product:ProductID, timestamp=1735669811231, value=12
1000742 column=product:ProductName, timestamp=1735669811231, value=Gaming Console
1000742 column=shipment:ShippingAddress, timestamp=1735669811231, value=54515 Rodriguez Glens Su
ite 695 Denisebury NE 27642
1000742 column=shipment:ShippingDate, timestamp=1735669811231, value=2021-08-18 12:00:00
1000743 column=customer:Age, timestamp=1735669811231, value=22
1000743 column=customer:Country, timestamp=1735669811231, value=Australia
1000743 column=customer:CustomerID, timestamp=1735669811231, value=42919
1000743 column=customer:Name, timestamp=1735669811231, value=Lawrence Coleman
1000743 column=customer:RegistrationDate, timestamp=1735669811231, value=2015-07-03 12:00:00
1000743 column=order:OrderDate, timestamp=1735669811231, value=2021-08-29 12:00:00
1000743 column=order:Quantity, timestamp=1735669811231, value=4
1000743 column=order:TotalAmount, timestamp=1735669811231, value=1456.52
1000743 column=product:Category, timestamp=1735669811231, value=Electronics
1000743 column=product:Price, timestamp=1735669811231, value=364.13
1000743 column=product:ProductID, timestamp=1735669811231, value=17
1000743 column=product:ProductName, timestamp=1735669811231, value=Desk Lamp
1000743 column=shipment:ShippingAddress, timestamp=1735669811231, value=9753 Roberts Hills Bryan
burgh OR 70954
1000743 column=shipment:ShippingDate, timestamp=1735669811231, value=2021-07-13 12:00:00
1000744 column=customer:Age, timestamp=1735669811231, value=79
1000744 column=customer:Country, timestamp=1735669811231, value=Australia
1000744 column=customer:CustomerID, timestamp=1735669811231, value=62641
1000744 column=customer:Name, timestamp=1735669811231, value=Ronald Greene
1000744 column=customer:RegistrationDate, timestamp=1735669811231, value=2023-12-07 12:00:00
1000744 column=order:OrderDate, timestamp=1735669811231, value=2022-06-22 12:00:00
1000744 column=order:Quantity, timestamp=1735669811231, value=2

```

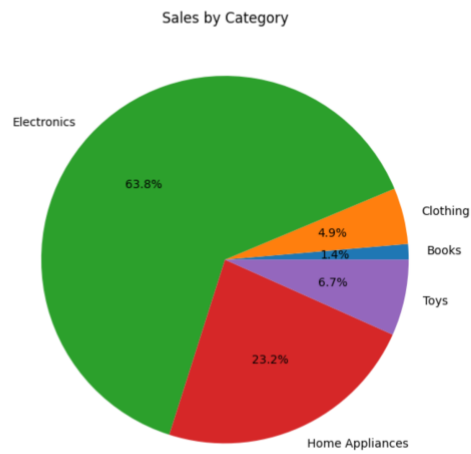
Setting up Flask Dashboard

```
C:\Windows\System32\cmd.exe X + v
C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_flask>docker compose up --build
time="2024-12-31T23:37:36+05:00" level=warning msg="C:\Users\Saad Ahmed Khan\Projects\BIG DATA\BigDataProject\BDA_flask\compose.yaml: the attribute 'version' is obsolete, it will be ignored, please remove it to avoid potential confusion"
[+] Building 6.4s (16/16) FINISHED
=> [web internal] load build definition from Dockerfile
=> => transferring dockerfile: 627B
=> [web] resolve image config for docker-image://docker.io/docker/dockerfile:1.4
=> [web auth] docker/dockerfile:pull token for registry-1.docker.io
=> CACHED [web] docker-image://docker.io/docker/dockerfile:1.4@sha256:9ba7531bd80fb0a858632727cf7a112fbfd19b17e9
=> [web internal] load .dockerignore
=> => transferring context: 2B
=> [web internal] load metadata for docker.io/library/python:3.10-alpine
=> [web auth] library/python:pull token for registry-1.docker.io
=> [web builder 1/6] FROM docker.io/library/python:3.10-alpine@sha256:748b5868188a58e05375eb70972cbdb338bae30c6e
=> [web internal] load build context
=> => transferring context: 39.95kB
=> CACHED [web builder 2/6] WORKDIR /app
=> CACHED [web builder 3/6] RUN apk add --no-cache gcc musl-dev python3-dev libffi-dev
=> CACHED [web builder 4/6] COPY requirements.txt /app
=> CACHED [web builder 5/6] RUN --mount=type=cache,target=/root/.cache/pip pip3 install -r requirements.txt
=> [web builder 6/6] COPY . /app
=> [web] exporting to image
=> => exporting layers
=> => writing image sha256:ae297f717650f9c6ed3775253481e618966a016d3a645f799ee1dcb3dccc0f5
=> => naming to docker.io/library/bda_flask-web
=> [web] resolving provenance for metadata file
time="2024-12-31T23:37:42+05:00" level=warning msg="a network with name bda_network exists but was not created for project \"bda_flask\".\nSet 'external: true' to use an existing network"
[+] Running 1/1
✔ Container bda_flask-web-1 Created
Attaching to web-1
web-1 | * Debug mode: on
web-1 | * WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
web-1 | * Running on all addresses (0.0.0.0)
web-1 | * Running on http://127.0.0.1:8000
web-1 | * Running on http://172.21.0.12:8000
web-1 | Press CTRL+C to quit
web-1 | * Restarting with stat
web-1 | * Debugger is active!
web-1 | * Debugger PIN: 121-162-808
View in Docker Desktop View Config Enable Watch
```

Dashboard:



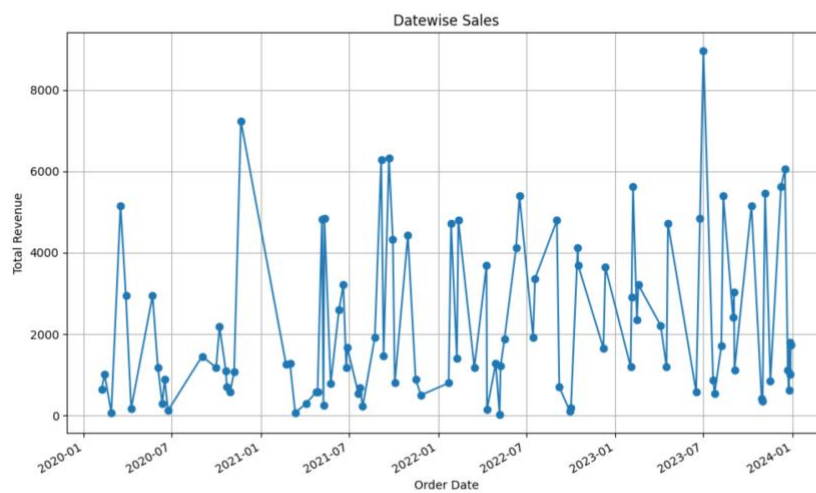
Sales by Category



Sales by Country

6235	Veronica Marshall	64	India	1
------	-------------------	----	-------	---

Datewise Sales



Sales by Category

localhost:8000/dashboard

Watch PAK vs Irelan... | Gmail | YouTube | Maps | Sakai : INTRODUC... | (no subject) - saada... | New Tab | Microsoft Office: W... | Download and insta... | Seller Dashboard | All Bookmarks

Welcome to the Data Analytics Dashboard

Top Customers

	CustomerName	Age	Country	TotalAmount
customer:CustomerID				
21160	Terry Morrison	44	UK	7227.18
78487	James Perry	65	UK	6330.15
72344	Mr. Matthew Gregory Jr.	67	Pakistan	6286.40
75929	James Sutton	36	France	6050.73
6655	Linda Ramirez	25	UK	5626.80

Top Products

	ProductName	Category	Price	TotalAmount
product:ProductID				
14	Printer	Electronics	737.01	28743.39
2	Headphones	Electronics	803.02	26499.66
16	Backpack	Electronics	599.01	24559.41
9	Smartphone	Home Appliances	606.62	23051.56
13	Washing Machine	Home Appliances	785.80	22002.40

Lowest Sales by Country

	TotalAmount
customer:Country	
India	14347.34
Canada	18443.16
USA	22731.98

12:08 am
31/12/2024

localhost:8000/product-search

Watch PAK vs Irelan... | Gmail | YouTube | Maps | Sakai : INTRODUC... | (no subject) - saada... | New Tab | Microsoft Office: W... | Download and insta... | Seller Dashboard | All Bookmarks

Product Search

Product Name:

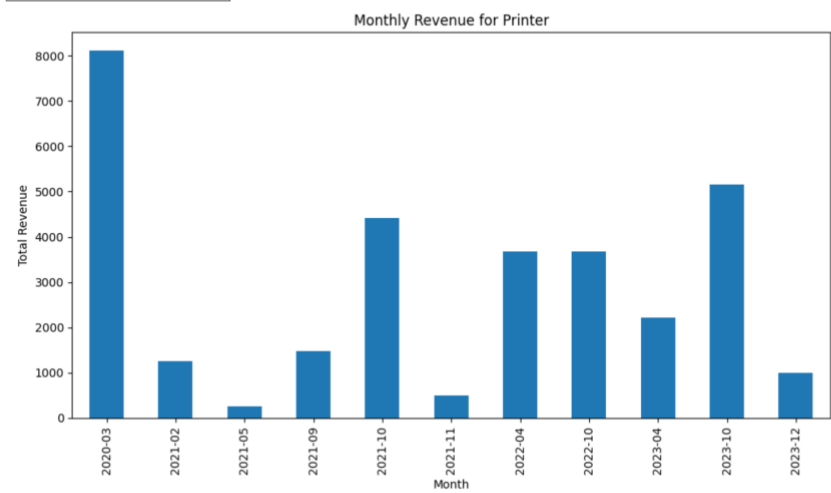
Monthly Revenue for Printer

	order:TotalAmount
order:OrderDate	
2020-03	8107.11
2021-02	1256.50
2021-05	251.30
2021-09	1474.02
2021-10	4422.06
2021-11	502.60
2022-04	3685.05
2022-10	3685.05
2023-04	2211.03
2023-10	5159.07
2023-12	1005.20

Monthly Revenue for Printer

12:09 am
31/12/2024

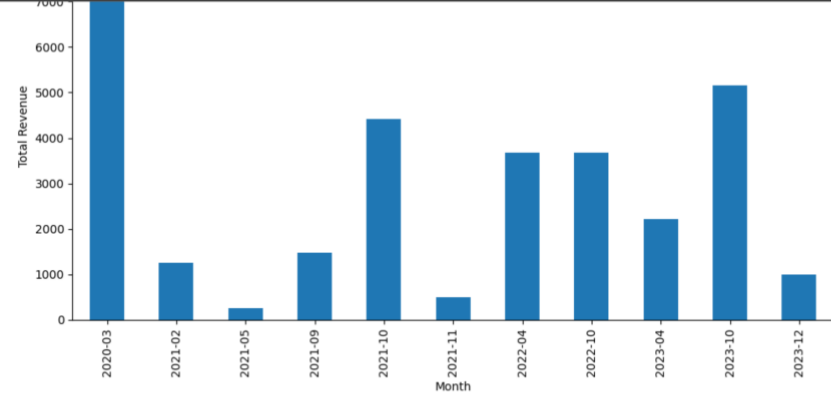
2023-10	5159.07
2023-12	1005.20



Top 5 Countries that Order Printer

order:TotalAmount

2023-10	5159.07
2023-12	1005.20



Top 5 Countries that Order Printer

	order:TotalAmount
customer:Country	
France	9581.13
Germany	5896.08
Australia	5159.07
Canada	4941.55
UK	2948.04