NETAJI SUHBAS UNIVERSITY OF TECHNOLOGY

AI Hardware and tools (CACSC09)
Lab File – VI Semester

Unit -4

Submitted by –

Paras Gunjyal (2021UCA1849)
Abhi Kansal (2021UCA1863)
Shivam Gupta (2021UCA1816)

Task 1: Explore RDD in spark

Resilient Distributed Dataset (RDD) is the fundamental data structure of Spark. They are immutable Distributed collections of objects of any type

1 Characteristics of RDDs:

- Resilient: RDDs are resilient to failure. They can be rebuilt from lineage information in case of data loss.
- Distributed: RDDs are distributed across multiple nodes in a cluster, enabling parallel processing.
- Immutable: Once created, RDDs are immutable. Their content cannot be changed.
- Lazy Evaluation: RDDs support lazy evaluation, meaning transformations on RDDs are not executed immediately. They are evaluated only when an action is performed.
- Partitioned: RDDs are divided into partitions, which are units of parallelism processed on individual nodes in the cluster.
- Typed and Untyped: RDDs can hold any type of Python, Java, or Scala objects, or they can be typed to hold specific types of objects.

2 Creation of RDDs:

- RDDs can be created in multiple ways:
- From existing data in memory
- From files in HDFS or other file systems
- By parallelizing an existing collection in the driver program
- By transforming existing RDDs through operations like map, filter, flatMap, etc.

3 Operations on RDDs:Transformations: Transformations create a new RDD from an existing one. Examples include map, filter, flatMap, groupByKey, reduceByKey, etc.

- Actions: Actions compute a result or write data to an external storage system. Examples include collect, count, reduce, saveAsTextFile, etc.

RDD Lineage:

- RDDs maintain a lineage graph, which is a directed acyclic graph (DAG) representing the sequence of transformations applied to the base dataset. This lineage information enables Spark to reconstruct lost partitions in case of failure.

Persistence:

- RDDs can be persisted in memory for faster access in subsequent operations. Persistence is useful when RDDs are reused across multiple computations.

Fault Tolerance:

- RDDs achieve fault tolerance through lineage information. If a partition of an RDD is lost, Spark can reconstruct it using the lineage graph.

Types of RDDs:

- Parallelized Collections: RDDs created by parallelizing an existing collection in the driver program.
- Hadoop Datasets: RDDs created by loading files from HDFS, supporting Hadoop InputFormats and OutputFormats.

- Transformed RDDs: RDDs created by applying transformations to existing RDDs.

Use Cases:

- RDDs are suitable for low-level programming and fine-grained control over data processing.
- They are useful when you need to perform custom, complex transformations on distributed datasets.

Limitations:

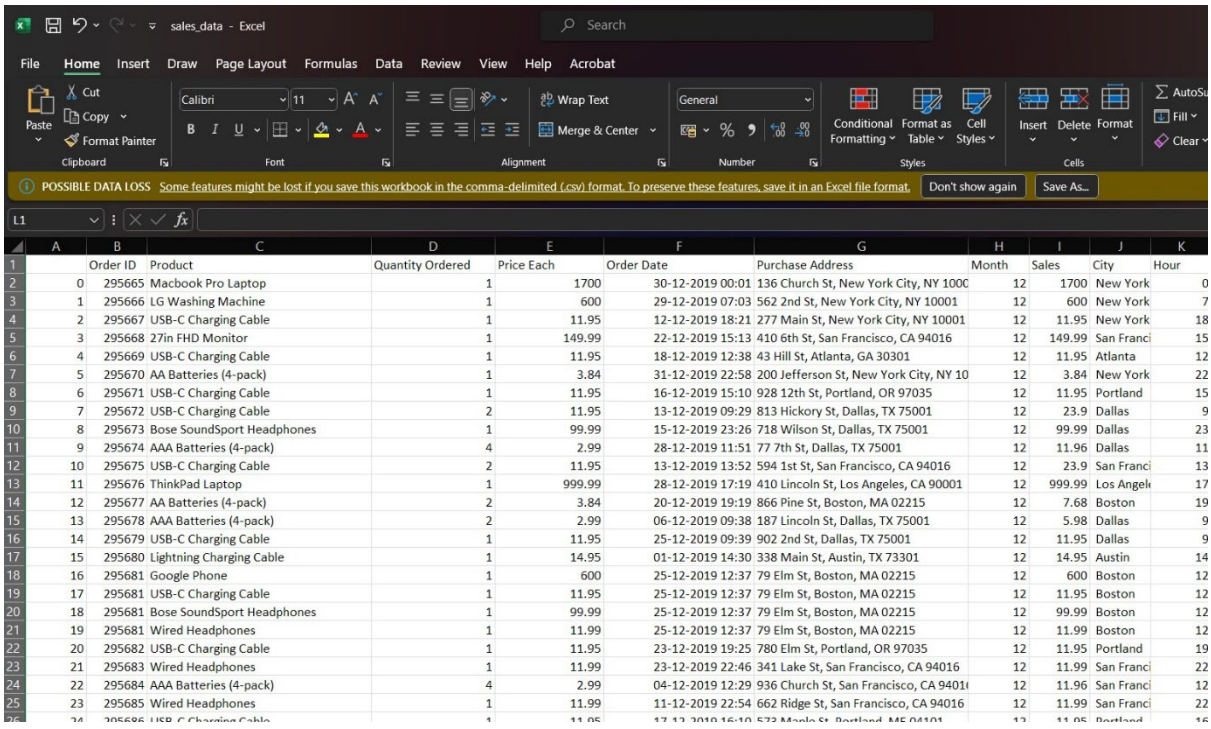
- RDDs lack optimizations present in higher-level abstractions like DataFrames and Datasets.
- They require manual optimization for performance.

Migration to DataFrames/Datasets:

- While RDDs provide flexibility, DataFrames and Datasets offer higher-level abstractions with optimizations for better performance. In many cases, migrating from RDDs to DataFrames or Datasets is recommended for improved productivity and performance.

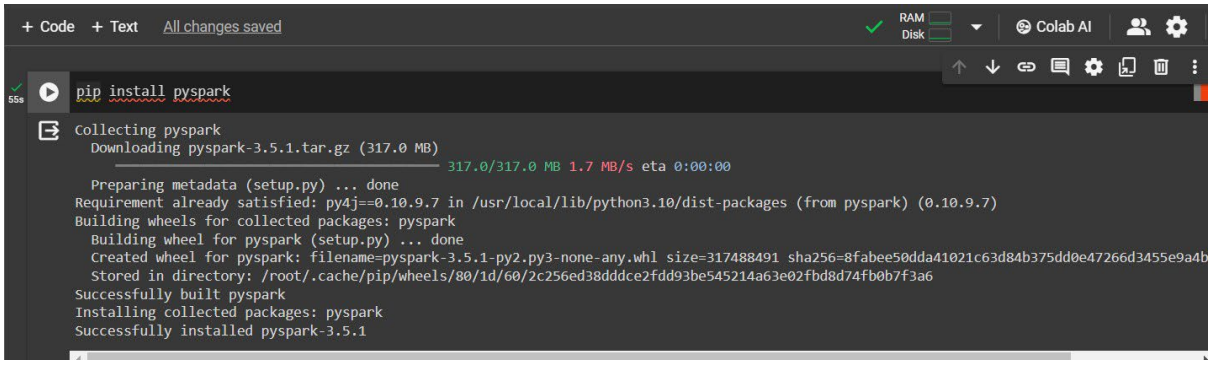
Task 2: In PySpark, create a program that reads a CSV file containing sales data, performs data cleaning by handling missing values and removing duplicates, calculates the total sales amount for each product, and finally, outputs the results to a new CSV file. Ensure to use transformations and actions in your PySpark script

Orignal files :

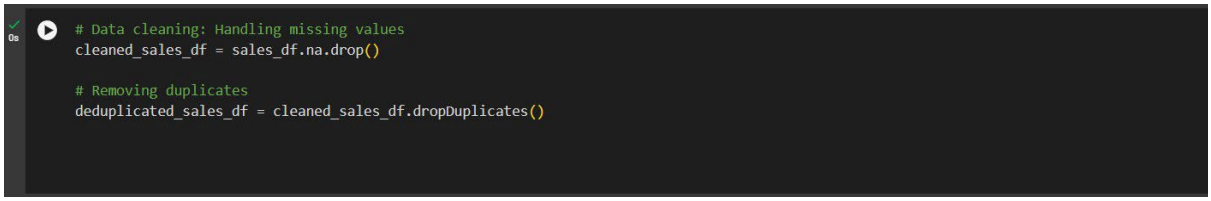


	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	Hour
0	295665	Macbook Pro Laptop	1	1700	30-12-2019 00:01	136 Church St, New York City, NY 1000	12	1700	New York	0
1	295666	LG Washing Machine	1	600	29-12-2019 07:03	562 2nd St, New York City, NY 10001	12	600	New York	7
2	295667	USB-C Charging Cable	1	11.95	12-12-2019 18:21	277 Main St, New York City, NY 10001	12	11.95	New York	18
3	295668	27in FHD Monitor	1	149.99	22-12-2019 15:13	410 6th St, San Francisco, CA 94016	12	149.99	San Franci	15
4	295669	USB-C Charging Cable	1	11.95	18-12-2019 12:38	43 Hill St, Atlanta, GA 30301	12	11.95	Atlanta	12
5	295670	AA Batteries (4-pack)	1	3.84	31-12-2019 22:58	200 Jefferson St, New York City, NY 10	12	3.84	New York	22
6	295671	USB-C Charging Cable	1	11.95	16-12-2019 15:10	928 12th St, Portland, OR 97035	12	11.95	Portland	15
7	295672	USB-C Charging Cable	2	11.95	13-12-2019 09:29	813 Hickory St, Dallas, TX 75001	12	23.9	Dallas	9
8	295673	Bose SoundSport Headphones	1	99.99	15-12-2019 23:26	718 Wilson St, Dallas, TX 75001	12	99.99	Dallas	23
9	295674	AAA Batteries (4-pack)	4	2.99	28-12-2019 11:51	77 7th St, Dallas, TX 75001	12	11.96	Dallas	11
10	295675	USB-C Charging Cable	2	11.95	13-12-2019 13:52	594 1st St, San Francisco, CA 94016	12	23.9	San Franci	13
11	295676	ThinkPad Laptop	1	999.99	28-12-2019 17:19	410 Lincoln St, Los Angeles, CA 90001	12	999.99	Los Angel	17
12	295677	AA Batteries (4-pack)	2	3.84	20-12-2019 19:19	866 Pine St, Boston, MA 02215	12	7.68	Boston	19
13	295678	AAA Batteries (4-pack)	2	2.99	06-12-2019 09:38	187 Lincoln St, Dallas, TX 75001	12	5.98	Dallas	9
14	295679	USB-C Charging Cable	1	11.95	25-12-2019 09:39	902 2nd St, Dallas, TX 75001	12	11.95	Dallas	9
15	295680	Lightning Charging Cable	1	14.95	01-12-2019 14:30	338 Main St, Austin, TX 73301	12	14.95	Austin	14
16	295681	Google Phone	1	600	25-12-2019 12:37	79 Elm St, Boston, MA 02215	12	600	Boston	12
17	295681	USB-C Charging Cable	1	11.95	25-12-2019 12:37	79 Elm St, Boston, MA 02215	12	11.95	Boston	12
18	295681	Bose SoundSport Headphones	1	99.99	25-12-2019 12:37	79 Elm St, Boston, MA 02215	12	99.99	Boston	12
19	295681	Wired Headphones	1	11.99	25-12-2019 12:37	79 Elm St, Boston, MA 02215	12	11.99	Boston	12
20	295682	USB-C Charging Cable	1	11.95	23-12-2019 19:25	780 Elm St, Portland, OR 97035	12	11.95	Portland	19
21	295683	Wired Headphones	1	11.99	23-12-2019 22:46	341 Lake St, San Francisco, CA 94016	12	11.99	San Franci	22
22	295684	AAA Batteries (4-pack)	4	2.99	04-12-2019 12:29	936 Church St, San Francisco, CA 9401	12	11.96	San Franci	12
23	295685	Wired Headphones	1	11.99	11-12-2019 22:54	662 Ridge St, San Francisco, CA 94016	12	11.99	San Franci	22

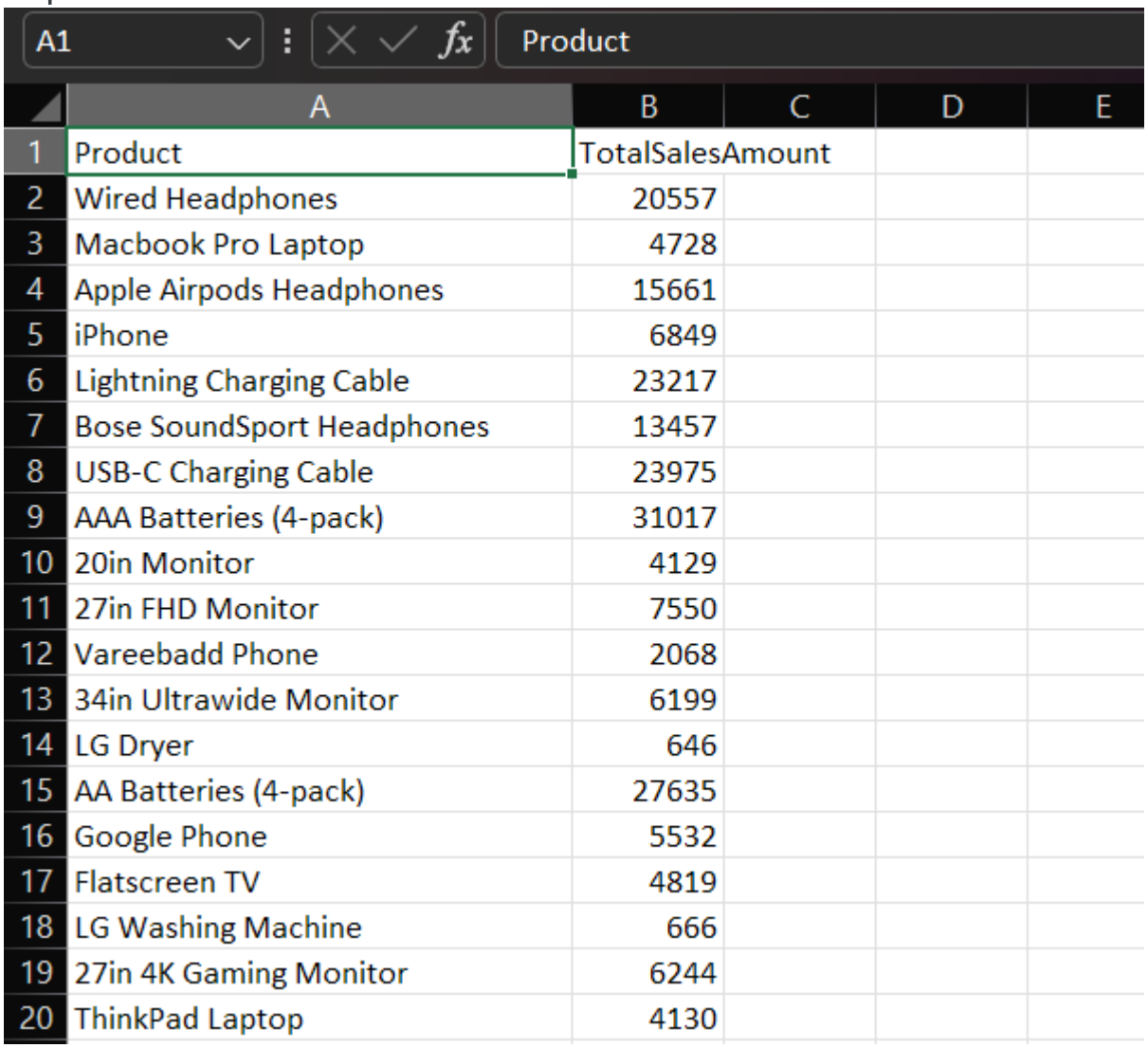
Installing pyspark library0



Initializing the library



Output total sales:



A1 | Product

	A	B	C	D	E
1	Product	TotalSalesAmount			
2	Wired Headphones	20557			
3	Macbook Pro Laptop	4728			
4	Apple Airpods Headphones	15661			
5	iPhone	6849			
6	Lightning Charging Cable	23217			
7	Bose SoundSport Headphones	13457			
8	USB-C Charging Cable	23975			
9	AAA Batteries (4-pack)	31017			
10	20in Monitor	4129			
11	27in FHD Monitor	7550			
12	Vareebadd Phone	2068			
13	34in Ultrawide Monitor	6199			
14	LG Dryer	646			
15	AA Batteries (4-pack)	27635			
16	Google Phone	5532			
17	Flatscreen TV	4819			
18	LG Washing Machine	666			
19	27in 4K Gaming Monitor	6244			
20	ThinkPad Laptop	4130			