

# BIG DATA – SAT 5165

1. Oscar Odera
2. Miltone Awiti
3. Eric Fosu-Kwabi

## Spark for Big Data Preprocessing

We focused on key preprocessing tasks to ensure the dataset containing 292,364 records was clean and ready for future analysis. The main tasks involved handling missing values, encoding categorical variables, and selecting relevant features for gender-based mental health treatment patterns.

The dataset was loaded into Spark using PySpark's DataFrame API, which allows for distributed data processing across multiple virtual machines. This initial exploration allowed us to inspect the structure of the dataset and ensure that the appropriate data types were inferred. One of the key preprocessing steps was to handle missing values in the dataset. We filled missing numerical values, such as age, using mean imputation and filled categorical values with "Unknown." This ensured that the dataset did not have gaps that could lead to issues during analysis.

Categorical variables were encoded using *StringIndexer* to assign a numerical index to each unique category. Following this, *OneHotEncoder* was applied to convert these indices into binary vectors. This transformation was necessary to convert string-based categorical data into a format suitable for machine learning models.

The final step in preprocessing was assembling the features into a single vector column. This assembled the encoded categorical variables and the numerical Age variable into a feature vector.

## Performance Comparison

The preprocessing performance was tested under different conditions to assess how efficiently the task could be completed in a distributed environment:

```
sat3812@hadoop1:/home/sat3812
[sat3812@hadoop1 ~]$ su
Password:
[root@hadoop1 sat3812]# /opt/spark/sbin/start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark-sat3812-org.apache.spark.deploy.maste
r.Master-1-hadoop1.out
[root@hadoop1 sat3812]# /opt/spark/sbin/start-master.sh
[root@hadoop1 sat3812]# /opt/spark/bin/spark-submit --master spark://hadoop1:7077 /opt/spark_script.py
```

## Using the Master Node Only

We started with master node to check on the performance

ooodera - hadoop1 - VMware Remote Console

VMRC

Activities Firefox Oct 23 17:50

Spark Master at spark://192.168.13.149:7077

URL: spark://192.168.13.149:7077

Alive Workers: 0

Cores in use: 0 Total, 0 Used

Memory in use: 0.0 B Total, 0.0 B Used

Resources in use:

Applications: 1 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (0)

Worker id	Address	State	Cores	Memory	Resources
-----------	---------	-------	-------	--------	-----------

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241023174821-0000	(vlt) StudentsPerformanceFactors	0	1024.0 MB		2024/10/23 17:48:21	root	WAITING	2.0 min

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

- **Execution Time:** 2 minutes
- **Observation:** Running the preprocessing tasks on a single master node resulted in a relatively longer execution time due to the workload being handled by one machine.

ooodera - hadoop1 - VMware Remote Console

VMRC ▾

Activities Firefox Oct 23 17:50

Spark Master at spark://192.168.13.149:7077

URL: spark://192.168.13.149:7077  
 Alive Workers: 0  
 Cores in use: 0 Total, 0 Used  
 Memory in use: 0.0 B Total, 0.0 B Used  
 Resources in use:  
 Applications: 1 Running, 0 Completed  
 Drivers: 0 Running, 0 Completed  
 Status: ALIVE

Workers (0)

Worker id	Address	State	Cores	Memory	Resources
-----------	---------	-------	-------	--------	-----------

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241023174821-0000	(vll) StudentPerformanceFactors	0	1024.0 MB		2024/10/23 17:48:21	root	WAITING	2.0 min

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

## Using the Worker Node

**Execution Time:** 8 seconds

**Observation:** The task was completed significantly faster when distributed across the worker node, showing a drastic reduction in execution time. This demonstrates the clear benefit of distributed computing in handling large datasets and heavy preprocessing tasks.

Spark Master at spark://192.168.13.149:7077

URL: spark://192.168.13.149:7077  
 Alive Workers: 0  
 Cores in use: 0 Total, 0 Used  
 Memory in use: 0.0 B Total, 0.0 B Used  
 Resources in use:  
 Applications: 2 Running, 0 Completed  
 Drivers: 0 Running, 0 Completed  
 Status: ALIVE

Workers (0)

Worker id	Address	State	Cores	Memory	Resources
-----------	---------	-------	-------	--------	-----------

Running Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241023175621-0001	(vll) StudentPerformanceFactors	0	1024.0 MB		2024/10/23 17:56:21	root	WAITING	8 s
app-20241023174821-0000	(vll) StudentPerformanceFactors	0	1024.0 MB		2024/10/23 17:48:21	root	WAITING	8.1 min

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

## Limitations:

- We encountered connection issues that prevented us from fully utilizing all six virtual machines. Errors occurred during the setup of some of the VMs, restricting us from running the job across multiple nodes simultaneously. Despite this, the speed improvement observed with two VMs shows the potential for even greater efficiency with more robust connections.

```
sat3812@hadoop1:/home/sat3812 — /opt/jdk/bin/java -cp /opt/spark/conf:/opt/spark/jars/* -Xmx1g -XX:+IgnoreUnrecognizedVMOptions --add-opens=java.base/java.lang=ALL-UNNAMED --add-opens=java.base/java.lang.invoke=ALL-UNNAMED
[Sat3812@hadoop1 ~]$ su
Password:
[root@hadoop1 sat3812]# start-all.sh
Starting namenodes on [hadoop1]
hadoop1: namenode is running as process 6972. Stop it first and ensure /tmp/hadoop-root-namenode.pid file is empty before retry.
Starting datanodes
hadoop6: root@hadoop6: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
192.168.13.108: root@192.168.13.108: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop4: root@hadoop4: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop5: root@hadoop5: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
192.168.13.107: root@192.168.13.107: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
192.168.13.117: root@192.168.13.117: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
192.168.13.116: root@192.168.13.116: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop3: root@hadoop3: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop2: mv: cannot stat '/opt/hadoop/logs/hadoop-root-datanode-hadoop2.out.4': No such file or directory
hadoop2: mv: cannot stat '/opt/hadoop/logs/hadoop-root-datanode-hadoop2.out.3': No such file or directory
hadoop2: mv: cannot stat '/opt/hadoop/logs/hadoop-root-datanode-hadoop2.out.2': No such file or directory
hadoop2: mv: cannot stat '/opt/hadoop/logs/hadoop-root-datanode-hadoop2.out.1': No such file or directory
Starting secondary namenodes [hadoop1]
hadoop1: secondarynamenode is running as process 7318. Stop it first and ensure /tmp/hadoop-root-secondarynamenode.pid file is empty before retry.
Starting resource manager
Starting nodemanagers
192.168.13.107: root@192.168.13.107: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
192.168.13.108: root@192.168.13.108: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop4: root@hadoop4: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop6: root@hadoop6: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
192.168.13.117: root@192.168.13.117: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
192.168.13.116: root@192.168.13.116: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop3: root@hadoop3: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
hadoop5: root@hadoop5: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
[root@hadoop1 sat3812]# /opt/spark/sbin/start-master.sh
org.apache.spark.deploy.master.Master running as process 3704. Stop it first.
[root@hadoop1 sat3812]# /opt/spark/sbin/stop-master.sh
stopping org.apache.spark.deploy.master.Master
[root@hadoop1 sat3812]# /opt/spark/sbin/start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark-sat3812-org.apache.spark.deploy.master.Master-1-hadoop1.out
[root@hadoop1 sat3812]# /opt/spark/sbin/start-slave.sh master spark://hadoop1:7077 /opt/spark_script.py
24/10/23 21:22:04 INFO SparkContext: Running Spark version 3.5.0
24/10/23 21:22:04 INFO SparkContext: OS Info Linux, 6.5.12-100.fc37.x86_64, amd64
24/10/23 21:22:04 INFO SparkContext: Java version 21.0.1
24/10/23 21:22:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/10/23 21:22:05 INFO ResourceUtils: =====
24/10/23 21:22:05 INFO ResourceUtils: No custom resources configured for spark.driver.
24/10/23 21:22:05 INFO ResourceUtils: =====
24/10/23 21:22:10 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 36221.
24/10/23 21:22:10 INFO NettyBlockTransferService: Server created on hadoop1:36221
24/10/23 21:22:10 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/10/23 21:22:10 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, hadoop1, 36221, None)
24/10/23 21:22:10 INFO BlockManagerMasterEndpoint: Registering block manager hadoop1:36221 with 413.9 MIB RAM, BlockManagerId(driver, hadoop1, 36221, None)
24/10/23 21:22:10 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, hadoop1, 36221, None)
24/10/23 21:22:10 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, hadoop1, 36221, None)
24/10/23 21:22:11 INFO StandaloneSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
24/10/23 21:22:12 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir.
24/10/23 21:22:12 INFO SharedState: Warehouse path is 'file:/home/sat3812/spark-warehouse'.
24/10/23 21:22:15 INFO InMemoryFileIndex: It took 166 ms to list leaf files for 1 paths.
24/10/23 21:22:16 INFO InMemoryFileIndex: It took 12 ms to list leaf files for 1 paths.
24/10/23 21:22:24 INFO FileSourceStrategy: Pushed Filters:
24/10/23 21:22:24 INFO FileSourceStrategy: Post-Scan Filters: (length(trim(value@, None))) > 0)
24/10/23 21:22:27 INFO CodeGenerator: Code generated in 560.181814 ms
24/10/23 21:22:27 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 190.5 KiB, free 413.7 MiB)
24/10/23 21:22:27 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 34.3 KiB, free 413.7 MiB)
24/10/23 21:22:27 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on hadoop1:36221 (size: 34.3 KiB, free: 413.9 MiB)
24/10/23 21:22:27 INFO SparkContext: Created broadcast_0 from csv at NativeMethodAccessorImpl.java:0
24/10/23 21:22:27 INFO FileSourceScanExec: Planning scan with bin packing, max size: 17646540 bytes, open cost is considered as scanning 4194304 bytes.
24/10/23 21:22:27 INFO SparkContext: Starting job: csv at NativeMethodAccessorImpl.java:0
24/10/23 21:22:28 INFO DAGScheduler: Got job 0 (csv at NativeMethodAccessorImpl.java:0) with 1 output partitions
24/10/23 21:22:28 INFO DAGScheduler: Final stage: ResultStage 0 (csv at NativeMethodAccessorImpl.java:0)
24/10/23 21:22:28 INFO DAGScheduler: Parents of final stage: List()
24/10/23 21:22:28 INFO DAGScheduler: Missing parents: List()
24/10/23 21:22:28 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[3] at csv at NativeMethodAccessorImpl.java:0), which has no missing parents
24/10/23 21:22:28 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 13.5 KiB, free 413.7 MiB)
24/10/23 21:22:28 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 6.4 KiB, free 413.7 MiB)
24/10/23 21:22:28 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on hadoop1:36221 (size: 6.4 KiB, free: 413.9 MiB)
24/10/23 21:22:28 INFO SparkContext: Created broadcast_1 from broadcast at DAGScheduler.scala:1580
24/10/23 21:22:29 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[3] at csv at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))
24/10/23 21:22:29 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks resource profile 0
24/10/23 21:22:44 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:22:59 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:23:14 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:23:29 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:23:44 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:23:59 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:24:14 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:24:29 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:24:44 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:24:59 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
24/10/23 21:25:14 WARN TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
```

Activate Windows  
Go to Settings to activate Windows.

Activate Windows  
Go to Settings to activate Window

The project codes can be accessed on;