

workspace - SalesReportGenerated/src/main/scala/bibek/training/ReportGenerator.scala - Scala IDE

File Edit Refactor Navigate Search Project Scala Run Window Help

Package Explorer

- SalesReportGenerated
 - Scala Library container [2.11.11]
 - JRE System Library [JavaSE-1.8]
- src
 - online.bibek.training
 - ReportGenerator.scala
- Maven Dependencies
- bin
- target
- pom.xml

SalesReportGenerated/pom.xml ReportGenerator.scala

```
package online.bibek.training

import org.apache.spark.SparkConf
import Undo Typing Ctrl+Z
import Revert File
import Save Ctrl+S
import Open Declaration F3
import Open Call Hierarchy Ctrl+Alt+H
import Show in Breadcrumb Alt+Shift+B
import Quick Outline Ctrl+O
import Open With >
import Show In Alt+Shift+W >
import Cut Ctrl+X
import Copy Ctrl+C
import Copy Qualified Name Ctrl+V
import Paste Ctrl+V
import Source Alt+Shift+S >
import Refactor >
import Local History >
import References >
import Declarations >
import Show Type Ctrl+Shift+Q
import Open Implicit Alt+F3
import Find occurrences Ctrl+Shift+G
import Open Type Hierarchy

Run As >
Debug As >
Validate >
Team >
Compare With >
Replace With >
Preferences...
```

Outline

- online.bibek.training
 - import declarations
 - ReportGenerator

Problems

0 items

Resource Path Location Type

1 Scala Application Alt+Shift+X, S

Run Configurations...

688M of 982M

Group E - Notepad

File Edit Format View Help

Data Technology Solutions

Group E

Bibek Shah Shankhar
Davinderjit Singh
Kirandeep Kaur
Kamaldeep kaur Dhami
Anirudh Siddula

Ln 1, Col 1 140% Windows (CRLF) UTF-8



workspace - SalesReportGenerated/src/main/scala/bibek/training/ReportGenerator.scala - Scala IDE

File Edit Refactor Navigate Search Project Scala Run Window Help

Quick Access

SalesReportGenerated

- Scala Library container [2.11.11]
- JRE System Library [JavaSE-1.8]
- src
 - online.bibek.training
 - ReportGenerator.scala
- Maven Dependencies
- bin
- target
- pom.xml

SalesReportGenerated/pom.xml ReportGenerator.scala

```
package online.bibek.training

import org.apache.spark.SparkConf
import org.apache.spark.SparkContext

object ReportGenerator {

    def main(args: Array[String]) {
        val conf = (new SparkConf)
            .setAppName("Sales Report Generator")
            .setMaster("local[*]")
        val sc = new SparkContext(conf)

        val rdd = sc.textFile("C:/Users/Lenovo/Documents/SalesJan2009.csv")

        val pairRdd = rdd.map(line => {
            val elements = line.split(",")
            (elements(2).trim, elements(1).trim.toInt)
        })

        val aggRdd = pairRdd.reduceByKey((x, y) => x + y)

        aggRdd.foreach(element => println(element._1 + " = " + element._2))
    }
}
```

Problems Tasks Console

```
<terminated> ReportGenerator$ [Scala Application] C:\Program Files\Java\jre1.8.0_311\bin\javaw.exe (Nov 25, 2021, 9:10:10 AM)
21/11/25 09:10:12 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
21/11/25 09:10:12 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 2 ms
21/11/25 09:10:12 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 2 ms
Mastercard = 458450
Amex = 188900
Diners = 133800
Visa = 849350
21/11/25 09:10:12 INFO Executor: Finished task 0.0 in stage 1.0 (TID 2). 1052 bytes result sent to driver
21/11/25 09:10:12 INFO Executor: Finished task 1.0 in stage 1.0 (TID 3). 1052 bytes result sent to driver
21/11/25 09:10:12 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 28 ms on localhost (e
```

Writable Smart Insert 10 : 44 698M of 982M

Group E - Notepad

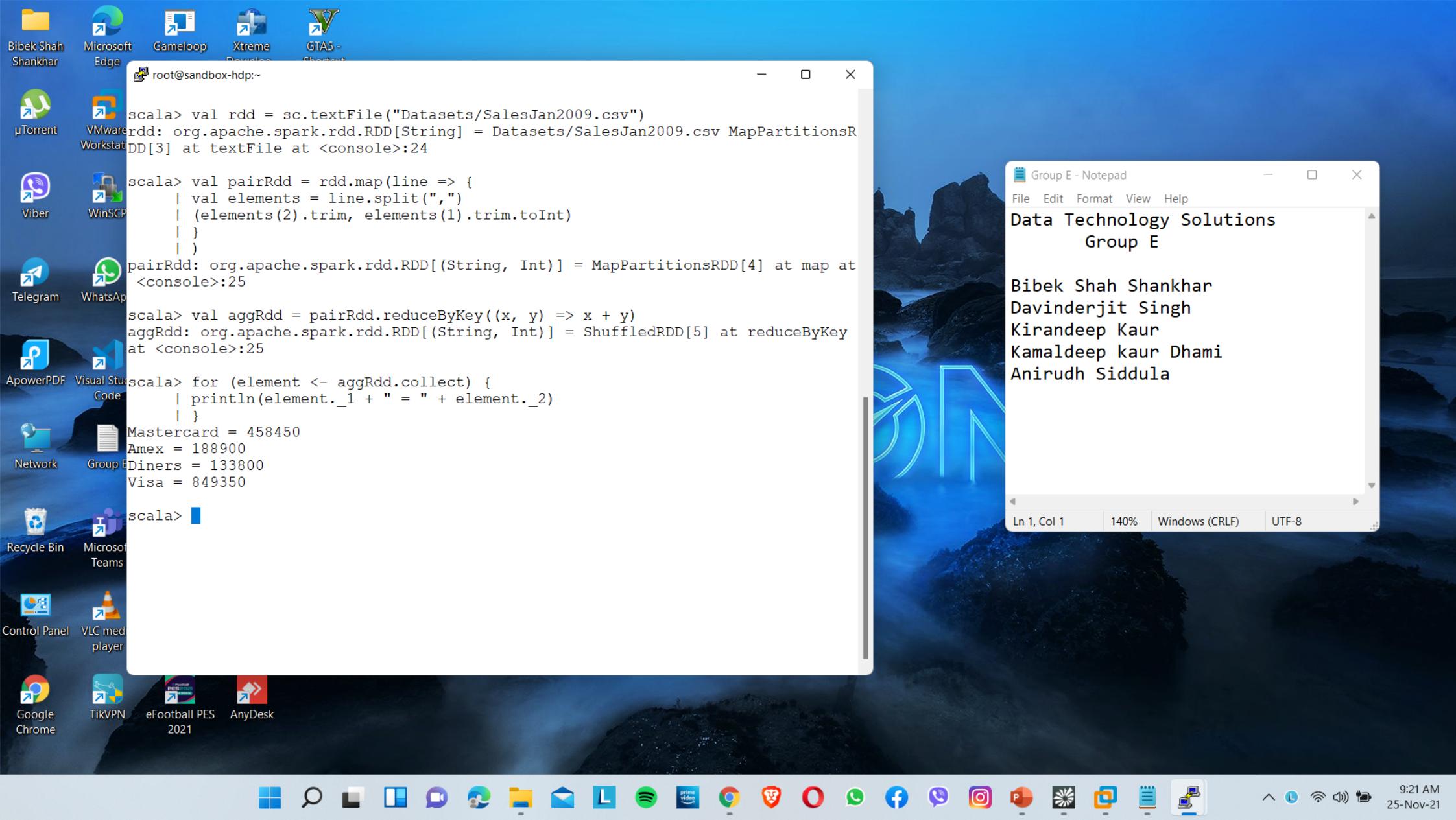
File Edit Format View Help

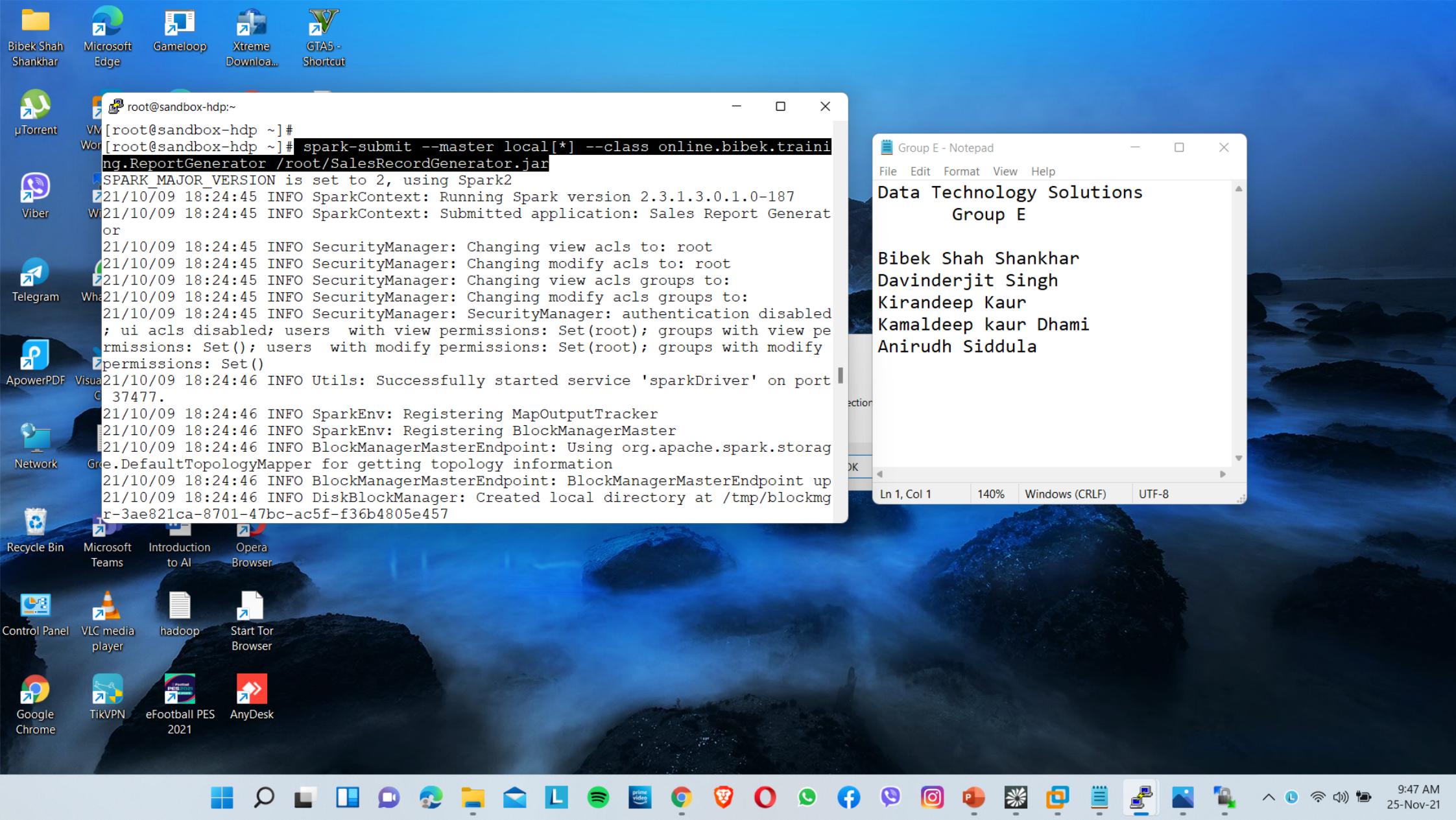
Data Technology Solutions

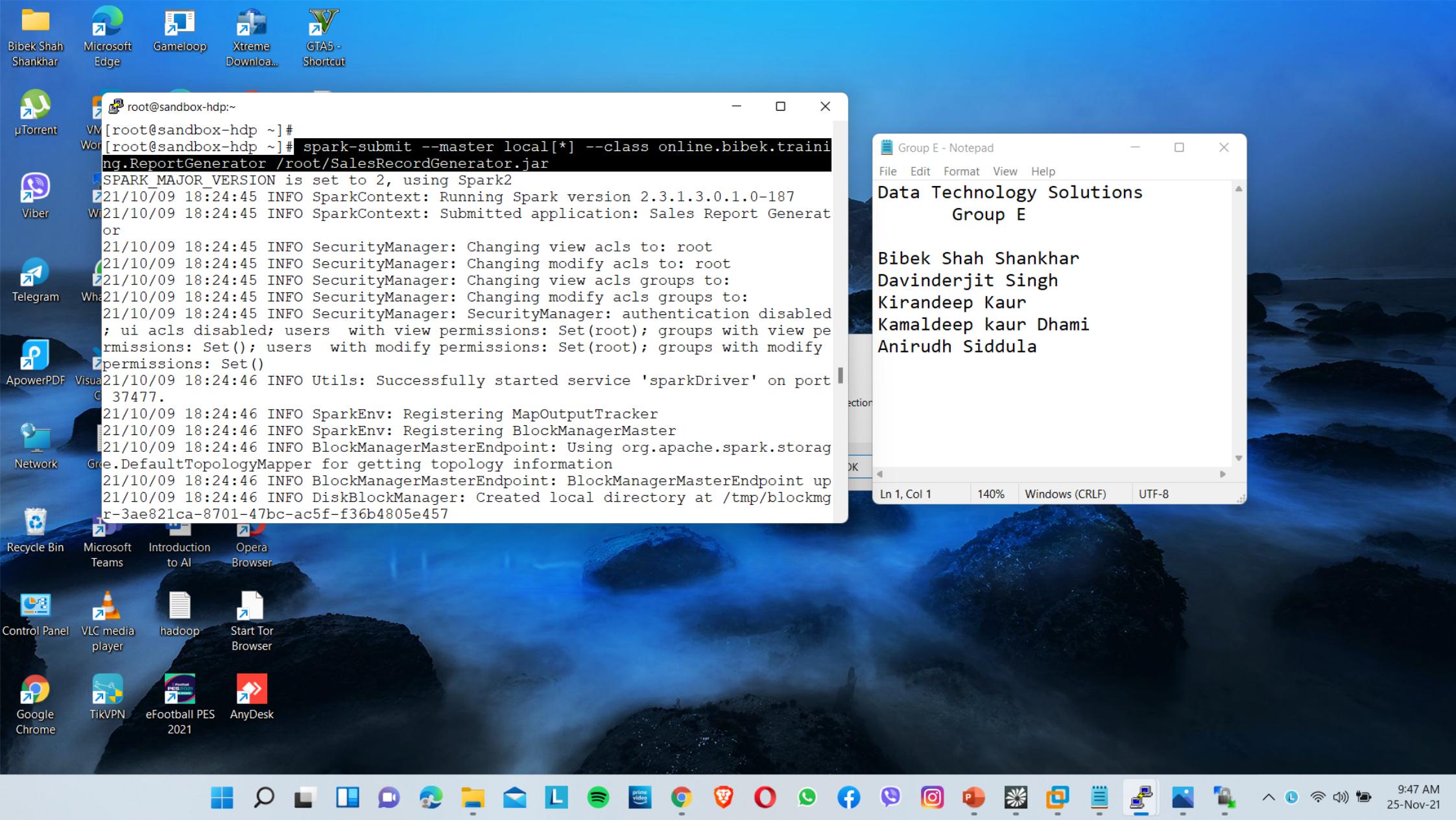
Group E

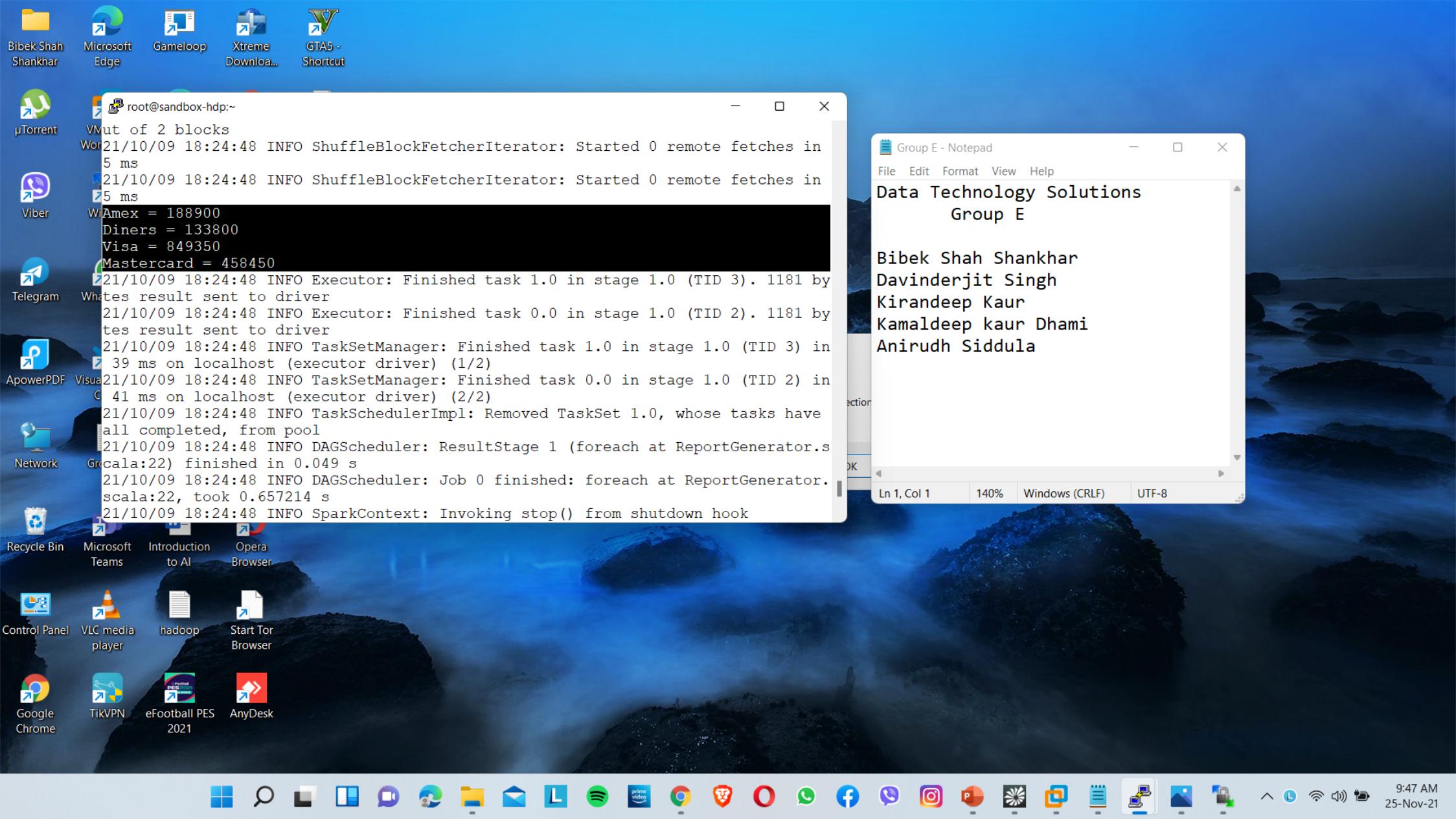
Bibek Shah Shankhar
Davinderjit Singh
Kirandeep Kaur
Kamaldeep kaur Dhami
Anirudh Siddula

Ln 1, Col 1 140% Windows (CRLF) UTF-8









[hadoop@ip-172-31-7-85:~

```
[hadoop@ip-172-31-7-85 ~]$ spark-submit --master yarn --deploy-mode cluster --num-executors 2 --executor-cores 2 --executor-memory 4G --driver-cores 1 --driver-memory 1G --class online.bibek.training.ReportGenerator SalesRecordGenerator.jar
21/11/25 17:22:50 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
21/11/25 17:22:51 INFO RMProxy: Connecting to ResourceManager at ip-172-31-7-85.us-east-2.compute.internal/172.31.7.85:8032
21/11/25 17:22:51 INFO Client: Requesting a new application from cluster with 2 NodeManagers
21/11/25 17:22:51 INFO Configuration: resource-types.xml not found
21/11/25 17:22:51 INFO ResourceUtils: Unable to find 'resource-types.xml'.
21/11/25 17:22:51 INFO ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/11/25 17:22:51 INFO ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/11/25 17:22:51 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
21/11/25 17:22:51 INFO Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
21/11/25 17:22:51 INFO Client: Setting up container launch context for our AM
21/11/25 17:22:51 INFO Client: Setting up the launch environment for our AM container
21/11/25 17:22:51 INFO Client: Preparing resources for our AM container
21/11/25 17:22:51 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
21/11/25 17:22:53 INFO Client: Uploading resource file:/mnt/tmp/spark-c4fceb25-53e4-40cf-a84e-441358abfbda/_spark_libs_8868275990520044970.zip -> hdfs://ip-172-31-7-85.us-east-2.compute.internal:8020/user/hadoop/.sparkStaging/application_1637858694574_0005/_spark_libs_8868275990520044970.zip
21/11/25 17:22:54 INFO Client: Uploading resource file:/home/hadoop/SalesRecordGenerator.jar -> hdfs://ip-172-31-7-85.us-east-2.compute.internal:8020/user/hadoop/.sparkStaging/application_1637858694574_0005/SalesRecordGenerator.jar
21/11/25 17:22:54 INFO Client: Uploading resource file:/mnt/tmp/spark-c4fceb25-53e4-40cf-a84e-441358abfbda/_spark_conf_6385150863197752417.zip -> hdfs://ip-172-31-7-85.us-east-2.compute.internal:8020/user/hadoop/.sparkStaging/application_1637858694574_0005/_spark_conf_.zip
21/11/25 17:22:55 INFO SecurityManager: Changing view acls to: hadoop
21/11/25 17:22:55 INFO SecurityManager: Changing modify acls to: hadoop
21/11/25 17:22:55 INFO SecurityManager: Changing view acls groups to:
21/11/25 17:22:55 INFO SecurityManager: Changing modify acls groups to:
21/11/25 17:22:55 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoop); groups with v
```

```
Group E - Notepad
File Edit Format View Help
Data Technology Solutions
Group E

Bibek Shah Shankhar
Davinderjit Singh
Kirandeep Kaur
Kamaldeep kaur Dhami
Anirudh Siddula

Ln 8, Col 16 | 140% | Windows (CRLF) | UTF-8
```



```
hadoop@ip-172-31-7-85:~
```

```
21/11/25 17:23:00 INFO Client: Application report for application_163785869457  
4_0005 (state: ACCEPTED)  
21/11/25 17:23:01 INFO Client: Application report for application_163785869457  
4_0005 (state: RUNNING)  
21/11/25 17:23:01 INFO Client:  
    client token: N/A  
    diagnostics: N/A  
    ApplicationMaster host: ip-172-31-10-76.us-east-2.compute.internal  
    ApplicationMaster RPC port: 37715  
    queue: default  
    start time: 1637860976023  
    final status: UNDEFINED  
    tracking URL: http://ip-172-31-7-85.us-east-2.compute.internal:20888/  
proxy/application_1637858694574_0005/  
user: hadoop  
21/11/25 17:23:02 INFO Client: Application report for application_163785869457  
4_0005 (state: RUNNING)  
21/11/25 17:23:03 INFO Client: Application report for application_163785869457  
4_0005 (state: RUNNING)  
21/11/25 17:23:04 INFO Client: Application report for application_163785869457  
4_0005 (state: RUNNING)  
21/11/25 17:23:05 INFO Client: Application report for application_163785869457  
4_0005 (state: RUNNING)  
21/11/25 17:23:06 INFO Client: Application report for application_163785869457  
4_0005 (state: RUNNING)  
21/11/25 17:23:07 INFO Client: Application report for application_163785869457  
4_0005 (state: FINISHED)  
21/11/25 17:23:07 INFO Client:  
    client token: N/A  
    diagnostics: N/A  
    ApplicationMaster host: ip-172-31-10-76.us-east-2.compute.internal  
    ApplicationMaster RPC port: 37715  
    queue: default  
    start time: 1637860976023  
    final status: SUCCEEDED  
    tracking URL: http://ip-172-31-7-85.us-east-2.compute.internal:20888/  
proxy/application_1637858694574_0005/  
user: hadoop  
21/11/25 17:23:07 INFO ShutdownHookManager: Shutdown hook called  
21/11/25 17:23:07 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-  
c4fce255-53e4-40cf-a84e-441358abfbda  
21/11/25 17:23:07 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-  
c1f99df2-b843-444e-bcc2-70754cddfb0  
[hadoop@ip-172-31-7-85 ~]$
```

Group E - Notepad

File Edit Format View Help

Data Technology Solutions Group E

Bibek Shah Shankhar
Davinderjit Singh
Kirandeep Kaur
Kamaldeep kaur Dhami
Anirudh Siddula

Ln 8, Col 16 | 140% | Windows (CRLF) | UTF-8



hadoop@ip-172-31-7-85:~

```
[hadoop@ip-172-31-7-85 ~]$ spark-submit --master yarn --deploy-mode cluster --num-executors 2 --executor-cores 2 --executor-memory 4G --driver-cores 1 --driver-memory 1G --class online.bibek.training.ReportGenerator SalesRecordGenerator.jar
21/11/25 17:28:29 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
21/11/25 17:28:30 INFO RMProxy: Connecting to ResourceManager at ip-172-31-7-85.us-east-2.compute.internal/172.31.7.85:8032
21/11/25 17:28:30 INFO Client: Requesting a new application from cluster with 2 NodeManagers
21/11/25 17:28:30 INFO Configuration: resource-types.xml not found
21/11/25 17:28:30 INFO ResourceUtils: Unable to find 'resource-types.xml'.
21/11/25 17:28:30 INFO ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/11/25 17:28:30 INFO ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/11/25 17:28:30 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
21/11/25 17:28:30 INFO Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
21/11/25 17:28:30 INFO Client: Setting up container launch context for our AM
21/11/25 17:28:30 INFO Client: Setting up the launch environment for our AM container
21/11/25 17:28:30 INFO Client: Preparing resources for our AM container
21/11/25 17:28:30 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
21/11/25 17:28:32 INFO Client: Uploading resource file:/mnt/tmp/spark-05cd9708-b655-44ee-86d5-fcf64062cc72/_spark_libs_3167492234282159933.zip -> hdfs://ip-172-31-7-85.us-east-2.compute.internal:8020/user/hadoop/.sparkStaging/application_1637858694574_0006/_spark_libs_3167492234282159933.zip
21/11/25 17:28:33 INFO Client: Uploading resource file:/home/hadoop/SalesRecordGenerator.jar -> hdfs://ip-172-31-7-85.us-east-2.compute.internal:8020/user/hadoop/.sparkStaging/application_1637858694574_0006/SalesRecordGenerator.jar
21/11/25 17:28:33 INFO Client: Uploading resource file:/mnt/tmp/spark-05cd9708-b655-44ee-86d5-fcf64062cc72/_spark_conf_3174268660788438931.zip -> hdfs://ip-172-31-7-85.us-east-2.compute.internal:8020/user/hadoop/.sparkStaging/application_1637858694574_0006/_spark_conf_.zip
21/11/25 17:28:33 INFO SecurityManager: Changing view acls to: hadoop
21/11/25 17:28:33 INFO SecurityManager: Changing modify acls to: hadoop
21/11/25 17:28:33 INFO SecurityManager: Changing view acls groups to:
21/11/25 17:28:33 INFO SecurityManager: Changing modify acls groups to:
21/11/25 17:28:33 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoop); groups with view permissions: Set(); users with modify permissions: Set(hadoop); groups with
```

Group E - Notepad
File Edit Format View Help
Data Technology Solutions
Group E

Bibek Shah Shankhar
Davinderjit Singh
Kirandeep Kaur
Kamaldeep kaur Dhami
Anirudh Siddula

Ln 8, Col 16 | 140% | Windows (CRLF) | UTF-8



```
hadoop@ip-172-31-7-85:~
```

```
21/11/25 17:28:39 INFO Client: Application report for application_163785869457  
4_0006 (state: ACCEPTED)  
21/11/25 17:28:40 INFO Client: Application report for application_163785869457  
4_0006 (state: RUNNING)  
21/11/25 17:28:40 INFO Client:  
    client token: N/A  
    diagnostics: N/A  
    ApplicationMaster host: ip-172-31-10-76.us-east-2.compute.internal  
    ApplicationMaster RPC port: 34583  
    queue: default  
    start time: 1637861314986  
    final status: UNDEFINED  
    tracking URL: http://ip-172-31-7-85.us-east-2.compute.internal:20888/  
proxy/application_1637858694574_0006/  
    user: hadoop  
21/11/25 17:28:41 INFO Client: Application report for application_163785869457  
4_0006 (state: RUNNING)  
21/11/25 17:28:42 INFO Client: Application report for application_163785869457  
4_0006 (state: RUNNING)  
21/11/25 17:28:43 INFO Client: Application report for application_163785869457  
4_0006 (state: RUNNING)  
21/11/25 17:28:44 INFO Client: Application report for application_163785869457  
4_0006 (state: RUNNING)  
21/11/25 17:28:45 INFO Client: Application report for application_163785869457  
4_0006 (state: RUNNING)  
21/11/25 17:28:46 INFO Client: Application report for application_163785869457  
4_0006 (state: FINISHED)  
21/11/25 17:28:46 INFO Client:  
    client token: N/A  
    diagnostics: N/A  
    ApplicationMaster host: ip-172-31-10-76.us-east-2.compute.internal  
    ApplicationMaster RPC port: 34583  
    queue: default  
    start time: 1637861314986  
    final status: SUCCEEDED  
    tracking URL: http://ip-172-31-7-85.us-east-2.compute.internal:20888/  
proxy/application_1637858694574_0006/  
    user: hadoop  
21/11/25 17:28:46 INFO ShutdownHookManager: Shutdown hook called  
21/11/25 17:28:46 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-  
b467571a-905f-4963-b613-9746c37a1bb9  
21/11/25 17:28:46 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-  
05cd9708-b655-44ee-86d5-fcf64062cc72  
[hadoop@ip-172-31-7-85 ~]$
```

Group E - Notepad

File Edit Format View Help

Data Technology Solutions Group E

Bibek Shah Shankhar
Davinderjit Singh
Kirandeep Kaur
Kamaldeep kaur Dhami
Anirudh Siddula

Ln 8, Col 16 | 140% | Windows (CRLF) | UTF-8





History Server

Event log directory: s3a://prod.us-east-2.appinfo/src/j-3444UU1V5QU8R/sparklogs

Last updated: 2021-11-25 12:42:40

Client local time zone: America/New_York

Search:

App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
application_1637858694574_0006	Sales Report Generator	2021-11-25 12:28:38	2021-11-25 12:28:45	7 s	hadoop	2021-11-25 12:42:16	<button>Download</button>
application_1637858694574_0005	Sales Report Generator	2021-11-25 12:22:59	2021-11-25 12:23:06	7 s	hadoop	2021-11-25 12:42:16	<button>Download</button>
application_1637858694574_0004	Sales Report Generator	2021-11-25 12:15:35	2021-11-25 12:15:44	9 s	hadoop	2021-11-25 12:42:16	<button>Download</button>

Showing 1 to 3 of 3 entries

Show incomplete applications