T15 - Hands-on

→ To open pdf file go to applications → file Manager → desktop → project → wings T15 → instructions → drag that wings T15 . file .

→ Already exist data in S3 → loan-data check whether existing given data(pdf) present or not .

* Select the region as : US-East -1 → rightside top corner .

* First Create Redshift cluster go to Console → type as Redshift → Click on Amazon Redshift → next go down → Click on Create Cluster .

(2)

give name as emr-spark-redshift. (copy & paste this name)
Choose as all-choose
Next select dc2 - large.
Number of node : 1
→ Database configurations
Manually add the admin password
password is 'Awsuser1'. click on (copy & paste)
show password.
→ a Newly created is database
Encryption in that select → Aws (KMS)
Manage Key. → go down → click on
create cluster.
→ Backend it will load.
Next we have to create EMR cluster
→ Click on create cluster
Name: My Cluster.
Select Amazon EMR release is emr-7.1.0 (select)

③ → Select → * Hadoop 3.3.6 * Sqoop 1.4.7
* Spark 3.5.0 * Hive 3.1.3
→ go to primary → M4.large (select)
→ Remove instance group (2 times)
→ go to cluster termination → Manually terminate the cluster
→ go to security configuration & EC2 key pair. → click on create key pair
→ give key pair name as emr-spark
Click on create key pair.
→ Now go Back to previous tag
Click on Browse → e → selected click on
⇒ Choose → Next select service role as
Aws-s3-rds-redshift-emr → Now go down select create an instance → select all S3, click on create cluster
go to coding part.
→ Application → development →
open visual studio code

④

→ go to file → open folder → Alt D + a → desktop → project folder → select wings +15 then open:

→ go to left side click on python

→ Next click on challenge.py

→ How to take Bucket

go to console → in search bar search for S3 inside → load data click that next copy that Bucket name.

→ go to coding part → control 'F' type Bucket_Name remove that space in double Quotes. (28 & 68 lines) (2 places)

→ get the jdbc URL from Redshift.

go to Console → Amazon Redshift → click on Clusters emr-spark-redshift after clicking Right side i will get jdbc URL

After writing the coding part click on control 'S'. save the code.

(5) coding part.

1) read the data.

df = spark.read.csv(s3_input_path, header=True, schema=customSchema).

2) clean data lo.

df = input_df.dropna().dropDuplicates()
df = df.filter(df.purpose != 'null')

3) load data lo.

data.coalesce(1).write.csv(output_path, header=True, mode="overwrite")

4) def result_1(input_df):

df = input_df.filter((col("purpose") == "Educational") | (col("purpose") == "small_business"))

```
df = df.withColumn("income_to_installment_ratio", col("log_annual_inc") / col("installment"))
df = df.withColumn("int_rate_category"
    .when(col("int_rate") < 0.1, "low")
    .when((col("int_rate") >= 0.1) & (col("int_rate") < 0.15), "medium")
    .otherwise("high"))
df = df.withColumn("high_risk_borrower",
    when((col("dti") > 20) | (col("fico") < 700) | (col("revol_util") > 80, 1)
    .otherwise(0))
```

result_2

```
df = input_df.groupBy("purpose").agg(
    (sum(col("not_fully_paid")) / count("*")).alias("default_rate"))
df = df.withColumn("default_rate",
    round(col("default_rate"), 2))
```

(7)

redshift - load data.

→ jdbcURL = "(copy jdbc URL)"
username = "awsuser"
password = "Awsuser1"
table_name = "result_2"

→ data.write \
.format("jdbc") \
.option("url", jdbcUrl) \
.option("dbtable", table_name) \
.option("user", username) \
.option("password", password) \
.mode("overwrite") \
.save()

After writing the code we will push the code to EMR cluster.

first, we go to EC2 → instances → select that instance ID → go to Actions → security → change the Security groups → search for my security group → Add the security group → next we have to save. → next in instances select connect → leave as it is then next click on connect. then in new tab EMR Cluster will open → go to Visual code → on the top click on Terminal → open New terminal.

→ go to application → file manager → downloads drag that file (emr-spark-pem) drag to (emr_copy-py).

→ Next it will open Beside challenge.py it will open in emr-spark-perm next copy that `python emr_copy.py` paste in emr-spark-perm then click on run. It will run Successfully

→ Next go to EMR cluster.

→ give command as `cd /home/hadoop` Enter → `ls` Enter → `cd python` Enter → `ls` enter → (Now path is in challenge.py)

→ `cd ..` (to go back)

→ Enter `ls` → goto `cd setup` → Next Enter `bash setup.sh` → It will load for sometime

Now to go back enter → `cd ..` Enter `ls` → Next go to python folder enter `cd python` → enter ls → Now we have submit the code EMR type as `spark-submit challenge.py` it will execute.

⇒ How to check loading data into Redshift. (10)

→ go to Amazon Redshift → click on that

→ click on Emr-spark-redshift → RS side click on Query data → Query data editor V2

→ check in S3.

Click on Emr-Spark-redshift.

→ then select database, username & pswd.

database → dev, Username → awsuser →

password → Awsuser1 → ☑ Showpassword.

click on create connection after creating connection

→ once go to S3 → click on load-data ~~check the~~ click on that → Check the o/p

Now goto left Emr-spark-redshift → dev → public → tables → result_2

check o/p data result_1 clean data

Now we can Query as

Select * from result_2. click on run.

O/p as:

purpose	default rate
educational	0.29
small_business	0.35
all_other	0.2
debt_consolidation	0.17
credit_card	0.14
Main Purchase	0.12