Purpose Of Glue Job : To read data from Epsilon Oracle Database using JDBC connectivity with these things in focus

**Basic Check:**

Successful to read data from db

Read incremental range of data with currentdate-4 range

Count check

Preview 10 rows

**Tune Job with Glue JDBC API:**

Hash field Hash partition Hash expression

Customize fetch size

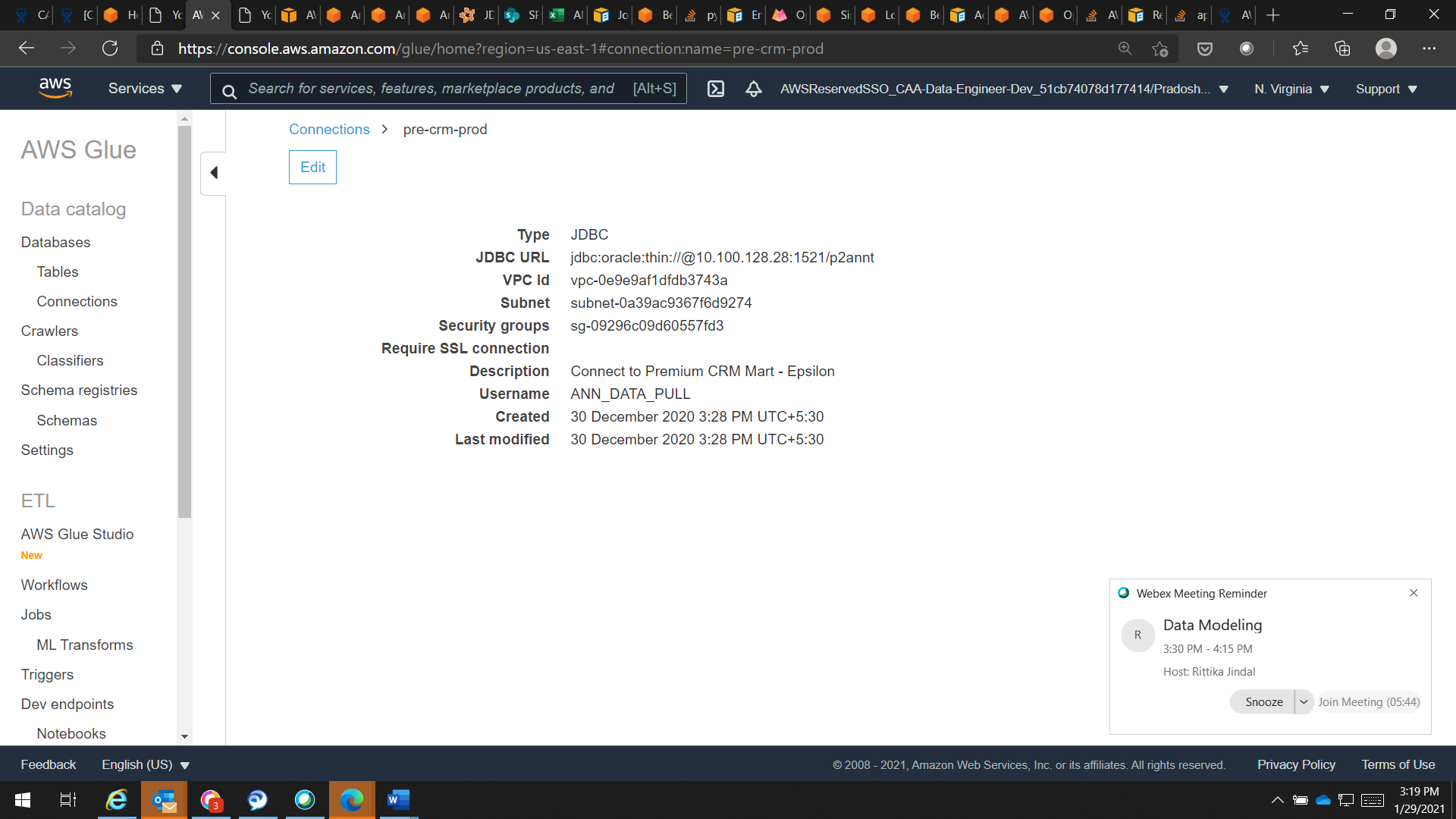
Tune job with Spark API :

Partition column , lower bount , upper bound , num partition , fetch size

JobName: pre-demo-job-1

Glue Connection used : pre-crm-prod

* We will pass the connection details from a connection.py file
* We will use the glue connection what we created in our job
* We are trying out Glue and Spark API to figure out which works better for our need .



**Initial Work**

Passing connection details through job parameter

dbname=args['dbname']

username=args['username']

password=args['password']

hostname=args['hostname']

port=args['port']

jdbcUrl = "jdbc:oracle:thin:@//"+hostname+":"+port+"/"+dbname

Method -1

connection\_oracle11\_options = {

"url": jdbcUrl,

"dbtable": "DM\_OWNER.LOFT\_F\_TXN\_ITEM\_VW",

"user": username,

"password": password}

transaction\_item\_df = glueContext.create\_dynamic\_frame.from\_options(connection\_type="oracle",

connection\_options=connection\_oracle11\_options)

Method -2

productdf = spark.read \

.format("jdbc") \

.option("url", jdbcUrl) \

.option("dbtable", "DM\_OWNER.lu\_product\_vw") \

.option("user", username) \

.option("password", password) \

.option("driver", "oracle.jdbc.driver.OracleDriver") \

.load()

**First issue reported with the AWS support team**

JobName:pre-demo-job-1

Error : JobRunId:jr\_9b2a8efd47f52cbc9b24b2ac79e8413a8a5ba65f14d1620d6fb79417eeba9df0 failed to execute with exception Number of IP addresses on subnet is 0: subnet-0a39ac9367f6d9274

JobRunId: jr\_00c063bf7359af08ec63e7cc4e0d6f99c9ba2489e6b7aebb46ad0c6498228818

**Initial Diagnosis  and findings:**

We are designing this  glue job to connect to  epsilon oracle database with jdbc connectivity to pull data  , but this is job is failing .  I did some initial diagnosis  and then  contacted AWS support regarding this to have further discussion  . It looks like , for this  subnet ‘subnet-0a39ac9367f6d9274’ , I could see that the IPv4 CIDR Block for this subnet is 172.24.128.0/25 , with 0 of those IP Addresses currently free. It seems , Glue job is not getting addresses to sets up ENIs that enable  the job to connect securely to other resources .  It looks like , we need a bigger IPv4 CIDR Block to get more available IP Addresses for running of glue job smoothly  or need  some solution to fix this on a long run.

**First reply from AWS support team**

You were getting below error while running the glue job

+++++++++++++++

jr\_00c063bf7359af08ec63e7cc4e0d6f99c9ba2489e6b7aebb46ad0c6498228818 failed to execute with exception Number of IP addresses on subnet is 0: subnet-0a39ac9367f6d9274

+++++++++++++++

We discussed that If a job needs to run in your VPC subnet, AWS Glue sets up elastic network interfaces that enable your jobs to connect securely to other resources. Each elastic network interface is assigned a private IP address from the IP address range within the subnet you specified[1]. This issue occurs when the subnet in which you are running the job does not have enough IP's to launch the job.

Upon checking subnet-0a39ac9367f6d9274, I could see that the IPv4 CIDR Block for this subnet is 172.24.128.0/25. This gives it a total of 128 Available IP Addresses, with 0 of those IP Addresses currently free. Since you are using 10 DPUs in glue job , you will need at least 10 available IP address to launch the job.

=============

Subnet : subnet-0a39ac9367f6d9274

IPv4 CIDR Block 172.24.128.0/25

Number of allocated ips : 128

Free address count : 0 ----------> (There are no available IPs to launch glue job)

=============

Also, I checked your previous job runs and find out that they were also failing with same issue. To solve this issue, Create a new subnet and give it a bigger IPv4 CIDR Block to get more Available IP Addresses. This will allow you to launch an glue job with out any issues.

**Second reply from AWS support team**

Here is a summary of our discussion -

1. Number of ENIs

Number of ENIs created for an ETL job depends on the number of workers configured.

If the job is configured with 10 workers and using a glue connection, glue creates 10 ENIs in the subnet configured in glue connection.

2. Slow running job -

Job took most (~90%) of the execution time to read data from JDBC source.

----------------------------------

g-2660a3a6dcfc09eb668eb4ed4150ef71671afc58:2021-01-27 15:26:14,166 INFO [dag-scheduler-event-loop] scheduler.DAGScheduler (Logging.scala:logInfo(54)): ShuffleMapStage 0 (count at NativeMethodAccessorImpl.java:0) finished in 1366.829 s

----------------------------------

This can be improved by increasing numPartitions in the spark jdbc read. Also by default fetchsize would be low, you can try with different value and check the performance of your job.[1]

3. We discussed about how number of partitions map to number of tasks and number of vcores per executor.

4. We then discussed about using Glue dynamic frame for JDBC read as you do not have to specify lower and upper bound parameters in that case.[2]

Equivalent of your current configuration would be to use PRODUCT\_ID as hashfield and set appropriate number of hashpartitions based on number of workers configured.

**Stats**

Job Name : pre-demo-job-1

|  |  |
| --- | --- |
| Job Run Id | jr\_6eeafacea527071632509aa518f051af9699d03d2e44f4d2cbaa5f5d33ef2c62 |
| Worker Type | G1.X |
| Number of workers | 40 |
|  |  |
|  |  |
|  |  |
|  |  |
| FETCHSIZE | 1000 |

Ran for long time - More than two hours

Job Name : pre-demo-job-1

|  |  |
| --- | --- |
| Job Run Id | jr\_2ac8d6575a77031165dc81a259b7c59177acca52e03c80f14a1b965a67146001 |
| Worker Type | G1.X |
| Number of workers | 40 |
| partitionColumn | TXN\_ITEM\_ID |
| lowerBound | 1255534593 |
| upperBound | 1829324174 |
| numPartitions | 40 |
| FETCHSIZE | 1000 |

Duration : 7 Mins

Job Name : pre-demo-job-1

|  |  |
| --- | --- |
| Job Run Id | jr\_75281a3fcb666af36da28bef393df7bf9b950035281098c11e927b0fef7a8416 |
| Worker Type | G1.X |
| Number of workers | 40 |
| partitionColumn | TXN\_ITEM\_ID |
| lowerBound | 1255534593 |
| upperBound | 1829324174 |
| numPartitions | 80 |
| FETCHSIZE | 1000 |

Duration : 4 Mins

Customized query

---------------------

transaction\_item\_df = spark.read \

.format("jdbc") \

.option("url", jdbcUrl) \

.option("dbtable", "(SELECT \* FROM DM\_OWNER.LOFT\_F\_TXN\_ITEM\_VW) subq") \

.option("user", username) \

.option("password", password) \

.option("driver", "oracle.jdbc.driver.OracleDriver") \

.option("partitionColumn", "TXN\_ITEM\_ID") \

.option("lowerBound", 1255534593) \

.option("upperBound", 1829324174) \

.option("numPartitions", 40) \

.option("fetchsize",1000) \

.load()

Job name : [pre\_spark\_jdbc\_tune\_pk](https://console.aws.amazon.com/glue/home?region=us-east-1#etl:tab=jobs)

|  |  |
| --- | --- |
| Job Run Id | [jr\_144baf04b3bd5b0790b5fc9dcc9760e6e2ef2227e8f1b32ffc3a70270a6ed6e7](https://console.aws.amazon.com/glue/home?region=us-east-1#jobRun:jobName=pre_spark_jdbc_tune_pk;jobRunId=jr_144baf04b3bd5b0790b5fc9dcc9760e6e2ef2227e8f1b32ffc3a70270a6ed6e7) |
| Worker Type | G1.X |
| Number of workers | 40 |
| partitionColumn | TXN\_ITEM\_ID |
| lowerBound | 1255534593 |
| upperBound | 1829324174 |
| numPartitions | 40 |
| FETCHSIZE | 1000 |

Duration : 7 mins

tableName="DM\_OWNER.LOFT\_F\_TXN\_ITEM\_VW"

pushdown\_query="(SELECT TXN\_ITEM\_ID,TXN\_ID,TXN\_NBR,MODIFIED\_DT FROM "+ tableName +" where MODIFIED\_DT>=CURRENT\_DATE-4) subq"

transaction\_item\_df = spark.read \

.format("jdbc") \

.option("url", jdbcUrl) \

.option("dbtable", pushdown\_query) \

.option("user", username) \

.option("password", password) \

.option("driver", "oracle.jdbc.driver.OracleDriver") \

.option("partitionColumn", "TXN\_ITEM\_ID") \

.option("lowerBound", 1255534593) \

.option("upperBound", 1829324174) \

.option("numPartitions", 40) \

.option("fetchsize",1000) \

.load()

Job name : [pre\_spark\_jdbc\_tune\_pk](https://console.aws.amazon.com/glue/home?region=us-east-1#etl:tab=jobs)

|  |  |
| --- | --- |
| Job Run Id | jr\_3f633f4219f344d7cb298eeccc8dcd4840749c65d58a02f261c05443ac118570 |
| Worker Type | G1.X |
| Number of workers | 10 |
| partitionColumn | TXN\_ITEM\_ID |
| lowerBound | 1255534593 |
| upperBound | 1829324174 |
| numPartitions | 40 |
| FETCHSIZE | 1000 |

Duration : 7 mins

Job name : [pre\_spark\_jdbc\_tune\_pk](https://console.aws.amazon.com/glue/home?region=us-east-1#etl:tab=jobs)

|  |  |
| --- | --- |
| Job Run Id | jr\_5e97df5a66347e8449514d3a38147b77946aea103dabda5d4b04b2800550d50d |
| Worker Type | G1.X |
| Number of workers | 10 |
| partitionColumn | TXN\_ITEM\_ID |
| lowerBound | 1255534593 |
| upperBound | 1829324174 |
| numPartitions | 40 |
| FETCHSIZE | 1000 |

Duration : 10 mins

Job name :

 jr\_11631093be17bdd857929a8433b36013fa8e1aa09834bdc5b6339a0bed01b98b

|  |  |
| --- | --- |
| Job Run Id | jr\_11631093be17bdd857929a8433b36013fa8e1aa09834bdc5b6339a0bed01b98b |
| Worker Type | G1.X |
| Number of workers | 40 |
| partitionColumn | TXN\_ITEM\_ID |
| lowerBound | 1255534593 |
| upperBound | 1829324174 |
| numPartitions | 40 |
| FETCHSIZE | 1000 |

Job failed : Due to unavailability of private ip address in the subnet

tableName="DM\_OWNER.LOFT\_F\_TXN\_ITEM\_VW"

pushdown\_query="(SELECT TXN\_ITEM\_ID,TXN\_ID,TXN\_NBR,MODIFIED\_DT FROM "+ tableName +" where MODIFIED\_DT>=CURRENT\_DATE-4) subq"

transaction\_item\_df = spark.read \

.format("jdbc") \

.option("url", jdbcUrl) \

.option("dbtable", pushdown\_query) \

.option("user", username) \

.option("password", password) \

.option("driver", "oracle.jdbc.driver.OracleDriver") \

.option("partitionColumn", "TXN\_ITEM\_ID") \

.option("lowerBound", 1255534593) \

.option("upperBound", 1829324174) \

.option("numPartitions", 40) \

.option("fetchsize",10000) \

.load()

|  |  |
| --- | --- |
| Job Run Id | jr\_89f85b2046988a04009c254633b0cad23b2e0f2e109c614b028899f30bcf396d |
| Worker Type | G1.X |
| Number of workers | 10 |
| partitionColumn | TXN\_ITEM\_ID |
| lowerBound | 1255534593 |
| upperBound | 1829324174 |
| numPartitions | 40 |
| FETCHSIZE | 10000 |

Duration – 2 mins