**Lab Exercise- Using AWS Glue Notebooks for Amazon S3 to AWS Glue Catalog**

To create a database in AWS Glue, you generally need to do this either via the AWS Management Console, using AWS SDKs, or with the boto3 library in Python.

Here’s how you can correctly create a database and catalog your data in AWS Glue using boto3 along with your AWS Glue Notebook.

import boto3

glue\_client = boto3.client('glue')

**Create the Database:**

**Use boto3 to create a new database:**

database\_name = "your\_database\_name"

response = glue\_client.create\_database(

DatabaseInput={

'Name': database\_name,

'Description': 'This is a test database created via boto3.'

}

)

print("Database created:", response)

**Step 3: Read Data from S3**

Import Necessary Libraries:

from awsglue.context import GlueContext

from pyspark.context import SparkContext

from awsglue.dynamicframe import DynamicFrame

glueContext = GlueContext(SparkContext.getOrCreate())

spark = glueContext.spark\_session

**Specify the S3 Path:**

s3\_input\_path = "s3://your-bucket-name/folder/your-data-file.csv"

**Read Data from S3:**

df = spark.read.format("csv").option("header", "true").load(s3\_input\_path)

df.show()

**Step 4: Create a Glue Catalog Table**

Convert to DynamicFrame:

dyf = DynamicFrame.fromDF(df, glueContext, "dyf")

**Write Data to Glue Catalog:**

Write the data into S3 and register the table in the Glue Data Catalog:

output\_s3\_path = "s3://your-bucket-name/processed-data/"

glueContext.write\_dynamic\_frame.from\_options(

frame = dyf,

connection\_type = "s3",

connection\_options = {"path": output\_s3\_path},

format = "parquet",

format\_options = {"compression": "snappy"}

)

glue\_client.create\_table(

DatabaseName=database\_name,

TableInput={

'Name': 'your\_table\_name',

'StorageDescriptor': {

'Columns': [{"Name": "col1", "Type": "string"},

{"Name": "col2", "Type": "int"}],

'Location': output\_s3\_path,

'InputFormat': 'org.apache.hadoop.mapred.TextInputFormat',

'OutputFormat': 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat',

'Compressed': False,

'NumberOfBuckets': -1,

'SerdeInfo': {

'SerializationLibrary': 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe',

'Parameters': {

'serialization.format': '1'

}

},

'BucketColumns': [],

'SortColumns': [],

'StoredAsSubDirectories': False

},

'TableType': 'EXTERNAL\_TABLE',

'Parameters': {

'EXTERNAL': 'TRUE'

}

}

)

**Step 5: Verify the Output**

* Navigate to the AWS Glue Console to check that the database and catalog have been created successfully.
* Also, check the S3 bucket to ensure that the data has been stored in the specified format.