Recon Data between RDS and Redshift （CN）

Redshift 不强制其主键的唯一性。 由于 DMS 中的错误，与源 RDS 表相比，迁移到 Redshift 中的数据可能会丢失或重复数据。

此方案提供了一种检测 RDS表和 Redshift 表之间数据不一致性的情况。

- 将 RDS 表中的 ID 导出到 S3 存储桶中。
- 将 Redshift 表中的 ID 导出到 S3 存储桶中。
- 比较 RDS 和 Redshift 导出的数据并找出不同的数据。

Steps

1. 为导出的数据创建/选择一个 S3 存储桶。

```
REGION="ap-southeast-1"
S3_BUCKET = 'temp-460453255610'
```

2. 设置 RDS 表的配置。 为了安全起见，考虑将用户名和密码存储在 AWS Secret Manager 中。

```
# Export data from RDS to S3 Bucket
RDS_ENDPOINT="database-3.cluster-c3bottoj4h9o.ap-southeast-1.rds.amazonaws.com"
RDS_DB="stackoverflow"
RDS_USER="admin"
RDS_PASSWD = "Qwer!234"
RDS_PREFIX = f'rds_{datetime_str}'
RDS_QUERY = 'select id from posts'
s3_path_rds = f's3-{REGION}://{S3_BUCKET}/{RDS_PREFIX}'
```

3. 设置 Redshift 表的配置。

```
# Export data from Redshift to S3 Bucket
REDSHIFT_HOST = 'redshift-cluster-1.cs21sivqdtax.ap-southeast-
1.redshift.amazonaws.com'
REDSHIFT_DB = 'dev'
REDSHIFT_USER = 'awsuser'
REDSHIFT_PASSWD = 'Qwer!234'
REDSHIFT_PREFIX = f'redshift_{datetime_str}'
REDSHIFT_QUERY = 'select id from stackoverflow.posts'
REDSHIFT_IAM_ROLE = 'arn:aws:iam::460453255610:role/RedshiftImportFromS3'
s3_path_redshift = f's3://{S3_BUCKET}/{REDSHIFT_PREFIX}'
```

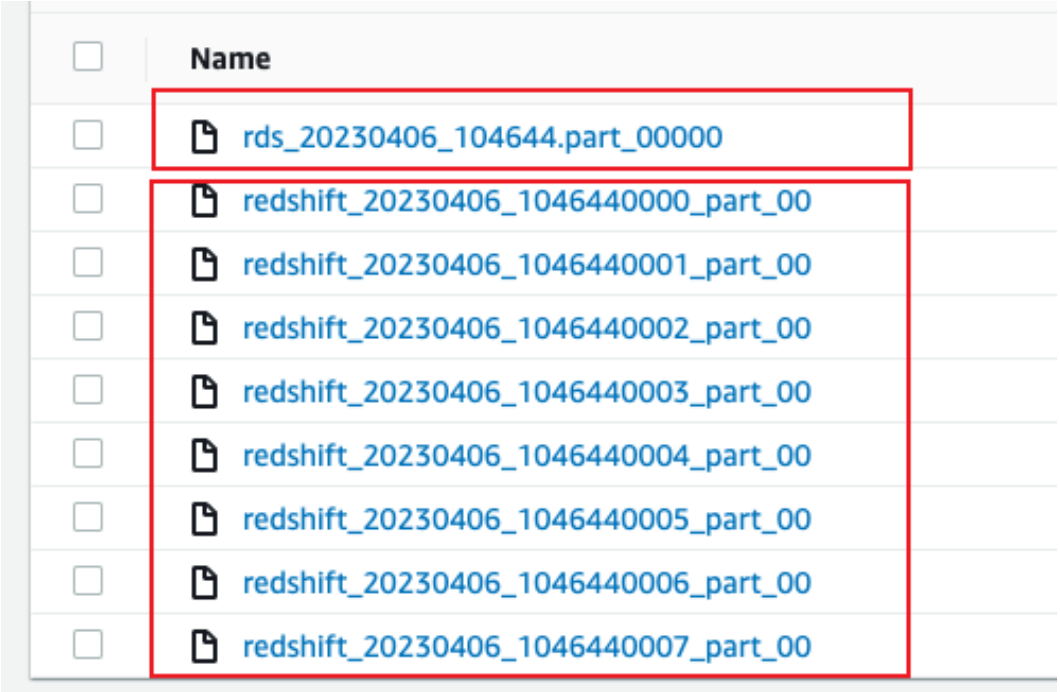
4. 使用 SELECT INTO S3 导出 RDS 表。

```
print('Select from RDS into S3...')
select_rds_into_s3(RDS_ENDPOINT, RDS_USER, RDS_PASSWD, RDS_DB, RDS_QUERY,
s3_path_rds)
```

5. 使用 UNLOAD 导出 Redshift 表。

```
print('Unload from Redshift into S3...')
unload_redshift_to_s3(REDSHIFT_HOST, REDSHIFT_DB, REDSHIFT_USER, REDSHIFT_PASSWD,
REDSHIFT_QUERY, s3_path_redshift, REDSHIFT_IAM_ROLE, REGION)
```

6. 导出的数据保存在 S3 存储桶中。
    - 可以考虑在`RDS_PREFIX`和`REDSHIFT_PREFIX`中添加文件夹，这样文件就可以放到文件夹中了。



7. 从 S3 下载文件并将它们分别读入 2 个单独的列表。

```
# Find matching files in S3 and download them
files = list_s3_files_by_prefix(bucket_name=S3_BUCKET, prefix=RDS_PREFIX)
create_folder('data_rds', True)
print('Download RDS data from S3...')
rds_files = download_s3_files(S3_BUCKET, [file['Key'] for file in files],
'data_rds')

rds_data = []
for file in rds_files:
    with open(file) as f:
        rds_data.extend(f.readlines())

# Find matching files in S3 and download them
files = list_s3_files_by_prefix(bucket_name=S3_BUCKET, prefix=REDSHIFT_PREFIX)
create_folder('data_redshift', True)
```

```
print('Download Redshift data from S3...')
redshift_files = download_s3_files(S3_BUCKET, [file['Key'] for file in files],
'data_redshift')

redshift_data = []
for file in redshift_files:
    with open(file) as f:
        redshift_data.extend(f.readlines())
```

8. 将它们转换成集合并找出它们之间的区别。

```
print(len(rds_data), len(redshift_data))

only_in_rds = set(rds_data).difference(set(redshift_data))
print('Only in RDS:', len(only_in_rds))

only_in_redshift = set(redshift_data).difference(set(rds_data))
print('Only in Redshift:', len(only_in_redshift))
```

9. 检查输出。

```
Download RDS data from S3...
Downloading file  rds_20230406_104905.part_00000
/home/ubuntu/environment/recon_rds_redshift/data_rds/rds_20230406_104905.part_00000
Download Redshift data from S3...
Downloading file  redshift_20230406_1049050000_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050000_part_00
Downloading file  redshift_20230406_1049050001_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050001_part_00
Downloading file  redshift_20230406_1049050002_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050002_part_00
Downloading file  redshift_20230406_1049050003_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050003_part_00
Downloading file  redshift_20230406_1049050004_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050004_part_00
Downloading file  redshift_20230406_1049050005_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050005_part_00
Downloading file  redshift_20230406_1049050006_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050006_part_00
Downloading file  redshift_20230406_1049050007_part_00
/home/ubuntu/environment/recon_rds_redshift/data_redshift/redshift_20230406_1049050007_part_00
920821 920821
Only in RDS: 0
Only in Redshift: 0
```

10. 如果数据有差异，脚本会抛出异常。这使用于部署在lambda上。

```
Only in RDS: 1
Only in Redshift: 1
Only in RDS: {'DUMMY in RDS'}
Only in Redshift: {'DUMMY in REDSHIFT'}
Traceback (most recent call last):
  File "main.py", line 171, in <module>
    raise Exception('Data is different between RDS and Redshift')
Exception: Data is different between RDS and Redshift
(venv) Admin:~/environment/recon_rds_redshift $
```

Considerations

1. 如果查询中包含where语句，可以考虑将where语句中的列设置为key，这样可以加快性能。
2. **SELECT INTO S3**可能会对RDS产生影响。 如果您担心性能影响，您可以考虑为 RDS 添加只读副本。