

Data Engineering – Search a file in S3

Problem Statement – A business group received 100+ data feeds from different source systems and use Glue or lambda and process the files. To ensure at end of the day, all files are available in the S3 or any source system did not transfer the file for that day, it is important to have a reconciliation process in place. The process will take the require files as input and generate report and sent a list which files are missing

In this blog, we will make try to achieve a one part of the use case and develop a function which will take input the file name, s3 bucket and prefix and validate whether the file is there in that prefix of the bucket or not. This program can be generalized by taking the file pattern and generate a report and sent via email.

Note:

1. If you are running this function via lambda, you need to configure additional details like put the function in handler and the function have right policies.
2. If you are running this function locally, ensure you have configured your aws credentials and the user should have right S3 permission. I developed this function locally and tested it.
3. import boto3
4. def S3FileCheck(file\_name,client,bucket\_name,bucket\_prefix):
5. my\_bucket = client.Bucket(bucket\_name)
6. file\_obj = []
7. for objects in my\_bucket.objects.filter(Prefix= bucket\_prefix):
8. file\_obj.append(objects.key)
9. s3\_file\_list = []
10. for file in file\_obj:
11. object\_key = str(file)
12. s3\_file\_name = file.split('/')[-1]
13. s3\_file\_list.append(s3\_file\_name)
14. file\_check\_status = {}
15. if file\_name in s3\_file\_list:
16. file\_check\_status['file\_name'] =file\_name
17. file\_check\_status['file\_available\_S3'] ='Y'
18. else:
19. file\_check\_status['file\_name'] =file\_name
20. file\_check\_status['file\_available\_S3'] ='N'
21. return file\_check\_status
23. client = boto3.resource('s3')
24. bucket\_name = 'sanjeeb-poc-lab-001'
25. bucket\_prefix = 'file\_validation'
26. file\_name='test\_sanjeeb.csv'
27. file\_check\_status = S3FileCheck(file\_name,client,bucket\_name,bucket\_prefix)
28. print(file\_check\_status)

Output:

Test case: File is not available in S3



Test case: File is available in S3 (changed the file name in the code and test it). 