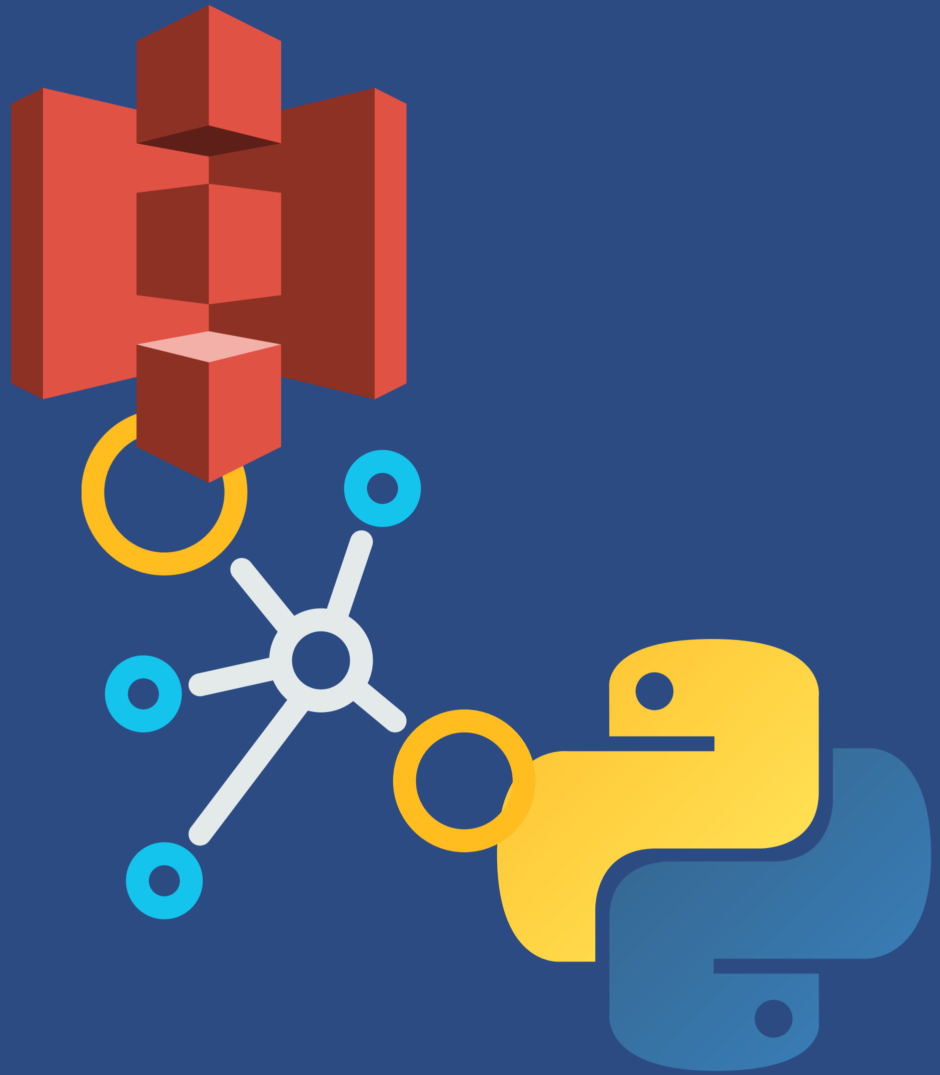


S3: Making it your Data Warehouse

Steps to super charge how you use AWS S3

Made By

Kamalraj M M

How to Make S3 your Data Warehouse

1

STEP

Connect to S3 using AWS Wrangler

Keep those Access Keys Handy... We will take AWS for Ride

2

STEP

Read, Pull and Push files in S3 buckets like Pro

Beware Free tier S3 has 5GB space and 2000 I/O limits

3

STEP

BTW
What is a Data warehouse

Know this to become AWS Ninja

4

STEP

Hands on Athena+ Glue with S3 = Data Pipleine

Understand Athena & Glue to use S3 efficiently

Connecting S3 : AWS Wrangler

What is AWS Wrangler?
It is a Python library written by AWS Professional Services team, to make the life AWS users super charged. They are calling it now aws-sdk-pandas, I will be calling it awswrangler.

```
# install using pip
pip install awswrangler
# lets import it
import awswrangler as wr
```

```
# Import boto3 and get a Session
import boto3
your_session =
boto3.Session(aws_access_key=aws_key,
aws_access_secret=aws_secret,
aws_region_name = aws_region)
```

```
# Use the session when you are calling the services
using awswrangler
Ex:
wr.s3.upload_file(file, boto3_session = your_session)
```

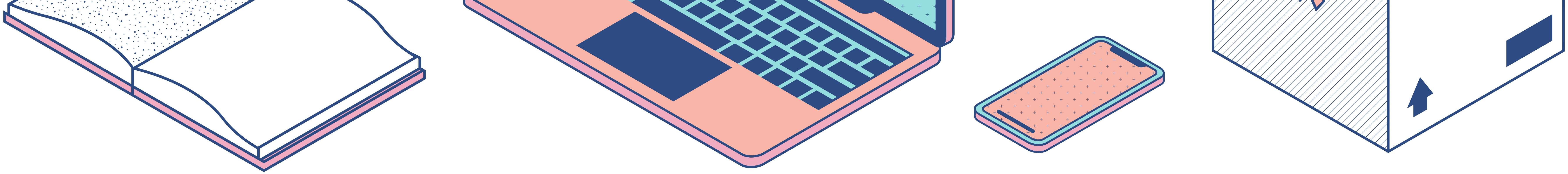
Key S3 Methods

Following methods are used for downloading and uploading files to S3 using AWSWrangler (wr is used as placeholder)

Methods for file Ops

Bucket_path = f"s3://{bucket}/csv/file1.csv"
Your_file = os.path.join(local_file_dir, "file1.csv")

- wr.s3.download(path=path1, local_file=local_file)
- wr.s3.upload(local_file=local_f, path=path2)
- wr.s3.list_objects(bucket_path)
- wr.s3.delete_objects(Bucket_path)



Methods for Reading in Files

bucket_path = 's3://{bucket}/folder/'

- var_name = wr.s3.read_fwf(bucket_path)
- var_name = wr.s3.read_json(bucket_path)
- var_name = wr.s3.read_csv(bucket_path)
- var_name = wr.s3.read_parquet(bucket_path)

Above functions read "all" the files in the bucket_path. No need for us to loop over them. This is how the Data warehouse programs have to function.

=> **wr.s3.write_* method is Special**

will be explained after explaining Data warehouse

What is a DWH?

DATA WAREHOUSE IS A CONNECTED NETWORK OF MACHINES WHICH HAVE THE FOLLOWING

1. Distributed File System, for example Hadoop/ EMRFS
2. Database engine to query HDFS. Example HIVE / Athena
3. A Metastore / Data catalog. Example is Hive Metastore /Glue catalog
4. Data transformation applications. Scala's Spark 3/ Pyspark
5. Has a resource manager or negotiator. Example Yarn

What Challenge DWH Solves?

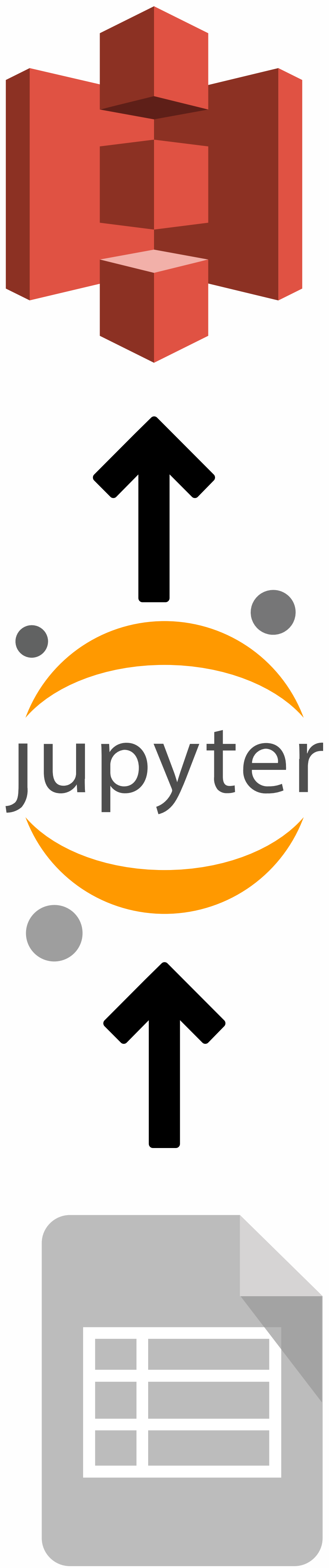
DATA WAREHOUSE (DWH) SOLVES TWO CHALLENGES

1. It makes it possible to work with files that are much bigger than your system's RAM
2. It provides Data redundancy by distributing the files in small parts on many systems on Network.

Why you need one?

HAVING ACCESS TO DWH PROVIDES TWO ADVANTAGES

1. Many repetitive activities like data loading from multiple folders, splitting files are automated.
2. Files ranging and terabytes are read as if they are just CSV files. These files are special format compressed files, optimising storage



File Read Methods

0 - Creating and connecting to AWS Session

1 - wr.s3.upload(bucket_location, your_local_file)

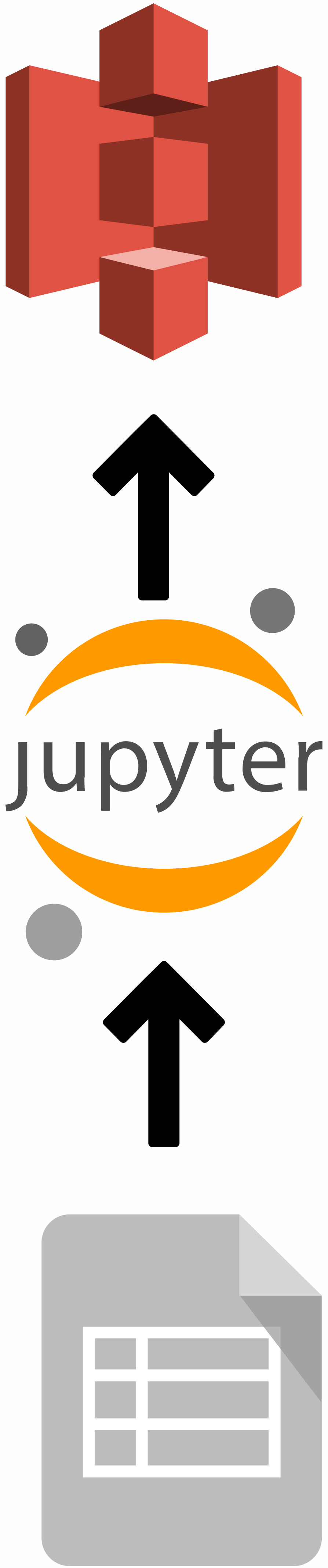
2 - Using wr.s3.list_objects(bucket_location)

3 - Using wr.s3.download(bucket_location, bucket_file)

4 - Reading json file using wr.s3.read_json(s3_bucket/folder)

5 - Reading csv file using wr.s3.read_csv(s3_bucket/folder)

6 - Reading parquet file using wr.s3.read_parquet(your_local_file)



File/Data Write Method

0 - Creating and connecting to AWS Session

1 - Writing csv file using wr.s3.to_csv(df, path=s3_destination/file.csv)

2 - Reading csv file using wr.s3.to_parquet(df, path=s3_destination/file.parquet)

3 - Reading fwf file using wr.s3.to_json(df, path=s3_destination/file.json)

4 - Reading fwf file using wr.s3.to_excel(df, path=s3_destination/file.xlsx)

Whats Next

HANDS ON CODING WITH AWS WRANGLER AND PYTHON

We will dive deeper into each method
and see the magic happen

Ingest 4 types of files in to S3 Bucket

Write the catalog directly into AWS Glue

Query the data written into S3 Using Athena

Do you have any questions?

Send it to us! We hope you learned something new.