Data ingestion and Exploration

1. Structureed relational data : CSV & TSV files

2. Unstructured data : images, audio and media files

3. Semi-Stuctured data : JSON & XML file

AWS data wrangler:

1. open sourse library
2. load/unload data from:
    - *data lakes
    - * data warehouse
    - * Databases

Code:

```
!pip install awswrangler

import pandas as pd

import awswrangler as wr

#Retrieving the data directly from amazon S3

df = wr.s3.read_csv(path = ' ')
```

AWS Glue Data Catalog:

This data catalog service is used to register or catalog the data stored in S3.

How can you register the data?

You can use the AWS Data Wrangler tool just as I introduced. The first step is to create an AWS Glue Data Catalog database. To do that, import the AWS Wrangler Python library as shown here, and then call the catalog.create_database function, providing a name for the database to create. AWS Data Wrangler also offers a convenience function called catalog.create_CSV_table that you can use to register the CSV data with the AWS Glue Data Catalog. The function will only store the schema and the metadata in the AWS Glue Data Catalog table that you specify. The actual data again remains in your S3 bucket.

Code:

```
import awswrangler as wr

# Create a database in the AWS Glue Data Catalog
```

wr.catalog.create.database(

name = ...)

Create CSV table(metadata onle) in the AWS Glue Data Catalog

wr.catalog.create_csv_table(

table = ...,

column_types = ...,

)

You can query the data stored in S3, using a tool called Amazon Athena. Athena is an interactive queries service that lets you run standard SQL queries to explore your data.

SQL Code:

"SELECT product_category FROM reiews"

Let's list all product categories from the AWS Glue table called reviews, and imagine this table points to the dataset stored in S3.

To run this query, you can use the again previously introduced AWS Data Wrangler tool again.

Code:

import aswrangler as wr

#Create Amazon Athena S3 bucket

wr.athena.create_athena_bucket()

#Execute SQL Query on Amazon Athena

df = wr.athena.read_sql_query(

sql = ...,

database = ...)

From the Python environment you're working in, just use the Data Wrangler, Athena, read_SQL_query function.

Pass in the SQL statement you have written and point to the AWS Glue database, which contains the table you'll reference here in the SQL query.