CESTE
Escuela Internacional de Negocios

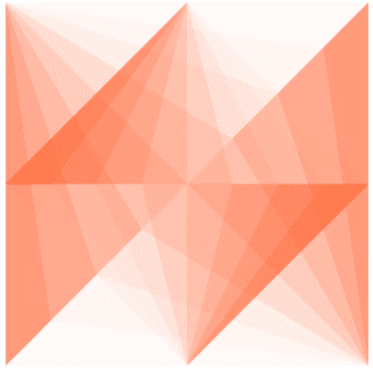
AWS SDK Data Wrangler

Intro

Sesión 1

AWS Data Wrangler

AWS, conexión de **DataFrames** y servicios de análisis y datos de AWS



AWS SDK for pandas

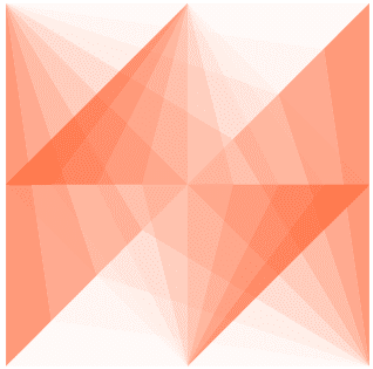
AWS Data Wrangler: descripción general

- ¿Qué es Data Wrangler?
- ¿ Cómo utilizarlo en el desarrollo de pipelines de datos ?
- La diferencia entre SageMaker Data Wrangler y AWS Data Wrangler
- ¿Qué servicios de AWS pueden utilizarse?
- ¿Qué servicios de AWS pueden ejecutar la libreria Python de AWS Data Wrangler ?

Qué es Data Wrangler?

- Libreria Python de código abierto
- Desarrollado sobre Pandas, Pyarrow y Boto 3

+



AWS SDK for pandas

¿Qué es Data Wrangler?

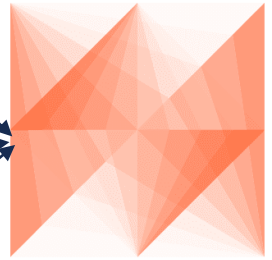
Data Lake

Amazon Athena

AWS Glue

Datawarehouse

Amazon Redshift



AWS SDK for pandas

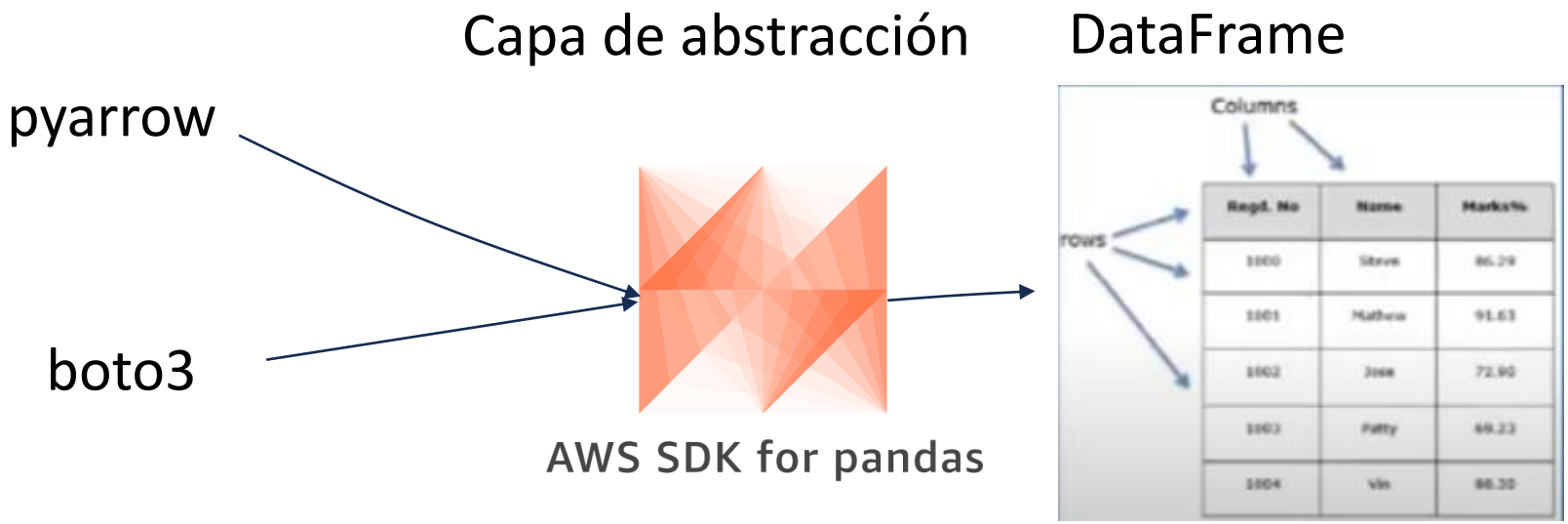
Databases

Amazon RDS

Amazon DynamoDB

¿ Cómo utilizarlo en el desarrollo de pipelines de datos ?

- Simplificando el proceso de desarrollo



El diferencia entre SageMaker Data Wrangler y AWS Data Wrangler

AWS Data Wrangler

PyPI (pip)

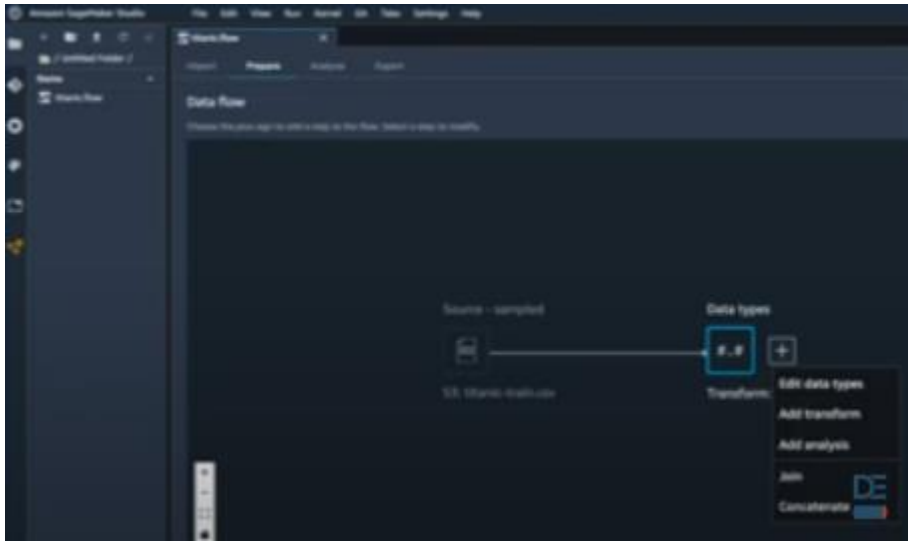
```
>>> pip install awswrangler
```

```
>>> # Optional modules are installed with:
>>> pip install 'awswrangler[redshift]'
```

Conda

```
>>> conda install -c conda-forge awswrangler
```

SageMaker Data Wrangler



CESTE

¿Qué servicios de AWS puede utilizar?

AWS SDK for pandas 3.10.0

Amazon S3
AWS Glue Catalog
Amazon Athena
Amazon Redshift
PostgreSQL
MySQL
Microsoft SQL Server
Oracle
Data API Redshift
Data API ROS
AWS Glue Data Quality
OpenSearch
Amazon Neptune
DynamoDB

Amazon Timestream
AWS Clean Rooms
Amazon EMR
Amazon EMR Serverless
Amazon CloudWatch Logs
Amazon QuickSight
AWS STS
AWS Secrets Manager
Amazon Chime
Typing
Global Configurations
Engine and Memory Format
Distributed - Ray

Trabajar con Data Lakes

- Admite lectura de Excel, archivos con formato de ancho fijo, JSON, Parquet desde S3
- Admite escritura de CSV, Excel, JSON y Parquet en S3
- Interactúa con datos y metadatos a través del catálogo de AWS Glue

Amazon Athena

AWS Glue

Trabajar con Data Warehouses

- Puede leer y escribir datos en Redshift
  - conectar
  - consultas SQL

Amazon Redshift

Trabajar con RDS

- Puede leer y escribir datos en Postgres, MySQL, Microsoft SQL
  - conectar
  - consultas SQL

Trabajar con Amazon Athena

- Puede leer y escribir datos con Athena
- Ejecutar Spark

```
>>> import awswrangler as wr
>>> df = wr.athena.read_sql_query(
...     sql="SELECT * FROM my_table WHERE name=:name AND city=:city",
...     params={"name": "filtered_name", "city": "filtered_city"}
... )
```

```
>>> import awswrangler as wr
>>> df = wr.athena.run_spark_calculation(
...     code="print(spark)",
...     workgroup="...",
... )
```

CESTE

Trabajar con Amazon EMR

- Administrar clúster de EMR
- Pueden crear Jobs de EMR


```
>>> import awswrangler as wr
>>> step_id = wr.emr.submit_spark_step(
>>>     cluster_id="cluster-id",
>>>     path="s3://bucket/emr/app.py"
>>> )
```

Trabajar con registros de Amazon Cloudwatch

- Consulta los registros y los devuélve como pandas DataFrame


```
>>> import awswrangler as wr
>>> df = wr.cloudwatch.read_logs(
...     log_group_names=["loggroup"],
...     query="fields @timestamp, @message | sort @timestamp desc | limit 5",
... )
```

Trabajar con Amazon DynamoDB

- Admite "puts" desde csv, dataframe o json
- Leer “items” en 3.10
- Eliminar “items” de la tabla


```
>>> import awswrangler as wr
>>> df = wr.dynamodb.read_items(
...     table_name='my-table',
...     partition_values=['pv_1', 'pv_2'],
...     sort_values=['sv_1', 'sv_2']
... )
```

Trabajar con Amazon Secrets Manager

- Recuperar secretos en Python
- Útil si se necesita acceso programático a las credenciales

AWS **Secrets Manager**

```
>>> import awswrangler as wr
>>> value = wr.secretsmanager.get_secret("my-secret")
```

Ejecución local de AWS Data Wrangler

- Pip

  “pip install awswrangler "

- Conda

  "conda install -c conda -forge awswrangler"

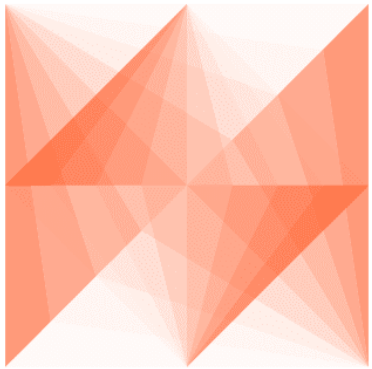
¿Qué servicios de AWS pueden ejecutar la libreria Python de AWS Data Wrangler ?

Iniciar Data Wrangler

- Lambda
  - Subir la librería AWS DataWrangler como “layer” de AWS Lambda
- AWS Glue PySpark jobs
  - Agregar nuevo parámetro de Job par -> clave / valor
  - Subir wheel de la librería a s3 para usar con el python shell
- Amazon Sage Maker Notebook
  - !pip install awswrangler
- EMR
  - sudo pip-3.6 instalar awswrangler

Resumen

- Simplifica el proceso de trabajo con Pandas con AWS
- Se integra bien con muchos servicios de AWS
- Puede ejecutarse localmente o en varios servicios de AWS



AWS SDK for pandas

EJERCICIOS

INSTALACION:
https://aws-sdk-pandas.readthedocs.io/en/stable/install.html

SERVICIOS

www.ceste.es