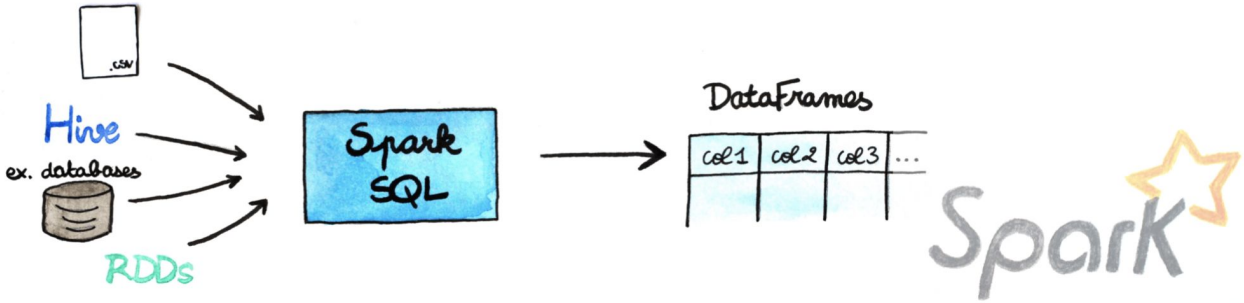
PySpark | PyConAR 2021

https://www.github.com/zilohumberto

Humberto Rodríguez | zilohumberto@gmail.com

Objetivos

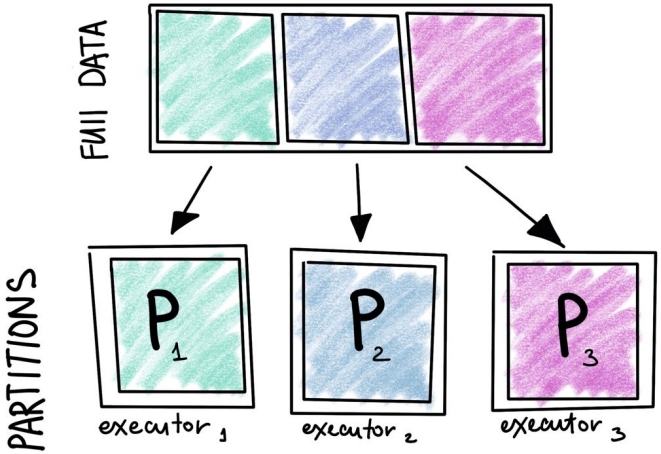
- Conocer los principios de Apache Spark
- Operaciones básicas con PySpark - SQL DataFrame
- Limpiar y preparar datos con PySpark + Airflow + DataBricks + SnowFlake



https://aspgems.com/migrando-de-pandas-a-spark-dataframes/

- Framework
- Computación distribuida
- Ejecuta cargas de trabajo 100x más rápido
- RDD
- R, SQL, Java, Scala y **Python**
- >100GB
- Fácil de usar

https://spark.apache.org

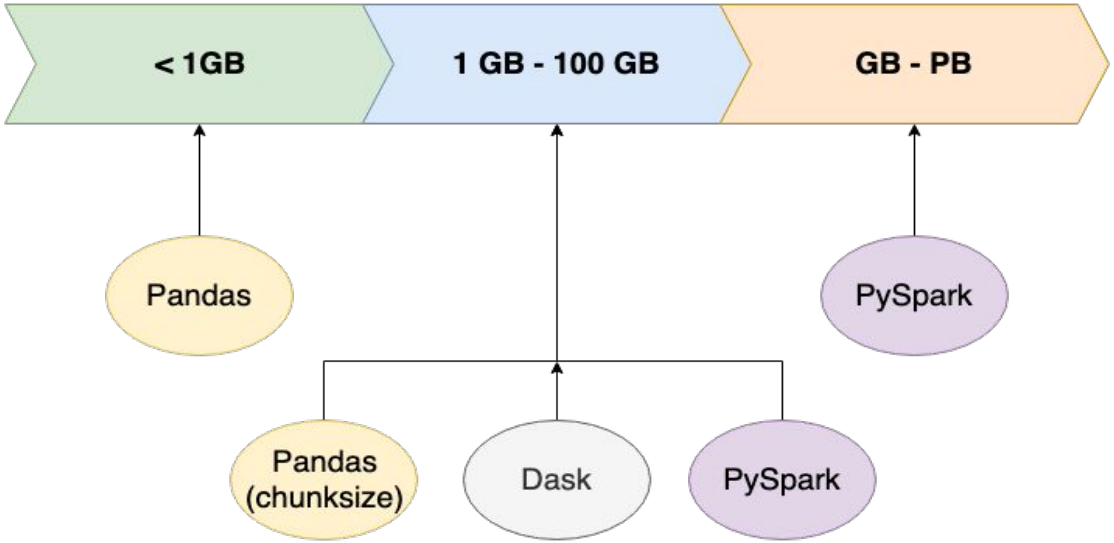


@luminousmen.com

- Framework
- Computación distribuida
- Ejecuta cargas de trabajo 100x más rápido
- RDD
- R, SQL, Java, Scala y **Python**
- >100GB
- Fácil de usar

https://spark.apache.org

Pandas vs Spark



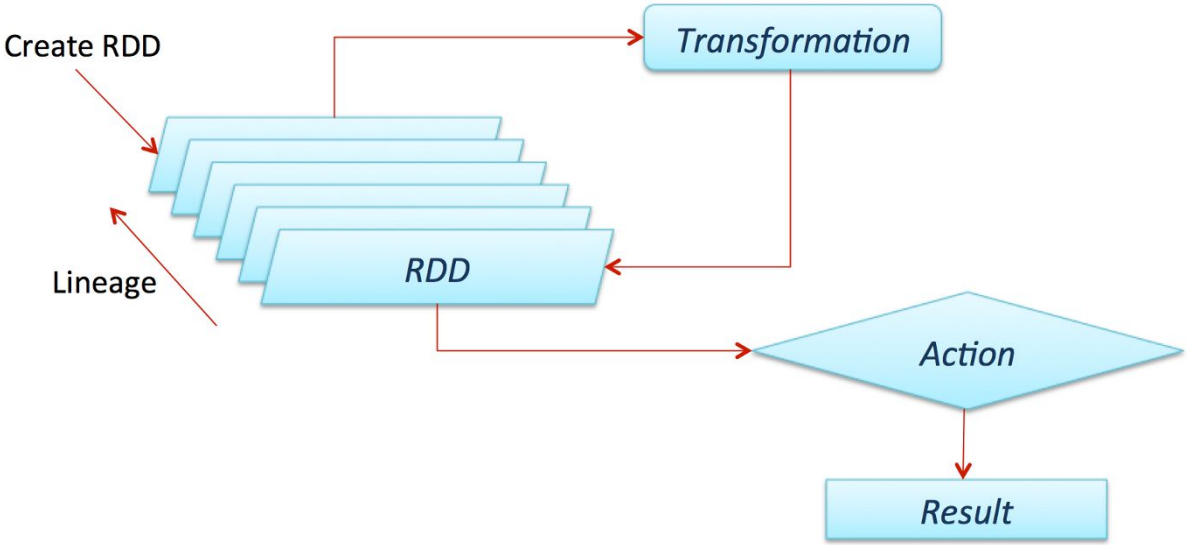
Pandas vs Spark

Es simple	Escalable
+Recursos en internet	No limitaciones en términos de data
Excelente rendimiento con datos <1GB	Base en Scala
Integración	Muchos tipos conectores

spark-vs-pandas-part-4-recommendations

- Interface/librería de Apache Spark escrita en Python
- DataFrame
- Características
  - Procesamiento en memoria
  - Ejecución ‘Lazy’
  - Particionado
  - Persistencia
  - Inmutabilidad

spark.apache.org/docs/latest/api/python/index.html


Create RDD
Transformation
RDD
Lineage
Action
Result

Pasos

Airflow	**PySpark**	Databricks	SnowFlake

https://github.com/zilohumberto/pyspark-pyconar-2021

Datos

- Son ~33GB, repartido en 1000 archivos
- Vamos a determinar qué canciones tiene de duración más de dos minutos y formatear la fecha
- lectura + procesamiento + escritura en 18 minutos (~66M canciones)

https://www.kaggle.com/adityak80/spotify-millions-playlist

PySpark

Create Cluster

New Cluster

Cancel | Create Cluster | DBU / hour: 3.75 | 4 Workers:56 GB Memory, 16 Cores 1 Driver:14 GB Memory, 4 Cores

UI | JSON

Cluster Name

PyConAr-2021

Cluster Mode

Standard

Databricks Runtime Version Learn more

Runtime: 8.3 (Scala 2.12, Spark 3.1.1)

Note Databricks Runtime 8.x and later use Delta Lake as the default table format. Learn more

Autopilot Options

- [ ] Enable autoscaling
- [x] Terminate after 120 minutes of inactivity

Worker Type

Standard_DS3_v2 — 14 GB Memory, 4 Cores

Workers

04

- [ ] Spot instances

New Configure separate pools for workers and drivers for flexibility. Learn more

Driver Type

Standard_DS3_v2 — 14 GB Memory, 4 Cores

DBU / hour: 3.75 | Standard_DS3_v2

```
example_cluster_config = {
    "new_cluster": {
        "num_workers": 4,
        "spark_version": "7.5.x-scala2.12",
        "spark_conf": {},
        "azure_attributes": {
            "first_on_demand": 1,
            "availability": "ON_DEMAND_AZURE",
            "spot_bid_max_price": -1
        },
        "node_type_id": "Standard_DS3_v2",  # 14 GB Memory, 4 cores
        "driver_node_type_id": "Standard_DS3_v2",
        "ssh_public_keys": [],
        "custom_tags": {},
        "spark_env_vars": {
            "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
        },
        "enable_elastic_disk": True,
        "cluster_source": "UI",
        "init_scripts": []
    },
    "libraries": [
    ],
    "spark_python_task": {"python_file": "dbfs:/FileStore/spotify_analyzer.py", "parameters": []}
}
```

```
copy_pyfile = BashOperator(
    task_id="copy_py",
    bash_command=f"source {venv_path}/bin/activate && "
                 f"databricks fs cp {home}/jobs/{py_file} dbfs:/FileStore/{py_file} --overwrite",
    dag=dag,
)
```

```
run_job = DatabricksSubmitRunOperator(
    task_id="spotify_task", json=example_cluster_config, dag=dag
)
```

Airflow DAGs Security Browse Admin Docs 00:32 UTC AA

DAGs

All 1 Active 1 Paused 0

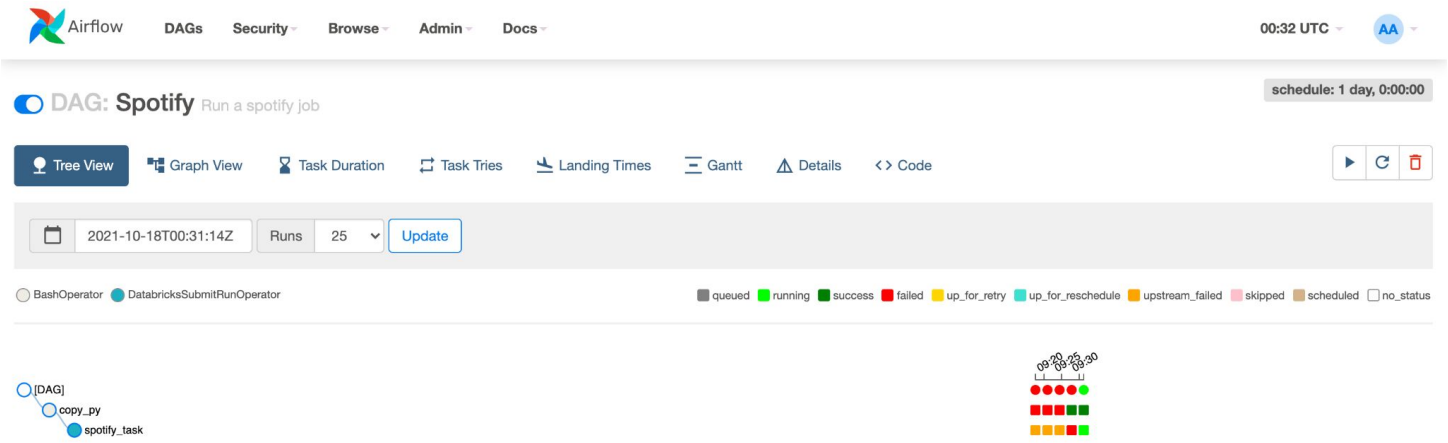
Filter DAGs by tag

spo

DAG	Owner	Runs	Schedule	Last Run	Recent Tasks	Actions	Links
Spotify	airflow	1 4	1 day, 0:00:00	2021-10-18, 00:31:14	2 1 1		

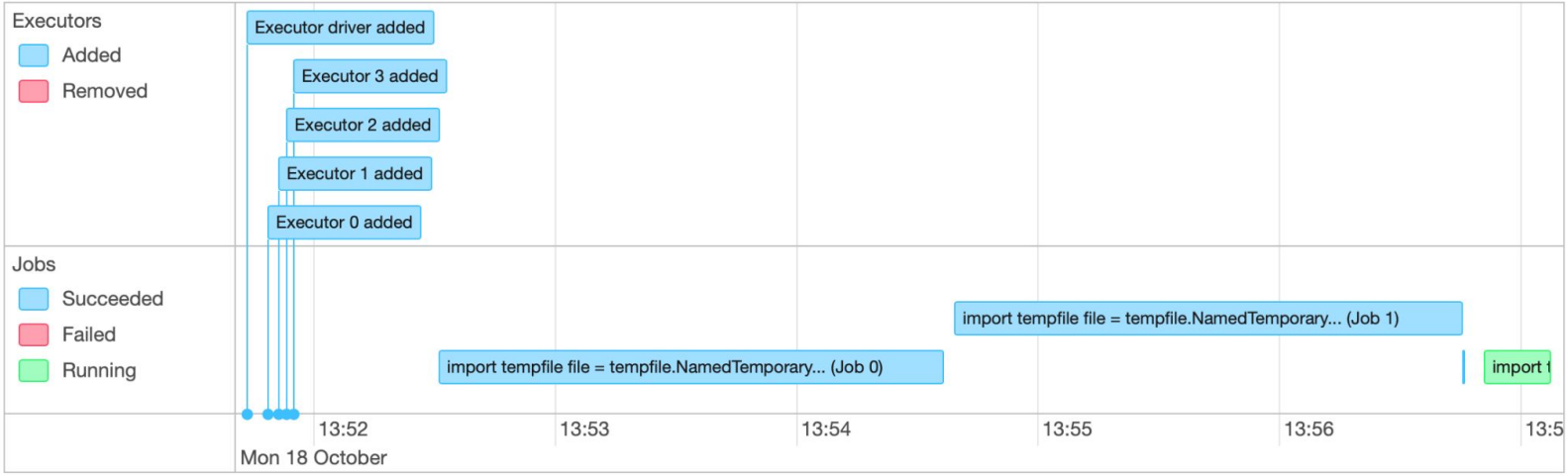
« ‹ 1 › »

Showing 1-1 of 1 DAGs


Airflow
DAGs
Security
Browse
Admin
Docs
00:32 UTC
AA
DAG: Spotify
Run a spotify job
schedule: 1 day, 0:00:00
Tree View
Graph View
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
2021-10-18T00:31:14Z
Runs
25
Update
BashOperator
DatabricksSubmitRunOperator
queued
running
success
failed
up_for_retry
up_for_reschedule
upstream_failed
skipped
scheduled
no_status
[DAG]
copy_py
spotify_task
09:20
09:25
09:30

Event Timeline

Enable zooming



```
show tables;
select * from SPOTIFY_PROCESSED limit 10
```

Results Data Preview Open History

Query ID SQL 519ms 10 rows

Filter result... Copy Columns

NAME	PID	ALBUM_NAME	ALBUM_URI	ARTIST_NAME	ARTIST_URI	DURATION_MS	POS	TRACK_NAME	TRACK_URI	PYAR_MORE_TW	PYAR_GENERATI
ride	227981	Cannonball (...	spotify:albu...	Singer's Edg...	spotify:artist...	204475	177	Cannonball (...	spotify:track...	TRUE	2017-12-03
vibin	227305	Excuse My F...	spotify:albu...	French Mont...	spotify:artist...	181920	63	Freaks	spotify:track...	TRUE	2017-12-03
Bounce Back	227583	Coloring Book	spotify:albu...	Chance The ...	spotify:artist...	141542	11	All Night (fe...	spotify:track...	TRUE	2017-12-03
Classic Rock	227360	Boston	spotify:albu...	Boston	spotify:artist...	285133	4	More Than a...	spotify:track...	TRUE	2017-12-03
Mood swings	227146	Manners	spotify:albu...	Passion Pit	spotify:artist...	174760	12	Sleepyhead	spotify:track...	TRUE	2017-12-03
better days	227323	A Wild Hunger	spotify:albu...	Ben Rosenb...	spotify:artist...	326480	45	This Fire	spotify:track...	TRUE	2017-12-03

Agradecimientos

https://enlyft.bamboohr.com/jobs/

https://www.linkedin.com/company/enlyft/