

Procesamiento de Datos de COVID-19 en Colombia usando PySpark en JupyterHub

1. Introducción Este laboratorio se centró en el procesamiento de datos de casos positivos de COVID-19 en Colombia utilizando PySpark en un entorno de JupyterHub con un clúster EMR. El objetivo fue realizar diversas operaciones de manipulación y análisis de datos sobre el dataset proporcionado por el Ministerio de Salud.

2. Descripción del Dataset

- **Fuente de Datos:** Casos positivos de COVID-19 en Colombia
- **Origen:** Ministerio de Salud
- **Archivo:** Casos_positivos_de_COVID-19_en_Colombia-100K.csv

3. Procesamiento de Datos con PySpark Operaciones Realizadas:

1. Análisis de Columnas:

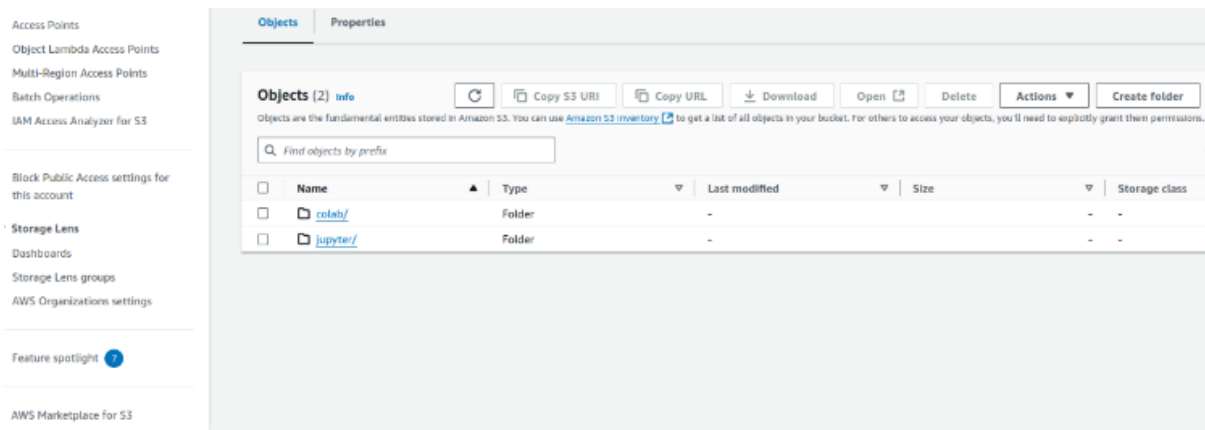
- Identificación de las columnas presentes en el dataset
- Evaluación de los tipos de datos de cada columna

2. Transformaciones de Datos:

- Selección de columnas específicas
- Renombrado de columnas
- Adición de nuevas columnas
- Eliminación de columnas no relevantes

3. Técnicas de Procesamiento:

- Filtrado de datos
- Implementación de:
 - Función UDF (User-Defined Function) para crear una nueva columna
 - Función lambda para generar columnas adicionales



Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3

Block Public Access settings for this account

▼ Storage Lens

Dashboards
Storage Lens groups
AWS Organizations settings

Feature spotlight ?

► AWS Marketplace for S3

Objects

Properties

Objects (2) [Info](#)



🔄

📄 Copy S3 URI

📄 Copy

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to

🔍 Find objects by prefix

<input type="checkbox"/>	Name	Type
<input type="checkbox"/>	 _SUCCESS	-
<input type="checkbox"/>	 part-00000-68a5ba2e-7626-4311-94b0-f95503c2fde2-c000.csv	csv