

Evidencias creación del clúster y DataLake

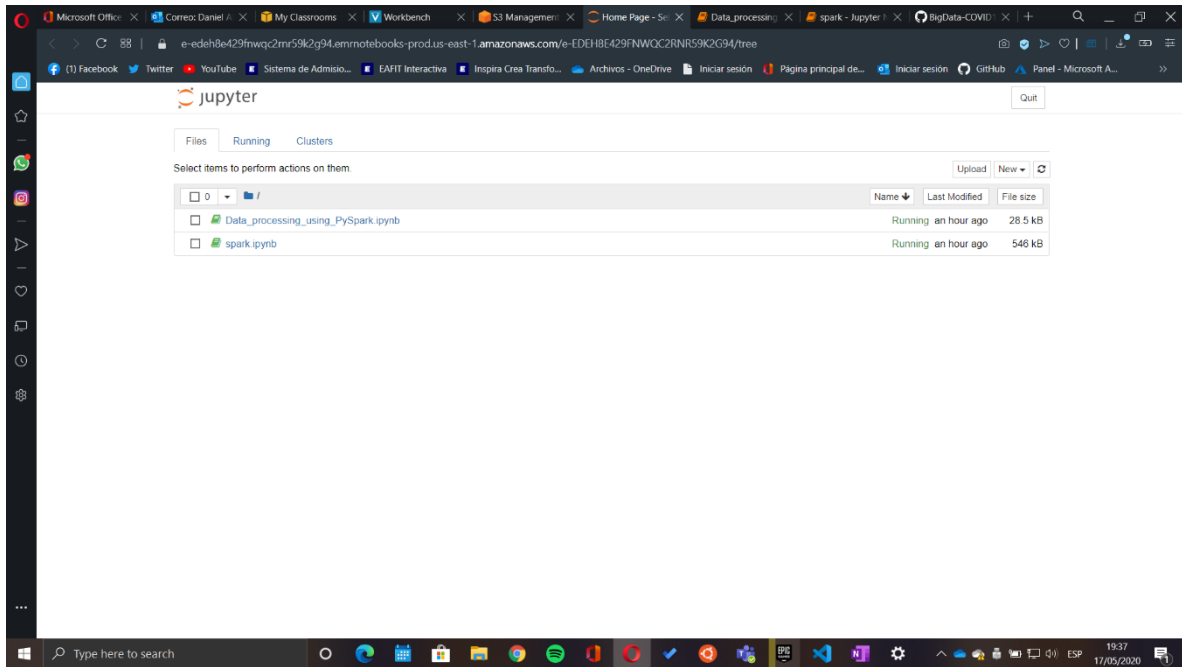
Clúster en AWS

The screenshot shows the AWS Management Console for an Amazon EMR cluster. The cluster is named "My cluster big data 3" and is in the "Waiting" state. The console displays various tabs for cluster management, including Summary, Application history, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. The Summary tab is active, showing details such as the cluster ID (j-3JB0PMY4W4RJ1), creation date (2020-05-17 14:55 UTC-5), and elapsed time (4 hours, 43 minutes). It also lists the master public DNS, history service, and tags. The Configuration details section shows the release label (emr-5.26.0), Hadoop distribution (Amazon 2.8.5), and applications (Hive 2.3.5, Hue 4.4.0, Spark 2.4.3, Sqoop 1.4.7, Oozie 5.1.0). The Network and hardware section indicates the availability zone (us-east-1a) and subnet (subnet-275cd06a). The Security and access section shows the key name (2bigData) and EC2 instance profile (EMR_EC2_DefaultRole).

DataLake

The screenshot shows the AWS S3 console for a bucket named "bigdata2020damesaa/Datasets/". A list of files is displayed, including "data.csv", "time_series_covid19_confirmed_global_iso3_regions.csv", "time_series_covid19_confirmed_global_narrow.csv", "time_series_covid19_deaths_global.csv", "time_series_covid19_deaths_global_iso3_regions.csv", "time_series_covid19_deaths_global_narrow.csv", "time_series_covid19_recovered_global.csv", "time_series_covid19_recovered_global_iso3_regions.csv", and "time_series_covid19_recovered_global_narrow.csv". A properties dialog box is open for the file "data.csv", showing its key, size (2.4 MB), expiration date, ETag, last modified date, and object URL. The dialog also displays the storage class (Standard), encryption (None), metadata (1), tags (0), and object lock (Disabled) settings.

Notebook en AWS-EMR



Importación de datos desde s3

