Data Cleaning & Transformation

...with PySpark

Timi O.

Objectives

Understand Spark as a Big Data Processor, and PySpark as its Python API.

Understand the concept of data partitioning, and why it is a necessary Data Engineering concept

Understand the proper storage systems to use as it relates to the user(s)' needs and data types.

Get familiar with more cleaning and transformation techniques

Data Cleaning & Transformation

Handling Null Values

Data type Conversion

Handling multi-values in a single cell

Column Names

Null Values

Partitioning & Sorting

Date Format Transformation

Result Representation

10Alytics
data & strategy

CASE STUDY

Case Study - Rheeza Pharmaceuticals

Rheeza Clinics & Pharmaceuticals conducted a research on Naproxen - a drug that has been claimed to normalize the blood pressure of teens and young adults.

The trial was carried out in three of their clinic branches on over 2000 individuals of mixed genders from ages 14 - 22 between February and May 2021.

This trial involved two groups - The in-active (Placebo) and active (Naproxen) groups to test the effect of the actual drug (Naproxen).

Records of all procedures were kept; and extracted from the storage for the ML Engineers to develop algorithms for the next stages of the experiment.

It was discovered that the dataset needed some Engineering to be performed on it for easier access by the Clinicians & ML Engineers.

You have been hired as a Data Engineer specifically for this purpose.

10Alytics
data & strategy

Case Study - Rheeza Pharmaceuticals

Clinicians' Needs

The clinicians will mostly perform analyses on their local spreadsheets based on the conclusion of the results as recorded by the clinician. In addition, they'll want to know:

- The total number of participants per drug per side-effect.
- The total number of participants per medication with Normalized Blood Pressure.

ML Engineers' Needs

The ML Engineers need easily retrievable & analysable records dedicated to them only. Their models will be developed based on each starting month of the experiment.

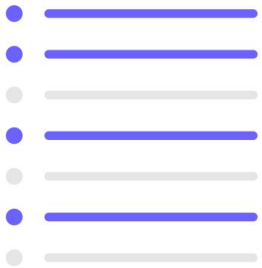
10Alytics
data & strategy

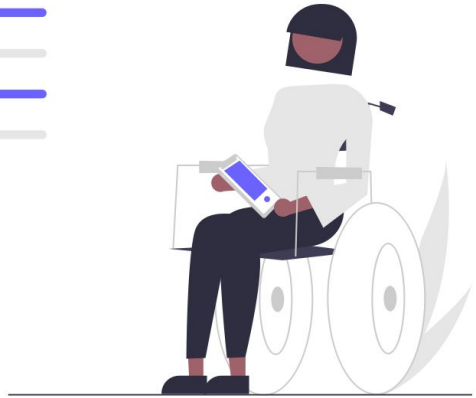
Case Study - Rheeza Pharmaceuticals

Your Task:

1. Perform transformation on this dataset, satisfying the requests of these data users.

2. Deliver the results using appropriate partitions/groupings and file formats as per their requests





10Alytics
data & strategy

Property	Description
ageofparticipant	The age of the participant
clinician.branch	Branch where the experiment was carried out
clinician.name	The name of the head clinician
clinician.role	The role of the assisting clinician
drug_used	The drug administered to the participant (Naproxen, Placebo)
experimentenddate	The date the experiment ended (Unix timestamp)
experimentstartdate	The date the experiment started (Unix timestamp)
noofhourspassedatfirstreaction	First record of reaction (measured in hours from start date)
result.conclusion	Experiment's conclusion as written by clinician
result.sideeffectsonparticipant	Patient's side effect(s) as observed by clinician

Case Study - Schema (JSON File Format)

10Alytics
data & strategy

Happy Engineering