**Glue job training**

Problem: “When we create a Licensed Secondary Database with Full Text (LSDwFT) we must determine how to map the LSD publication titles to MFS title information in order to produce consistent publication data for our customers. Title mapping (matching) is an exercise in determining which titles, represented in two lists, are the same.”

**Building an AWS Glue ETL pipeline locally without an AWS account**

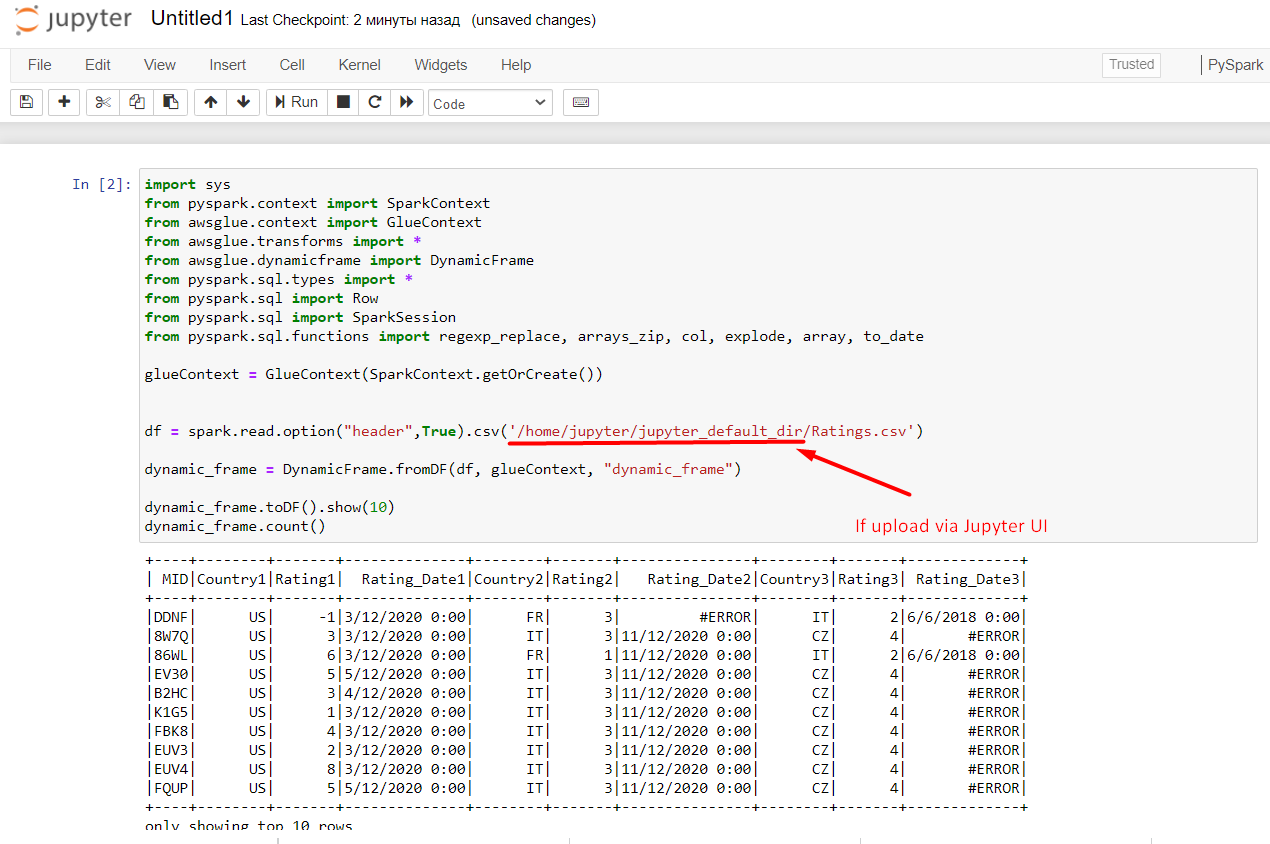
<https://aws.amazon.com/blogs/big-data/building-an-aws-glue-etl-pipeline-locally-without-an-aws-account/>

**Task #1**

Create Glue Job with “Normalization LSD Titles” name to normalized **“**Medline\_Title” field from LSD-titles.csv to make the title consistent with MFS style. Changes include:

* Replace “and“ with “&“
* Remove any leading articles such as “A “ or “The“
* Create a new field called “Normalized\_Title”
* Save all transformations in “Normalized\_Title” field
* “Medline\_Title” field should be kept without changes
* Save normalized dataset on S3 in a folder named “LSD-Titles”

Example:



**Task #2**

Create Glue Job with “Journals creation” name.

* Join LSD-titles.csv and MFS-publist.csv by ISSN
* If ISSN is empty, use “title” field
* Save joined dataset on S3 in a folder named “Journals”

**Task #3**

Create Glue Job with “Report by Country” name to make reports by country.

* Join Journals with Ratings by MID field
* Filter out invalid records
  + if rating is not in the range from 1 to 5
  + if there is an error
  + if rating date is not in range from “Full Text Start Date” to “Full Text End Date” (Journals)
* Accumulate all invalid records in a folder named “Errors”
* Sort records by rating from highest to lowest
* Results shall be saved on S3 and split by country

Expected

* Report\_by\_UK
* Report\_by\_IT
* Report\_by\_FR
* …

**IMPORTANT!!!**

AWS pricing - <https://aws.amazon.com/ru/glue/pricing/>

In order to minimize costs, it is desirable to set 2 workers.

