**Question**

During consumption of large datasets, it’s very common to encounter malformed/invalid lines, however, we don’t want to stop the pipeline every time that a bad line is found, we need to react to this issue by storing the malformed records in a different path without stopping the pipeline. Without implementing any code, explain how you would approach a solution to this problem.

**Answer**

Spark’s solution to encountering malformed lines while processing is to use .option(‘mode’, ‘permissive’) in combination with .option(‘columnNameOfCorruptRecord’, ‘corrupt\_record’). The former function allows you to skip malformed lines and store the value as a string in a new column; the latter function lets you rename that function. As shown above, I would use this approach, renaming the latter column to ‘corrupt\_record’ or something similar.

After the read of the batch data is complete, I would then filter the result for records where df[“malformed\_string”].isNotNull() to locate all malformed records. From here I would save these records to a directory for further manual processing.

We will also then need to filter the df for all records where df[“malformed\_string”].isNull(); these records were properly formed and are immediately suitable for further processing in our pipeline.

**Question**

The resulting tables have to be stored in an appropriate format, for example storing the data in a relational database or CSV. You must describe why you chose the storage format you have chosen (How/why is this format helpful, what are the advantages/disadvantages)

**Answer**

Chosen: CSV

**Advantages:**

Size: At this point the number of records in any of the individual tables is no greater than the length of the number of countries in the world. In other words, we are working with quite small data at this point.

Flexibility: CSV files are plain text files that are human-readable and easily editable with any text editor or spreadsheet software (e.g., Excel).

Simplicity: Using CSV avoids the need for a database server, additional libraries, or special configurations, simplifying the deployment and maintenance of the solution.

**Disadvantages:**

Schema: With CSV files there is no schema enforcement, as would be obtained if we used a relational database or parquet.