### Spark readups and task:

**Use Case: (Spark - big data processing)**

In Qualdo all metrics computation will be running as a spark job in spark execution engines depending on the cloud environment. For aws- EMR, azure- databricks, gcp - dataproc.

**Read-ups:**

Spark architecture

Rdd,jobs,stages,tasks, sparksession, sparkcontext

Understand spark properties and its usage

**Data**:

Get any random tabular data from kaggle around 200mb or greater than that. Data to have numeric columns, string columns.

example dataset:

<https://www.kaggle.com/datasets/ankurnapa/brewery-operations-and-market-analysis-dataset>

Store the selected data in cloud storage.

Pyspark tasks:

Read the dataset as dataframe from cloud storage and do the following :

1. Do computation - std,avg for a integer column
2. Top 5 frequency in any one of the string based column
3. To get the minimum, 25th percentile, median, 75th percentile , and max of a numeric column.
4. Rename a column name
5. To calculate the number of characters in each word in a string column

Separate tasks:

1.df = spark.createDataFrame([ ("A", 1, None), ("B", None, "123" ), ("B", 3, "456"), ("D", None, None), ], ["Name", "Value", "id"])

How to Drop rows with NA values specific to a particular column?

Expected Output:

|Name|Value| id|

+----+-----+----+

| A| 1|null|

| B| null| 123|

| B| 3| 456|

| D| null|null|

