Lab 3: Spark SQL

On Google Colab, install PySpark:

```
! pip install pyspark
```

Let's take an example where we create and use a DataFrame. We will use the Google N-Gram dataset[1]. In this example, the file **ngram.csv** contains data on bigrams with the following columns: **ngram:** String, **Year:** int, **Count:** int, **Pages:** int, **Books:** int.

Each row contains:

- **Ngram:** The bigram itself.
- **Year:** The year in which the bigram appeared.
- **Count:** The number of times the bigram appeared in books for the corresponding year.
- **Pages:** The number of pages on which the bigram appeared in that year (page-count).
- **Books:** The number of distinct books in which the bigram appeared in that year (book-count).

1) Create a DataFrame from the file **ngram.csv.**
2) Register the created DataFrame as a temporary table.
3) Answer the following queries using two methods: (1) SQL language, and (2) SparkSQL API methods.
   3.1) Return all bigrams where the Count is greater than five.
   3.2) Return the total number of bigrams for each year.
   3.3) Return the bigrams with the highest Count for each year.
   3.4) Return all bigrams that appeared in 20 different years.
   3.5) Return all bigrams that contain the character '!' in the first part and the character '9' in the second part (the two parts are separated by a space).
   3.6) Return the bigrams that appeared in all the years present in the dataset.
   3.7) Return the total number of pages and books in which each bigram appeared for each available year, sorted in alphabetical order.
   3.8) Return the total number of distinct bigrams for each year, sorted in descending order of the year.

Execute this lab in two different ways: (1) on Google Colab, (2) in the Spark cluster.

[1] Google NGram Dataset: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html