Лабораторная 2. Формирование отчётов в Apache Spark

Задание

1. Преобразовать любой файл набора данных stackoverflow в parquet формат с помощью Apache Spark.
2. Сформировать отчёт с информацией о частоте обсуждения 10 наиболее популярных языков программирования в каждом году с 2010 года по сегодняшний день. Используйте теги входящие в список языков перечисленных в википедии https://en.wikipedia.org/wiki/List_of_programming_languages.

Исходный набор данных: https://archive.org/details/stackexchange
В архиве с заданиями в папке data доступен пример данных:

- posts_sample.xml

Для выполнения задания используйте Dataset API и/или SQL API.
Рекомендуется отлаживать код на небольшой выборке данных.

Пример кода для получения тестовой выборки:

```
sc.textFile("/user/mapr/posts.xml").mapPartitions(_.take(100))
```

Для парсинга xml строк используйте метод scala.xml.XML.loadString, для получения значений из полученного объекта myxml.attribute('UserId') или myxml.attributes.asAttrMap('UserId').

Ссылки на источники:

1. https://spark.apache.org/docs/latest/sql-programming-guide.html
2. http://timepasstechies.com/spark-dataset-api-examples-tutorial-20/
3. https://jaceklaskowski.gitbooks.io/mastering-spark-sql/
4. https://en.wikipedia.org/wiki/OLAP_cube
5. http://homepage.cs.latrobe.edu.au/zhe/ZhenHeSparkRDDAPIExamples.html