**Подготовка**

1. Установка Apache Spark
2. Установка Apache Maven
3. IDEA Ultimate + Scala (Scala sdk-2.12.8)
4. JDK
5. Winutils (Если Windows)
6. Переменные среды (Если Windows)

HADOOP\_HOME = C:\spark

SPARK\_HOME = C:\spark

В Path добавляем: C:\spark\bin

1. Набор данных (Можно брать с kaggle.com)

**Датасеты**

1. <https://www.kaggle.com/vitaliymalcev/russian-passenger-air-service-20072020>
2. <https://www.kaggle.com/dwdkills/russian-demography>
3. <https://www.kaggle.com/trolukovich/russian-schools-geodata>

**Основы Scala**

<https://docs.scala-lang.org/ru/tour/basics.html>  
<https://twitter.github.io/scala_school/ru/index.html>

Ему требуется виртуальная машина Java (JVM). Тоже объектная модель: любое значение – объект, любая операция – вызов метода. Функциональный язык (функции – это значения).

Средство сборки - Sbt.

При создании проекта надо создать объект Scala (синглтон фактически). Если Scala класс не доступен, то “**Add Framework support’=> (2.12.8). Добавить из maven репозитория библиотеки: spark-core 2.4.3 и spark-sql 2.4.3**

**Разбор примеров на Scala**

1. Первый – подсчет количества слов в текстовом файле
2. Второй – статистика из первой ссылки (статистика по пассажирским авиаперевозкам в России за 2007-2020 гг)

Пример: <https://github.com/pavlentytest/FileReadDemo>

**Задание**

1. Выбрать любой датасет на kaggle.com
2. Cделать 10 выборок данных на ваше усмотрение