Цель работы

Проанализовать набор данных (Частота смерти и средняя продолжительность жизни) с помощью Spark SQL (работа на одной машине с несколькими потоками без полноценной среды Hadoop) и визуализировать данные используя библиотеку JFree.Chart.

Краткая теория

Spark SQL – это модуль Apache Spark, интегрирующий реляционную обработку данных и процедурный API Spark. Spark SQL является частью ядра Spark с версии 1.0. Он может работать совместно с Hive (HiveQL/SQL) или замещать его.

Благодаря Spark SQL, функционал фреймворка получает два ключевых дополнения. Во-первых, модуль обеспечивает тесную интеграцию между реляционной и процедурной обработкой данных посредством интеграции декларативного DataFrame API и процедурного API Spark. Во-вторых, он включает в себя расширяемый оптимизатор, созданный на языке Scala, обладающем широкими возможностями сопоставления с образцом (pattern matching), что позволяет легко формировать правила, управлять генерацией кода и создавать расширения.

Spark SQL и DataFrame

DataFrame – это распределенная коллекция данных, организованных посредством именованных столбцов. Данная абстракция предназначена для выборки, фильтрации, агрегации и визуализации структурированных данных.

DataFrame поддерживает глубокую реляционную/процедурную интеграцию в рамках программ Spark и позволяет манипулировать данными как с помощью процедурного API Spark, так и посредством нового реляционного API, обеспечивающего более эффективную оптимизацию. DataFrame может быть создан непосредственно из RDD, что обеспечивает возможность реляционной обработки уже имеющихся данных.

DataFrame предоставляет более удобные и эффективные средства обработки данных, чем процедурный API Spark. В частности, можно вычислить несколько агрегаций за один проход с мощью SQL-инструкции, что достаточно сложно реализовать посредством традиционного процедурного API.

В отличие от RDD, DataFrame отслеживает свою схему и поддерживает различные реляционные операции, что обеспечивает более оптимизированное выполнение. DataFrame формирует схему посредством отражения (reflection).

DataFrame является «ленивой» структурой данных, то есть содержит логический план для вычисления набора данных, при этом вычисления не выполняются до тех пор, пока пользователь не запросит специальную «операцию вывода», например, сохранение. Такой подход обеспечивает эффективную оптимизацию всех операций.

Концепция DataFrame расширяет модель RDD. В результате, благодаря упрощенным методам фильтрации и агрегации, Spark-разработчики получают возможность быстрее и эффективнее работать с большими наборами структурированных данных.

Ход работы

В процессе работы мы рассмотрим набор данных, состоящий из данных о смертности и средней продолжительности жизни. Прежде всего мы ознакомимся с структурой данной информации. Набор данных из 1044 строк представляет из себя файл с расширением CSV (Comma-Separated Values — значения, разделённые запятыми). Его заголовок содержит следующие поля:

1. Year (Год)
2. Sex (Пол)
3. Race (Раса)
4. Average Life Expectancy (Средняя продолжительность жизни в годах)
5. Age-adjusted Death Rate (Смертей на 100 000)

Набор данных имеет следующую структуру:

Year	Race	Sex	Average	Death_Rate
2015	All Races	Both Sexes	null	733.1
2014	All Races	Both Sexes	78.9	724.6
2013	All Races	Both Sexes	78.8	731.9
2012	All Races	Both Sexes	78.8	732.8
2011	All Races	Both Sexes	78.7	741.3
2010	All Races	Both Sexes	78.7	747.0
2009	All Races	Both Sexes	78.5	749.6
2008	All Races	Both Sexes	78.2	774.9
2007	All Races	Both Sexes	78.1	775.3
2006	All Races	Both Sexes	77.8	791.8

Некоторые данные пропущены. Для чистоты анализа, отфильтруем строки с пропущенными данными. После фильтрации размер датафрэйма снизился до 1035 строк.

Рассчитаем стандартные метрики для очищенного датафрэйма. Столбцы Пол и Раса не берем в рассмотрение, так как там всего 3 возможных варианта.

summary	Average	Death_Rate
count	1035	1035
mean	64.1	1621.3
stddev	11.8	676.4
min	29.1	616.7
max	81.4	3845.7

Данные о смертности не столь интересны, так как имеют не столь высокую дисперсию относительно минимального и максимального значения. В то время, как Средняя продолжительность жизни имеет бОльшую изменчивость.

Согласно теореме о 2/3 сигмах, за пределами интервала 2 сигм от среднего значения находится около 5% выборки. Рассмотрим года, в которых средняя продолжительность жизни ниже чем среднее значение – 2*сигма.

Полученный датафрэйм имеет следующие характеристики:

summary	Average	Death_Rate
count	57	57
mean	35.1	3148.5
stddev	2.9	275.3
min	29.1	2477.7
max	40.5	3845.7

При просмотре получившегося датафрэйма, можно пронаблюдать, что почти все строки принадлежат расе Black.

Race	Count	Percent
Black	53	92.9
All Races	2	3.5
White	2	3.5

В тоже время, распределение по Полу выглядит более равномерным.

Sex	Count	Percent
Male	21	36.8
Both Sexes	20	35
Female	16	28

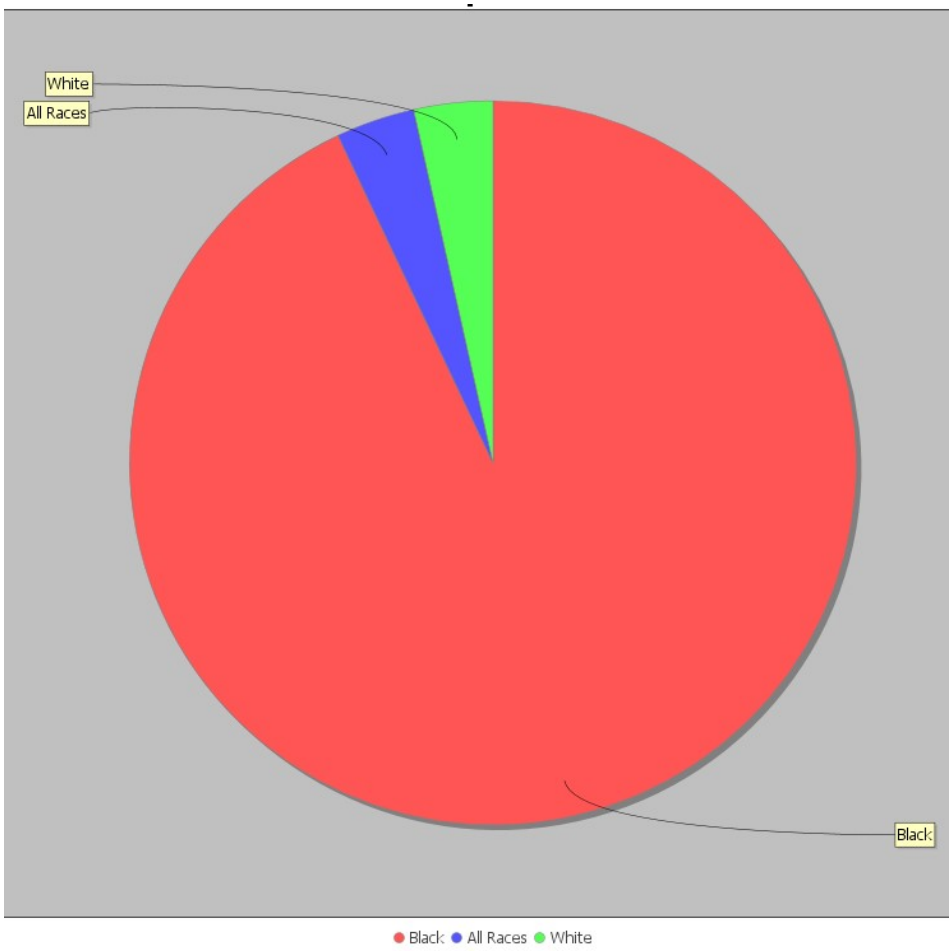

Рис 1. Круговая диаграмма распределения по Расе

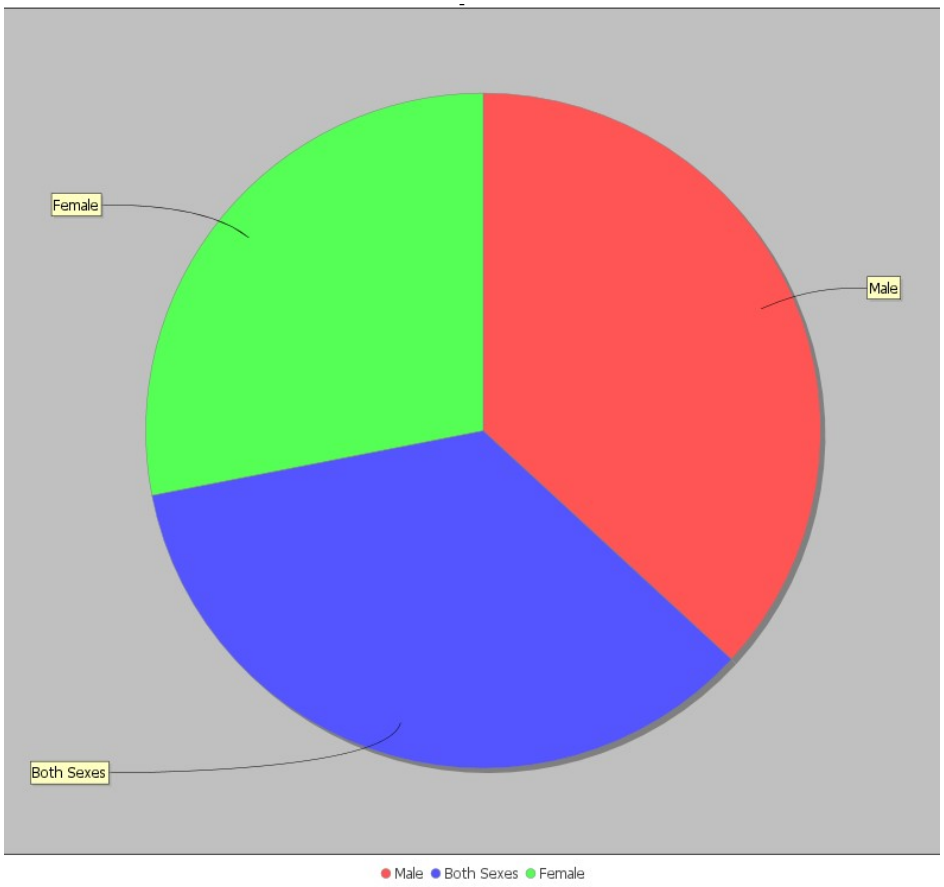

Рис 2. Круговая диаграмма распределения по Полу.

Рассмотрим более подробно распределение по Полу и Расе.

Средняя продолжительность жизни

Race/Sex	Male	Both Sexes	Female
Black	55.8	58.4	60.9
All Races	64	66.5	69.2
White	64.8	67.3	70

Средняя смертность

Race/Sex	Male	Both Sexes	Female
Black	2122	1889	1703
All Races	1689	1484	1310
White	1657	1453	1280

В графическом виде данное распределение выглядит следующим образом.

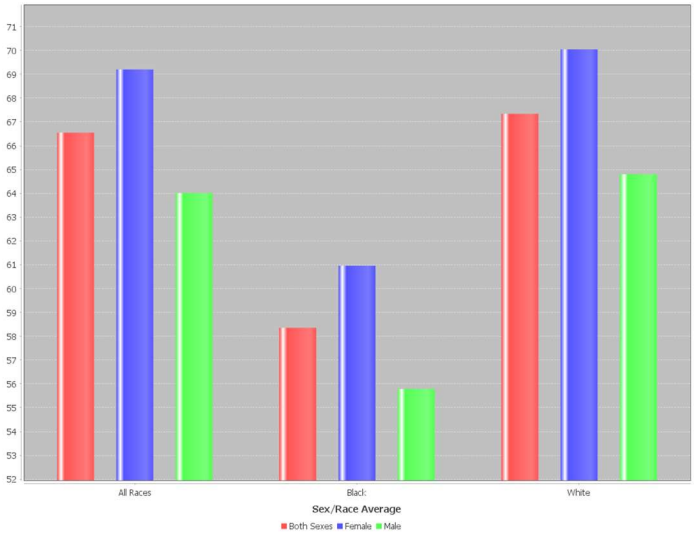

Рис 3. Средняя смертность по Расе и Полу

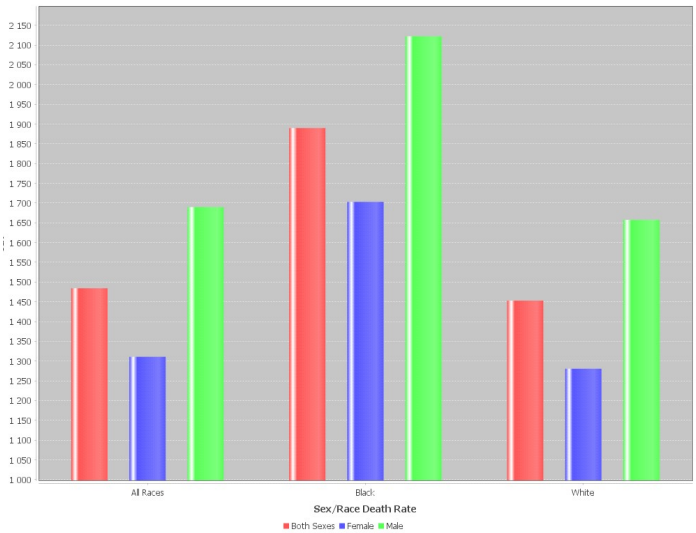

Рис 4. Продолжительность по Расе и Полу

В целом, можно заметить, что в группе Black Male средняя продолжительность жизни гораздо ниже, чем в группе White Female. Так же, можно отметить, что при различных комбинациях признаков, сохраняется зависимость, такая, что у группы Female продолжительность жизни выше группы Male, та же ситуация с группами White и Black. В обоих случаях, группы Both Sexes и All Races оказываются по середине.

Стоит отметить, что в средние показатели группы All Races ближе к показателям группы White. Это может свидетельствовать о том, что в среднем представители расы White преобладают. С группами Male/Both Sexes/Female распределение не смещено, скорее из-за того, что количество мужчин и женщин приблизительно одинаково.

Рассмотрим стандартные отклонения для разных групп.

Race	STD(Average)
All Races	9.8
Black	13.3
White	9.8

Sex	STD(Average)
Both Sexes	11.6
Female	12.3
Male	10.9

Судя по всему, признак Race сильнее влияет на Среднюю продолжительность жизни.

Race	STD(Death_Rate)		Sex	STD(Death_Rate)
All Races	566.2		Both Sexes	662.6
Black	788.5		Female	686.2
White	557.9		Male	623.0

С средней смертностью, ситуация аналогичная.

Рассмотрим среднюю продолжительность жизни и смертность в каждый год.

Рис 5. Средняя продолжительность жизни за год

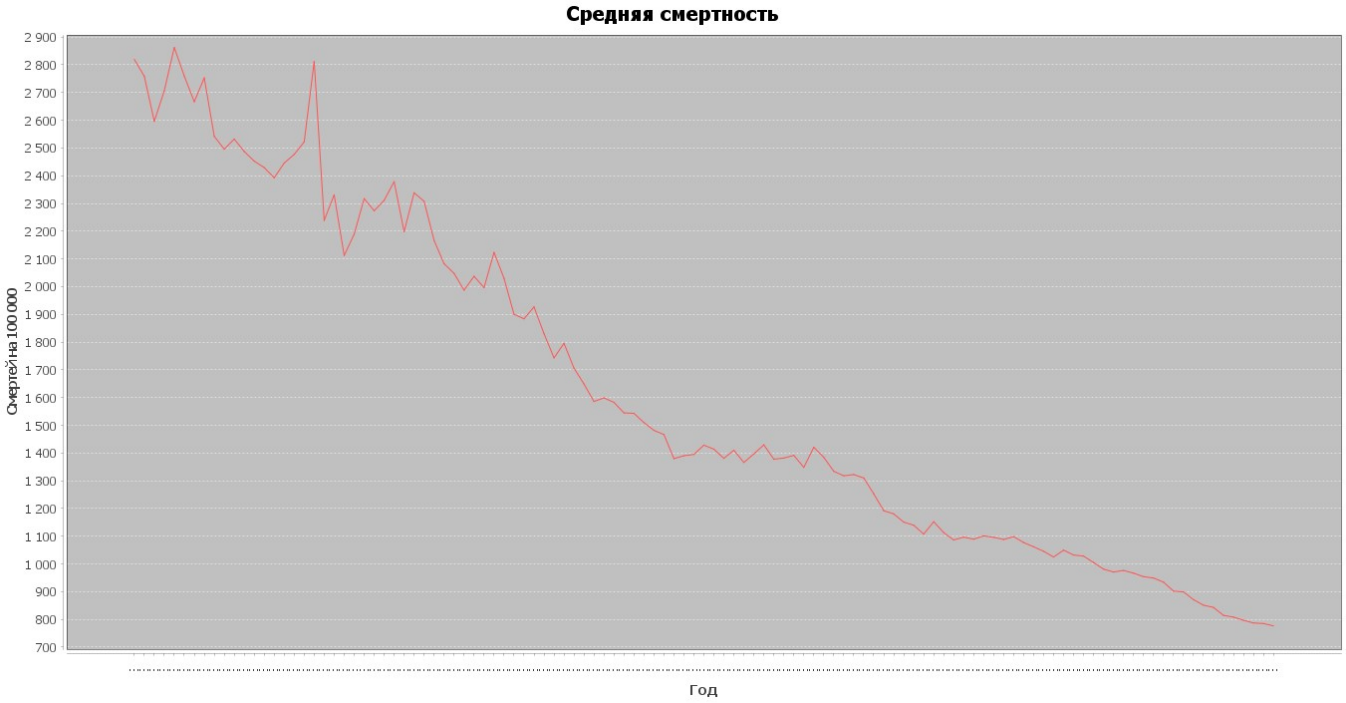

Рис 6. Средняя продолжительность жизни за год

В целом, по графикам видно, что рассматриваемые величины обратно пропорциональны и изменение одной величины отображается на другой. Что в целом, очевидно.

В целом по графику видно ярко выраженный тренд увеличения средней продолжительности жизни

На графике, можно заметить резкий провал в 1917-1918 годах, который можно связать с вступление США в первую мировую войну. Дальнейшие нестабильность в росте можно списать на годы великой депрессии.