Работа с RDD в Apache Spark

1. Подготовка к работе

Подготовьте файл `Работа с RDD - отчёт.docx`. Помещайте в него команды и вставляйте скриншоты выводимых на экран результатов.

Скачайте и распакуйте архив `bigd_course_spark.zip`.

Перейдите в папку `bigd_course_hadoop` и откройте файл `docker-compose.yaml` в простом текстовом редакторе, например, в Блокноте. Изучите конфигурацию кластера. Создайте в отчёте таблицу по образцу ниже и запишите в неё информацию об узлах кластера (за исключением IP адресов). При необходимости найдите описание предназначения узлов в Интернете.

Локальное доменное имя (hostname)	**IP адрес в сети кластера**	**Отображаемые порты**	**Монтируемые тома и файлы хоста**	**Предназначение**
x	x	x	x	x
x	x	x	x	x

Запишите в отчёт ответы на следующие вопросы:

- Сколько в кластере рабочих узлов?
- Сколько вычислительных ядер на каждом рабочем узле?
- Сколько оперативной памяти на каждом рабочем узле?

Проанализируйте монтируемые тома и запишите в отчёт путь к общей («расшареной») папке данных, которая будет доступна всем узлам кластера. Через эту папку узлы будут обмениваться файлами.

Чтобы Ваши файлы стали доступные Spark'у, помещайте их в общую папку данных. Результаты своей работы Spark будет сохранять также в этой папке.

Запустите кластер.

Выведите на экран информацию о структуре сети кластера, найдите IP адреса узлов и поместите их в таблицу выше.

Для корректной работы кластера требуется, чтобы на всех узлах были установлены одинаковые версии Java, Python и самого Spark. Создайте в отчёте таблицу, заходите поочередно в консоли узлов кластера, находит версии и записывайте версии. Так как конфигурации всех рабочих узлов идентичны, запишите версии ПО только для первого узла.

Локальное доменное имя (hostname)	**Версия Spark**	**Версия Java**	**Версия Python**
x	x	x	x
x	x	x	x

Найдите в папке `work` хоста файл `flights.json.gz`. Распакуйте и скопируйте его в общую папку данных кластера.

Откройте в браузере Jupyter из состава кластера (найдите в `docker-compose.yaml` на какой порт хоста он отображается) и откройте в нём файл `work/SparkRDD.ipynb`.

Обратите внимание, что в рабочей папке Jupyter для удобства создана ссылка на общую папку данных кластера и создаваемые в ней файлы можно открывать и просматривать непосредственно в нём.

Ниже приведены задания, которые нужно выполнить. При этом нужно придерживаться следующих ограничений:

- Все вычисления должны выполняться через операции с RDD. DataFrame и DataSet использовать запрещено.
- Результаты должны визуализироваться через matplotlib/seaborn

Возможно при выполнении заданий Вам будут полезны следующие операции RDD:

- `map()`, `filter()`, `reduceByKey()`, `groupBy()`
- `sortBy()`, `take()`, `count()`, `distinct()`
- `sum()`, `mean()`, `min()`, `max()`

В каждом задании имеется вопрос. По результатам выполнения задания на него нужно дать развёрнутый ответ.

Сопоставляйте свои выводы с реальными ожиданиями от авиаперевозок.

Для повышения эффективности, при создании цепочек операций не забывайте кэшировать RDD, для которых выполняются по несколько действий.

2. Задания

2.1 Исследование структуры данных

Вопрос: Какие характеристики рейсов содержатся в dataset и как они распределены? Проанализируйте набор данных и запишите ответ в отчёт.

Что нужно сделать:

- Загрузите данные из JSON-файла в RDD
- Исследуйте структуру данных (поля, типы значений)
- Определите диапазоны значений для числовых характеристик
- Найдите аномалии в данных (отрицательные задержки, нулевые расстояния и т.д.)

2.2 Анализ временных паттернов

Вопрос: Как меняются задержки рейсов в зависимости от времени суток и дня недели?

Что нужно сделать:

- Сгруппируйте рейсы по времени вылета (утро/день/вечер/ночь)
- Рассчитайте средние задержки вылета и прибытия для каждой группы
- Определите, в какие дни недели наблюдается наибольшая вариативность задержек
- Выявите сезонные паттерны (если данные покрывают несколько месяцев)

2.3 Корреляционный анализ

Вопрос: Существует ли зависимость между различными параметрами рейсов?

Что нужно сделать:

- Исследуйте взаимосвязь между задержкой вылета и задержкой прибытия
- Проанализируйте, как расстояние перелета влияет на время в пути
- Определите, существует ли зависимость между временем вылета и величиной задержки
- Рассчитайте коэффициент корреляции между ключевыми параметрами

2.4 Классификация рейсов по "пунктуальности"

Вопрос: Какие рейсы можно считать "пунктуальными", а какие - "проблемными"?

Что нужно сделать:

- Разработайте критерии классификации рейсов (на основе задержек)
- Создайте категории: "досрочные", "пунктуальные", "незначительные задержки", "существенные задержки"
- Рассчитайте распределение рейсов по категориям
- Определите характеристики наиболее "пунктуальных" рейсов

2.5 Пространственный анализ

Вопрос: Как расстояние перелета связано с другими параметрами рейсов?

Что нужно сделать:

- Сгруппируйте рейсы по дистанциям (короткие/средние/длинные)
- Сравните характеристики задержек для разных групп расстояний
- Определите оптимальную скорость для каждой категории расстояний
- Выявите аномалии - рейсы с непропорциональным временем в пути для их расстояния

3. Завершение работы

После того как выполнение задания окончено остановите кластер.

При необходимости удалите тома и сами образы кластера. Это освободит несколько гигабайт свободного места на диске.

ВАЖНО: Если Вы удалите тома, то все результаты работы будут потеряны. Если есть такая возможность, не выполните полную очистку до конца семестра.