Assignment 5 Big Data

Lavandier Théo (s1103617)

> **Question 1:** Which of the materials is (on average) most expensive? (Also write down your added code.)

The most expensive material on average is the Dragon material with an average value of around 70000 GP.

Here is the code that I added to find this result:

```
val avg_price_material = spark.sql("SELECT material, AVG(price) AS avg_price " +
    "FROM sales " +
    "GROUP BY material")


    val query = avg_price_material
      .writeStream
      .outputMode("complete")
      .format("console")
      .start()
```

> **Question 2:** Which of the weapons is (on average) most expensive? (Also write down your added code.)

On average the most expensive weapon is the Halberd with an average price of around 30500 GP.
Here is the code that I added to find this result:

```
val avg_price_weapons = spark.sql("SELECT tpe, AVG(price) AS avg_price " +
    "FROM sales " +
    "GROUP BY tpe")


    val query = avg_price_weapons
      .writeStream
      .outputMode("complete")
      .format("console")
      .start()
```

> **Question 3:** There is one item that costs roughly 9850 GP. Which item is this? (Also write down your added code.)

the item that costs roughly 9850 GP is the Warhammer Adamant.

Here is the code that I added to find this result:

```
val item9850 = spark.sql("SELECT tpe, material, price " +
      "FROM sales " +
      "WHERE price >= 9840 AND price <= 9850 " +
      "GROUP BY tpe, material, price " +
      "HAVING price = 9850")


    val query = item9850
      .writeStream
      .outputMode("complete")
      .format("console")
      .start()
```

> **Question 4:** How much gold is (on average) spent on Mithril swords? (Also write down your added code.) Note that there are multiple types of swords. For the purposes of this assignment, we will consider every item type that contains the string "sword"

The average price of mythril sword is around 6400 GP.

Here is the code that I added to find this result:

```
package org.rubigdata

import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types._
import org.apache.spark.sql.functions._
import org.apache.spark.sql.expressions.scalalang.typed


object RUBigDataApp {
  def main(args: Array[String]) {
    val spark = SparkSession.builder.appName("RUBigDataApp").getOrCreate()
    import spark.implicits._
    spark.sparkContext.setLogLevel("WARN")
    val regex = "^([A-Z].+) ([A-Z].+) was sold for (\\d+)gp$"
    val socketDF = spark.readStream
      .format("socket")
      .option("host", "localhost")
      .option("port", 9999)
      .load()

    val sales = socketDF
      .select(
        regexp_extract($"value", regex, 2) as "tpe",
        regexp_extract($"value", regex, 3).cast(IntegerType) as "price",
        regexp_extract($"value", regex, 1) as "material"
      )
      .as[RuneData]

    sales.createOrReplaceTempView("sales")
    val goldMithrilSword = spark.sql("SELECT material, AVG(price) as avg_price " +
      "FROM sales " +
      "WHERE tpe LIKE '%sword%' AND material = 'Mithril' " +
      "GROUP BY material")


    val query = goldMithrilSword
      .writeStream
      .outputMode("complete")
      .format("console")
      .start()

    query.awaitTermination()
    spark.stop()
  }
}

case class RuneData(tpe: String, price: Int, material: String)
```

(I put the full code to have a save of the code somewhere)

> **Question 5:** Which weapons are, on average, most expensive? One-handed or two-handed weapons? (Also write down your added code.)

On average, two handed weapons are most expensive (23500 GP on average for two handed weapons agains 9600 GP for one handed weapons).

Here is the code that I added to find this result:

```
val weapons = spark.read
      .option("header", "true")
      .option("inferSchema", "true")
      .csv("/opt/hadoop/rubigdata/weapons.csv")
      .as[Weapon]
    weapons.createOrReplaceTempView("weapons")

    val combined = spark.sql("SELECT * FROM sales " +
                             "INNER JOIN weapons " +
                             "ON sales.tpe = weapons.name")
    combined.createOrReplaceTempView("combined")

    val avg_price_num_hands = spark.sql("SELECT num_hands, AVG(price) as avg_price "+
                                        "FROM combined "+
                                        "GROUP BY num_hands")

    val query = avg_price_num_hands
      .writeStream
      .outputMode("complete")
      .format("console")
      .start()
```

> **Question 6:** On which weapons are, in total, most gold pieces spent? One-handed or two-handed weapons? (Also write down your added code.)

In total, most gold are spend on One-handed weapons.

I found this result using the following code:

```
val sum_price_num_hands = spark.sql("SELECT num_hands, SUM(price) as sum_price "+
                                        "FROM combined "+
                                        "GROUP BY num_hands")

    val query = sum_price_num_hands
      .writeStream
      .outputMode("complete")
```

```
        .format("console")
        .start()
```

> **Question 7:** Suppose we have the fixed weapons dataset, and the streaming sales dataset. We want to enrich the weapons dataset by including the total amount of sales for each weapon type. Will the following query work? Briefly explain why/why not.

```
SELECT weapons.*, sales_agg.total_sales
FROM weapons
LEFT OUTER JOIN
 (
 SELECT tpe, SUM(price) AS total_sales
 FROM sales
 GROUP BY tpe
 ) AS sales_agg
ON weapons.name = sales_agg.tpe
```

The `sales` dataset is a streaming dataset, while the `weapons` dataset is a fixed dataset. Streaming datasets and fixed datasets have different characteristics and cannot be joined directly in this manner.

Here we are using `LEFT OUTER JOIN`, our left table is `static` while our right table is a `streaming` dataset. The problem is that this operation is not supported by spark, so it is not going to work.