# 山东大学 计算机科学与技术 学院

## 数据分析 课程实验报告

学号: 202100130222 | 姓名: 贺若桓 | 班级: 数据

实验题目: spark

本实验旨在介绍学习者如何配置和运行 Apache Spark,以及如何使用 Spark 进行简单的数据处理和分析。实验将涵盖以下内容:

- 1) 安装和配置 Apache Spark。
- 2)运行一个简单的 Spark 应用程序,以理解 Spark 的基本概念。 3)使用 Spark 进行数据处理和分析

### 实验环境:

虚拟机, 在VMware上运行Ubuntu 22.04,在上学期大数据管理与分析中配置过Hadoop和Spark的环境 以下是虚拟机配置。



#### 实验步骤与内容:

3、数据集 sales.csv https://pan.baidu.com/s/18eqffdkDTP8cIbZpE8G10w?pwd=xg2p

#### 4、实验步骤

- 1) 安装和配置 Apache Spark
- 2) 运行一个简单的 Spark 应用程序
  - a. 创建一个新的 Spark 应用程序,使用 Python 编写。
  - b. 在应用程序中初始化 SparkSession 对象, 它是 Spark 应用程序的入 口点。
  - c. 加载示例数据集到 Spark 中。
  - d. 对数据进行一些简单的转换操作,如筛选、映射或聚合。 e. 运行 Spark 应用程序并查看结果。

```
from pyspark.sql import SparkSession

# 初始化 SparkSession
spark = SparkSession.builder.appName("SimpleSparkApp").getOrCreate()

# 加穀数据
data = spark.read.csv("sales_data.csv", header=True, inferSchema=True)

# 执行一些数据处理操作
result = data.filter(data["product_category"] == "Electronics").groupBy("date").sum("revenue")

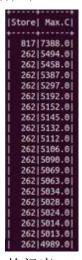
# 显示结果
result.show()
```

- 5、相关Spark程序
  - 1) 根据文档运行 word count:

your: 1
your: 8
yes,: 1
yes,: 1
yes,: 1
yes,: 1
yes,: 1
yes,: 1
your: 2
york: 2
york: 2
york: 2
york: 2
york: 2
york: 1
york: 1
with: 11
winds: 1
york: 1
y

2) 使用 Spark 求工作日销量最大的门店 ID:

sparksal('selectstore,custonersvgasNfrowglobaltepppeoplexhereDayofNeek>landDayofkeek<-5orderbyNdesc')show()



3) 使用 Spark 求工作日销量最大的门店 ID,每个门店出现一次 :

```
|Store| Max.C|
   817 | 7388.0 |
   262 | 5494.0 |
  1114 | 4911.0 |
   586 | 4762.0 |
   733 | 4645.0 |
   251 | 4635.0 |
   769 | 4582.0 |
   756 | 4180.0 |
   562 | 4099.0 |
   336 | 3961.0 |
   335|3929.0|
  1097 | 3804.0 |
   259 | 3648.0 |
   320 | 3507.0 |
   580 | 3360.0 |
   467 | 3351.0 |
   513 | 3302.0 |
   827 | 3241.0 |
   983 | 3200.0 |
   470 | 3178.0 |
```

3) 使用 Spark 求平均销量最大的门店 ID,每个门店出现一次 :

```
datasets(Row> = spark.sql('select Store, avg(Customers+0) as Afron
Store | Avg.C|
  733|3403.5|
  262 | 3402.0 |
  562|3105.1|
769|3081.1|
 1114 2664.1
 817 2605.5
1097 2420.9
  335 | 2385.3 |
  259 | 2347.1 |
  251|2026.5|
586|1945.3|
  756 | 1940.5 |
  423 | 1886.1 |
  383|1826.2|
682|1758.8|
  513 | 1744.7 |
  320|1717.1|
948|1687.0|
  676 1644.6
  336 | 1623.2 |
```