Scala でも データ分析したい！！

2024-05-12

Affiliation: JAIST Ph.D. Student

Name: ADACHI Yuya

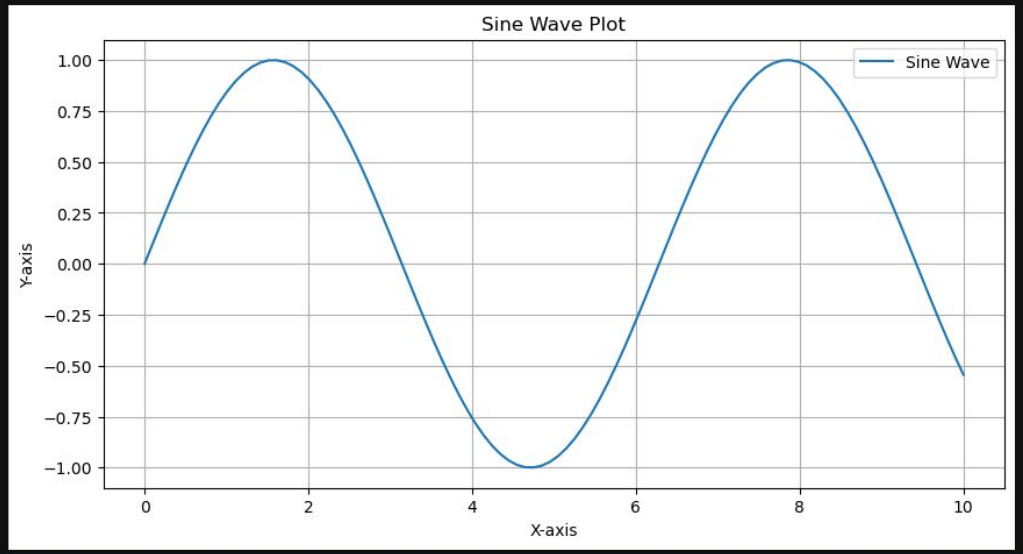
E-mail: s2120001@jaist.ac.jp

はじめに

- 現状、データ分析のエコシステムはPython が強い
- スタンダードなPython エコシステム
  - pandas
  - NumPy
  - Matplotlib
  - scikit-learn
  - JupyterLab
- R や Julia も頑張っているけれどPython が強すぎる

JupyterLab 実行例

[1]:
```python
import numpy as np
import matplotlib.pyplot as plt

# データの生成
x = np.linspace(0, 10, 100)  # 0から10までの100ポイント
y = np.sin(x)                # 正弦波

# プロットの作成
plt.figure(figsize=(10, 5))
plt.plot(x, y, label='Sine Wave')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.title('Sine Wave Plot')
plt.legend()
plt.grid(True)
plt.show()
```



Scala でも データ分析したい！！

- 一番、簡単なのはDatabricksを使う方法
- Databricksは、Apache Sparkの創設者であるMatei Zahariaと他の共同創業者によって設立されました。ビッグデータ解析および機械学習のための統合プラットフォームで、Apache Sparkを基盤としています
- 特徴
  - スケーラビリティ: 大規模データ処理に対する優れたスケーラビリティ
  - コラボレーション: ノートブックを通じたデータサイエンティストやデータエンジニア間の協力が容易
  - 自動化 : ワークフローの自動化とジョブスケジューリング機能

Databricks 実行例 (1)

Databricks 実行例 (2)

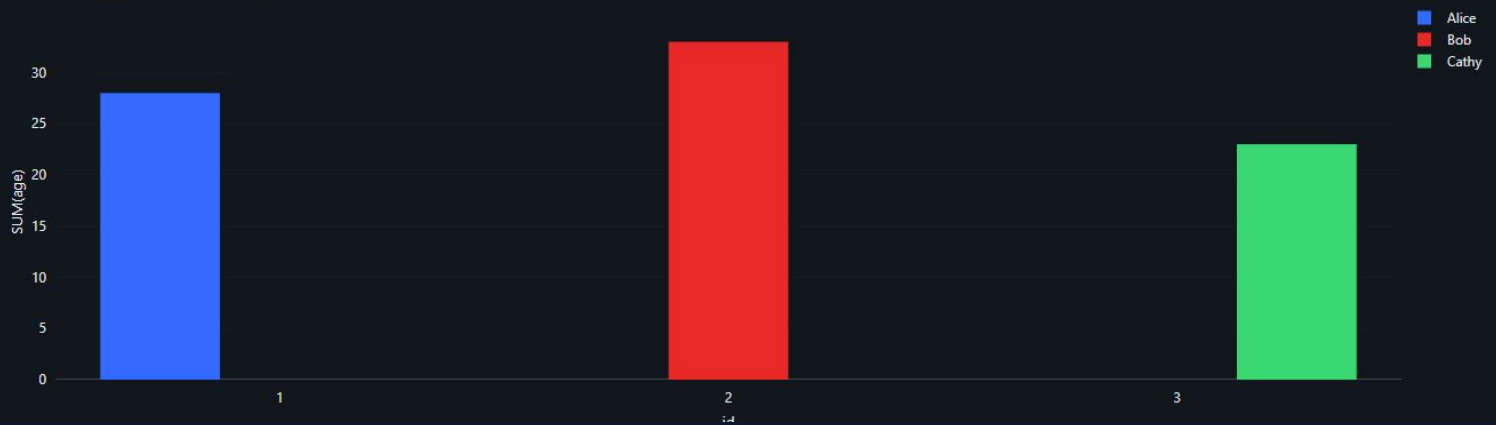
07:53 (2秒)

Scala

```scala
// サンプルデータを定義
val data = Seq(
  (1, "Alice", 28),
  (2, "Bob", 33),
  (3, "Cathy", 23)
)

// スキーマを定義
val schema = Seq("id", "name", "age")

// データフレームを作成
val df = spark.createDataFrame(data).toDF(schema: _*)

// データフレームの内容を表示
display(df)
```

df: org.apache.spark.sql.DataFrame = [id: integer, name: string ... 他に1件のフィールドあり]

テーブル　可視化 1　+



Databricks 不満点

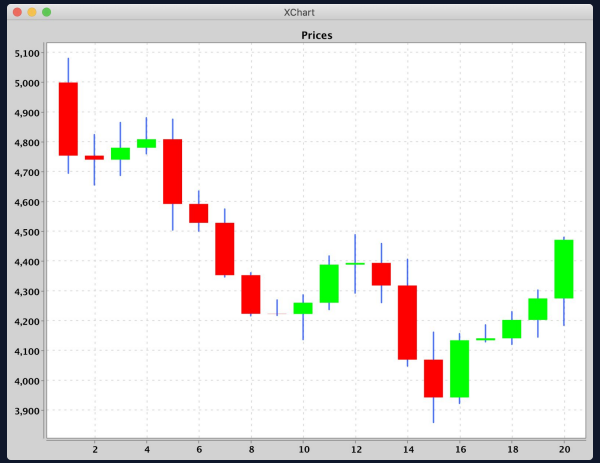
- Community Edition だと 1 〜 2 時間放置するとクラスターが停止する
- クラスターを起動するのに数分かかる(地味にストレス)
- インターネット環境が必須
- ローカルのちょっとしたデータを分析するには過剰

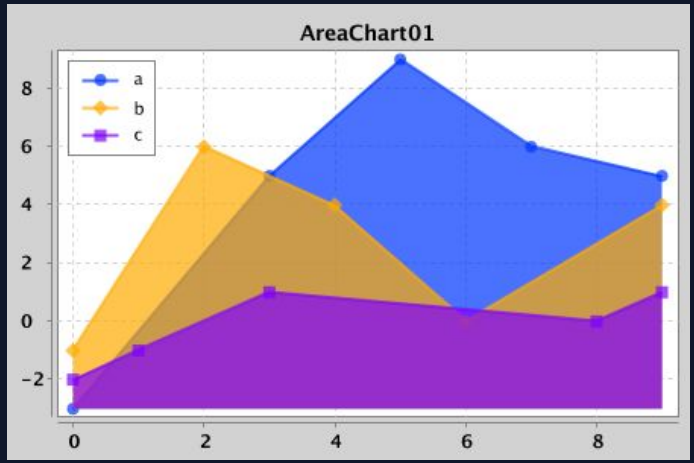
Almond + Apache Spark + XChart 構成

- ライブラリをかき集めて魔改造構成も一応考えられる..
- pandas, NumPy, scikit-learn → Apache Spark
  - https://spark.apache.org/
- Matplotlib → XChart
  - https://knowm.org/open-source/xchart/
- JupyterLab → Almond で Scala に対応させる
  - https://almond.sh/

Almond + Apache Spark + XChart 実行例

```
//> using scala 2.13
//> using dep "org.apache.spark::spark-core:3.5.1"
//> using dep "org.knowm.xchart:xchart:3.8.8"

import org.apache.spark.*
import org.apache.spark.sql.*
import org.knowm.xchart.{CategoryChartBuilder, BitmapEncoder}
import org.knowm.xchart.style.Styler

object SparkXChartExample {
  def main(args: Array[String]): Unit = {
    // Sparkセッションの作成
    val spark = SparkSession.builder
      .appName("Spark XChart Example")
      .master("local[*]")
      .getOrCreate()

    import spark.implicits._

    // サンプルデータの作成
    val data = Seq(
      ("Category A", 10),
      ("Category B", 15),
      ("Category C", 8),
      ("Category D", 20)
    ).toDF("Category", "Value")

  }
}
```





Python の方が幸せになれる！！