Pyspark の初歩

アジェンダ

- やること、やらないこと
- Pysparkとは?
- 使い方は?

やること

- DataFrameをSparkSQLからごにょごにょするー!
- 雰囲気を味わう!

やらないこと

- 難しいこと全部!
- sparkとHadoopとの違い

Pysparkとは?

簡単に言えばApache SparkをPythonで使うよーって事!
Apache Spark(https://spark.apache.org/documentation.html)
現在はversion3.xまで出ている

そのため、Pysparkはなんなのかということは、sparkを理解すれば
OK！

Apache Sparkとは?(wiki君より)

Apache Sparkとは**オープンソース**のクラスタコンピューティングフレームワークである。暗黙のデータ並列性と対象外性を備えたクラスタ全体をプログラミングできる(wiki参照)

…

……

wikiから見たのがいけないんだ!公式を見てみよう!

Apache Sparkとは?(こーしき)

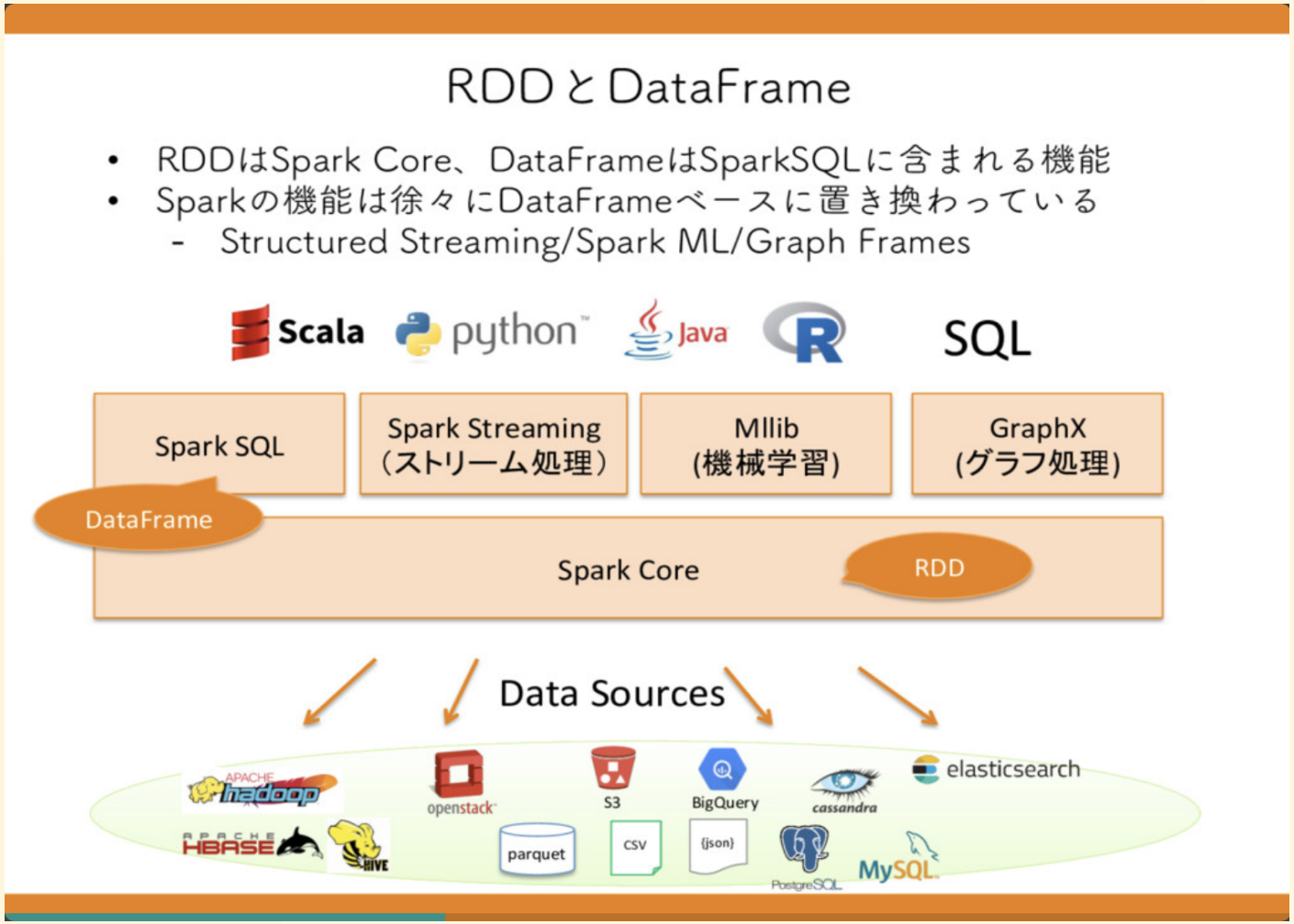
大規模なデータ処理のための統合分析エンジン。java, scale, PythonおよびRのAPIを提供!
また、SQLおよび構造化データ処理のためのSparkSQL,機械学習のためのMLlib、グラフ処理のためのGraphX、逐次計算およびストリーム処理のための構造化ストリーミングを含む高レベルの充実ツールがあるよ!

.....


…あまり難しい言葉を遣うなよ
話についていけない

Apache Sparkとは?

- 大量のデータを高速に処理したいときに使うよ
- サーバを跨いだ並列処理が行えるよ
- sparkは以下のプログラム言語から使用できるよ
  - java, Scala, Python, R
- sparkは以下の機能が提供されてるよ
  - 構造化データ(csvとか)をSQLみたいに処理できるSpark SQL
  - 機械学習を行いたい人向けに、アルゴリズムが用意されているMLlib
  - グラフを扱って計算とかしたい人向けGraphX
  - ストリームデータをリアルタイム処理したい人向けSpark Streaming



引用(https://speakerdeck.com/chie8842/pythondeda-liang-detachu-li-pysparkwoyong-itadetachu-li-tofen-xi-falsekihon?slide=14)

RDDとDataFrame

RDD

横に伸びていくデータ構造
[(値1, 値1), (値2, 値2), (値3, 値3)...]
RDDのまま処理を行うには、lambda関数,mapなどを駆使する

DataFrame

RDDを使いやすくするために、テーブル構造にラッパーしてあげた
列を付与した形(そのため、RDDから変換時に列側の概念設定は必要)
.〇〇でORマッパーみたいにデータを扱える

spark機能を使うために

spark.〇〇 の"spark"部分の作り方ですねー、大体の記事でspark使うなら当たり前やろ?的な感じで省かれてます...(公式すら省いてます)

- SparkContext()
  - sparkの機能との入り口
  - 引数は沢山あるけど、主にmasterとappNameが使われる
    - master: 接続先のクラスター指定
    - appName: ジョブの名前
  - 内部でhttpサーバとしての機能が起動するので、最後にstop()で終了してあげる必要がある(http://mogile.web.fc2.com/spark/spark200/monitoring.html)

```
from pyspark.context import SparkContext

sc = SparkContext('master', 'appName')
sc.stop()
```

- SparkSession
  - dataframeを利用するにはSparkSessionが必要になる
  - SparkContextの後続の書き方
  - SparkSessionの内部にはSparkContext, SparkConfがいる

つまりSparkSessionを定義すればSparkContextはいらんよってこと
基本はSparkSessionでいいと思われ

```
from pyspark.sql.session import SparkSession

spark = SparkSession.builder.master('master').appName('appName').getOrCreate()
spark.stop()
```

また、1jvm上で起動できるSparkSession, Contextは一つだけ
https://spark.apache.org/docs/2.3.0/api/java/org/apache/spark/SparkContext.html

では、DataFrameを定義して、SparkSQLでいじってみよう

データセット紹介

kaggleからポケモン画像データセット
(https://www.kaggle.com/vishalsubbiah/pokemon-images-and-types)
についてきたcsvを使おうと思います

ポケモン名, type1, type2の３列のみとなっています

使い方

環境構築はいろいろとめんどそうだった(jdkいれて、spark入れてパス通してうんぬんかんぬん)ので、docker-hubで公開されている**jupyter公式**のイメージを拝借したいと思います

url(https://hub.docker.com/r/jupyter/pyspark-notebook)

```
$ docker pull jupyter/pyspark-notebook
$ docker run -itdp 8888:8888 jupyter/pyspark-notebook
```

Jupyterとは

データ分析、研究機構当でよく利用されています(知りませんでいた...)
ブラウザ上でコードを実行できたり、ドキュメントを作成できたり便利そう!

~画面共有で操作~

count(sql詳しくなくて...type1と2合わせたcount取りたい!)

```
from pyspark.sql.types import *
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession

sc = SparkContext('local')
spark = SparkSession(sc)
# spark = SparkSession.builder.master("local")でも同じ!

struct = StructType([
  StructField('name', StringType(), False),
  StructField('type1', StringType(), False),
  StructField('type2', StringType(), True),
  # StructField(カラム名, 型, nullの可否)
])

df = spark.read.csv('pokemon.csv', schema=structm, header=True)  # 元から定義したstructを指定, headerを読み込まない
df.show(5)
df.groupBy('type1').count().sort('count', ascending=False).show(5)
#  グループして       数え上げて、並び替えて    昇順、降順         表示している
# ここからSpark SQL
df.registerTempTable('pokemon')  # dfをpokemonってテーブルとして認識
spark.sql('select * from pokemon').show(5)
spark.sql('select type1, count(*) as count from pokemon group by type1 order by count desc').show(5)
```

join、出力

```
# さっきの続きで
df2 = spark.read.csv('pokemon.csv', schema=structm, header=True)
df2.show(5)
df2.registerTempTable('pokemon2')
joined_df = spark.sql('select p.name, p2.name as name2 from pokemon as p join pokemon2 as p2 on p.name = p2.name')
joined_df.write.csv("joined_pokemon.csv")
```

jupyterのホーム画面にjoined_pokemon.csvができている

参考資料

apache spark(https://spark.apache.org/documentation.html)

pyspark(https://spark.apache.org/docs/latest/api/python/index.html)

Apache Sparkの初心者が環境構築とPySparkでのデータ集計までやってみる(https://qiita.com/mkyz08/items/0c1d8fa47179933c3a56)

ps:土日全く資料作らずなにしてたの?

Among usという、人狼ににたゲーム