Start coding or generate with AI.

Spark

Spark is an open-source, distributed, unified analytics engine used for real-time data processing and acts as a faster cluster computing framework. Spark is popular due to its in-memory computation power, which increases the data processing speed and makes it capable to handle huge amounts of data.

Apache spark is an advanced version of Hadoop because Hadoop is a framework that uses map-reduce for the processing, which reads the data from disk and forms key-value pair so if we read data from disk, process it, and write it again to disk so it is very time-consuming and spark does all things in main memory means data store in RAM was to compute time reduces and operations happen very fast.

Spark is built on Scala, an advanced Java version that runs on JVM. Spark provides high-level APIs through which we can code and use spark in any language, including Java, Python, Scala, R, etc. And working with spark through Python is known as Pyspark

What is Spark MLLIB?

MLLIB stands for Machine learning library in Spark. This library aims to make practical machine learning scalable and easy to implement. It provides tools to implement all machine learning algorithms, including Regression, classification, dimensionality reduction tools, transformation, feature extraction, pipelines (tunning), save and load algorithm, and utilities for linear algebra and statistics.

When we talk about spark MLLIB so, it has a dataframe-based API, and as of spark 2 onwards, the RDD-based API entered the maintenance phase, and the primary ML API is now a dataframe-based API that is a spark. ml.0

Spark MLLIB Tools

Spark provides a different set of machine learning tools to perform different tasks and take different actions.

Machine Learning algorithms – It provides tools and techniques to implement Regression, Classification, clustering, and collaborative filtering.

Featurization – Tools for feature extraction, transformation, dimensionality reduction, and feature selection.

Pipelines – tools for constructing, evaluating, and tuning ML pipelines.

Persistence – save and load algorithms, models, and pipelines.

Utilities – Linear algebra, statistics, data handling

Spark MLLIB Data Types

Spark supports different data types. Spark MLLIB supports local vectors and Matrices stored on a single machine and distributed matrices. So it supports many data types packed with one or many RDDs.

Local Vector – MLLIB supports two types of local vectors, which are dense and sparse. A labeled point is a local vector, either dense or sparse, that is associated with a label or response. For example, in binary classification, the label should be either 0 (negative) or 1 (positive).

Local matrix – It has integer type row, column indices, and double type values stored in a single machine.

Distributed matrix – It has long-type row and column indices and double-type values. It is stored in a distributed manner in one or more RDD.

Machine learning Pipelines

When we talk about ML pipelines, it is all about understanding different stages, including estimator, evaluator, transformer, etc. Machine learning pipelines provide uniform high-level APIs built on top of data frames. It is used to create and tune practical machine learning pipelines. It is mainly used with structured data.

Dataframe – A dataframe from spark SQL is used as a machine learning dataset.

It holds a variety of data types text, feature vectors, labels, etc.

Transformer – A transformer is an algorithm that transforms one algorithm into another dataframe.

Estimator – An estimator is an algorithm that can be fit on a dataframe to produce a transformer.

Pipeline – A pipeline integrate multiple transformers and estimators to specify a machine learning workflow.

Evaluator – It will evaluate the outcome of the model.

Running Pyspark on Jupyter Notebook and Google Colab

We now know about spark and why today it is used by each organization to process their data. To get hands-on practical knowledge about spark let up first install and set up complete spark on our system. First, we are installing Pyspark on the Jupyter notebook.

- Step-1) Install Anaconda – If you have Anaconda, then good else, then download and install it using the official link.

Step-2) Install JDK 1.8.0 – As we studied, spark works on top of JDK and JVM, you need to have JDK installed in your system if you want to work with a spark on a local Jupyter notebook.

Step-3) Download Spark – From the official spark website we have to download a tar file of spark and extract it.

step-4) set environment variable – You have to set a Path and environment variable as shown below, and you will be good to go with spark.

```
[ ] ! pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.10/dist-packages (3.5.0)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
```

```
[ ] from pyspark import SparkContext, SparkConf
    conf = SparkConf().setAppName("pyspark_practice").setMaster("local")
    sc = SparkContext(conf=conf)
```

```
[ ] from pyspark.sql import SparkSession
    spark = SparkSession.builder.master("yarn").appName("MyApp").getOrCreate()
```

```
[ ] names = sc.parallelize(['Shubham','rishi','prayag','shivam','rahul','Madhav','Nihal',
    print(type(names))
```

```
<class 'pyspark.rdd.RDD'>
```

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ] #Reading any csv file
    csv_file = spark.read.csv('/content/drive/My Drive/Colab Notebooks/Dataset/spam.csv',
```

```
[ ] csv_file
```

```
DataFrame[Category: string, Message: string]
```

RDD Actions

RDD stands for Resilient Distributed Dataset. In PySpark, RDD is the fundamental data structure representing an immutable distributed collection of objects. RDDs are fault-tolerant, allowing for distributed processing and parallel execution across a cluster. They support various transformations (map, filter, etc.)

Actions are used to execute the scheduled task on the dataset because when we apply the transformation, It only creates a DAG and when we act then tasks task tasks tasked display an output. We will study some popular actions used on the dataset.

1. Collect

This is the first action that will display all values right away. It has created a list. If you will perform the transformation, then nothing will be displayed.

```
[ ] csv_file.collect()
```

```
[Row(Category='ham', Message='Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...'),
 Row(Category='ham', Message='Ok lar... Joking wif u oni...'),
 Row(Category='spam', Message="Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply
```

```
08452810075over18's"),
 Row(Category='ham', Message='U dun say so early hor... U c already then say...'),
 Row(Category='ham', Message="Nah I don't think he goes to usf, he lives around here though"),
 Row(Category='spam', Message="FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv"),
 Row(Category='ham', Message='Even my brother is not like to speak with me. They treat me like aids patent.'),
 Row(Category='ham', Message="As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune"),
 Row(Category='spam', Message='WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.'),
 Row(Category='spam', Message='Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030'),
 Row(Category='ham', Message="I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today."),
 Row(Category='spam', Message='SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info'),
 Row(Category='spam', Message='URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18'),
 Row(Category='ham', Message="I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times."),
 Row(Category='ham', Message='I HAVE A DATE ON SUNDAY WITH WILL!!'),
 Row(Category='spam', Message='XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL'),
 Row(Category='ham', Message="Oh k...i'm watching here:)"),
 Row(Category='ham', Message='Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.'),
 Row(Category='ham', Message='Fine if that\x92s the way u feel. That\x92s the way its gota b'),
 Row(Category='spam', Message='England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ú1.20 POBOXox36504W45WQ 16+'),
 Row(Category='ham', Message='Is that seriously how you spell his name?'),
 Row(Category='ham', Message='I‘m going to try for 2 months ha ha only joking'),
 Row(Category='ham', Message='So ü pay first lar... Then when is da stock comin...'),
 Row(Category='ham', Message='Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?'),
 Row(Category='ham', Message='Ffffffffff. Alright no way I can meet up with you sooner?'),
 Row(Category='ham', Message="Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows I'm sick when I turn down pizza. Lol"),
 Row(Category='ham', Message='Lol your always so convincing.'),
 Row(Category='ham', Message="Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?"),
 Row(Category='ham', Message="I'm back &amp;amp; we're packing the car now, I'll let you know if there's room"),
 Row(Category='ham', Message='Ahhh. Work. I vaguely remember that! What does it feel like? Lol'),
 Row(Category='ham', Message="Wait that's still not all that clear, were you not sure about me being sarcastic or that that's why x doesn't want to live with us"),
 Row(Category='ham', Message="Yeah he got in at 2 and was v apologetic. n had fallen out and she was actin like spoilt child and he got caught up in that. Till 2! But we won't go there! Not doing too badly cheers. You?"),
 Row(Category='ham', Message='K tell me anything about you.'),
 Row(Category='ham', Message='For fear of fainting with the of all that housework you just did? Quick have a cuppa'),
 Row(Category='spam', Message='Thanks for your subscription to Ringtone UK your mobile will be charged £5/month Please confirm by replying YES or NO. If you reply NO you will not be charged'),
 Row(Category='ham', Message='Yup... Ok i go home look at the timings then i msg ü again... Xuhui going to learn on 2nd may too but her lesson is at 8am'),
 Row(Category='ham', Message="Oops, I'll let you know when my roommate's done"),
 Row(Category='ham', Message='I see the letter B on my car'),
 Row(Category='ham', Message='Anything lor... U decide...'),
 Row(Category='ham', Message="Hello! How's you and how did saturday go? I was just texting to see if you'd decided to do anything tomo. Not that i'm trying to invite myself or anything!"),
 Row(Category='ham', Message='Pls go ahead with watts. I just wanted to be sure. Do have a great weekend. Abiola'),
 Row(Category='ham', Message='Did I forget to tell you ? I want you , I need you, I crave you ... But most of all ... I love you my sweet Arabian steed ... Mmmmmm ... Yummy'),
 Row(Category='spam', Message='07732584351 - Rodger Burns - MSG = We tried to call you re your reply to our sms for a free nokia mobile + free camcorder. Please call
```

2. count By Value

If you want a count of a particular value in data, then you can use this action. The alternative to this function you can also use a simple count function which is also one action.

```
[ ] csv_file.printSchema()
```

```
root
 |-- Category: string (nullable = true)
 |-- Message: string (nullable = true)
```

DataFrame Manipulations

How to See DataType of Columns?

To see the types of columns in DataFrame, we can use the printSchema, dtypes. Let’s apply printSchema() on train which will Print the schema in a tree format.

```
[ ] csv_file.printSchema()
```

```
root
 |-- Category: string (nullable = true)
 |-- Message: string (nullable = true)
```

```
[ ] Start coding or generate with AI.
```

How to Show First n Observations?

```
csv_file.head(5)
```

```
[Row(Category='ham', Message='Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...'),
 Row(Category='ham', Message='Ok lar... Joking wif u oni...'),
 Row(Category='spam', Message="Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply
08452810075over18's"),
 Row(Category='ham', Message='U dun say so early hor... U c already then say...'),
 Row(Category='ham', Message="Nah I don't think he goes to usf, he lives around here though")]
```

How to Count the Number of Rows in DataFrame?

We can use count operation to count the number of rows in DataFrame. Let's apply count operation on train & test files to count the number of rows.

```
csv_file.count()
```

```
5574
```

```
dataframes = csv_file.randomSplit([0.6, 0.2, 0.2], seed=26)
```

```
dataframes[0].count()
```

```
3349
```

```
dataframes[1].count()
```

```
1096
```

```
dataframes[2].count()
```

```
1129
```

```
csv_file.crosstab('Category','Message').show()
```

```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.

Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```

how to create Dataframe using pyspark

```
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('Pyspark data frames to pandas').getOrCreate()
data = [("James","","Smith","36636","M",60000),("Michael","Rose","","40288","M",70000
columns = ["first_name","middle_name","last_name","dob","gender","salary"]
pysparkDF = spark.createDataFrame(data = data, schema = columns)
pysparkDF.printSchema()
pysparkDF.show(truncate=False)
```

```
root
 |-- first_name: string (nullable = true)
 |-- middle_name: string (nullable = true)
 |-- last_name: string (nullable = true)
 |-- dob: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- salary: long (nullable = true)

+----------+-----------+---------+-----+------+------+
|first_name|middle_name|last_name|dob  |gender|salary|
+----------+-----------+---------+-----+------+------+
|James     |           |Smith    |36636|M     |60000 |
|Michael   |Rose       |         |40288|M     |70000 |
+----------+-----------+---------+-----+------+------+
```

```
pandasDF = pysparkDF.toPandas()
print(pandasDF)
```

```
  first_name middle_name last_name    dob gender  salary
0      James                Smith  36636      M   60000
1    Michael        Rose           40288      M   70000
```

How to Get the DataFrame Which Don't Have Duplicate Rows of a Given DataFrame?

We can use dropDuplicates operation to drop the duplicate rows of a DataFrame and get the DataFrame which won't have duplicate rows. To demonstrate that I am performing this on two columns Age and Gender of train and get the all unique rows for these columns.