Advanced Institute for Artificial Intelligence

RDDs e DataFRames

Advanced Institute for Artificial Intelligence – AI2

https://advancedinstitute.ai

Background

“Devemos acreditar que somos talentosos para algumas coisas, e que essa coisa, a qualquer custo, deve ser alcançada”.

“ *Marie Curie*

Referências

- RDD Programming Guide (Link) ou
- Learning Apache Spark with Python (.pdf)
- Spark with Python (PySpark) Tutorial

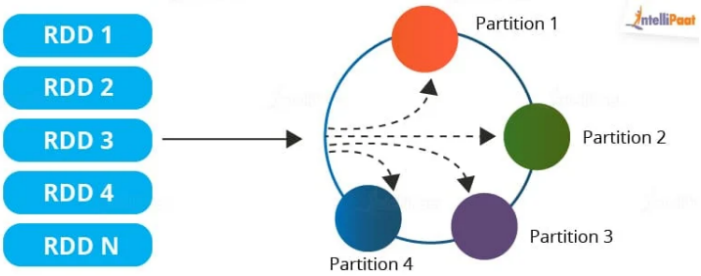
RDD - Resilient Dristributed Datasets (Conjunto de Dados Resilientes e Distribuídos)

Apresentação

RDD - Resilient Distributed Dataset

- Trata-se de um bloco de construção fundamental para o **PySpark**. Uma coleção de objetos imutáveis e tolerantes a falhas ou seja, uma vez que você cria um RDD, você não pode alterá-lo.
- Cada registro no RDD é dividido em partições lógicas, que podem ser calculadas em diferentes nós do *cluster*.
- Em outras palavras, podemos dizer que se trata de uma coleção de objetos similares a uma lista em Python.
- Ou ainda... como uma tabela do banco de dados que pode guardar qualquer tipo de dado.

RDD


RDD 1
RDD 2
RDD 3
RDD 4
RDD N
Partition 1
Partition 2
Partition 3
Partition 4
IntelliPaat

Recursos de um RDD no Spark

- Resiliência:
  - RDDs rastreiam informações de linhagem de dados para recuperar dados perdidos, automaticamente em caso de falha. Também é chamado de tolerância a falhas.
- Distribuído:
  - Os dados presentes em um RDD residem em vários nós. Ele é distribuído em diferentes nós de um cluster.
- Avaliação Lenta:
  - Os dados não são carregados em um RDD, mesmo que você os defina. As transformações são realmente calculadas quando você chama a ação, como contar ou coletar, ou salva a saída em um sistema de arquivos.

Desvantagens de um RDD no Spark

- Estrutura de baixo nível;
- Complexo e verboso, ou seja, para executarmos as tarefas, será necessário digitar muito código;
- Otimização difícil pelo Spark (Lembram-se do plano de execução do Spark?)

- Dataset e Dataframe
  - Datasets disponíveis em Java e Scala
  - Não disponíveis em R e Python
  - Então, iremos ver RDD porém, voltado ao Dataframe.

SEMELHANTE A UMA TABELA DE BANCO DE DADOS

COMPATÍVEL COM OBJETOS DATAFRAME DO R E PYTHON

DataFrames

- **Já conhecemos as características de um Dataframe:**
  - Tabelas com linhas e colunas
  - Imutáveis
  - Com schema conhecido
  - Spark preserva as etapas que os dados foram alterados
  - Possui análises comuns, como Agrupar, ordenar, filtrar
  - Spark pode otimizar estas analises através de planos de execução

Schema

- **Estamos nos referindo aos metadados da tabela:**
  - Nome das tabelas e o tipo dos dados
  - O Spark pode inferir a partir de parte dos dados, ou
  - Nós mesmos podemos definir o schema
- **Vantagens em Definirmos o tipo dos dados:**
  - **Tipo Correto:** Certeza que o Spark irá importar o tipo correto de dados
  - **Sem Overhead:** O Spark analizando os dados para buscar (inferir) o tipo dos dados das colunas.

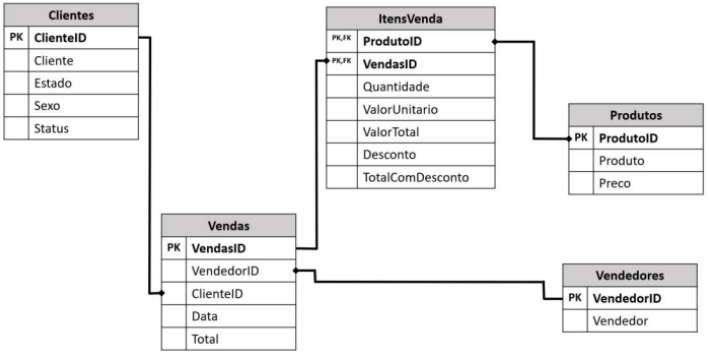
- Tipos de Dados suportados pelo DataFrame

Tipo
ByteType
ShortType
IntegerType
LongType
FloatType
DoubleType
DecimalType
StringType
BinaryType
BooleanType
TimestampType
DateType
ArrayType
MapType
StructType
StructField

Nosso primeiro DataFrame no Pyspark

- Criando um DataFrame sem as informações das colunas:

Baseado no Diagrama abaixo, resolva os exercícios:



Temos as seguintes tabelas:

- **Clientes:**
  - 250 amostras;
  - 3 status: Silver, Gold e Platinum

ClienteID	Cliente	Estado	Sexo	Status
1	Adelina Buenaventura	RJ	M	Silver
2	Adelino Gago	RJ	M	Silver
3	Adolfo Patrício	PE	M	Silver
4	Adriana Guedelha	RO	F	Platinum
5	Adélio Lisboa	SE	M	Silver
6	Adérito Bahia	MA	M	Silver
7	Aida Dorneles	RN	F	Silver
8	Alarico Quinterno	AC	M	Silver
9	Alberto Cezimbra	AM	M	Silver
10	Alberto Monsanto	RN	M	Gold
11	Albino Canela	AC	M	Silver
12	Alceste Varanda	RR	F	Silver
13	Alcides Carvalhais	RO	M	Silver

Exercícios

Temos as seguintes tabelas:

- **Vendedores:**
  - 10 amostras;

VendedorID	Vendedor
1	Armando Lago
2	Capitolino Bahía
3	Daniel Pirajá
4	Godo Capiperibe
5	Hélio Liberato
6	Iberê Lacerda
7	Jéssica Castelão
8	Napoleão Méndez
9	Simão Rivero
10	Tobias Furtado

Temos as seguintes tabelas:

- **Produtos:**
  - 10 amostras;

ProdutoID	Produto	Preco
1	Bicicleta Aro 29 Mountain Bike Endorphine 6.3 - 24 Marchas - Shimano - Alumínio	R$ 8.852,00
2	Bicicleta Altools Stroll Aro 26 Freio À Disco 21 Marchas	R$ 9.201,00
3	Bicicleta Gts Advanced 1.0 Aro 29 Freio Disco Câmbio Traseiro Shimano 24 Marchas	R$ 4.255,00
4	Bicicleta Trinc Câmbios Shimano Aro 29 Freio A Disco 24v	R$ 7.658,00
5	Bicicleta Gometws Endorphine 7.3 - Shimano Alumínio Aro 29 - 24 Marchas	R$ 2.966,00
6	Bicicleta Gometws Endorphine 6.1 Shimano Alumínio- Aro 26 - 21 Marchas	R$ 2.955,00
7	Capacete Gometws Endorphine 2.0	R$ 155,00
8	Luva De Ciclismo - Gometws Sports	R$ 188,00
9	Bermuda Predactor 3Xu Pro	R$ 115,00
10	Camiseta Predactor 3Xu Multiplied	R$ 135,00

Temos as seguintes tabelas:

- **Vendas:**
  - 400 amostras;

VendasID	VendedorID	ClienteID	Data	Total
1	1	91	01/01/2016	R$ 8.053,60
2	6	185	01/01/2016	R$ 150,40
3	7	31	02/01/2016	R$ 6.087,00
4	5	15	02/01/2016	R$ 13.828,60
5	8	120	03/01/2016	R$ 26.096,66
6	9	74	04/01/2016	R$ 18.402,00
7	9	191	06/01/2016	R$ 7.524,20
8	6	186	06/01/2016	R$ 12.036,60
9	7	91	06/01/2016	R$ 2.804,75
10	2	202	06/01/2016	R$ 8.852,00
11	1	58	08/01/2016	R$ 16.545,25
12	7	32	09/01/2016	R$ 11.411,88
13	4	136	10/01/2016	R$ 15.829,70
14	3	249	12/01/2016	R$ 6.154,36
15	4	249	12/01/2016	R$ 3.255,08

Temos as seguintes tabelas:

- **Itens Vendas:**
  - 940 amostras;

ProdutoID	VendasID	Quantidade	ValorUnitario	ValorTotal	Desconto	TotalComDesconto
2	400	2	R$ 9.201,00	R$ 18.402,00	R$ 6.256,68	R$ 12.145,32
2	385	2	R$ 9.201,00	R$ 18.402,00	R$ 5.704,62	R$ 12.697,38
4	395	2	R$ 6.892,20	R$ 13.784,40	R$ 5.100,23	R$ 8.684,17
4	367	2	R$ 6.509,30	R$ 13.018,60	R$ 4.816,88	R$ 8.201,72
2	380	2	R$ 7.038,77	R$ 14.077,54	R$ 4.364,04	R$ 9.713,50
2	346	2	R$ 8.280,90	R$ 16.561,80	R$ 4.140,45	R$ 12.421,35
2	339	2	R$ 8.280,90	R$ 16.561,80	R$ 3.312,36	R$ 13.249,44
2	397	1	R$ 9.201,00	R$ 9.201,00	R$ 3.312,36	R$ 5.888,64
1	346	2	R$ 7.966,80	R$ 15.933,60	R$ 3.186,72	R$ 12.746,88
2	264	2	R$ 8.280,90	R$ 16.561,80	R$ 3.146,74	R$ 13.415,06
4	355	2	R$ 5.858,37	R$ 11.716,74	R$ 3.046,35	R$ 8.670,39
2	376	1	R$ 8.280,90	R$ 8.280,90	R$ 2.981,12	R$ 5.299,78
2	374	1	R$ 9.201,00	R$ 9.201,00	R$ 2.944,32	R$ 6.256,68
1	397	1	R$ 7.524,20	R$ 7.524,20	R$ 2.859,20	R$ 4.665,00
2	303	2	R$ 8.280,90	R$ 16.561,80	R$ 2.815,51	R$ 13.746,29
4	358	2	R$ 6.509,30	R$ 13.018,60	R$ 2.733,91	R$ 10.284,69
4	374	1	R$ 7.658,00	R$ 7.658,00	R$ 2.527,14	R$ 5.130,86
3	336	2	R$ 4.255,00	R$ 8.510,00	R$ 2.467,90	R$ 6.042,10
1	399	1	R$ 6.771,78	R$ 6.771,78	R$ 2.437,84	R$ 4.333,94
1	292	2	R$ 6.771,78	R$ 13.543,56	R$ 2.437,84	R$ 11.105,72
4	382	1	R$ 6.892,20	R$ 6.892,20	R$ 2.412,27	R$ 4.479,93
1	379	1	R$ 7.524,20	R$ 7.524,20	R$ 2.407,74	R$ 5.116,46
1	377	1	R$ 8.852,00	R$ 8.852,00	R$ 2.390,04	R$ 6.461,96
1	389	1	R$ 8.852,00	R$ 8.852,00	R$ 2.390,04	R$ 6.461,96

Exercícios

Baseado nos DataFrames apresentados e disponibilizados em nosso diretório no Git, resolva os seguintes exercícios:

- Github Ai2