Universidade Federal do Rio Grande do Norte
Departamento de Engenharia de Computação e Automação
DCA0132 – Engenharia de Dados

# Análise de dados do Enem 2015


Evandro Carlos B. dos Santos
Giuliard Cosmo Rodrigues

# Introdução

O dataset escolhido para realização do trabalho foi extraído de um arquivo CSV do Enem 2015 que foi disponibilizado no Kaggle pelo usuário Gustavo Bonesso. Os dados fornecem informações detalhadas dos alunos dos estados do centro-oeste (Mato Grosso do Sul, Mato Grosso, Goiás e Distrito Federal).

# Colunas do arquivo

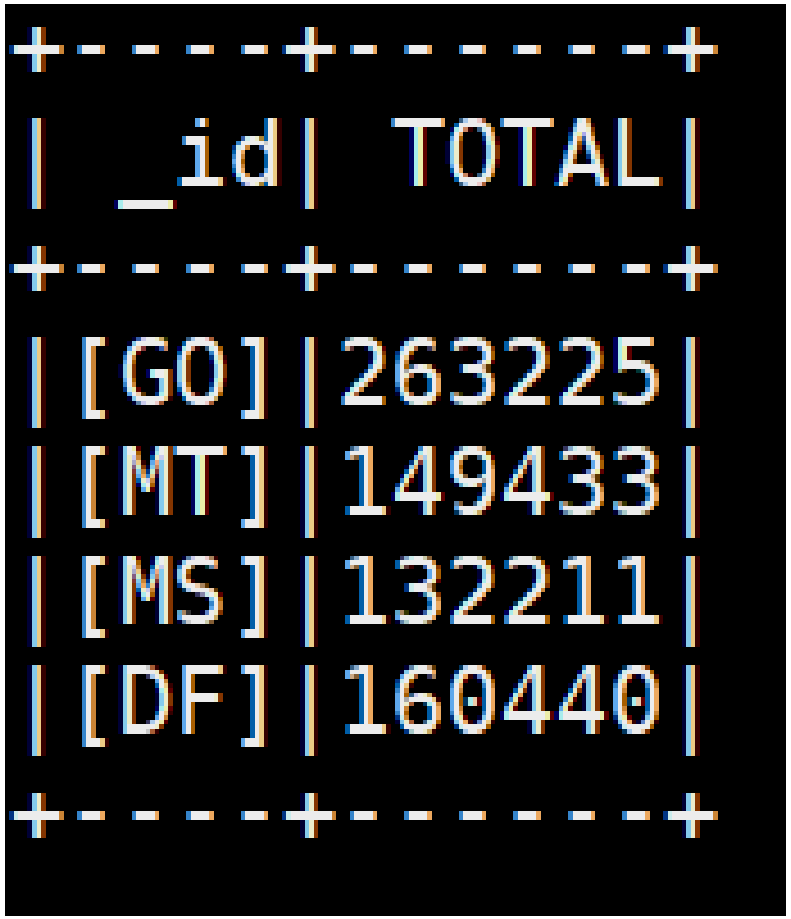
- NU_INSCRICAO
- NU_ANO
- CO_MUNICIPIO_RESIDENCIA
- CO_UF_RESIDENCIA
- SG_UF_RESIDENCIA
- IN_ESTUDA_CLASSE_HOSPITALAR
- IN_TREINEIRO
- CO_ESCOLA
- CO_MUNICIPIO_ESC
- CO_UF_ESC
- SG_UF_ESC
- TP_DEPENDENCIA_ADM_ESC
- TP_LOCALIZACAO_ESC
- TP_SIT_FUNC_ESC
- NU_IDADE
- TP_SEXO
- TP_NACIONALIDADE
- CO_MUNICIPIO_NASCIMENTO
- CO_UF_NASCIMENTO
- SG_UF_NASCIMENTO
- TP_ST_CONCLUSAO
- TP_ANO_CONCLUIU
- TP_ESCOLA
- TP_ENSINO
- TP_ESTADO_CIVIL
- TP_COR_RACA
- IN_MATERIAL_ESPECFICO
- IN_CERTIFICADO
- CO_UF_ENTIDADE_CERTIFICACAO
- SG_UF_ENTIDADE_CERTIFICACAO
- TP_PRESENCA_CN
- TP_PRESENCA_CH
- TP_PRESENCA_LC
- TP_PRESENCA_MT
- CO_PROVA_CN
- CO_PROVA_CH
- CO_PROVA_LC
- CO_PROVA_MT
- NU_NOTA_CN
- NU_NOTA_CH
- NU_NOTA_LC
- NU_NOTA_MT
- TP_LINGUA
- TP_STATUS_REDACAO
- NU_NOTA_COMP1
- NU_NOTA_COMP2
- NU_NOTA_COMP3
- NU_NOTA_COMP4
- NU_NOTA_COMP5
- NU_NOTA_REDACAO

# Temas Abordados

- Porcentagem de homens e mulheres que participaram do exame;
- Número de candidatos por estado;
- Ídade média dos candidatos;
- Nota média em cada área de conhecimento.

# Infraestrutura





| _id | TOTAL |
|---|---|
| [GO] | 263225 |
| [MT] | 149433 |
| [MS] | 132211 |
| [DF] | 160440 |

# Quantidade de Candidatos por Estado

# Quantidade de Candidatos por Estado

```
pipeline = [{"$group":{"_id":{"UF_RESIDENCIA":"$SG_UF_RESIDENCIA"}, "TOTAL":{"$sum":1}}}, {"$project":{"_id":1, "TOTAL":1}}]
```

| _id | IDADE_MEDIA |
|---|---|
| [F, DF] | 24.038835613149796 |
| [M, MS] | 24.14781008743687 |
| [M, MT] | 23.38069145107351 |
| [F, MT] | 23.666085507879497 |
| [F, MS] | 24.270836610125034 |
| [M, GO] | 21.972230520940236 |
| [F, GO] | 22.07721803652557 |
| [M, DF] | 23.489305807923067 |

# Idade Média dos Candidatos

# Idade Média dos Candidatos

```
pipeline = [{"$group":{"_id":{"UF_RESIDENCIA":"$SG_UF_RESIDENCIA" , "GENERO":"$TP_SEXO"},
                       "IDADE_MEDIA":{"$avg":"$NU_IDADE"}}}, {"$project":{"_id":1,"GENERO":1,"IDADE_MEDIA":1}}]
dfPipe = MYspark.read.format("com.mongodb.spark.sql.DefaultSource").option("pipeline", pipeline).load()

exprs = {'IDADE_MEDIA': 'avg'}
dfPipe = dfPipe.groupBy(["_id"]) \
    .agg(exprs) \
    .withColumnRenamed('AVG(IDADE_MEDIA)', 'IDADE_MEDIA')
dfPipe.write.format("com.mongodb.spark.sql.DefaultSource").mode("overwrite").save()
```

| _id | LC | CH | MT | CN | REDACAO |
|---|---|---|---|---|---|
| [F, DF] | 514.0974133478798 | 554.1584395529582 | 454.6103644656808 | 473.5002001694721 | 378.752666632818 |
| [M, MS] | 493.11169066327176 | 559.1500646856923 | 470.41744673836166 | 488.9521281101505 | 344.59764815686935 |
| [M, MT] | 485.33867455071834 | 553.2200269807312 | 461.46761071442995 | 482.98349233207125 | 337.56444826537785 |
| [F, MT] | 490.4037509497715 | 538.2799850639278 | 433.9210136273181 | 461.12064751704145 | 356.98937891399794 |
| [F, MS] | 496.5372266230984 | 543.2101637527442 | 442.7917741133718 | 465.7869115510136 | 361.3079135410155 |
| [M, GO] | 498.93714424780626 | 563.0412151688374 | 479.37938042828983 | 492.73045325161775 | 369.7298706778379 |
| [F, GO] | 503.2218304169285 | 547.2186549034527 | 449.6230556511772 | 467.88248061024643 | 384.5515764250762 |
| [M, DF] | 514.5852129179766 | 574.2663439826722 | 490.8597877050309 | 500.9345681831577 | 367.83226905753753 |

# Nota Média

# Nota Média

```
pipeline = [{"$group":{"_id":{"UF_RESIDENCIA":"$SG_UF_RESIDENCIA","GENERO":"$TP_SEXO"},
                                "CN":{"$avg":"$NU_NOTA_CN"},"CH":{"$avg":"$NU_NOTA_CH"},
                                "LC":{"$avg":"$NU_NOTA_LC"},"MT":{"$avg":"$NU_NOTA_MT"},
                                "REDACAO":{"$avg":"$NU_NOTA_REDACAO"}}},
            {"$project":{"_id":1,"GENERO":1,"CN":1, "CH":1, "LC":1, "MT":1, "REDACAO":1}}]

# 1 -Read e Aggregation no Pipeline
start_time = time.time()

dfPipe = MYspark.read.format("com.mongodb.spark.sql.DefaultSource").option("pipeline", pipeline).load()

dfPipe = dfPipe.groupBy(["_id"]) \
    .agg({'CN':'avg','CH':'avg','LC':'avg','MT':'avg','REDACAO':'avg'}).\
    withColumnRenamed('AVG(CN)', 'CN').withColumnRenamed('AVG(CH)', 'CH').withColumnRenamed('AVG(LC)','LC').\
    withColumnRenamed('AVG(MT)', 'MT').withColumnRenamed('AVG(REDACAO)', 'REDACAO')
dfPipe.write.format("com.mongodb.spark.sql.DefaultSource").mode("overwrite").save()
```

# Quantidade Homens x Mulheres

| _id | TOTAL |
|---|---|
| [F, DF] | 92324 |
| [M, MS] | 58021 |
| [M, MT] | 62793 |
| [F, MT] | 86640 |
| [F, MS] | 74190 |
| [M, GO] | 113117 |
| [F, GO] | 150108 |
| [M, DF] | 68116 |

- DF:
  - Homens: 42,5%
  - Mulheres: 57,5%
- GO:
  - Homens: 42,97%
  - Mulheres: 57,03%
- MS:
  - Homens: 43,9%
  - Mulheres:56,1%
- MT:
  - Homens: 40,03%
  - Mulheres: 57,97%

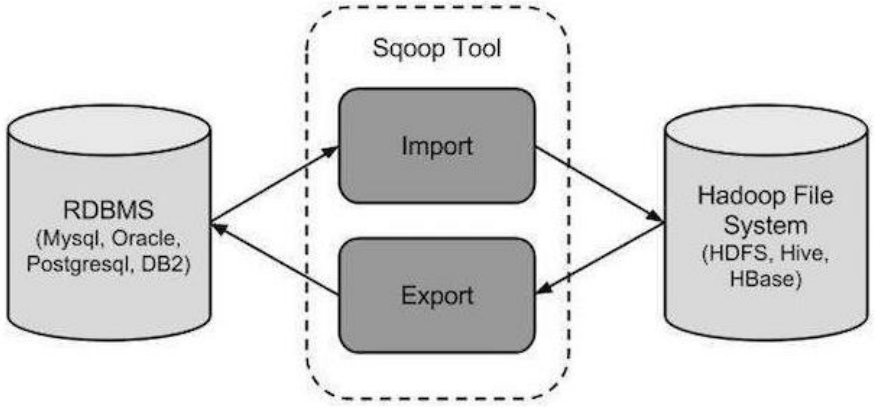
# Candidatos Homem x Mulher

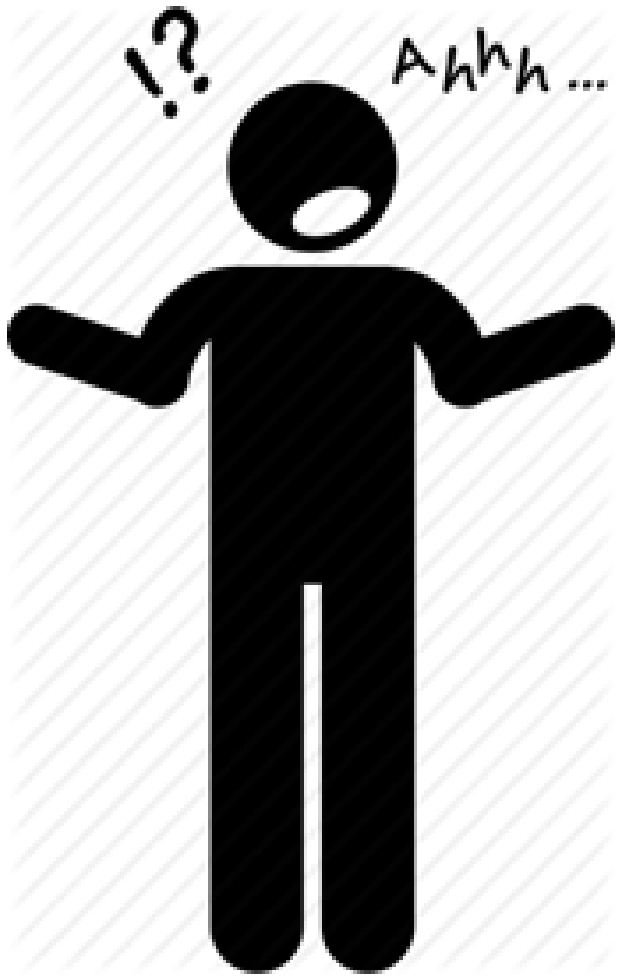
```
MYspark = SparkSession \
    .builder \
    .appName("Media_Pipeline_App") \
    .config("spark.mongodb.input.uri", "mongodb://10.7.40.136/enem2015.candidatos") \
    .config("spark.mongodb.output.uri", "mongodb://10.7.40.136/enem2015.candidatos_estado") \
    .getOrCreate()

pipeline = [{"$group":{"_id":{"UF_RESIDENCIA":"$SG_UF_RESIDENCIA" , "GENERO":"$TP_SEXO"}, "TOTAL":{"$sum":1}}}, {"$project":{"_id":1,"GENERO":1,"TOTAL":1}}]
```

```
dfPipe = MYspark.read.format("com.mongodb.spark.sql.DefaultSource").option("pipeline", pipeline).load()

exprs = {'TOTAL': 'sum'}
dfPipe = dfPipe.groupBy(["_id"]) \
    .agg(exprs) \
    .withColumnRenamed('SUM(TOTAL)', 'TOTAL')
dfPipe.write.format("com.mongodb.spark.sql.DefaultSource").mode("overwrite").save()
dfPipe.show()
```

# Apache Sqoop

- Desde 2012 é um dos projetos top-level da Apache Software Foundation.
- Abreviação de “SQL para Hadoop”
- Tem como objetivo executar a transferência eficiente e bidirecional de dados entre o Hadoop e diversos serviços de armazenamento externo de dados estruturados.






# Dificuldades

Conteúdo de instalção e configuração do Apache Sqoop

Desenvolver os primeiros scripts PySpark + Mongo