Configuração do Spark

In [4]:

```python
import findspark
findspark.init()

import pyspark
import random

sc = pyspark.SparkContext(appName="Pi")
num_samples = 100000000

def inside(p):
  x, y = random.random(), random.random()
  return x*x + y*y < 1

count = sc.parallelize(range(0, num_samples)).filter(inside).count()

pi = 4 * count / num_samples
print(pi)

sc.stop()
```

```
3.14131664
```

Carregando as bases

In [5]:

```python
from pyspark.sql import SparkSession
from pyspark import SparkConf, SparkContext
from pyspark.sql.types import *
from pyspark.sql.functions import *

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

#df = spark.read.json("/files/passei/BASE-A/courses.json")

# - Usuários que acessaram
dfStudents = spark.read.format("json").load("/files/passei/BASE-A/students.json")

# - Visitas que os usuários fizeram ao longo do mês
dfSessions = spark.read.format("json").load("/files/passei/BASE-A/sessions.json")

# - Assinaturas dos usuários que aderiram ao Plano Premium
dfSubscriptions = spark.read.format("json").load("/files/passei/BASE-A/subscriptions.j

# - Lista de Universidades cadastradas
dfUniversities = spark.read.format("json").load("/files/passei/BASE-A/universities.js

# - Lista de Cursos cadastrados
dfCourses = spark.read.format("json").load("/files/passei/BASE-A/courses.json")

# - subjects.json - Lista de Disciplinas
dfSubjects = spark.read.format("json").load("/files/passei/BASE-A/subjects.json")

# - Disciplinas que cada usuário está seguindo
dfStudentFollow_subject = spark.read.format("json").load("/files/passei/BASE-A/student
```

Conhecendo as bases

In [6]:

```python
# - Usuários que acessaram
dfStudents.printSchema()
dfStudents.show()
```

```
root
 |-- City: string (nullable = true)
 |-- CourseId: long (nullable = true)
 |-- Id: string (nullable = true)
 |-- RegisteredDate: string (nullable = true)
 |-- SignupSource: string (nullable = true)
 |-- State: string (nullable = true)
 |-- StudentClient: string (nullable = true)
 |-- UniversityId: long (nullable = true)

+----------------+--------+--------------------+--------------------+------------+----
------------+-------------+------------+
|            City|CourseId|                  Id|      RegisteredDate|SignupSource|
       State|StudentClient|UniversityId|
+----------------+--------+--------------------+--------------------+------------+----
------------+-------------+------------+
|  Rio de Janeiro| 1199555|0cade9bf00234e378...|2012-09-06 17:27:...|    Facebook|  Ri
o de Janeiro|         null|      664704|
|  Rio de Janeiro| 1199521|8a501cab6c0a5a7e9...|2012-09-05 15:31:...|    Facebook|  Ri
o de Janeiro|         null|      664704|
|            null| 1199517|b8a39150d98d74685...|2012-08-17 10:00:...|    Facebook|
        null|         null|      661625|
|  Rio de Janeiro| 1199491|0f6c90f966a70b84b...|2012-10-22 21:16:...|    Facebook|  Ri
o de Janeiro|         null|      664768|
|            null| 1199573|388bf84cffb321938...|2012-10-24 01:48:...|    Facebook|
Minas Gerais|         null|      663106|
|   Florianópolis| 1199521|5764c7b7c75f30474...|2012-11-07 17:23:...|    Facebook|  Sa
nta Catarina|         null|      664138|
|        Mesquita| 1199517|4522694e98b51ba30...|2012-08-21 15:13:...|    Facebook|  Ri
o de Janeiro|         null|      664742|
|        Londrina| 1199553|0360730b79c8ecd04...|2012-11-22 23:51:...|    Facebook|
      Paraná|         null|      663609|
|  Rio de Janeiro| 1199491|668514e190e792018...|2012-10-22 21:16:...|       Email|  Ri
o de Janeiro|         null|      664768|
|      Ouro Preto| 1199536|7c743c3365d9b9696...|2012-10-24 16:28:...|    Facebook|
Minas Gerais|         null|      663054|
|          Viçosa| 1199725|09e1497357395b825...|2012-10-15 21:34:...|    Facebook|
Minas Gerais|         null|      663065|
|            null| 1199453|29037b0a52c5b576d...|2012-10-30 17:02:...|       Email|
Minas Gerais|         null|      664565|
|   Volta Redonda| 1199701|9f3a457a426cae314...|2012-08-21 15:27:...|    Facebook|  Ri
o de Janeiro|         null|      665419|
|            null| 1199461|f0ec48700197a23c4...|2012-10-15 21:34:...|    Facebook|
        null|         null|      663065|
|São João del Rei| 1199555|57e35f4dbe1f277de...|2012-09-18 09:53:...|    Facebook|
Minas Gerais|         null|      663106|
|        Brasilia| 1199536|5247dbaa0d6cac436...|2012-09-10 20:26:...|    Facebook|Dist
rito Federal|         null|      662926|
|        Sorocaba| 1199553|603aec86462715af4...|2012-12-03 19:55:...|    Facebook|
   São Paulo|         null|      664240|
|            null| 1199741|7e7d6eb03294f5a7f...|2012-09-10 20:26:...|       Email|Dist
rito Federal|         null|      662926|
| Poços de Caldas| 1199517|735218971fa78ff17...|2012-09-26 02:29:...|    Facebook|
Minas Gerais|         null|      663134|
|       Uberlândia| 1199532|3c912e19c1203ff04...|2012-10-15 21:34:...|    Facebook|
Minas Gerais|         null|      663075|
+----------------+--------+--------------------+--------------------+------------+----
------------+-------------+------------+
only showing top 20 rows
```

In [7]:

```python
# - Visitas que os usuários fizeram ao longo do mês
## - Quantidade de visitas?
dfSessions.printSchema()
dfSessions.show()
```

```
root
 |-- SessionStartTime: string (nullable = true)
 |-- StudentClient: string (nullable = true)
 |-- StudentId: string (nullable = true)

+-------------------+-------------+--------------------+
|   SessionStartTime|StudentClient|           StudentId|
+-------------------+-------------+--------------------+
|2017-11-18 15:47:33|      Website|0cade9bf00234e378...|
|2017-11-20 22:21:13|      Website|0cade9bf00234e378...|
|2017-11-20 22:35:31|      Website|0cade9bf00234e378...|
|2017-11-20 23:35:46|      Website|0cade9bf00234e378...|
|2017-11-23 21:24:00|      Website|0cade9bf00234e378...|
|2017-11-24 02:08:08|      Website|0cade9bf00234e378...|
|2017-11-13 13:12:49|      Website|8a501cab6c0a5a7e9...|
|2017-11-06 21:34:45|      Website|b8a39150d98d74685...|
|2017-11-20 15:26:09|      Website|b8a39150d98d74685...|
|2017-11-20 21:40:32|      Website|b8a39150d98d74685...|
|2017-11-28 21:40:03|      Website|b8a39150d98d74685...|
|2017-11-02 11:04:17|      Website|0f6c90f966a70b84b...|
|2017-11-03 14:19:56|      Website|0f6c90f966a70b84b...|
|2017-11-06 11:59:36|      Website|0f6c90f966a70b84b...|
|2017-11-09 15:09:56|      Website|0f6c90f966a70b84b...|
|2017-11-15 14:37:10|      Website|0f6c90f966a70b84b...|
|2017-11-20 22:27:53|      Website|0f6c90f966a70b84b...|
|2017-11-21 17:38:22|      Website|388bf84cffb321938...|
|2017-11-22 14:35:46|      Website|388bf84cffb321938...|
|2017-11-06 20:33:29|      Website|5764c7b7c75f30474...|
+-------------------+-------------+--------------------+
only showing top 20 rows
```

In [9]:

```python
# - Lista de Universidades cadastradas
dfUniversities.printSchema()
dfUniversities.show()
```

```
root
 |-- Id: long (nullable = true)
 |-- Name: string (nullable = true)

+------+----------+
|    Id|      Name|
+------+----------+
|664704|      UERJ|
|661625|   PUC-RIO|
|664768|    UNIRIO|
|663106|      UFSJ|
|664138|      UFSC|
|664742|      UFRJ|
|663609|    PUC-PR|
|663054|      UFOP|
|663065|       UFV|
|664565|ESTÁCIO EAD|
|665419|     UNIFOA|
|662926|       UNB|
|664240|     UFSCAR|
|663134| PUC-MINAS|
|663075|       UFU|
|664623|   ESTÁCIO|
|665677|       UVV|
|662981|       UFG|
|663165|      UFMG|
|662944|      UFES|
+------+----------+
only showing top 20 rows
```

In [10]:

```python
# - Lista de Cursos cadastrados
dfCourses.printSchema()
dfCourses.show()
```

```
root
 |-- Id: long (nullable = true)
 |-- Name: string (nullable = true)

+-------+--------------------+
|     Id|                Name|
+-------+--------------------+
|1199555| Engenharia Elétrica|
|1199521|Economia / Ciênci...|
|1199517|             Direito|
|1199491| Ciências Ambientais|
|1199573|  Engenharia Quimica|
|1199553|Engenharia de Pro...|
|1199536|Engenharia Ambiental|
|1199725|             Quimica|
|1199453|      Administração|
|1199701|         Odontologia|
|1199461|           Agronomia|
|1199741|      Serviço Social|
|1199532|Engenharia Aeroná...|
|1199544|    Engenharia Civil|
|1199724|Comunicação Socia...|
|1199704|           Pedagogia|
|1199699|            Nutrição|
|6495411|      Gestão Pública|
|1199687|            Medicina|
|1199734|Relações Internac...|
+-------+--------------------+
only showing top 20 rows
```

In [11]:

```python
# - subjects.json - Lista de Disciplinas
dfSubjects.printSchema()
dfSubjects.show()
```

```
root
 |-- Id: long (nullable = true)
 |-- Name: string (nullable = true)

+-------+--------------------+
|     Id|                Name|
+-------+--------------------+
| 682889|      Eletrônica II|
| 684377|Circuitos Elétric...|
| 698583|          Eletrônica|
| 757882|        Circuitos II|
|1910475|Circuitos Elétric...|
| 673956|Teoria Microeconô...|
| 674245|       Econometria I|
| 681478|     Microeconomia 2|
| 681925|     Microeconomia 1|
| 693487|Métodos Quantitat...|
| 670492|Direito Administr...|
| 671126|Direito Constituc...|
| 686806|             Direito|
| 689481|     Direito Civil I|
|1663821|Extras (livros, C...|
| 670937|   Direito Ambiental|
| 671197|         Estatistica|
| 673846|             Ecologia|
| 674226|    Geoprocessamento|
| 695790|Sistemas de Infor...|
+-------+--------------------+
only showing top 20 rows
```

In [12]:

```python
# - Disciplinas que cada usuário está seguindo
dfStudentFollow_subject.printSchema()
dfStudentFollow_subject.show()
```

```
root
 |-- FollowDate: string (nullable = true)
 |-- StudentId: string (nullable = true)
 |-- SubjectId: long (nullable = true)

+--------------------+--------------------+---------+
|          FollowDate|           StudentId|SubjectId|
+--------------------+--------------------+---------+
|2015-09-07 15:49:...|0cade9bf00234e378...|   682889|
|2015-09-07 15:49:...|0cade9bf00234e378...|   684377|
|2015-09-07 15:49:...|0cade9bf00234e378...|   698583|
|2015-09-07 15:49:...|0cade9bf00234e378...|   757882|
|2015-09-07 15:49:...|0cade9bf00234e378...|  1910475|
|2015-12-07 16:17:...|8a501cab6c0a5a7e9...|   673956|
|2015-12-07 16:17:...|8a501cab6c0a5a7e9...|   674245|
|2015-12-07 16:17:...|8a501cab6c0a5a7e9...|   681478|
|2015-12-07 16:17:...|8a501cab6c0a5a7e9...|   681925|
|2015-12-07 16:17:...|8a501cab6c0a5a7e9...|   693487|
|2016-05-30 10:31:...|b8a39150d98d74685...|   670492|
|2016-05-30 10:31:...|b8a39150d98d74685...|   671126|
|2016-05-30 10:31:...|b8a39150d98d74685...|   686806|
|2016-05-30 10:31:...|b8a39150d98d74685...|   689481|
|2016-05-30 10:31:...|b8a39150d98d74685...|  1663821|
|2017-11-03 12:23:...|0f6c90f966a70b84b...|   670937|
|2016-04-01 20:22:...|0f6c90f966a70b84b...|   671197|
|2017-11-03 12:23:...|0f6c90f966a70b84b...|   673846|
|2016-04-01 20:22:...|0f6c90f966a70b84b...|   674226|
|2016-04-01 20:22:...|0f6c90f966a70b84b...|   695790|
+--------------------+--------------------+---------+
only showing top 20 rows
```

Montando a Tabela Cadastro

Tabela base para análise de perfil dos usuários.

Table Name: bi.cadastro

Qual região é mais acessada?

Por qual canal é mais acessado?

Qual universidade/ curso tem mais acessos?

In [23]:

```python
#Table Name: bi.cadastro
dfCadastro = dfStudents.alias("students")\
    .join(dfUniversities.alias("universities"),
          trim(col("students.UniversityId")) == trim(col("universities.Id")),
          how = "inner"
         )\
    .join(dfCourses.alias("courses"),
          trim(col("students.CourseId")) == trim(col("courses.Id")),
          how = "inner"
         )\
.select(
        col("students.Id").alias("students_id"),
        col("students.State").alias("students_state"),
        col("students.City").alias("students_city"),
        col("students.SignupSource").alias("students_signup_source"),
        col("students.StudentClient").alias("students_student_client"),
        col("students.RegisteredDate").alias("students_registered_date"),
        col("universities.Id").alias("universities_id"),
        col("universities.Name").alias("universities_name"),
        col("students.CourseId").alias("courses_id"),
        col("courses.Name").alias("courses_name"),
        current_timestamp().alias("ts_processamento"),
        lit("2017-11").alias("pt_ano_mes_referencia")
       ).distinct()
```

In [24]:

```python
dfCadastro.printSchema()
```

```
root
 |-- students_id: string (nullable = true)
 |-- students_state: string (nullable = true)
 |-- students_city: string (nullable = true)
 |-- students_signup_source: string (nullable = true)
 |-- students_student_client: string (nullable = true)
 |-- students_registered_date: string (nullable = true)
 |-- universities_id: long (nullable = true)
 |-- universities_name: string (nullable = true)
 |-- courses_id: long (nullable = true)
 |-- courses_name: string (nullable = true)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = false)
```

In [25]:

```python
dfCadastro.show()
```

```
+--------------------+-------------------+--------------+----------------------+------
-----------------+------------------------+---------------+-----------------+--------
-+--------------------+--------------------+---------------------+
|         students_id|     students_state| students_city|students_signup_source|studen
ts_student_client|students_registered_date|universities_id|universities_name|courses_i
d|        courses_name|    ts_processamento|pt_ano_mes_referencia|
+--------------------+-------------------+--------------+----------------------+------
-----------------+------------------------+---------------+-----------------+--------
-+--------------------+--------------------+---------------------+
|32a67b23002c18c32...|              Ceará|     Fortaleza|              Facebook|
             null|    2013-06-23 21:59:...|         664623|          ESTÁCIO|   119945
3|       Administração|2021-03-28 18:40:...|              2017-11|
|27d36212707731bbe...|       Minas Gerais|Belo Horizonte|              Facebook|
             null|    2013-07-11 18:13:...|         665162|     NEWTON PAIVA|   119955
3|Engenharia de Pro...|2021-03-28 18:40:...|              2017-11|
|0bedd2549c73c94df...|              Bahia|      Salvador|              Facebook|
             null|    2013-07-11 18:13:...|         662801|             UFBA|   119972
7|Quimica e Engenha...|2021-03-28 18:40:...|              2017-11|
|439b8aa1192423c04...|          São Paulo|     São Paulo|              Facebook|
             null|    2013-09-14 12:26:...|         664331|         UNIMONTE|   119951
7|             Direito|2021-03-28 18:40:...|              2017-11|
|dda5851d67df9d5bb...|              Bahia|      Camaçari|                 Email|
             null|    2013-10-01 15:26:...|         662801|             UFBA|   119951
7|             Direito|2021-03-28 18:40:...|              2017-11|
|c47a99e8259a4ff67...|             Paraná|          null|              Facebook|
             null|    2013-10-09 22:39:...|         664060|            UDESC|   119955
5| Engenharia Elétrica|2021-03-28 18:40:...|              2017-11|
|ab53ffac361e7bbe8...|  Rio Grande do Sul|          null|                 Email|
             null|    2013-10-15 23:37:...|         663928|              URI|   119954
4|    Engenharia Civil|2021-03-28 18:40:...|              2017-11|
|4a04a3ddd18a19f84...|               null|          null|              Facebook|
             null|    2013-10-24 15:24:...|         663985|         FCMS/JF|   119968
7|            Medicina|2021-03-28 18:40:...|              2017-11|
|2ba9b6aa32c6df8b5...|       Minas Gerais|          null|                 Email|
             null|    2013-11-06 20:58:...|         663205|            UFVJM|   151052
5|Ciências Exatas e...|2021-03-28 18:40:...|              2017-11|
|a90017700be972f6c...|           Maranhão|          null|                 Email|
             null|    2013-11-15 19:30:...|         662269|             UNDB|   119951
7|             Direito|2021-03-28 18:40:...|              2017-11|
|e2ed116d20c02ecb9...|Rio Grande do Norte|       Mossoró|              Facebook|
             null|    2013-11-16 18:33:...|         663775|              UNP|   119955
3|Engenharia de Pro...|2021-03-28 18:40:...|              2017-11|
|2ca63d7ce3eaf135d...|     Rio de Janeiro|Rio de Janeiro|              Facebook|
             null|    2013-11-25 12:40:...|         664822|             IFRJ|   194492
7|Quimica de Alimentos|2021-03-28 18:40:...|              2017-11|
|207f1fca4f28ded16...|     Rio de Janeiro|Rio de Janeiro|              Facebook|
             null|    2013-11-26 23:34:...|         664623|          ESTÁCIO|   119970
1|         Odontologia|2021-03-28 18:40:...|              2017-11|
|584dd1972e5f26744...|               null|          null|              Facebook|
             null|    2013-12-12 19:12:...|         664418|          UNINOVE|   119951
7|             Direito|2021-03-28 18:40:...|              2017-11|
|ed4bdb83d3ae49ae7...|               null|          null|              Facebook|
             null|    2014-01-14 10:29:...|         663650|           UNIPAR|   119951
7|             Direito|2021-03-28 18:40:...|              2017-11|
|f0192601dba7a9c4e...|       Minas Gerais|Belo Horizonte|                 Email|
          Website|    2014-03-01 11:55:...|         663165|             UFMG|   119945
3|       Administração|2021-03-28 18:40:...|              2017-11|
|c35096d0f8ac8abb9...|          São Paulo|Ribeirão Pires|                 Email|
          Website|    2014-03-04 14:14:...|         663246|        TIJUCUSSU|   119951
7|             Direito|2021-03-28 18:40:...|              2017-11|
|110aaceb47191de96...|               null|          null|                 Email|
          Website|    2014-03-12 14:23:...|         663946|            ULBRA|   119954
4|    Engenharia Civil|2021-03-28 18:40:...|              2017-11|
|4458bef99de0c938f...|           Maranhão|      São Luis|              Facebook|
          Website|    2014-03-18 22:40:...|         664623|          ESTÁCIO|   119951
7|             Direito|2021-03-28 18:40:...|              2017-11|
|5d64bea4031cb8475...|Rio Grande do Norte|         Natal|              Facebook|
          Website|    2014-03-28 13:40:...|         663754|             UFRN|   119954
4|    Engenharia Civil|2021-03-28 18:40:...|              2017-11|
+--------------------+-------------------+--------------+----------------------+------
-----------------+------------------------+---------------+-----------------+--------
-+--------------------+--------------------+---------------------+
only showing top 20 rows
```

In [44]:

```python
#Qual região é mais acessada, agrupada por ano e mês.
#Table Name: bi.qtd_acesso_regiao
dfQtdRegiao = dfCadastro\
.groupBy("students_state",
        "pt_ano_mes_referencia")\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy(desc("qtd"))\
.select("students_state",
       "qtd",
       "pt_ano_mes_referencia")
```

In [45]:

```python
dfQtdRegiao.printSchema()
```

```
root
 |-- students_state: string (nullable = true)
 |-- qtd: long (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = false)
```

In [46]:

```python
dfQtdRegiao.show()
```

```
+-------------------+-----+---------------------+
|     students_state|  qtd|pt_ano_mes_referencia|
+-------------------+-----+---------------------+
|               null|35393|              2017-11|
|     Rio de Janeiro| 5168|              2017-11|
|          São Paulo| 3870|              2017-11|
|       Minas Gerais| 2417|              2017-11|
|             Paraná| 1438|              2017-11|
|              Bahia| 1354|              2017-11|
|              Ceará| 1236|              2017-11|
|  Rio Grande do Sul| 1192|              2017-11|
|         Pernambuco|  954|              2017-11|
|     Santa Catarina|  876|              2017-11|
|              Goiás|  815|              2017-11|
|     Espirito Santo|  595|              2017-11|
|               Pará|  526|              2017-11|
|           Maranhão|  507|              2017-11|
|            Sergipe|  427|              2017-11|
| Mato Grosso do Sul|  421|              2017-11|
|   Distrito Federal|  385|              2017-11|
|Rio Grande do Norte|  367|              2017-11|
|            Paraiba|  345|              2017-11|
|           Amazonas|  342|              2017-11|
+-------------------+-----+---------------------+
only showing top 20 rows
```

In [51]:

```python
#Por qual canal é mais acessado, por região?
#Table Name: bi.qtd_acesso_canal_regiao

dfQtdCanal = dfCadastro\
.groupBy("students_state",
         "students_student_client",
        "pt_ano_mes_referencia"
        )\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy(desc("qtd"))\
.select(
    "students_state",
    "students_student_client",
       "qtd",
       "pt_ano_mes_referencia")
```

In [53]:

```python
dfQtdCanal.printSchema()
```

```
root
 |-- students_state: string (nullable = true)
 |-- students_student_client: string (nullable = true)
 |-- qtd: long (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = false)
```

In [54]:

```python
dfQtdCanal.show()
```

```
+-----------------+-----------------------+-----+---------------------+
|   students_state|students_student_client|  qtd|pt_ano_mes_referencia|
+-----------------+-----------------------+-----+---------------------+
|             null|                Website|23286|              2017-11|
|   Rio de Janeiro|                Website| 2947|              2017-11|
|        São Paulo|                Website| 2561|              2017-11|
|             null|                 Webapp| 2405|              2017-11|
|             null|                   null| 2100|              2017-11|
|             null|   Android | 6.0.1 |...| 1896|              2017-11|
|     Minas Gerais|                Website| 1481|              2017-11|
|   Rio de Janeiro|                   null| 1412|              2017-11|
|             null|   Android | 7.0 | s...|  942|              2017-11|
|           Paraná|                Website|  928|              2017-11|
|             null|   Android | 6.0 | s...|  845|              2017-11|
|            Bahia|                Website|  823|              2017-11|
|        São Paulo|                   null|  811|              2017-11|
|Rio Grande do Sul|                Website|  784|              2017-11|
|            Ceará|                Website|  712|              2017-11|
|     Minas Gerais|                   null|  695|              2017-11|
|   Santa Catarina|                Website|  585|              2017-11|
|       Pernambuco|                Website|  553|              2017-11|
|             null|   Android | 5.1.1 |...|  517|              2017-11|
|            Goiás|                Website|  510|              2017-11|
+-----------------+-----------------------+-----+---------------------+
only showing top 20 rows
```

Montando a Dimensão subject_user

Tabelas: bi.subject_user e bi.qtd_subject_user

Table name: bi.subject_user

Poderá ser analisado as disciplinas mais seguidas por cada usuário

Alimentação diária por data do arquivo

In [59]:

```python
#Table Name: bi.subject_user
dfSubjectsUser = dfSubjects.alias("disc")\
    .join(dfStudentFollow_subject.alias("disc_user"),
          trim(col("disc.Id")) == trim(col("disc_user.SubjectId")),
          how = "left"
         )\
.select(
        col("disc.Id").alias("subjects_id"),
        col("disc.Name").alias("subjects_name"),
        col("disc_user.FollowDate").alias("subjects_follow_date"),
        col("disc_user.StudentId").alias("student_id"),
        current_timestamp().alias("ts_processamento"),
        lit("2017-11").alias("pt_ano_mes_referencia") #Data da partição eu usaria data
       )
```

In [56]:

```python
dfSubjectsUser.printSchema()
```

```
root
 |-- subjects_id: long (nullable = true)
 |-- subjects_name: string (nullable = true)
 |-- subjects_follow_date: string (nullable = true)
 |-- student_id: string (nullable = true)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = false)
```

In [60]:

```python
#Conhecendo a base
dfSubjectsUser\
.show()
```

```
+-----------+--------------------+--------------------+--------------------+----------
----------+---------------------+
|subjects_id|       subjects_name|subjects_follow_date|          student_id|    ts_pro
cessamento|pt_ano_mes_referencia|
+-----------+--------------------+--------------------+--------------------+----------
----------+---------------------+
|    1190352|Piscicultura e Ra...|2017-11-30 01:09:...|0cb72964a442c2f08...|2021-03-28
19:17:...|              2017-11|
|    1214010|Bases da Cultura ...|2015-09-13 21:18:...|6836b06beaaa7cc1e...|2021-03-28
19:17:...|              2017-11|
|    1214010|Bases da Cultura ...|2017-06-23 14:40:...|888b761f9be4f9e7c...|2021-03-28
19:17:...|              2017-11|
|    1485777|Exames Hematológi...|2016-04-27 00:47:...|db681c210ffb150c9...|2021-03-28
19:17:...|              2017-11|
|    1485777|Exames Hematológi...|2017-05-14 22:56:...|869f5faae5dc132d2...|2021-03-28
19:17:...|              2017-11|
|    1485777|Exames Hematológi...|2017-09-01 19:41:...|42ca8756348976afd...|2021-03-28
19:17:...|              2017-11|
|    1485777|Exames Hematológi...|2015-09-27 11:24:...|754ca68a4414af257...|2021-03-28
19:17:...|              2017-11|
|    1505509|  Higiene e Inspeção|2017-05-31 13:17:...|8075da88376651f9a...|2021-03-28
19:17:...|              2017-11|
|    1505509|  Higiene e Inspeção|2017-10-27 12:25:...|6a80d1b9f1ade3b0c...|2021-03-28
19:17:...|              2017-11|
|    1505509|  Higiene e Inspeção|2017-08-14 19:59:...|1f9b30a864061321d...|2021-03-28
19:17:...|              2017-11|
|    1507765|          Volumetria|2017-10-10 12:47:...|66d44f3e5b0a4f725...|2021-03-28
19:17:...|              2017-11|
|    1507781|Fundamentos de Pr...|2016-03-28 20:08:...|334c55fde8f9e38ae...|2021-03-28
19:17:...|              2017-11|
|    1507781|Fundamentos de Pr...|2017-08-16 14:00:...|7cc2cd70f6252ea5b...|2021-03-28
19:17:...|              2017-11|
|    1507781|Fundamentos de Pr...|2017-09-19 08:44:...|6b9e8fb4225e152ba...|2021-03-28
19:17:...|              2017-11|
|    1510540|            Farmácia|2017-07-21 15:04:...|21e4bac8714edf712...|2021-03-28
19:17:...|              2017-11|
|    1510540|            Farmácia|2016-10-07 12:51:...|640a80e55245f64fb...|2021-03-28
19:17:...|              2017-11|
|    1510540|            Farmácia|2016-10-16 22:25:...|b8e262aa3756e3011...|2021-03-28
19:17:...|              2017-11|
|    1510540|            Farmácia|2016-09-25 15:10:...|c952d27aac489af40...|2021-03-28
19:17:...|              2017-11|
|    1510540|            Farmácia|2017-11-16 22:58:...|38cd46815aada7efd...|2021-03-28
19:17:...|              2017-11|
|    1510540|            Farmácia|2016-03-25 19:48:...|9fa9b43a469e889cc...|2021-03-28
19:17:...|              2017-11|
+-----------+--------------------+--------------------+--------------------+----------
----------+---------------------+
only showing top 20 rows
```

Table name: bi.qtd_subject_user

Quais são as disciplinas mais acessadas ao longo do mês?

Alimentação diária por ano e mês

In [74]:

```python
#Table Name: bi.qtd_subject_user
dfQtdDisc = dfSubjectsUser\
.groupBy("pt_ano_mes_referencia",
         "subjects_id",
         "subjects_name")\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy("pt_ano_mes_referencia",
         desc("qtd"))\
.withColumn("ts_processamento", current_timestamp())\
.select("subjects_id",
       "subjects_name",
        "qtd",
        "ts_processamento",
       "pt_ano_mes_referencia")
```

In [75]:

```python
dfQtdDisc.printSchema()
```

```
root
 |-- subjects_id: long (nullable = true)
 |-- subjects_name: string (nullable = true)
 |-- qtd: long (nullable = false)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = false)
```

In [76]:

```python
#Conhecendo a base
dfQtdDisc.show()
```

```
+-----------+--------------------+----+--------------------+---------------------+
|subjects_id|       subjects_name| qtd|    ts_processamento|pt_ano_mes_referencia|
+-----------+--------------------+----+--------------------+---------------------+
|     671154|          Bioquímica|4223|2021-03-28 19:24:...|              2017-11|
|     669699|            Cálculo I|4053|2021-03-28 19:24:...|              2017-11|
|     670780|     Anatomia Humana|3950|2021-03-28 19:24:...|              2017-11|
|     671126|Direito Constituc...|3926|2021-03-28 19:24:...|              2017-11|
|     669660|Matemática Financ...|3405|2021-03-28 19:24:...|              2017-11|
|     669796|            Fisica I|3023|2021-03-28 19:24:...|              2017-11|
|     676833|      Direito Penal I|2956|2021-03-28 19:24:...|              2017-11|
|     680401|   Fisiologia Humana|2906|2021-03-28 19:24:...|              2017-11|
|     676780|Direito Constituc...|2857|2021-03-28 19:24:...|              2017-11|
|     672228|Direito Processua...|2574|2021-03-28 19:24:...|              2017-11|
|     669744|          Cálculo II|2537|2021-03-28 19:24:...|              2017-11|
|     670592|Cálculo Diferenci...|2457|2021-03-28 19:24:...|              2017-11|
|     669454|Resistência dos M...|2365|2021-03-28 19:24:...|              2017-11|
|     671086|            Genética|2275|2021-03-28 19:24:...|              2017-11|
|     672547|        Farmacologia|2240|2021-03-28 19:24:...|              2017-11|
|     670384|      Mecânica Geral|2147|2021-03-28 19:24:...|              2017-11|
|     675399|Metodologia Cient...|2135|2021-03-28 19:24:...|              2017-11|
|     682100|       Administração|2128|2021-03-28 19:24:...|              2017-11|
|     689481|     Direito Civil I|2087|2021-03-28 19:24:...|              2017-11|
|     670819|Direito Administr...|2075|2021-03-28 19:24:...|              2017-11|
+-----------+--------------------+----+--------------------+---------------------+
only showing top 20 rows
```

Montando a Dimensão sessions

Tabelas: bi.sessions e bi.qtd_sessions

Table name: bi.sessions

Tabela base para análise comportamental dos usuários.

Qual StudentClient e data são mais acessados?

Alimentação diária por data do arquivo, partição ano_mês

In [96]:

```python
#Table Name: bi.sessions
dfSessions = dfSessions\
.withColumn("ts_processamento", current_timestamp())\
 .withColumn("pt_ano_mes_referencia", substring(to_date("SessionStartTime"),  1, 7))
```

In [97]:

```python
dfSessions.printSchema()
```

```
root
 |-- SessionStartTime: string (nullable = true)
 |-- StudentClient: string (nullable = true)
 |-- StudentId: string (nullable = true)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = true)
```

In [98]:

```python
dfSessions.show()
```

```
+-------------------+-------------+--------------------+--------------------+---------
------------+
|   SessionStartTime|StudentClient|           StudentId|    ts_processamento|pt_ano_me
s_referencia|
+-------------------+-------------+--------------------+--------------------+---------
------------+
|2017-11-18 15:47:33|      Website|0cade9bf00234e378...|2021-03-28 19:34:...|
      2017-11|
|2017-11-20 22:21:13|      Website|0cade9bf00234e378...|2021-03-28 19:34:...|
      2017-11|
|2017-11-20 22:35:31|      Website|0cade9bf00234e378...|2021-03-28 19:34:...|
      2017-11|
|2017-11-20 23:35:46|      Website|0cade9bf00234e378...|2021-03-28 19:34:...|
      2017-11|
|2017-11-23 21:24:00|      Website|0cade9bf00234e378...|2021-03-28 19:34:...|
      2017-11|
|2017-11-24 02:08:08|      Website|0cade9bf00234e378...|2021-03-28 19:34:...|
      2017-11|
|2017-11-13 13:12:49|      Website|8a501cab6c0a5a7e9...|2021-03-28 19:34:...|
      2017-11|
|2017-11-06 21:34:45|      Website|b8a39150d98d74685...|2021-03-28 19:34:...|
      2017-11|
|2017-11-20 15:26:09|      Website|b8a39150d98d74685...|2021-03-28 19:34:...|
      2017-11|
|2017-11-20 21:40:32|      Website|b8a39150d98d74685...|2021-03-28 19:34:...|
      2017-11|
|2017-11-28 21:40:03|      Website|b8a39150d98d74685...|2021-03-28 19:34:...|
      2017-11|
|2017-11-02 11:04:17|      Website|0f6c90f966a70b84b...|2021-03-28 19:34:...|
      2017-11|
|2017-11-03 14:19:56|      Website|0f6c90f966a70b84b...|2021-03-28 19:34:...|
      2017-11|
|2017-11-06 11:59:36|      Website|0f6c90f966a70b84b...|2021-03-28 19:34:...|
      2017-11|
|2017-11-09 15:09:56|      Website|0f6c90f966a70b84b...|2021-03-28 19:34:...|
      2017-11|
|2017-11-15 14:37:10|      Website|0f6c90f966a70b84b...|2021-03-28 19:34:...|
      2017-11|
|2017-11-20 22:27:53|      Website|0f6c90f966a70b84b...|2021-03-28 19:34:...|
      2017-11|
|2017-11-21 17:38:22|      Website|388bf84cffb321938...|2021-03-28 19:34:...|
      2017-11|
|2017-11-22 14:35:46|      Website|388bf84cffb321938...|2021-03-28 19:34:...|
      2017-11|
|2017-11-06 20:33:29|      Website|5764c7b7c75f30474...|2021-03-28 19:34:...|
      2017-11|
+-------------------+-------------+--------------------+--------------------+---------
------------+
only showing top 20 rows
```

Table name: bi.qtd_sessions_users

#Quais estudantes mais acessam? Agrupamento por data.

Alimentação diária por ano e mês

In [301…

```python
#Qual StudentClient e data são mais acessados?
#Table Name: bi.qtd_sessions_users
dfQtdSessionsUsers = dfSessions\
.groupBy(to_date("SessionStartTime").alias("date"),
        "StudentId")\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy(desc("qtd"))\
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_mes_referencia", substring(to_date("date"),  1, 7))
```

In [302…

```python
dfQtdSessionsUsers.printSchema()
```

```
root
 |-- date: date (nullable = true)
 |-- StudentId: string (nullable = true)
 |-- qtd: long (nullable = false)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = true)
```

In [303…

```python
dfQtdSessionsUsers.show()
```

```
+----------+--------------------+---+--------------------+---------------------+
|      date|           StudentId|qtd|    ts_processamento|pt_ano_mes_referencia|
+----------+--------------------+---+--------------------+---------------------+
|2017-11-22|efbb2450d87831fe3...|948|2021-03-28 23:49:...|              2017-11|
|2017-11-24|53a9a87d540030b2b...|918|2021-03-28 23:49:...|              2017-11|
|2017-11-22|1374456fffa9f1373...|802|2021-03-28 23:49:...|              2017-11|
|2017-11-22|8e12fc20b0e06589d...|606|2021-03-28 23:49:...|              2017-11|
|2017-11-22|96b110d5d80e81359...|575|2021-03-28 23:49:...|              2017-11|
|2017-11-22|1060f13f2ae2f1df2...|457|2021-03-28 23:49:...|              2017-11|
|2017-11-22|59b6af9a36f0fe891...|345|2021-03-28 23:49:...|              2017-11|
|2017-11-29|06038eb332fb0ec91...|330|2021-03-28 23:49:...|              2017-11|
|2017-11-22|06c59ff3c83314594...|292|2021-03-28 23:49:...|              2017-11|
|2017-11-22|e2dedde2c68fc63a6...|280|2021-03-28 23:49:...|              2017-11|
|2017-11-22|be64d66f3c3f7a057...|274|2021-03-28 23:49:...|              2017-11|
|2017-11-22|d4219eeee12f34dbf...|273|2021-03-28 23:49:...|              2017-11|
|2017-11-25|8a9d5014a0be4bf26...|247|2021-03-28 23:49:...|              2017-11|
|2017-11-22|4d62f38fae9bbfbca...|243|2021-03-28 23:49:...|              2017-11|
|2017-11-23|2e5dcabe957dffb24...|239|2021-03-28 23:49:...|              2017-11|
|2017-11-22|3b0f92f0a7a8c6751...|226|2021-03-28 23:49:...|              2017-11|
|2017-11-22|ce214c3f536ca17ab...|223|2021-03-28 23:49:...|              2017-11|
|2017-11-22|7a1893fafb2653245...|210|2021-03-28 23:49:...|              2017-11|
|2017-11-26|8a9d5014a0be4bf26...|210|2021-03-28 23:49:...|              2017-11|
|2017-11-26|e31c7cc0be0b41e2c...|207|2021-03-28 23:49:...|              2017-11|
+----------+--------------------+---+--------------------+---------------------+
only showing top 20 rows
```

Table name: bi.qtd_sessions

#Qual StudentClient são mais acessados? Agrupamento por data.

Alimentação diária por ano e mês

In [304…

```python
#Table Name: bi.qtd_sessions_StudentClient
dfQtdSessionsStudentClient = dfSessions\
.groupBy(to_date("SessionStartTime").alias("dateSession"),
        "StudentClient")\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy("dateSession",
         desc("qtd"))\
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_mes_referencia", substring(to_date("dateSession"), 1, 7)) #Partiçã
```

In [305…

```python
dfQtdSessionsStudentClient.printSchema()
```

```
root
 |-- dateSession: date (nullable = true)
 |-- StudentClient: string (nullable = true)
 |-- qtd: long (nullable = false)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = true)
```

In [306…

```python
dfQtdSessionsStudentClient.show()
```

```
+-----------+-------------+----+--------------------+---------------------+
|dateSession|StudentClient| qtd|    ts_processamento|pt_ano_mes_referencia|
+-----------+-------------+----+--------------------+---------------------+
| 2017-11-01|      Website|4072|2021-03-28 23:50:...|              2017-11|
| 2017-11-01|      Android|1609|2021-03-28 23:50:...|              2017-11|
| 2017-11-01|       Webapp| 300|2021-03-28 23:50:...|              2017-11|
| 2017-11-01|          iOS| 227|2021-03-28 23:50:...|              2017-11|
| 2017-11-02|      Website|3439|2021-03-28 23:50:...|              2017-11|
| 2017-11-02|      Android|1148|2021-03-28 23:50:...|              2017-11|
| 2017-11-02|          iOS| 185|2021-03-28 23:50:...|              2017-11|
| 2017-11-02|       Webapp| 180|2021-03-28 23:50:...|              2017-11|
| 2017-11-03|      Website|3938|2021-03-28 23:50:...|              2017-11|
| 2017-11-03|      Android|1432|2021-03-28 23:50:...|              2017-11|
| 2017-11-03|       Webapp| 274|2021-03-28 23:50:...|              2017-11|
| 2017-11-03|          iOS| 203|2021-03-28 23:50:...|              2017-11|
| 2017-11-04|      Website|3450|2021-03-28 23:50:...|              2017-11|
| 2017-11-04|      Android|1417|2021-03-28 23:50:...|              2017-11|
| 2017-11-04|       Webapp| 222|2021-03-28 23:50:...|              2017-11|
| 2017-11-04|          iOS| 209|2021-03-28 23:50:...|              2017-11|
| 2017-11-05|      Website|4463|2021-03-28 23:50:...|              2017-11|
| 2017-11-05|      Android|1634|2021-03-28 23:50:...|              2017-11|
| 2017-11-05|          iOS| 293|2021-03-28 23:50:...|              2017-11|
| 2017-11-05|       Webapp| 281|2021-03-28 23:50:...|              2017-11|
+-----------+-------------+----+--------------------+---------------------+
only showing top 20 rows
```

Montando a Dimensão subscriptions

Tabelas: bi.subscriptions

Table name: bi.subscriptions

Alimentação diária por ano e mês

In [102…

```
#bi.subscriptions (premium)
dfSubscriptions = dfSubscriptions\
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_referencia", year("PaymentDate"))\
.withColumn("pt_mes_referencia", month("PaymentDate"))
```

In [103…

```
dfSubscriptions.printSchema()
```

```
root
 |-- PaymentDate: string (nullable = true)
 |-- PlanType: string (nullable = true)
 |-- StudentId: string (nullable = true)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_referencia: integer (nullable = true)
 |-- pt_mes_referencia: integer (nullable = true)
```

In [104…

```
dfSubscriptions.show()
```

```
+--------------------+--------+--------------------+--------------------+-------------
----+-----------------+
|         PaymentDate|PlanType|           StudentId|    ts_processamento|pt_ano_refere
ncia|pt_mes_referencia|
+--------------------+--------+--------------------+--------------------+-------------
----+-----------------+
|2017-11-14 19:52:...|  Mensal|29037b0a52c5b576d...|2021-03-28 19:36:...|
2017|               11|
|2017-11-08 11:52:...|  Mensal|b2bace77d15c3dfaf...|2021-03-28 19:36:...|
2017|               11|
|2017-11-05 21:27:...|  Mensal|f423d6fe2f8964db6...|2021-03-28 19:36:...|
2017|               11|
|2017-11-15 14:36:...|  Mensal|55ccbe518d2edbbd5...|2021-03-28 19:36:...|
2017|               11|
|2017-11-12 22:19:...|  Mensal|b1b0f63fe3e4820cb...|2021-03-28 19:36:...|
2017|               11|
|2017-11-22 01:03:...|  Mensal|ed46832f6b716fb2e...|2021-03-28 19:36:...|
2017|               11|
|2017-11-24 19:03:...|  Mensal|69b7bee32821cf76b...|2021-03-28 19:36:...|
2017|               11|
|2017-11-11 21:01:...|  Mensal|6553923125fe6364e...|2021-03-28 19:36:...|
2017|               11|
|2017-11-12 16:41:...|  Mensal|3903a334d1af8ce83...|2021-03-28 19:36:...|
2017|               11|
|2017-11-21 11:52:...|  Mensal|4487f81a4ea9b3c3c...|2021-03-28 19:36:...|
2017|               11|
|2017-11-06 22:14:...|  Mensal|bde8436a92ab53ce2...|2021-03-28 19:36:...|
2017|               11|
|2017-11-17 19:30:...|  Mensal|3ae094228e1b9324e...|2021-03-28 19:36:...|
2017|               11|
|2017-11-10 22:33:...|  Mensal|5a0df378119b95d3c...|2021-03-28 19:36:...|
2017|               11|
|2017-11-08 22:06:...|  Mensal|fed3de568756e019f...|2021-03-28 19:36:...|
2017|               11|
|2017-11-01 18:14:...|  Mensal|68beec3e289e3316e...|2021-03-28 19:36:...|
2017|               11|
|2017-11-10 15:22:...|  Mensal|ecb4e29cd9585b080...|2021-03-28 19:36:...|
2017|               11|
|2017-11-09 01:10:...|  Mensal|2e34dfda3cca387a7...|2021-03-28 19:36:...|
2017|               11|
|2017-11-08 00:41:...|  Mensal|85529f64c1428b69f...|2021-03-28 19:36:...|
2017|               11|
|2017-11-30 12:57:...|  Mensal|a189c626b75ca9925...|2021-03-28 19:36:...|
2017|               11|
|2017-11-13 12:20:...|  Mensal|fe9f61ef8539dfbdb...|2021-03-28 19:36:...|
2017|               11|
+--------------------+--------+--------------------+--------------------+-------------
----+-----------------+
only showing top 20 rows
```

Table name: bi.qtd_disc_premium

Quais disciplinas são mais acessadas por premium dentro do mês?

Alimentação diária por ano e mês

In [105…

```
#Table name: bi.qtd_disc_premium
dfQtdDiscPremium = dfSubscriptions.alias("sub")\
.join(dfSubjectsUser.alias("subUser"),
      trim(col("sub.StudentId")) == trim(col("subUser.student_id")),
      how = "inner"
     )\
.groupBy(
    substring(col("sub.PaymentDate"), 1, 7).alias("payment_year_month"),
    col("subUser.subjects_id"),
    col("subUser.subjects_name")
)\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy(desc("qtd"))\
.withColumn("ts_processamento", current_timestamp())\
.select(
    col("subUser.subjects_id"),
    col("subUser.subjects_name"),
    col("qtd"),
    col("ts_processamento"),
    col("payment_year_month")
)
```

In [106…

```
dfQtdDiscPremium.printSchema()
```

```
root
 |-- subjects_id: long (nullable = true)
 |-- subjects_name: string (nullable = true)
 |-- qtd: long (nullable = false)
 |-- ts_processamento: timestamp (nullable = false)
 |-- payment_year_month: string (nullable = true)
```

In [107…

```
dfQtdDiscPremium.show()
```

```
+-----------+--------------------+---+--------------------+------------------+
|subjects_id|       subjects_name|qtd|    ts_processamento|payment_year_month|
+-----------+--------------------+---+--------------------+------------------+
|     669660|Matemática Financ...| 70|2021-03-28 19:38:...|           2017-11|
|     683360|Introdução à Admi...| 60|2021-03-28 19:38:...|           2017-11|
|     682100|       Administração| 48|2021-03-28 19:38:...|           2017-11|
|     692813|Contabilidade Básica| 46|2021-03-28 19:38:...|           2017-11|
|     669699|           Cálculo I| 46|2021-03-28 19:38:...|           2017-11|
|     670592|Cálculo Diferenci...| 41|2021-03-28 19:38:...|           2017-11|
|     669404|Administração Fin...| 35|2021-03-28 19:38:...|           2017-11|
|     669744|          Cálculo II| 35|2021-03-28 19:38:...|           2017-11|
|     669454|Resistência dos M...| 34|2021-03-28 19:38:...|           2017-11|
|     670384|      Mecânica Geral| 34|2021-03-28 19:38:...|           2017-11|
|     675399|Metodologia Cient...| 33|2021-03-28 19:38:...|           2017-11|
|     670196|            Didática| 32|2021-03-28 19:38:...|           2017-11|
|     689481|     Direito Civil I| 32|2021-03-28 19:38:...|           2017-11|
|     671126|Direito Constituc...| 31|2021-03-28 19:38:...|           2017-11|
|     671154|          Bioquimica| 31|2021-03-28 19:38:...|           2017-11|
|     670661|Cálculo Diferenci...| 31|2021-03-28 19:38:...|           2017-11|
|     669796|            Fisica I| 31|2021-03-28 19:38:...|           2017-11|
|    1109915|Planejamento de C...| 31|2021-03-28 19:38:...|           2017-11|
|     672147|Contabilidade / C...| 29|2021-03-28 19:38:...|           2017-11|
|     671776|      Álgebra Linear| 29|2021-03-28 19:38:...|           2017-11|
+-----------+--------------------+---+--------------------+------------------+
only showing top 20 rows
```

Table name: bi.qtd_disc_npremium

#Quais disciplinas são mais acessadas por não premium dentro do mês?

Alimentação diária por ano e mês

In [196…

```
#bi.qtd_disc_npremium
dfDiscGrat =  dfSubscriptions.alias("sub")\
.join(dfSubjectsUser.alias("SubUser"),
      trim(col("sub.StudentId")) == trim(col("SubUser.student_id")),
      how = "right"
     )\
.filter("sub.StudentId is null")\
.groupBy(
    col("SubUser.subjects_id"),
    col("SubUser.subjects_name")
)\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy(desc("qtd"))\
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_mes_referencia", substring("ts_processamento", 1, 7))
```

In [197…

```
dfDiscGrat.printSchema()
```

```
root
 |-- subjects_id: long (nullable = true)
 |-- subjects_name: string (nullable = true)
 |-- qtd: long (nullable = false)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = false)
```

In [198…

```
dfDiscGrat.show()
```

```
+-----------+--------------------+----+--------------------+---------------------+
|subjects_id|       subjects_name| qtd|    ts_processamento|pt_ano_mes_referencia|
+-----------+--------------------+----+--------------------+---------------------+
|     671154|          Bioquimica|4192|2021-03-28 21:45:...|              2021-03|
|     669699|           Cálculo I|4008|2021-03-28 21:45:...|              2021-03|
|     670780|     Anatomia Humana|3930|2021-03-28 21:45:...|              2021-03|
|     671126|Direito Constituc...|3895|2021-03-28 21:45:...|              2021-03|
|     669660|Matemática Financ...|3335|2021-03-28 21:45:...|              2021-03|
|     669796|            Fisica I|2993|2021-03-28 21:45:...|              2021-03|
|     676833|     Direito Penal I|2928|2021-03-28 21:45:...|              2021-03|
|     680401|   Fisiologia Humana|2884|2021-03-28 21:45:...|              2021-03|
|     676780|Direito Constituc...|2829|2021-03-28 21:45:...|              2021-03|
|     672228|Direito Processua...|2556|2021-03-28 21:45:...|              2021-03|
|     669744|          Cálculo II|2502|2021-03-28 21:45:...|              2021-03|
|     670592|Cálculo Diferenci...|2416|2021-03-28 21:45:...|              2021-03|
|     669454|Resistência dos M...|2331|2021-03-28 21:45:...|              2021-03|
|     671086|            Genética|2254|2021-03-28 21:45:...|              2021-03|
|     672547|        Farmacologia|2226|2021-03-28 21:45:...|              2021-03|
|     670384|      Mecânica Geral|2113|2021-03-28 21:45:...|              2021-03|
|     675399|Metodologia Cient...|2102|2021-03-28 21:45:...|              2021-03|
|     682100|       Administração|2080|2021-03-28 21:45:...|              2021-03|
|     670819|Direito Administr...|2058|2021-03-28 21:45:...|              2021-03|
|     689481|     Direito Civil I|2055|2021-03-28 21:45:...|              2021-03|
+-----------+--------------------+----+--------------------+---------------------+
only showing top 20 rows
```

Load base B

Table name bi.events

Tabela alimentada diariamente de acordo com o recebimento dos logs

O que acessam na nossa plataforma e como fazem isso.

In [348…

```
#Table name bi.events
#Obs: Criado a coluna student_id. Utilizando a coluna studentId_clientType, removendo
dfEvents = spark.read.format("json").load("/files/passei/BASE-B/*.json")\
.withColumn("students_id", regexp_replace(col("studentId_clientType"), "(@[^\\@]+)$",
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_referencia", year("at"))\
.withColumn("pt_mes_referencia", month("at"))
```

In [349…

```
dfEvents.printSchema()
```

```
root
 |-- Last Accessed Url: string (nullable = true)
 |-- Page Category: string (nullable = true)
 |-- Page Category 1: string (nullable = true)
 |-- Page Category 2: string (nullable = true)
 |-- Page Category 3: string (nullable = true)
 |-- Page Name: string (nullable = true)
 |-- at: string (nullable = true)
 |-- browser: string (nullable = true)
 |-- carrier: string (nullable = true)
 |-- city_name: string (nullable = true)
 |-- clv_total: long (nullable = true)
 |-- country: string (nullable = true)
 |-- custom_1: string (nullable = true)
 |-- custom_2: string (nullable = true)
 |-- custom_3: string (nullable = true)
 |-- custom_4: string (nullable = true)
 |-- device_new: boolean (nullable = true)
 |-- first-accessed-page: string (nullable = true)
 |-- install_uuid: string (nullable = true)
 |-- language: string (nullable = true)
 |-- library_ver: string (nullable = true)
 |-- marketing_campaign: string (nullable = true)
 |-- marketing_medium: string (nullable = true)
 |-- marketing_source: string (nullable = true)
 |-- model: string (nullable = true)
 |-- name: string (nullable = true)
 |-- nth: long (nullable = true)
 |-- os_ver: string (nullable = true)
 |-- platform: string (nullable = true)
 |-- region: string (nullable = true)
 |-- session_uuid: string (nullable = true)
 |-- studentId_clientType: string (nullable = true)
 |-- type: string (nullable = true)
 |-- user_type: string (nullable = true)
 |-- uuid: string (nullable = true)
 |-- students_id: string (nullable = true)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_referencia: integer (nullable = true)
 |-- pt_mes_referencia: integer (nullable = true)
```

In [350…

```
dfEvents.head()
```

Out[350…

```
Row(Last Accessed Url='/', Page Category='perfil', Page Category 1='perfil', Page Cate
gory 2='Undefined', Page Category 3='Undefined', Page Name='/perfil/22482764/materiai
s', at='2017-11-16 02:10:20', browser='Chrome 62', carrier='Telemar Norte Leste S.A.',
city_name=None, clv_total=None, country='br', custom_1='ESTÁCIO EAD', custom_2='Pedago
gia', custom_3=None, custom_4='Core User', device_new=False, first-accessed-page=None,
install_uuid='fdfff303505f8a18b17ee40587e785f6bb9c8374', language='pt', library_ver='w
eb_3.3.3', marketing_campaign=None, marketing_medium=None, marketing_source=None, mode
l='Linux armv7l', name='Page View', nth=17, os_ver='', platform='Linux', region=None,
session_uuid='188031bec37fc43b737c2c49349076700ae89128', studentId_clientType='34cbeaf
4a28c798de94cd9afb43d4e2e49ce80d6b52364e097371db586d4ea48@Website', type='e', user_typ
e='known', uuid='1b3ed1360694ceae79f6361ed11b03cf245311c8', students_id='34cbeaf4a28c7
98de94cd9afb43d4e2e49ce80d6b52364e097371db586d4ea48', ts_processamento=datetime.dateti
me(2021, 3, 29, 1, 25, 29, 281000), pt_ano_referencia=2017, pt_mes_referencia=11)
```

Table name bi_qtd_acess_users_all

O que os usuários acessam na plataforma e como acessam?

Agrupado o campo "platform" com diversas versões de Windows para "Windows"

In [331…

```
#Table name: bi_qtd_acess_users_all
#Obs: Agrupado diferentes versões de Windows 'N', para Windows.
from pyspark.sql import functions as F
dfQtdAcessUserAll = dfEvents\
.withColumn("platform",
            when(F.col("platform").like('Windows%'), "Windows")
            .otherwise(col("platform"))
           )\
.groupBy(
    to_date("at").alias("day_event"),
    "Page Name",
    "Page Category",
    "Page Category 1",
    "Page Category 2",
    "Page Category 3",
    "model",
    "platform")\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy(desc("qtd"),
        "Page Name")\
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_mes_referencia", substring("day_event", 1,7))
```

In [332…

```
dfQtdAcessUserAll.show()
```

```
+----------+--------------------+-----------------+-----------------+--------------
+--------------+------------+--------+-----+--------------------+--------------------
-+
| day_event|           Page Name|    Page Category|  Page Category 1|Page Category 2
|Page Category 3|       model|platform|  qtd|    ts_processamento|pt_ano_mes_referenci
a|
+----------+--------------------+-----------------+-----------------+--------------
+--------------+------------+--------+-----+--------------------+--------------------
-+
|2017-11-16|                   /|             home|        Undefined|      Undefined
|      Undefined|       Win32| Windows|51411|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|    /cadastro/passo1|         cadastro|         cadastro|      Undefined
|      Undefined|       Win32| Windows|18436|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|    /cadastro/passo2|         cadastro|         cadastro|      Undefined
|      Undefined|       Win32| Windows|17775|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|   /cadastro/premium|         cadastro|         cadastro|      Undefined
|      Undefined|       Win32| Windows|12454|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|             /listas|           listas|           listas|      Undefined
|      Undefined|       Win32| Windows| 9166|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|                   /|             home|             null|           null
|           null|       Win32| Windows| 8047|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|            /premium|          premium|             null|           null
|           null|       Win32| Windows| 6003|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|/configuracoes/conta|    configuracoes|    configuracoes|      Undefined
|      Undefined|       Win32| Windows| 4727|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|                   /|             home|             null|           null
|           null|Linux armv7l| Android| 3244|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16| /explorar-materiais|explorar-materiais|explorar-materiais|      Undefined
|      Undefined|       Win32| Windows| 3213|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|/disciplina/seguindo|       disciplina|       disciplina|       seguindo
|      Undefined|       Win32| Windows| 3086|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|/pagamento/premiu...|        pagamento|        pagamento|      Undefined
|      Undefined|       Win32| Windows| 2609|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|       /ranking/curso|          ranking|          ranking|      Undefined
|      Undefined|       Win32| Windows| 2350|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|                   /|             home|        Undefined|      Undefined
|      Undefined|       Win64| Windows| 2266|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|   /listas/historico|           listas|           listas|      Undefined
|      Undefined|       Win32| Windows| 2214|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|/pagamento/premiu...|        pagamento|        pagamento|      Undefined
|      Undefined|       Win32| Windows| 2118|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|    /cadastro/passo1|         cadastro|             null|           null
|           null|Linux armv7l| Android| 1949|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|              /busca|            busca|            busca|      Undefined
|      Undefined|       Win32| Windows| 1341|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|/configuracoes/me...|    configuracoes|    configuracoes|      Undefined
|      Undefined|       Win32| Windows| 1335|2021-03-29 01:12:...|              2017-1
1|
|2017-11-16|    /cadastro/passo2|         cadastro|             null|           null
|           null|Linux armv7l| Android| 1231|2021-03-29 01:12:...|              2017-1
1|
+----------+--------------------+-----------------+-----------------+--------------
+--------------+------------+--------+-----+--------------------+--------------------
-+
only showing top 20 rows
```

In [255…

```
# Consulta: Após o agrupamento, temos as seguintes plataformas.
dfQtdAcessUserAll\
.groupBy("platform")\
.count()\
.show()
```

```
+-------------+------+
|     platform| count|
+-------------+------+
|          iOS| 10499|
|        Linux|  2490|
|        Other|    34|
|       Fedora|    38|
|    Chrome OS|   535|
|BlackBerry OS|     2|
|       Ubuntu|  1051|
|      Android| 24822|
|     Mac OS X|  3925|
|      Windows|185689|
+-------------+------+
```

In [261…

```
dfEvents\
.groupBy("user_type")\
.count()\
.show()
```

```
+---------+------+
|user_type| count|
+---------+------+
|anonymous| 52676|
|    known|646266|
+---------+------+
```

Table name bi.events_premium

O que os usuários premium mais acessam? De qual região? Qual Universidade e Curso?

In [366…

```
dfQtEventsUserAcess = dfEvents\
.filter(col("user_type") == "known")\
.filter(F.col("students_id").isNotNull())\
.withColumn("platform",
            when(F.col("platform").like('Windows%'), "Windows")
            .otherwise(col("platform"))
           )\
.groupBy(
    to_date("at").alias("day_event"),
    "Page Name",
    "Page Category",
    "Page Category 1",
    "Page Category 2",
    "Page Category 3",
    "platform",
    "students_id")\
.count()\
.withColumnRenamed("count", "qtd")\
.orderBy(desc("qtd"),
        "Page Name")\
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_mes_referencia", substring("day_event", 1,7))
```

In [376…

```
#O que os usuários premium mais acessam? De qual região? Qual Universidade e Curso?
#Table name: bi.events_premium
dfEventsPremium = dfQtEventsUserAcess.alias("eventsUsers")\
    .join(dfCadastro.alias("cadastro"),
          trim(col("eventsUsers.students_id")) == trim(col("cadastro.students_id")),
          how = "inner")\
    .join(dfSubscriptions.alias("premium"),
         trim(col("cadastro.students_id")) == trim(col("premium.StudentId")),
          how = "inner"
         )\
.select(
    col("eventsUsers.day_event"),
    col("eventsUsers.Page Name").alias("page_name"),
    col("eventsUsers.Page Category").alias("page_categoria"),
    col("eventsUsers.Page Category 1").alias("page_categoria_1"),
    col("eventsUsers.Page Category 2").alias("page_categoria_2"),
    col("eventsUsers.Page Category 3").alias("page_categoria_3"),
    col("eventsUsers.platform"),
    col("eventsUsers.students_id"),
    col("cadastro.students_state"),
    col("cadastro.students_city"),
    col("cadastro.universities_id"),
    col("cadastro.universities_name"),
    col("cadastro.courses_id"),
    col("cadastro.courses_name"),
    to_date(col("premium.PaymentDate")).alias("payment_date"),
    col("premium.PlanType").alias("plan_type"),
    col("eventsUsers.qtd").alias("qtd_acesso")
).distinct()\
.orderBy(desc("qtd_acesso"))\
.withColumn("ts_processamento", current_timestamp())\
.withColumn("pt_ano_mes_referencia", substring("day_event", 1,7))
```

In [377…

```
dfEventsPremium.printSchema()
```

```
root
 |-- day_event: date (nullable = true)
 |-- page_name: string (nullable = true)
 |-- page_categoria: string (nullable = true)
 |-- page_categoria_1: string (nullable = true)
 |-- page_categoria_2: string (nullable = true)
 |-- page_categoria_3: string (nullable = true)
 |-- platform: string (nullable = true)
 |-- students_id: string (nullable = true)
 |-- students_state: string (nullable = true)
 |-- students_city: string (nullable = true)
 |-- universities_id: long (nullable = true)
 |-- universities_name: string (nullable = true)
 |-- courses_id: long (nullable = true)
 |-- courses_name: string (nullable = true)
 |-- payment_date: date (nullable = true)
 |-- plan_type: string (nullable = true)
 |-- qtd_acesso: long (nullable = false)
 |-- ts_processamento: timestamp (nullable = false)
 |-- pt_ano_mes_referencia: string (nullable = true)
```

In [378…

```
dfEventsPremium.show()
```

```
+----------+--------------------+--------------+----------------+--------------------+
----------------+--------+--------------------+----------------+-------------+------
---------+-----------------+----------+--------------------+------------+---------+---
-------+--------------------+---------------------+
| day_event|           page_name|page_categoria|page_categoria_1|    page_categoria_2|
page_categoria_3|platform|         students_id|  students_state| students_city|univer
sities_id|universities_name|courses_id|        courses_name|payment_date|plan_type|qtd
_acesso|    ts_processamento|pt_ano_mes_referencia|
+----------+--------------------+--------------+----------------+--------------------+
----------------+--------+--------------------+----------------+-------------+------
---------+-----------------+----------+--------------------+------------+---------+---
-------+--------------------+---------------------+
|2017-11-16|/disciplina/psico...|    disciplina|      disciplina|psicologia-instit...|
       Undefined| Windows|4dd2870e46a505511...|       Pernambuco|         null|
   664623|          ESTÁCIO|   1199704|           Pedagogia|  2017-11-13|   Mensal|
     93|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/pedag...|    disciplina|      disciplina|pedagogia-nas-ins...|
       Undefined| Windows|4dd2870e46a505511...|       Pernambuco|         null|
   664623|          ESTÁCIO|   1199704|           Pedagogia|  2017-11-13|   Mensal|
     38|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/metod...|    disciplina|      disciplina|metodologia-cient...|
       Undefined| Windows|235399080e00a71d6...|        São Paulo|         null|
   664623|          ESTÁCIO|   1199517|             Direito|  2017-11-06|   Mensal|
     26|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/gesta...|    disciplina|      disciplina| gestao-da-qualidade|
       Undefined| Windows|4dd2870e46a505511...|       Pernambuco|         null|
   664623|          ESTÁCIO|   1199704|           Pedagogia|  2017-11-13|   Mensal|
     22|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/matem...|    disciplina|      disciplina|matematica-financ...|
       Undefined| Windows|485429e1456118612...|   Rio de Janeiro|Rio de Janeiro|
   664623|          ESTÁCIO|   1199633|Gestão de Comérci...|  2017-11-08|   Mensal|
     16|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/ensin...|    disciplina|      disciplina|ensino-clinico-ii...|
       Undefined| Windows|2b084ab4c16977823...|            Ceará|         null|
   664623|          ESTÁCIO|   1199529|          Enfermagem|  2017-11-14|   Mensal|
     14|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/avali...|    disciplina|      disciplina|avaliacao-institu...|
       Undefined| Windows|4dd2870e46a505511...|       Pernambuco|         null|
   664623|          ESTÁCIO|   1199704|           Pedagogia|  2017-11-13|   Mensal|
     13|2021-03-29 10:36:...|              2017-11|
|2017-11-16|             /listas|        listas|          listas|           Undefined|
       Undefined| Windows|4dd2870e46a505511...|       Pernambuco|         null|
   664623|          ESTÁCIO|   1199704|           Pedagogia|  2017-11-13|   Mensal|
     13|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/direi...|    disciplina|      disciplina|   direito-penal-iii|
       Undefined| Windows|53d05d495983fd52d...|             null|         null|
   664623|          ESTÁCIO|   1199517|             Direito|  2017-11-07|   Mensal|
     13|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/calcu...|    disciplina|      disciplina|calculo-diferenci...|
       Undefined| Windows|d70d842713afce978...|             null|         null|
   664623|          ESTÁCIO|   1199553|Engenharia de Pro...|  2017-11-05|   Mensal|
     12|2021-03-29 10:36:...|              2017-11|
|2017-11-16|                   /|          home|       Undefined|           Undefined|
       Undefined| Windows|e574a0b44f0a00d23...|   Rio de Janeiro|         null|
   664623|          ESTÁCIO|   1199517|             Direito|  2017-11-16|   Mensal|
     11|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/logic...|    disciplina|      disciplina|logica-de-program...|
       Undefined| Windows|9f8a1157a0ded0df9...|            Goiás|       Goiânia|
  1614530|          UNINTER|   1199704|           Pedagogia|  2017-11-08|   Mensal|
     10|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/merca...|    disciplina|      disciplina|  mercado-financeiro|
       Undefined| Windows|5c04ed16fa75216c5...|        São Paulo|          Embu|
   664623|          ESTÁCIO|   1199644|   Gestão Financeira|  2017-11-04|   Mensal|
     10|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/model...|    disciplina|      disciplina|  modelagem-de-dados|
       Undefined| Windows|f68937b75647b3de6...|            Goiás|         null|
   664623|          ESTÁCIO|   1199745|Sistemas de Infor...|  2017-11-12|   Mensal|
     10|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/matem...|    disciplina|      disciplina|matematica-para-n...|
       Undefined| Windows|cfbb0c85eb4c8fdfc...|        São Paulo|         null|
   664623|          ESTÁCIO|   1199453|       Administração|  2017-11-16|   Mensal|
     10|2021-03-29 10:36:...|              2017-11|
|2017-11-16|                   /|          home|       Undefined|           Undefined|
       Undefined| Windows|6b4fc69dcf30183e1...|             null|         null|
   662052|       UNIASSELVI|   1199506|Contabilidade / C...|  2017-11-14|   Mensal|
     10|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/matem...|    disciplina|      disciplina|matematica-financ...|
       Undefined| Windows|26a81c536f61c79a5...|        São Paulo|         null|
  1614530|          UNINTER|  30724897|Gestão da Tecnolo...|  2017-11-14|   Mensal|
     10|2021-03-29 10:36:...|              2017-11|
|2017-11-16|                   /|          home|       Undefined|           Undefined|
       Undefined| Windows|cfbb0c85eb4c8fdfc...|        São Paulo|         null|
   664623|          ESTÁCIO|   1199453|       Administração|  2017-11-16|   Mensal|
      9|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/pagamento/premiu...|     pagamento|       pagamento|           Undefined|
       Undefined| Windows|e4b2f20b4afe40f43...|Rio Grande do Sul|    Rio Pardo|
  1614530|          UNINTER|   1199553|Engenharia de Pro...|  2017-11-27|    Anual|
      8|2021-03-29 10:36:...|              2017-11|
|2017-11-16|/disciplina/direi...|    disciplina|      disciplina|direito-constituc...|
       Undefined| Windows|4ed25e8e003f808f2...|   Santa Catarina| Florianópolis|
   664565|      ESTÁCIO EAD|   1199517|             Direito|  2017-11-03|   Mensal|
      7|2021-03-29 10:36:...|              2017-11|
+----------+--------------------+--------------+----------------+--------------------+
----------------+--------+--------------------+----------------+-------------+------
---------+-----------------+----------+--------------------+------------+---------+---
-------+--------------------+---------------------+
only showing top 20 rows
```