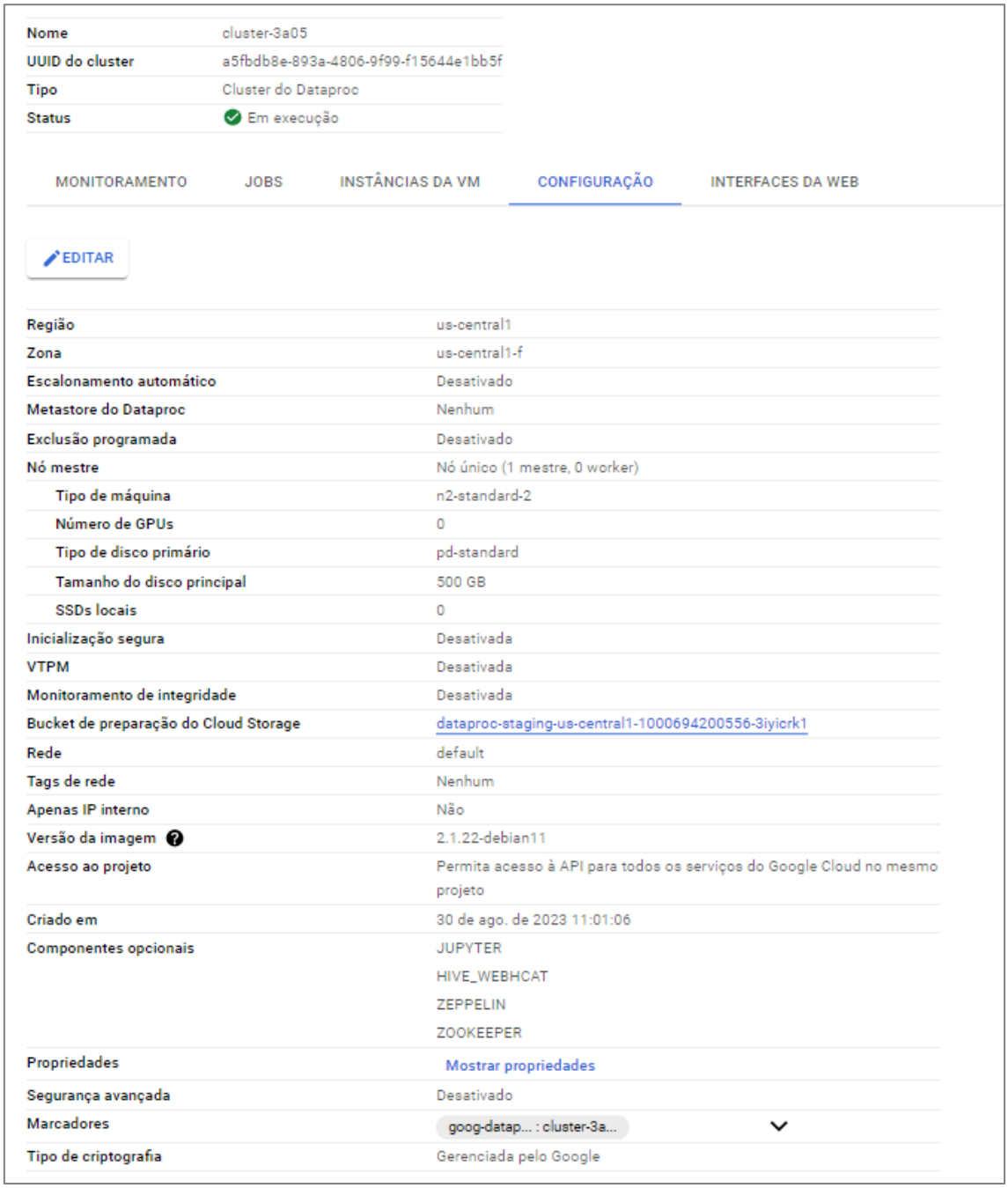
INSTITUTO INFNET
PROJETO FINAL DE MÓDULO – INFRAESTRUTURA HADOOP

ALUNO: FLÁVIO PRADO DE AQUINO
PROGESOR: ANDRÉ ORMASTRONI VICTOR

Tema: Análise de desempenho do ENEM do ano de 2022, em contrate com o IDH (Índice de Desenvolvimento Humano) por cidade.

1 – Definição e provisionamento de um Cluster Haddop no ambiente Google CDloud



Nome	cluster-3a05
UUID do cluster	a5fbdb8e-893a-4806-9f99-f15644e1bb5f
Tipo	Cluster do Dataproc
Status	Em execução

MONITORAMENTO | JOBS | INSTÂNCIAS DA VM | CONFIGURAÇÃO | INTERFACES DA WEB

EDITAR

Região	us-central1
Zona	us-central1-f
Escalonamento automático	Desativado
Metastore do Dataproc	Nenhum
Exclusão programada	Desativado
Nó mestre	Nó único (1 mestre, 0 worker)
Tipo de máquina	n2-standard-2
Número de GPUs	0
Tipo de disco primário	pd-standard
Tamanho do disco principal	500 GB
SSDs locais	0
Inicialização segura	Desativada
VTPM	Desativada
Monitoramento de integridade	Desativada
Bucket de preparação do Cloud Storage	dataproc-staging-us-central1-1000694200556-3iyicrk1
Rede	default
Tags de rede	Nenhum
Apenas IP interno	Não
Versão da imagem	2.1.22-debian11
Acesso ao projeto	Permita acesso à API para todos os serviços do Google Cloud no mesmo projeto
Criado em	30 de ago. de 2023 11:01:06
Componentes opcionais	JUPYTER HIVE_WEBHCAT ZEPPELIN ZOOKEEPER
Propriedades	Mostrar propriedades
Segurança avançada	Desativado
Marcadores	goog-datap... : cluster-3a...
Tipo de criptografia	Gerenciada pelo Google

2 – Subir os arquivos CSV do servidor local para o HDFS

```
hdfs dfs -put enem_2022_tratado.csv /user/flaviopradoaquino
hdfs dfs -put idh_2010_tratado.csv /user/flaviopradoaquino
```

3 – Provisionar o Beeline Hive

```
beeline -u jdbc:hive2://localhost:10000/default -n
flaviopradoaquino@cluster-3a05-m -d org.apache.hive.jdbc.HiveDriver
```

4 – Criar o banco de dados

```
create database if not exists db_enem
comment "Banco de dados Enem"
location "/user/flaviopradoaquino"
with DBPROPERTIES('Date' = '2023-08-30', 'Country' = 'BR' , 'Creator'
= 'Flavio Aquino');
```

```
0: jdbc:hive2://localhost:10000/default> show databases;
INFO  : Compiling command(queryId=hive_20230903202632_0f2649dc-6a2f-4d8a-8ac5-fcc9e61404d7): show databases
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)], properties:nu
ll)
INFO  : Completed compiling command(queryId=hive_20230903202632_0f2649dc-6a2f-4d8a-8ac5-fcc9e61404d7); Time taken: 0.019 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230903202632_0f2649dc-6a2f-4d8a-8ac5-fcc9e61404d7): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230903202632_0f2649dc-6a2f-4d8a-8ac5-fcc9e61404d7); Time taken: 0.016 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------+
| database_name  |
+----------------+
| db_enem        |
| default        |
+----------------+
2 rows selected (0.187 seconds)
0: jdbc:hive2://localhost:10000/default>
```

5 – Criar as tabelas que receberam os dados dos arquivos CSV

```
CREATE TABLE IF NOT EXISTS enem_2022(
NU_INSCRICAO BIGINT,
TP_SEXO STRING,
TP_COR_RACA INT,
TP_ESCOLA INT,
IN_TREINEIRO INT,
CO_MUNICIPIO_PROVA INT,
NO_MUNICIPIO_PROVA STRING,
SG_UF_PROVA STRING,
NU_NOTA_MT INT,
NU_NOTA_REDACAO INT
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```

```
CREATE TABLE IF NOT EXISTS idh_2010(
NM_FEDERACAO STRING,
NM_MUNICIPIO STRING,
IDHM DECIMAL,
IDH INT
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```

```
0: jdbc:hive2://localhost:10000/default> show tables;
INFO  : Compiling command(queryId=hive_20230903205314_59a72ba8-bd6a-441f-b4f8-f6c867afa28c): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230903205314_59a72ba8-bd6a-441f-b4f8-f6c867afa28c); Time taken: 0.022 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230903205314_59a72ba8-bd6a-441f-b4f8-f6c867afa28c): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230903205314_59a72ba8-bd6a-441f-b4f8-f6c867afa28c); Time taken: 0.011 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------+
|  tab_name  |
+------------+
| enem_2022  |
| idh_2010   |
+------------+
2 rows selected (0.056 seconds)
```

6 – Ler os dados dos HDFS para as tabelas do Hive

```
LOAD DATA INPATH '/user/flaviopradoaquino/idh_2010_tratado.csv'
overwrite into table idh_2010;
```

```
LOAD DATA INPATH '/user/flaviopradoaquino/enem_2022_tratado.csv'
overwrite into table enem_2022;
```

```
select count(*) from enem_2022
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.72 s
--------------------------------------------------------------------------------
+----------+
|   _c0    |
+----------+
| 1048576  |
+----------+
1 row selected (30.74 seconds)
```

```
select count(*) from idh_2010
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.49 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230903205835_f8369521-cfd0-4397-90c6-8dc9ed884a4b); Time taken: 12.816 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------+
|  _c0  |
+-------+
| 5565  |
+-------+
1 row selected (13.006 seconds)
```

7 – Consultas realizadas

>> Média das notas no Enem por cidade pareada com o IDH limitado a vinte saídas

```
select a.no_municipio_prova, AVG(a.nu_nota_redacao), b.idh
from enem_2022 a
join idh_2010 b on a.no_municipio_prova = b.nm_municipio
where a.nu_nota_redacao is not null
order by b.idh desc
group by a.no_municipio_prova, b.idh
limit 20;
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.62 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230831014836_4de4d2b3-a9f8-4068-8bfb-a3cd4ae232d1); Time taken: 14.4 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------+--------------------+--------+
| a.no_municipio_prova |        _c1         | b.idh  |
+----------------------+--------------------+--------+
| Alfenas              | 675.8847736625514  | 761    |
| Alvin�polis          | 642.2222222222222  | 676    |
| Al�m Para�ba         | 691.5189873417721  | 726    |
| Andrel�ndia          | 617.4242424242424  | 700    |
| Araguari             | 652.9347826086956  | 773    |
| Arax�                | 678.6111111111111  | 772    |
| Areado               | 677.1186440677966  | 727    |
| Baependi             | 665.0793650793651  | 681    |
| Almenara             | 607.6439790575917  | 642    |
| Alpin�polis          | 696.6666666666666  | 725    |
| Andradas             | 686.5384615384615  | 734    |
| Abaet�               | 661.7647058823529  | 698    |
| Aimor�s              | 649.8795180722891  | 684    |
| Ara�na�              | 645.1957295373666  | 663    |
| Arcos                | 710.8              | 749    |
| Arinos               | 588.0392156862745  | 656    |
| Bambu�               | 681.7142857142857  | 741    |
| Barbacena            | 674.5846153846154  | 769    |
| Bar�o de Cocais      | 664.9411764705883  | 722    |
| Belo Horizonte       | 692.0523639510981  | 810    |
+----------------------+--------------------+--------+
20 rows selected (14.994 seconds)
```

>> Top 10 dos melhores índices IDH

```
select NM_MUNICIPIO, IDH
from idh_2010
order by IDH desc
limit 10;
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.57 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230831013602_b286dbe0-d47e-4b7c-ae85-020b17d63333); Time taken: 9.914 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------------+------+
|    nm_municipio     | idh  |
+---------------------+------+
| S�O CAETANO DO SUL  | 862  |
| �GUAS DE S�O PEDRO  | 854  |
| FLORIAN�POLIS       | 847  |
| BALNE�RIO CAMBORI�  | 845  |
| VIT�RIA             | 845  |
| SANTOS              | 840  |
| NITER�I             | 837  |
| JOA�ABA             | 827  |
| BRAS�LIA            | 824  |
| CURITIBA            | 823  |
+---------------------+------+
10 rows selected (10.203 seconds)
```

>> Top 20 das melhores notas de redação por cidade e seu respectivo IDH

```
select a.no_municipio_prova, a.nu_nota_redacao, b.idh
from enem_2022 a
join idh_2010 b on a.no_municipio_prova = b.nm_municipio
where a.nu_nota_redacao is not null
order by a.nu_nota_redacao desc
limit 20;
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.09 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230831013214_ddc0cebb-1351-42c2-9847-85f8938587c2); Time taken: 14.721 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------+--------------------+--------+
| a.no_municipio_prova  | a.nu_nota_redacao  | b.idh  |
+-----------------------+--------------------+--------+
| Governador Valadares  | 1000               | 727    |
| S�o Louren�o          | 1000               | 759    |
| Montes Claros         | 1000               | 770    |
| S�o Jo�o del Rei      | 980                | 758    |
| Guaxup�               | 980                | 751    |
| Tim�teo               | 980                | 770    |
| Campo Belo            | 980                | 711    |
| S�o Jo�o del Rei      | 980                | 758    |
| Campo Belo            | 980                | 711    |
| S�o Jo�o del Rei      | 980                | 758    |
| Governador Valadares  | 980                | 727    |
| S�o Jo�o del Rei      | 980                | 758    |
| Nova Lima             | 980                | 813    |
| Guanh�es              | 980                | 686    |
| Tim�teo               | 980                | 770    |
| S�o Jo�o del Rei      | 980                | 758    |
| Nova Lima             | 980                | 813    |
| Campo Belo            | 980                | 711    |
| S�o Jo�o del Rei      | 980                | 758    |
| Nova Lima             | 980                | 813    |
+-----------------------+--------------------+--------+
20 rows selected (15.179 seconds)
```

>> Top 20 das melhores notas de redação por cidade e seu respectivo IDH

```
select a.SG_UF_PROVA, AVG(a.nu_nota_mt), AVG(b.idh)
from enem_2022 a
join idh_2010 b on a.no_municipio_prova = b.nm_municipio
where a.nu_nota_redacao is not null
group by a.SG_UF_PROVA
limit 20;
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.05 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230831015256_73d1f622-b68a-4d44-a7b0-8b32e562ded2); Time taken: 14.49 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------+--------------------+--------------------+
| a.sg_uf_prova  |        _c1         |        _c2         |
+----------------+--------------------+--------------------+
| BA             | 499.67487684729065 | 678.0              |
| PB             | 503.8965517241379  | 715.0              |
| PE             | 495.63793103448273 | 679.0              |
| PR             | 557.5698924731183  | 664.4203187250996  |
| SP             | 535.2739130434783  | 720.0              |
| AL             | 492.4607142857143  | 775.0              |
| MG             | 571.0769842704755  | 746.8417207792207  |
| MT             | 514.1058823529412  | 576.0              |
| RJ             | 516.3272727272728  | 656.0              |
| RO             | 506.8369565217391  | 672.0              |
| RR             | 452.2              | 637.0              |
| MA             | 506.26666666666665 | 715.0              |
+----------------+--------------------+--------------------+
12 rows selected (14.873 seconds)
```

>> Número de abstenção na prova de redação por cidade limitado a 20 saídas

```
select no_municipio_prova, count(*)
from enem_2022
where nu_nota_redacao is null
group by no_municipio_prova
limit 20;
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.52 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230903214329_d7605a25-ec83-47ab-8dd8-60175adc78a0); Time taken: 35.285 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------------------+------+
|   no_municipio_prova   | _c1  |
+------------------------+------+
| Acara?                 | 70   |
| Acari                  | 26   |
| Acopiara               | 60   |
| Adamantina             | 43   |
| Afogados da Ingazeira  | 49   |
| Afonso Cl?udio         | 27   |
| Agrestina              | 23   |
| Alagoa Grande          | 60   |
| Alagoinha              | 26   |
| Abaetetuba             | 398  |
| Abaet?                 | 31   |
| Abreu e Lima           | 172  |
| Acarape                | 34   |
| Acar?                  | 61   |
| Acrel?ndia             | 24   |
| Acre?na                | 81   |
| Afr?nio                | 54   |
| Agudos                 | 27   |
| Aimor?s                | 33   |
| Alagoa Nova            | 45   |
+------------------------+------+
20 rows selected (39.855 seconds)
```