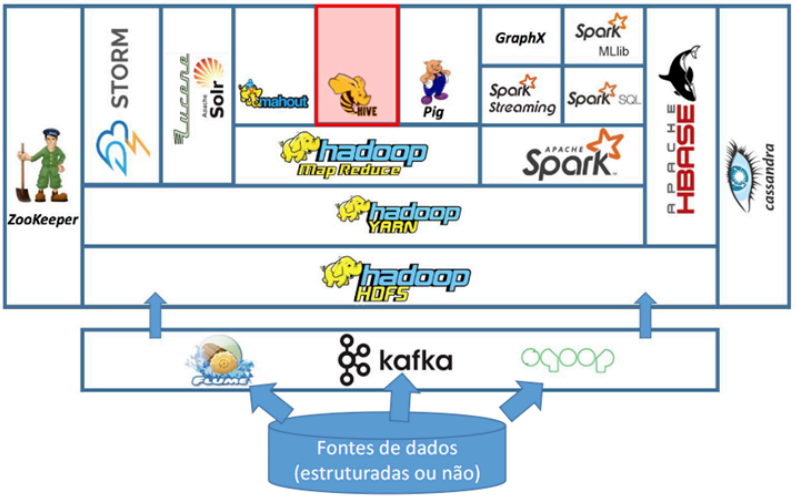
TRABALHO CONCLUSÃO DE CURSO

INFRAESTRUTURA HADOOP

Prof. Andre Victor | andre.victor@prof.infnet.edu.br

Aluno: Viviane de Sales Seródio | vivi.serodio@gmail.com

Visão geral da Arquitetura



Definição do Tema

O tema escolhido para a análise no curso foi 'COVID-19', com foco em dados relacionados a casos, óbitos, recuperações, vacinação e outros aspectos pertinentes.

Google Console: https://console.cloud.google.com/dataproc/clusters?project=infrahadoop
Fonte de dados: https://covid19.who.int/WHO-COVID-19-global-data.csv
Repositório Github:
Google Drive:
https://docs.google.com/document/d/1mRr9rHzZGyX2YsdXv_bDc-kLvP83vFOGofug0QsFtm0/edit?usp=sharing

Coleta e Preparação dos Dados

Fonte de Dados

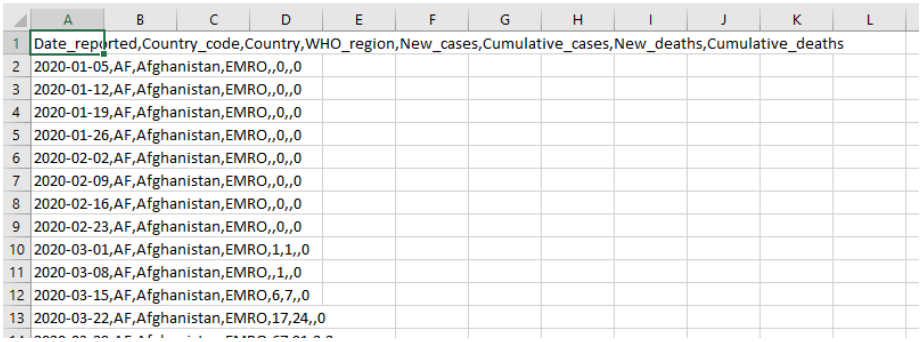
Os dados foram acessados por meio do repositório no GitHub 'Our World in Data' (OWID), que oferece uma ampla gama de informações sobre COVID-19 em formato CSV.
URL: https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

Preparação dos Dados

Efetuada a verificação e limpeza de dados para garantir que estivessem em um formato adequado para análise, incluindo a remoção de valores nulos, colunas irrelevantes, padronização de formatos, etc.
Essa preparação foi feita diretamente no Excel e em CSV para facilitar o carregamento no Hive.

Criação do arquivo

Os dados já estavam em formato estruturado (CSV).



	A	B	C	D	E	F	G	H	I	J	K	L
1	Date_reported,Country_code,Country,WHO_region,New_cases,Cumulative_cases,New_deaths,Cumulative_deaths											
2	2020-01-05,AF,Afghanistan,EMRO,,0,,0											
3	2020-01-12,AF,Afghanistan,EMRO,,0,,0											
4	2020-01-19,AF,Afghanistan,EMRO,,0,,0											
5	2020-01-26,AF,Afghanistan,EMRO,,0,,0											
6	2020-02-02,AF,Afghanistan,EMRO,,0,,0											
7	2020-02-09,AF,Afghanistan,EMRO,,0,,0											
8	2020-02-16,AF,Afghanistan,EMRO,,0,,0											
9	2020-02-23,AF,Afghanistan,EMRO,,0,,0											
10	2020-03-01,AF,Afghanistan,EMRO,1,1,,0											
11	2020-03-08,AF,Afghanistan,EMRO,,1,,0											
12	2020-03-15,AF,Afghanistan,EMRO,6,7,,0											
13	2020-03-22,AF,Afghanistan,EMRO,17,24,,0											

Análise de Dados

O dataset fornece um histórico completo de como a pandemia de COVID-19 se desenvolveu ao longo do tempo em diferentes países e regiões. Para cada data temos o número de novos casos e mortes, bem como os totais acumulados até aquele ponto.

A granularidade diária desses dados permite análises temporais detalhadas, como a observação de picos e quedas em casos ou mortes, além de possibilitar comparações entre diferentes regiões do mundo.

Entendendo os Dados

Tipos de Dados: Casos confirmados, mortes, recuperações, taxas de vacinação, variantes do vírus, dados demográficos, etc.

Esquema de Dados CSV - Dicionário de Dados

Caminho da fonte de dados: https://covid19.who.int/WHO-COVID-19-global-data.csv

Seq	Nome coluna	Descrição	Formato / Informações
1	Date_reported	Data da verificação	Data (YYYY-MM-DD)
2	Country_code	Código de país de duas letras	Exemplo: US para Estados Unidos BR para Brasil.
3	Country	Nome do país ou região.	-

4	WHO_region	Região da OMS a que o país pertence	Possíveis valores: AMRO (Américas) EMRO (Mediterrâneo Oriental) EURO (Europa) ARO (Sudeste AsiáticoWPRO (Pacífico Ocidental)
5	New_cases	Número de novos casos confirmados de COVID-19 relatados naquela data.	Número inteiro
6	Cumulative_cases	Número cumulativo de casos confirmados até aquela data.	Número inteiro
7	New_deaths	Número de novas mortes relatadas naquela data.	Número inteiro
8	Cumulative_deaths	Número cumulativo de mortes confirmadas até aquela data.	Número inteiro

Configuração do ambiente

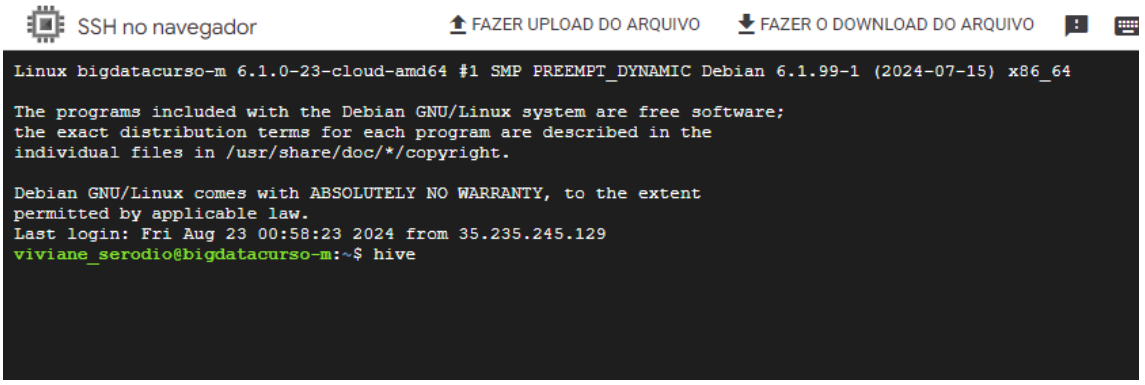
Configuração do Ambiente no Google Cloud

- Criado um projeto: Google Cloud Console
- Nome do projeto: InfraHadoop
- Criado cluster: bigdatacurso

Acessar o Cluster e o Hive

Acessar o Cluster via SSH

Acessar o Hive



Criar base de dados - Hive

Base de Dados que será criada: **dados_covid**

– Criar base de dados Dados_covid

```
CREATE DATABASE dados_covid;
USE dados_covid;
```

Criar Tabelas - Hive

Baseando-se na estrutura do arquivo CSV fornecido na URL, foram criadas duas tabelas.
Uma para armazenar dados diários de COVID-19 por país e outra para armazenar um resumo por país

Tabela 1: dados_diarios

Esta tabela armazenará os dados diários de casos, mortes e recuperados por país.

```
CREATE TABLE dados_diarios (
   data_reportada STRING,
   pais STRING,
   codigo_iso STRING,
   regiao STRING,
   novos_casos INT,
   acumulado_casos INT,
   novas_mortes INT,
   acumulado_mortes INT,
   novos_recuperados INT,
   acumulado_recuperados INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

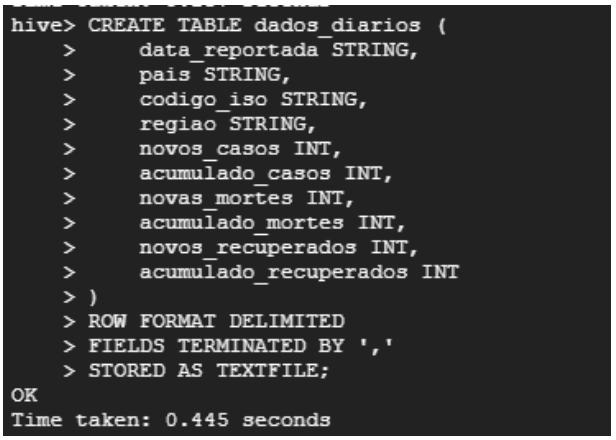


Tabela 2: resumo_pais

Esta tabela armazenará um resumo dos dados de COVID-19 por país, contém o resumo dos dados diários de COVID-19, agregados por país, com os totais máximos de casos, mortes e recuperados.

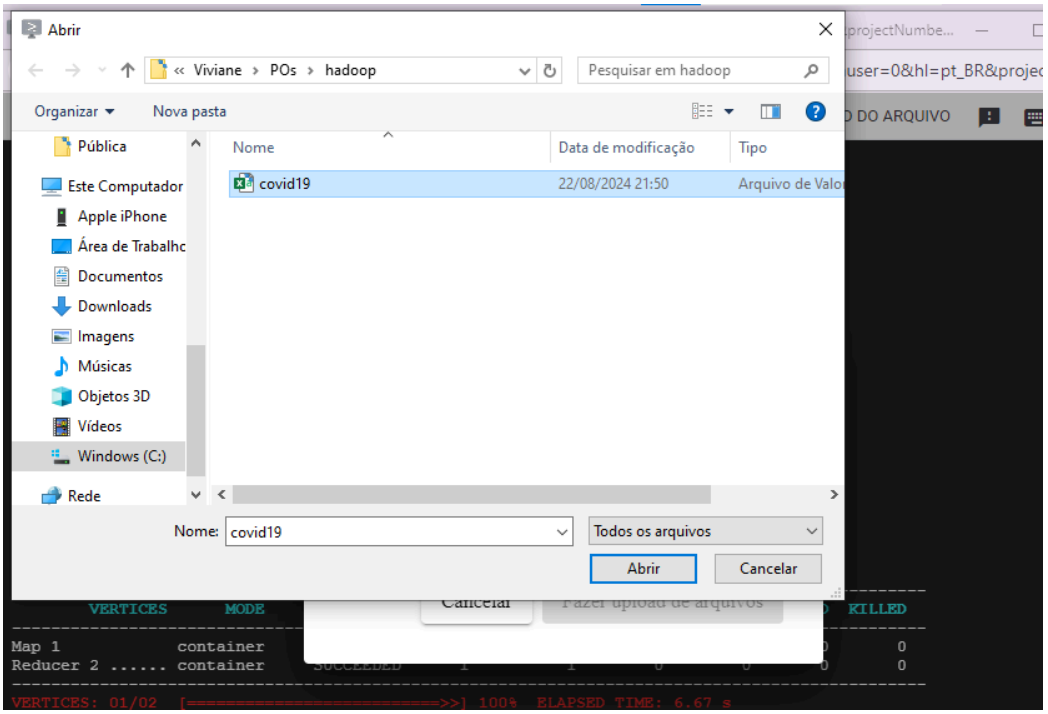
```
CREATE TABLE resumo_pais AS
SELECT
pais,
MAX(acumulado_casos) AS total_casos,
MAX(acumulado_mortes) AS total_mortes,
MAX(acumulado_recuperados) AS total_recuperados
FROM dados_diarios
GROUP BY pais;
```

```
hive> CREATE TABLE resumo_pais AS
    > SELECT
    >     pais,
    >     MAX(acumulado_casos) AS total_casos,
    >     MAX(acumulado_mortes) AS total_mortes,
    >     MAX(acumulado_recuperados) AS total_recuperados
    > FROM dados_diarios
    > GROUP BY pais;
Query ID = viviane_serodio_20240823014552_53238a18-e398-4e80-aba7-9cb6c30c2f70
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1724376868750_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1           container     SUCCEEDED      0          0        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/02  [==========================>>] 100%  ELAPSED TIME: 6.67 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdatacurso-m/user/hive/warehouse/dados_covid.db/resumo_pais
OK
Time taken: 18.889 seconds
hive>
```

Carregar os dados públicos do csv

A planilha já estava baixada na máquina com o tratamento mencionado no início do trabalho.



Análise de Dados - Perguntas Analíticas

Perguntas analíticas formuladas para serem respondidas com os dados disponíveis, via consultas:

1. **Qual o total de casos confirmados por país?**

Objetivo: Identificar o número total de casos confirmados em cada país.

Tipo de Análise: Agregação de dados por país.

2. **Quais os países com maior número de mortes?**

Objetivo: Determinar quais países registraram o maior número de mortes.

Tipo de Análise: Ordenação e filtragem dos dados de mortalidade por país.

3. **É possível analisar os novos casos por data específica?**

Objetivo: Explorar os novos casos relatados em uma data específica.

Tipo de Análise: Filtro temporal e agregação de dados.

4. **Qual a média de casos por país?**

Objetivo: Calcular a média de casos confirmados por país.

Tipo de Análise: Cálculo estatístico da média por país.

5. **Resumo de casos, mortes e recuperados por país**

Objetivo: Obter um resumo das estatísticas principais para cada país (casos, mortes, recuperações)

Tipo de Análise: Agregação múltipla de dados por país.

Consultas no Apache Hive

Consulta 1: Total de casos confirmados por país

```
SELECT pais, SUM(acumulado_casos) AS total_casos FROM dados_diarios GROUP BY pais
ORDER BY total_casos DESC;
```

Consulta 2: Países com o maior número de mortes

```
SELECT pais, SUM(acumulado_mortes) AS total_mortes FROM dados_diarios GROUP BY pais
ORDER BY total_mortes DESC;
```

Consulta 3: Novos casos por data específica / Ano (2022)

```
SELECT data_reportada, SUM(novos_casos) AS novos_casos_total FROM dados_diarios WHERE
data_reportada = '2024-08-01' GROUP BY data_reportada;
```

```
SELECT data_reportada, SUM(novos_casos) AS novos_casos_total FROM dados_diarios WHERE
```

YEAR(FROM_UNIXTIME(UNIX_TIMESTAMP(data_reportada, 'yyyy-MM-dd'))) = 2022 GROUP BY data_reportada;

Consulta 4: Média de novos casos por país
SELECT pais, AVG(novos_casos) AS media_novos_casos FROM dados_diarios GROUP BY pais ORDER BY media_novos_casos DESC;

Consulta 5: Resumo de casos, mortes e recuperados por país
SELECT * FROM resumo_pais ORDER BY total_casos DESC;

Proposta de Melhoria

A análise atual dos dados de COVID-19, proposta neste trabalho, fornece uma visão abrangente sobre casos confirmados, mortes e recuperações. No entanto, para um entendimento mais completo da situação pandêmica e entendimento da possibilidade de novos casos, é essencial integrar dados de vacinação ao estudo.
A inclusão de informações sobre taxas de vacinação por região, permitirá uma análise mais profunda que poderá prever a propagação do vírus e a gravidade dos casos.
Como proposta de melhoria, podemos investigar a correlação entre a vacinação e a redução de novos casos e mortes. Além disso, a análise de dados de vacinação pode ajudar a identificar disparidades regionais e avaliar a eficácia das campanhas de imunização.
Com mais essa abordagem, além de enriquecer a análise existente, proporcionará insights valiosos para a formulação de políticas de saúde pública mais eficazes.