PLANEJAMENTO DO PROJETO

SPEEDTEST BY OOKLA GLOBAL FIXED AND MOBILE NETWORK PERFORMANCE MAP

Integrantes do grupo:

Erik Assunção Figueiredo

Paulo Henrique de Souza Pereira Prazeres

Willian Alves Barboza

LISTA DE FIGURAS

SUMÁRIO

RESUMO

Este estudo aborda a análise de dados de testes de internet realizados por usuários ao longo de 2022, com o objetivo responder a três perguntas de pesquisa sobre o uso da internet no Brasil. Foram utilizadas várias ferramentas, como Python, Amazon S3, AWS Glue, Great Expectations GX e Microsoft PowerBI, para realizar a ingestão, processamento, armazenamento e visualização dos dados. Os dados foram distribuídos pela Ookla por meio de um bucket da AWS, e foram utilizados arquivos Shapefiles. O processo de ingestão foi automatizado e os dados foram tratados para garantir sua qualidade. Foi criado um modelo Star Schema, incluindo uma tabela de entrega na camada delivery, para facilitar a análise dos dados.

A análise revelou diferenças relevantes entre o uso de internet fixa e móvel, um aumento contínuo das métricas de desempenho ao longo dos trimestres e os estados com maior uso de internet foram identificados como DF, SP e RJ. O relatório e o painel de visualização dos dados estão disponíveis em um workspace pessoal do PowerBi com uma conta da USP, e os códigos utilizados estão no repositório do GitHub.

- GitHub: https://github.com/erikassuncao/Projeto-Integrador
- Painel PowerBi (Disponível por 30 dias): https://app.powerbi.com/links/UPtgMtwVpg?ctid=7e93e286-b29a-4454-a41a-e8419ec9deb5&pbi_source=linkShare

1. INTRODUÇÃO

A utilização da internet cresce em grande velocidade, uma prova disto é que em 2016, pelo menos 66% da população brasileira (10 anos ou mais de idade) teve acesso à internet, passando para 79% em 2019 e 84% em 2021 (IBGE,2022).

Dados de teste de internet, tais como latência, download e upload, são medidas de suma importância para avaliação da qualidade das conexões. Com essas informações, é possível identificar problemas, como velocidade lenta ou intermitência ou quedas de conexões. Provedores de internet utilizam esses dados constantemente para acompanhar a qualidade, identificar e até mesmo prever tais problemas, de forma a efetuar correções e manutenções preventivas.

O objetivo deste estudo é analisar os dados de testes de internet realizados por diversos usuários ao longo de 2022, no serviço online que fornece análise gratuita de métricas de desempenho de acesso à internet: Ookla, que pode ser acessado através do link: https://www.speedtest.net/pt. A fim de fazer a ingestão, processamento, armazenamento e apresentação dos dados de forma a propiciar visões analíticas que possam gerar insights, colocando em prática o conhecimento adquirido ao longo do curso “Engenharia de dados e Big Data na Escola Politécnica da USP”.

Fontes: (IBGE, PNAD Continua, 2022). disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101963_informativo.pdf

2. OBJETIVOS

O objetivo deste estudo é analisar dados de testes de internet realizados por diversos usuários ao longo de 2022 e distribuídos pela Ookla, a fim de responder 3 perguntas de pesquisa escolhidas em conjunto pelo grupo, através de um relatório criado após a ingestão, armazenamento e processamento dos dados, transformando-os em informação.

Após uma análise exploratória dos dados, inicialmente escolhemos 3 perguntas:

1. Existe uma diferença relevante de uso entre internet fixa e rede móvel?
2. O uso da internet muda dependendo do período do ano?
3. Quais são os estados com maior uso de internet e maior velocidade?

3. FERRAMENTAS ESCOLHIDAS

Optou-se por fazer o processo com ferramentas que estamos aprendendo, para aproveitar o trabalho e praticar mais. A seguir é exibida a lista com as ferramentas:

1. Ingestão de Dados
    - Python (https://www.python.org/)
    - Apache Spark (https://spark.apache.org/)

2. Processamento de Dados
    - Python (https://www.python.org/)
    - Apache Spark (https://spark.apache.org/)

3. Armazenamento de Dados
    - Amazon S3 (https://aws.amazon.com/pt/s3/)

4. Orquestração de Ferramentas
    - AWS Glue

5. Qualidade dos dados
    - Great Expectations GX (https://greatexpectations.io/gx-cloud)

6. Visualização
    - Microsoft PowerBI (https://powerbi.microsoft.com/pt-br/)

4. FONTES DE DADOS

Os dados são distribuídos pela Ookla através de um bucket da AWS, os dados são distribuídos com as opções de arquivos Parquet e Shapefiles. Optou-se por utilizar a segunda opção. Um arquivo shapefile consiste em vários arquivos com extensões diferentes que trabalham juntos para representar e armazenar os dados geoespaciais.

Os arquivos são disponibilizados trimestralmente, existindo 2 arquivos para cada trimestre: Um arquivo representando dados de internet Fixa e outro arquivo representando dados de internet Móvel.

Os arquivos utilizados no processo são:

- Ookla speedtest Mobile (Móvel)
- Ookla speedtest Fixed (Fixa)
- Malha territorial do Brasil (disponibilizado pelo IBGE, através do link: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html)

A seguir, encontra-se os detalhes das origens de dados com seus respectivos campos e descrições.

Speedtest Ookla (Móvel e Fixa)

Campo	Tipo	Descrição
avg_d_kbps	Integer	A velocidade média de download de todos os testes realizados no bloco, representada em kbps por segundo.
avg_u_kbps	Integer	A velocidade média de upload de todos os testes realizados no bloco, representada em kbps por segundo.
avg_lat_ms	Integer	A latência média de todos os testes realizados no bloco, representada em milissegundos
tests	Integer	O número de testes realizados no bloco.
devices	Integer	O número de dispositivos exclusivos que contribuem com testes no bloco.
quadkey	Text	Código Quadkey
Geometry	Geometry	conjunto de pontos espaciais que descrevem a forma e a localização dos objetos geográficos

Figura 1 - Descrição dos campos dos shapefiles "fixed e mobile"

Malha territorial do Brasil (IBGE)

Coluna	Tipo	Descrição
geometry	geometry	conjunto de pontos espaciais que descrevem a forma e a localização dos objetos geográficos
cd_mun	int	código do município
nm_mun	varchar	nome do município
sigla_uf	varchar	sigla do município
area_km2	int	tamanho da área em km2

Figura 2 - Descrição dos campos da malha territorial do Brasil

5. ARQUITETURA

Após uma análise exploratória das fontes de dados disponíveis necessárias para a construção dos painéis, optou-se por utilizar arquitetura AWS, orquestrando o processamento pela AWS Glue, com a arquitetura a seguir:

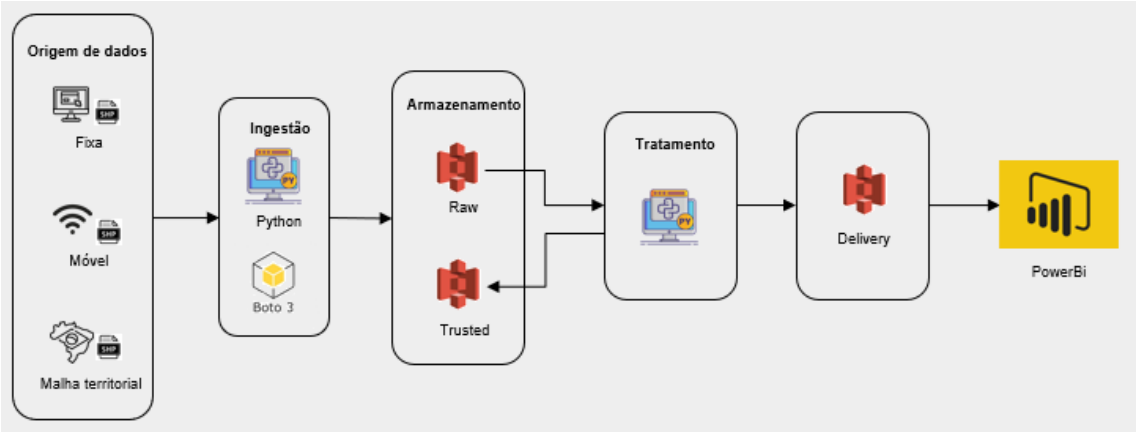

Figura 3 - Arquitetura de dados no AWS Glue

6. FLUXO DE DADOS E MODELO STAR SCHEMA

Ingestão Raw

O processo foi construído de forma com que seja capaz de identificar automaticamente se existem novos arquivos disponibilizados pela Ookla, com o auxílio de tabelas de logs, e importa somente o necessário. A fim de economizar recursos com o processamento de dados, a ingestão dos arquivos shapefiles, é realizada com o auxílio da malha territorial do brasil, para filtrar apenas dados necessários para o Bucket na S3 da AWS, pois a nossa análise requer apenas dados do Brasil.

Ingestão Raw

O processo foi construído de forma com que seja capaz de identificar automaticamente se existem novos arquivos disponibilizados pela Ookla, com o auxílio de tabelas de logs, e importa somente o necessário. A fim de economizar recursos com o processamento de dados, a ingestão dos arquivos shapefiles, é realizada com o auxílio da malha territorial do brasil, para filtrar apenas dados necessários para o Bucket na S3 da AWS, pois a nossa análise requer apenas dados do Brasil.

	quadkey	avg_d_kbps	avg_u_kbps	avg_lat_ms	tests	devices	geometry	CD_MUN	fonte
0	0323230233223102	15591	17249	26	4	2	POLYGON ((-60.75439 2.84978, -60.74890 2.84978...	1400100	https://ookla-open-data.s3-us-west-2.amazonaws...
1	0323230233232230	16998	20375	25	2	1	POLYGON ((-60.72144 2.82234, -60.71594 2.82234...	1400100	https://ookla-open-data.s3-us-west-2.amazonaws...
2	0323232011001023	17268	8038	29	2	2	POLYGON ((-60.77087 2.79491, -60.76538 2.79491...	1400100	https://ookla-open-data.s3-us-west-2.amazonaws...
3	0323230233232011	116923	4563	36	1	1	POLYGON ((-60.71594 2.85526, -60.71045 2.85526...	1400100	https://ookla-open-data.s3-us-west-2.amazonaws...
4	0323230233232033	5092	7704	31	2	2	POLYGON ((-60.71594 2.83880, -60.71045 2.83880...	1400100	https://ookla-open-data.s3-us-west-2.amazonaws...

Figura 4 - Camada Raw speedtest

CD_MUN	NM_MUN	SIGLA_UF	AREA_KM2	geometry
1100015	Alta Floresta D'Oeste	RO	7067.127	POLYGON ((-62.00806 -12.13379, -62.00784 -12.2...
1100023	Ariquemes	RO	4426.571	POLYGON ((-63.17933 -10.13924, -63.17746 -10.1...
1100031	Cabixi	RO	1314.352	POLYGON ((-60.52408 -13.32137, -60.37162 -13.3...
1100049	Cacoal	RO	3793.000	POLYGON ((-61.35502 -11.50452, -61.35524 -11.5...
1100056	Cerejeiras	RO	2783.300	POLYGON ((-60.82135 -13.11910, -60.81773 -13.1...

Figura 5 - Camada Raw Malha Territorial do Brasil

Ingestão Trusted

A ingestão de dados da camada Raw para a camada trusted contempla uma série de tratamento de dados e inclui algumas das principais etapas do processo de qualidade de dados, com o objetivo de manter a qualidade dos dados de forma constante durante a execução dos processos.

- Extração dos dados de latitude e longitude
- Verificação se os dados de latitude e longitude estão no raio do Brasil
- Considerando que as bases de dados disponibilizadas pela Ookla não possuem uma coluna para identificação de data, foi feito a criação da coluna de data possuindo trimestre de disponibilização do dado.
- Verificação do tipo de dados de geometria, pois diferentes arquivos shapefiles podem conter diferentes formatos de dados geométricos. Precisa-se garantir que todos encontram-se no mesmo formato.

	quadkey	avg_d_kbps	avg_u_kbps	avg_lat_ms	tests	devices	CD_MUN	geometry	latitude	longitude	date	quarter	network_type
0	0323230233223102	15591	17249	26	4	2	1400100	POLYGON ((-60.75439 2.84978, -60.74890 2.84978...	2.847033	-60.751648	2022-01-01	1	mobile
1	0323230233232230	16998	20375	25	2	1	1400100	POLYGON ((-60.72144 2.82234, -60.71594 2.82234...	2.819601	-60.718689	2022-01-01	1	mobile
2	0323232011001023	17268	8038	29	2	2	1400100	POLYGON ((-60.77087 2.79491, -60.76538 2.79491...	2.792168	-60.768127	2022-01-01	1	mobile
3	0323230233232011	116923	4563	36	1	1	1400100	POLYGON ((-60.71594 2.85526, -60.71045 2.85526...	2.852520	-60.713196	2022-01-01	1	mobile
4	0323230233232033	5092	7704	31	2	2	1400100	POLYGON ((-60.71594 2.83880, -60.71045 2.83880...	2.836060	-60.713196	2022-01-01	1	mobile

Figura 6 - Camada trusted speedtest

CD_MUN	NM_MUN	SIGLA_UF	AREA_KM2	geometry
1100015	Alta Floresta D'Oeste	RO	7067.127	POLYGON ((-62.00806 -12.13379, -62.00784 -12.2...
1100023	Ariquemes	RO	4426.571	POLYGON ((-63.17933 -10.13924, -63.17746 -10.1...
1100031	Cabixi	RO	1314.352	POLYGON ((-60.52408 -13.32137, -60.37162 -13.3...
1100049	Cacoal	RO	3793.000	POLYGON ((-61.35502 -11.50452, -61.35524 -11.5...
1100056	Cerejeiras	RO	2783.300	POLYGON ((-60.82135 -13.11910, -60.81773 -13.1...

Figura 7 - Camada Trusted Malha Territorial do Brasil

Ingestão Delivery e Star Schema

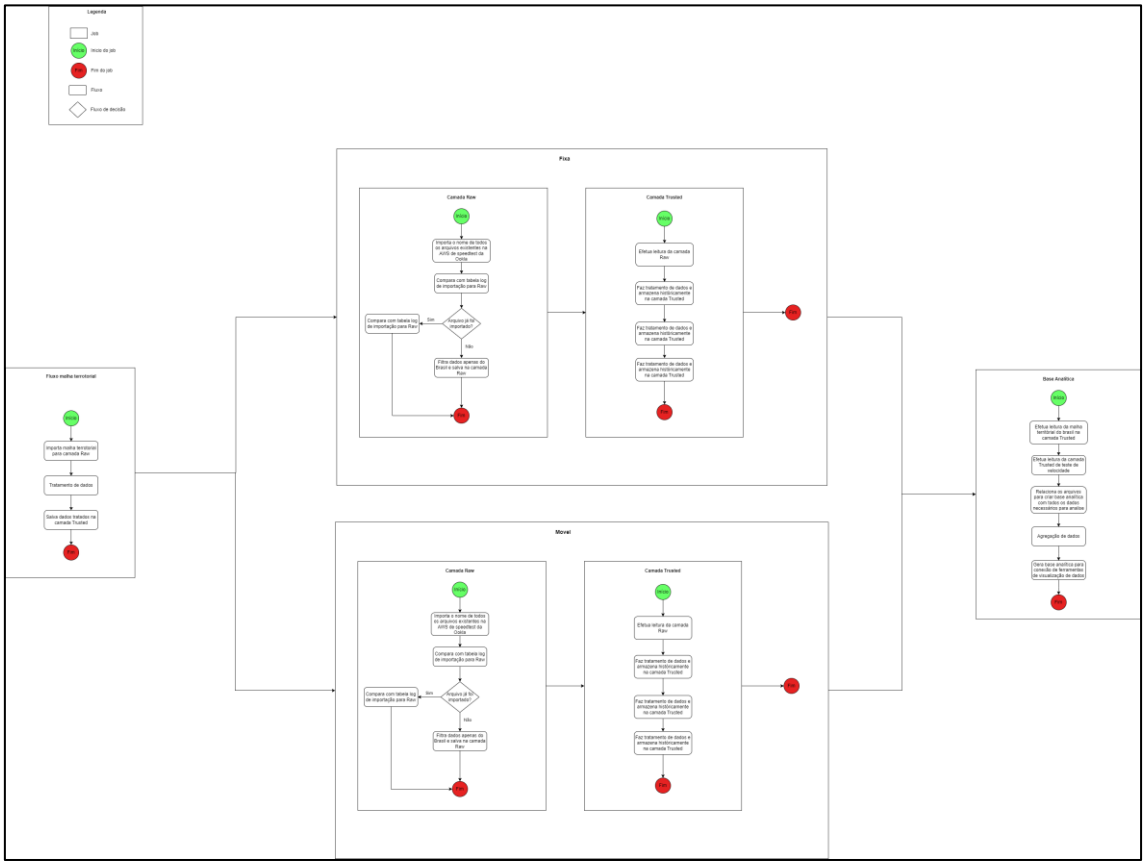

Figura 8 - Fluxo de dados

Embora tenha-se criado um modelo Star Schema que possui a fato speedtest que contém os dados numéricos das camadas trusted, referente a internet fixa e móvel, e as dimensões de calendário que possui dados categóricos de período, e a dimensão geográfica, possuindo dados dos estados e municípios brasileiros.

Com o objetivo de disponibilizar para os usuários uma base que seja fácil de analisar, caso necessário e minimizar o tempo de atualização e resposta dos painéis, foi criado uma tabela para a camada de Delivery que já possui os devidos relacionamentos, contemplando os dados da tabela fato e dimensões em um único lugar. Essa técnica é muito útil quando se sabe as análises que precisam ser feitas e que não necessitem de muitas colunas das tabelas dimensões. A imagem a seguir ilustra o modelo Star Schema e a camada delivery utilizada nos painéis:

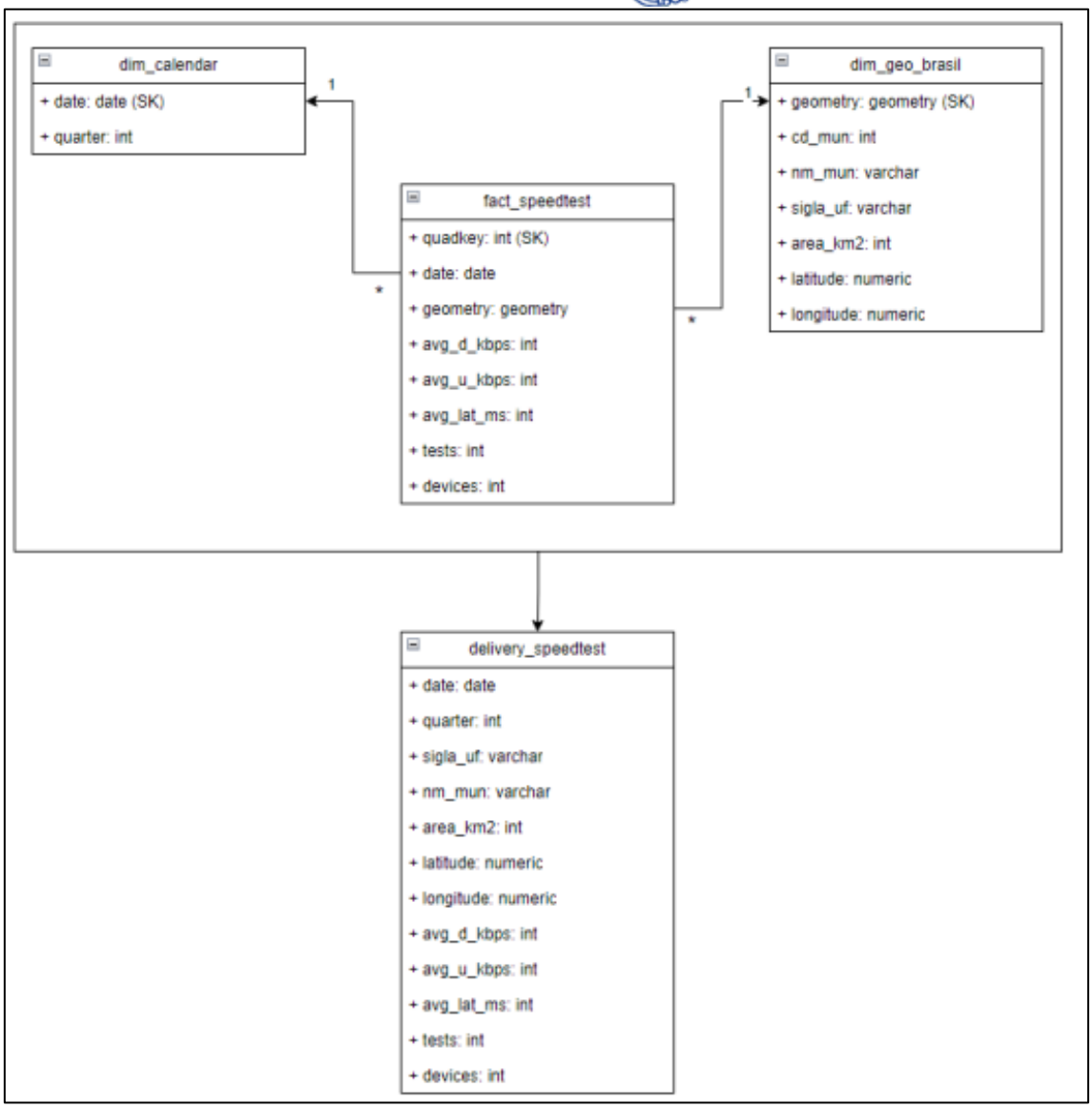

Figura 9 - Modelo Star Schema e camada Delivery

7. APRESENTAÇÃO DETALHADA DO PAINEL

O painel foi criado através da ferramenta de visualização de dados PowerBi. A premissa do painel foi responder as 3 questões bases do projeto, mencionadas no capítulo de Objetivos, são elas:

1. Existe uma diferença relevante de uso entre internet fixa e rede móvel?
2. O uso da internet muda dependendo do período do ano?
3. Quais são os estados com maior uso de internet e maior velocidade?

O painel recebeu o nome de “Painel de Acompanhamento trimestral de teste de velocidade” e possui uma capa a qual descreve brevemente, sobre do que se trata aquele relatório, assim como os nomes dos autores e possui botões que levam o usuário para a página que ele clicar.

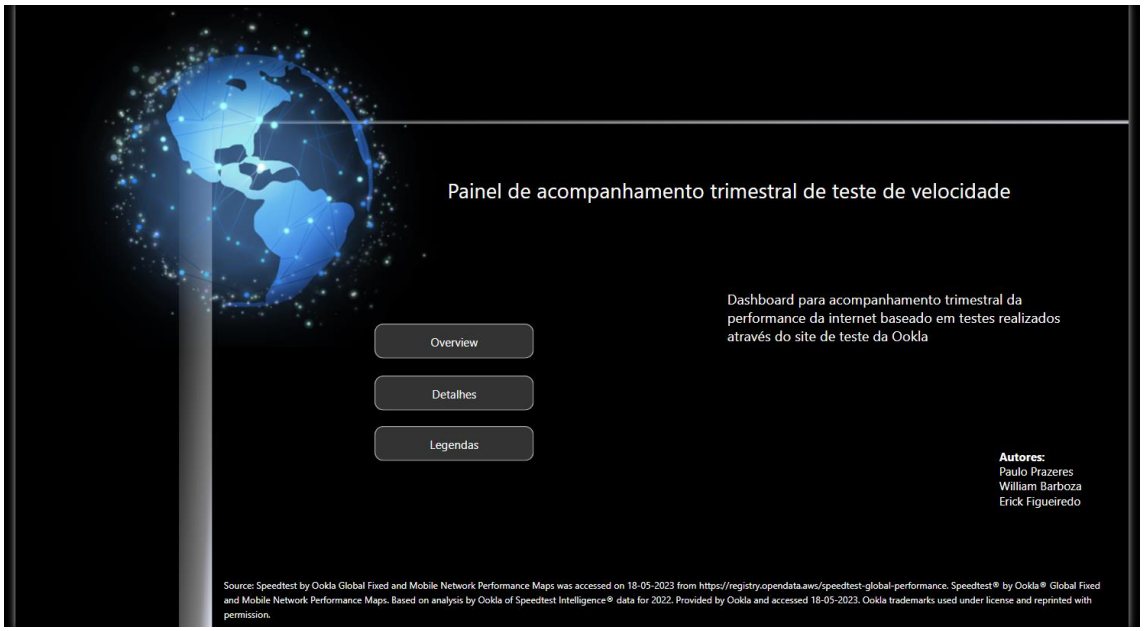

Figura 10 - Painel - Capa

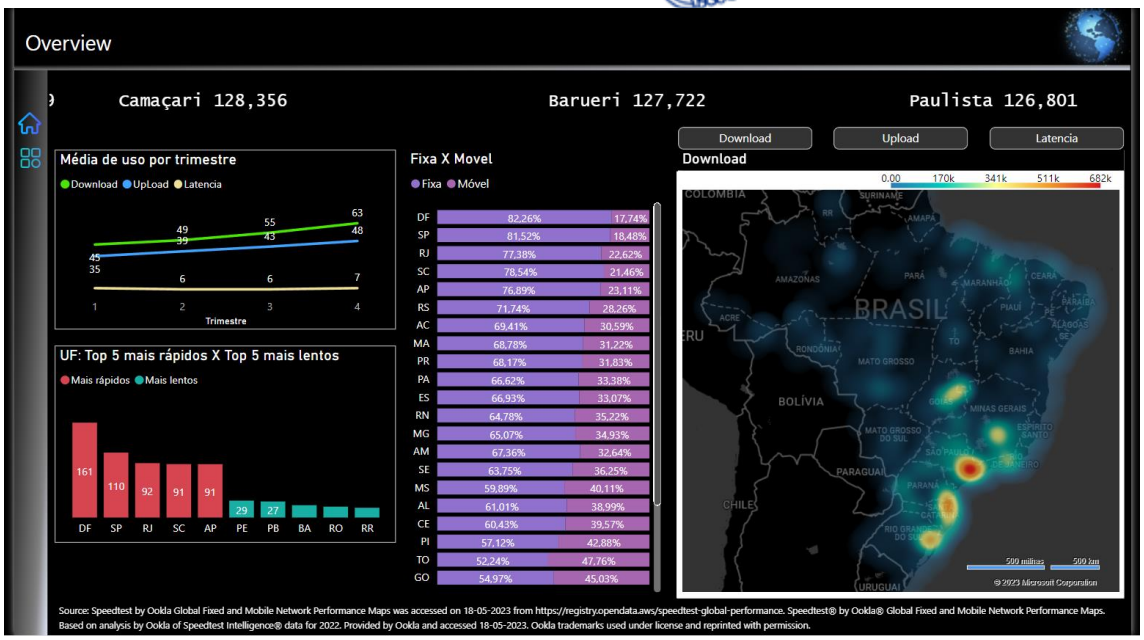

Figura 11 - Painel

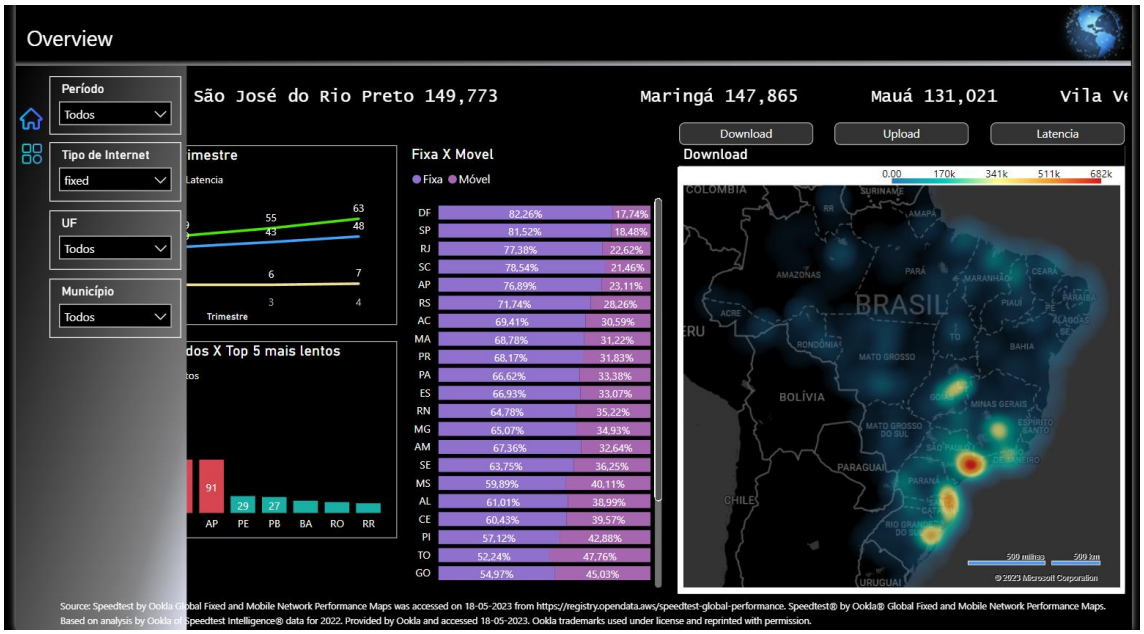

Figura 12 - Filtros do Painel

Os filtros do painel podem ser acessados através de um botão no canto esquerdo e são exibidos apenas se o usuário precisar, objetivando poupar espaço no painel.

Existem 4 principais gráficos criados, descritos a seguir:

- Gráfico de linhas, nomeado de: “Média de uso por trimestre”
  - Possui a média de uso de Download, Upload e Latencia, categorizado pelo trimestre do ano
- Gráfico de barras: “UF: Top 5 mais rápidos X Top 5 mais lentos”
  - Como o próprio nome diz, este gráfico possui os 5 estados mais rápidos exibidos em vermelho e os 5 estados mais lentos, exibidos na cor verde
- Gráfico de barras lateral: “Fixa X Movel”
  - Porcentagem de teste realizados com internet fixa em comparação a internet móvel no ano
- Mapa de calor: Download, Upload e Latencia
  - O mapa de calor pode ser analisado pelas 3 principais métricas do projeto: Download, Upload e Latencia
  - Foi utilizado degradê vermelho (valor alto) e azul (valor baixo) para criação do gráfico. Quanto mais alto é o valor, mais forte e próxima de vermelho é a cor, e consequentemente, quanto menor o valor, mais próximo do azul claro.

8. CONCLUSÃO

Ao analisar o painel criado, pode-se constatar que existe uma diferença considerável de uso entre a internet fixa e a internet móvel na maioria dos estados do Brasil.

Existe um aumento significativo e de forma contínua de Download, Upload e Latência ao longo dos trimestres no ano, não obstante, para uma análise mais completa, é necessário analisar dados de anos anteriores. Afinal, com o advento da tecnologia e crescimento constante do uso de internet, é de se esperar um aumento contínuo.

Por fim, os estados com maior uso de internet são DF, SP e RJ, os estados mais populosos e com maior infraestrutura.

9. REPOSITÓRIO GITHUB

https://github.com/erikassuncao/Projeto-Integrador