**Disciplina: Engenharia de Dados (40530010057\_20230\_01)**

**Aluno: Weverson Machado de Oliveira**

**Objetivo**

O objetivo principal deste trabalho é realizar uma análise abrangente das transferências de recursos públicos federais efetuadas no exercício de agosto de 2023, utilizando como fonte de dados o Portal da Transparência do Governo Federal. Buscamos, através desta investigação, proporcionar uma compreensão mais profunda das características, distribuições e destinações desses recursos, contribuindo para uma maior transparência e responsabilidade fiscal.

Antes de iniciar a busca pelos dados, propomo-nos a resolver os seguintes questionamentos:

1. Quais são os principais tipos de transferências realizadas e seus respectivos volumes de recursos?
2. Como essas transferências estão distribuídas geograficamente nas Unidades Federativas?
3. Quais são as áreas (funções e subfunções) mais beneficiadas pelas transferências de recursos?
4. Como os recursos são distribuídos entre os diferentes tipos de favorecidos?
5. Há correlação entre o valor transferido e a natureza ou localização do favorecido?

Traçando de maneira clara e objetiva essas indagações, estabelecemos um norte para nosso trabalho, garantindo que cada etapa subsequente da pesquisa seja alinhada ao propósito de elucidar essas questões. Uma vez definidos o objetivo e os questionamentos a serem respondidos, procedemos à busca e organização dos dados necessários para desenvolver nosso MVP (Produto Mínimo Viável).

**Busca pelos Dados**

O primeiro passo do nosso trabalho consistiu em realizar uma busca por dados que fossem relevantes e confiáveis. Para isso, recorremos ao Portal da Transparência do Governo Federal, que é uma iniciativa destinada a oferecer informações detalhadas sobre a execução financeira e orçamentária do governo.

O link específico utilizado para a obtenção dos dados foi: Portal da Transparência – Recursos Transferidos (https://portaldatransparencia.gov.br/download-de-dados/transferencias). Nesse portal, tivemos a oportunidade de acessar e baixar um conjunto de dados referente às transferências realizadas no exercício de agosto de 2023. Os dados disponíveis neste portal são detalhados e abrangem diversas categorias e subcategorias, oferecendo uma visão ampla das transações financeiras realizadas pelo governo.

Os campos disponíveis no conjunto de dados incluem, mas não se limitam a: Tipo de Transferência, Tipo de Favorecido, UF (Unidade Federativa), Código e Nome do Município, Códigos e Nomes de Função, Subfunção, Programa, e Ação, entre outros. Cada campo fornece informações específicas, por exemplo, o campo "Tipo de Transferência" pode ser classificado como 'Constitucionais e Royalties’, ‘Legais’, ‘Voluntárias e Específicas’, ‘Direta ao Cidadão’, ‘Múltiplo', enquanto o "Tipo de Favorecido" informa a natureza do beneficiário da transferência, podendo ser Pessoa Física, Administração Pública, Entidades sem Fins Lucrativos, etc (Figura 1).

Interface gráfica do usuário, Texto, Aplicativo, Word, Email

Descrição gerada automaticamente

Figura 1 – Dicionário de Dados – Recursos Transferidos

**Coleta**

Após o download do site, os dados sofreram um tratamento para remoção de acentos, visando aprimorar sua compatibilidade e manipulação. Posteriormente, foram carregados em um Bucket no Google Cloud, denominado "mvp\_bucket\_weverson", assegurando a integridade e acessibilidade das informações. O conjunto de dados tratado está contido no arquivo nomeado "202308\_TransferenciasSemAcentosPonto.csv (Figura 2).

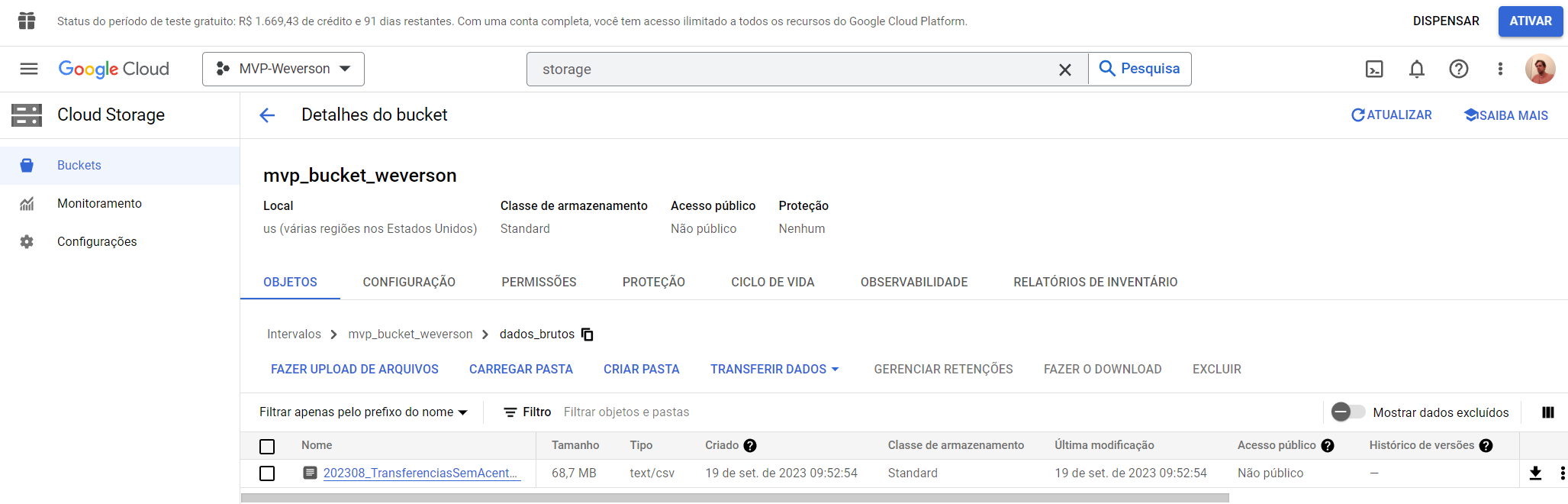


Figura 2 – Bucket com arquivos CSV

**Criação da Tabela**

Uma vez que os dados em formato CSV estavam devidamente armazenados no Bucket, procedeu-se à criação da tabela subsequente, destinada tanto à carga quanto à consulta dos dados. A estrutura dessa tabela pode ser visualizada na Figura 3.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Figura 3 - Criação da tabela no banco de dados

**Análise**

A etapa de análise é crucial para extrair insights significativos dos dados coletados e responder às perguntas formuladas na fase de definição dos objetivos. Essa fase será dividida em duas partes: análise da qualidade dos dados e solução do problema proposto.

a. Qualidade de Dados

Cada atributo do conjunto de dados deve passar por uma análise de qualidade rigorosa. Afinal, a precisão e a confiabilidade dos nossos resultados dependem da integridade dos dados com os quais trabalhamos. Precisamos identificar se existem problemas nos dados, como valores ausentes, inconsistentes ou outliers, e determinar como esses problemas podem ser resolvidos para que não afetem as respostas às perguntas formuladas.

Embora alguns conjuntos de dados estejam bem tratados antes de serem disponibilizados, é essencial realizar uma análise de valores por atributo. Se não forem encontrados problemas, isso também será devidamente documentado, evidenciando a robustez do conjunto de dados em uso.

b. Solução do Problema

Após assegurada a qualidade dos dados, passaremos à etapa de solução do problema previamente definido nos objetivos. Nessa fase, empregaremos técnicas analíticas apropriadas para buscar respostas para as perguntas estabelecidas, discutindo cada resultado e conectando os números obtidos às soluções propostas para o problema.

Através da análise dos dados e da interpretação cuidadosa dos resultados, visamos alcançar uma compreensão aprofundada das transferências de recursos públicos federais, contribuindo para a transparência e responsabilidade fiscal.

Cada descoberta técnica será acompanhada de uma discussão, oferecendo uma visão completa sobre como os dados analisados se traduzem em respostas significativas para as questões.

**Análise de Qualidade dos Dados - Valores Ausentes**

Durante nossa análise inicial da qualidade dos dados, focamos na verificação de valores ausentes que poderiam afetar a integridade e a precisão dos resultados da pesquisa. Um dos atributos essenciais examinados foi "Nome\_favorecido", que representa o nome dos beneficiários das transferências.

Para realizar essa verificação, utilizamos a seguinte consulta SQL:

SELECT \* FROM `mvp-weverson.mvp\_database.tb-tranferencia`

WHERE nome\_favorecido IS NULL;

Felizmente, a execução desta consulta não retornou nenhum registro, indicando que não existem valores ausentes na coluna "nome\_favorecido". Isso é uma indicação positiva da qualidade dos dados neste aspecto específico, uma vez que a ausência de informações críticas como o nome do favorecido poderia comprometer a análise subsequente das transferências (Figura 4).

A constatação de que não há valores ausentes em "nome\_favorecido" permite-nos prosseguir com a análise subsequente com maior confiança na integridade dos dados, reduzindo a necessidade de etapas adicionais de limpeza ou imputação de dados neste campo específico.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Figura 4 – Resutlado Valores Ausentes

**Análise de Qualidade dos Dados - Valores Inconsistentes**

Uma parte essencial da verificação de consistência dos dados foi a identificação de valores que, por sua natureza, não deveriam aparecer no conjunto de dados. Em particular, focamos na detecção de valores negativos, que são inconsistentes no contexto das transferências financeiras.

A seguinte consulta SQL foi utilizada para identificar tais registros:

SELECT \* FROM `mvp-weverson.mvp\_database.tb-tranferencia`

WHERE valor\_transferido < 0;

O resultado desta consulta mostrou a existência de 34 registros com valores de transferência negativos, representando uma inconsistência nos dados. Cada um destes registros precisará ser investigado para determinar a causa da inconsistência e aplicar as correções apropriadas (Figura 5 e 6).

A presença destes valores inconsistentes reforça a importância de uma análise de qualidade rigorosa antes de proceder com análises mais aprofundadas, assegurando que as conclusões derivadas dos dados sejam confiáveis e precisas.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Figura 5 – Valores inconsistentes

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Figura 6 – Valores inconsistentes (Dados originais - csv)

Para resolver esta inconsistência, uma das abordagens pode ser entrar em contato com a fonte dos dados ou os responsáveis pela coleta e reportar as discrepâncias encontradas, solicitando esclarecimentos ou correções. Alternativamente, dependendo da natureza e da quantidade dos dados inconsistentes, pode-se optar por excluí-los do conjunto de dados se isso não comprometer significativamente a análise geral.

Durante a fase de análise da qualidade dos dados, identificamos 34 registros na tabela **tb\_transferencia** que continham valores inconsistentes, especificamente valores negativos na coluna **valor\_transferido**. Dada a natureza desses dados e o fato de que o número de linhas afetadas é relativamente pequeno em comparação com o tamanho total do conjunto de dados, optamos por excluir esses registros inconsistentes. Esta decisão foi tomada para garantir a integridade e a confiabilidade dos resultados da nossa análise, ponderando-se que a remoção dessas linhas não introduziria viés significativo ou distorção nos resultados finais. O processo de exclusão foi realizado através da execução da seguinte consulta SQL:

DELETE FROM `mvp-weverson.mvp\_database.tb-tranferencia`

WHERE valor\_transferido < 0;

Após a execução dessa consulta, confirmamos que todos os registros com valores inconsistentes foram removidos com sucesso da tabela, permitindo-nos prosseguir com a análise dos dados restantes (Figura 7).

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Figura 7 – Excluindo valores inválido

**Solução do Problema - Quais são os principais tipos de transferências realizadas e seus respectivos volumes de recursos?**

Após a execução da consulta SQL com o intuito de identificar os diferentes tipos de transferências e o volume de recursos associado a cada um, constatamos um resultado singular: apenas um tipo de transferência foi registrado no período analisado. A consulta utilizada foi a seguinte:

SELECT

    TIPO\_TRANSFERENCIA,

    SUM(VALOR\_TRANSFERIDO) AS VOLUME\_TOTAL

FROM

    `mvp-weverson.mvp\_database.tb-tranferencia`

GROUP BY

    TIPO\_TRANSFERENCIA

ORDER BY

    VOLUME\_TOTAL DESC;

Este resultado é peculiar e realça a predominância de um único método de transferência no conjunto de dados analisado. O tipo de transferência identificado foi “Legais Voluntarias e Especificas” e o volume total de recursos associado a este método foi de R$ 23.300.619.491,63 (Figura 8).

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Figura 8 – Tipo de transferência

Este achado restringe a análise da distribuição de recursos entre diferentes tipos de transferências, pois evidencia a existência de uma única modalidade predominante no período estudado. Contudo, essa descoberta em si é relevante, pois indica uma concentração significativa de recursos em uma determinada forma de transferência, o que pode refletir prioridades, políticas ou necessidades específicas no contexto analisado.

**Solução do Problema - Como essas transferências estão distribuídas geograficamente nas Unidades Federativas?**

Após a execução da consulta SQL, foi possível observar como as transferências de recursos estão distribuídas geograficamente entre as Unidades Federativas (UFs) do Brasil. A consulta proporcionou uma visão clara do volume de recursos direcionados a cada UF (Figura 9).

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

Figura 9 - Transferências de recursos distribuídos

É importante ressaltar que, durante a análise dos resultados, identificamos a presença de um registro em que a Unidade Federativa estava representada como **null**. Isso pode indicar um caso de informação não registrada ou perdida no processo de transferência ou coleta de dados. A presença de **null** na coluna UF sugere que, para algumas transferências, o local de destino não foi devidamente especificado no conjunto de dados original. Essa lacuna nas informações reforça a importância de uma rigorosa validação e limpeza dos dados, garantindo a confiabilidade e precisão dos resultados obtidos em nossa análise.

**Solução do Problema - Quais são as áreas (funções e subfunções) mais beneficiadas pelas transferências de recursos?**

Após a execução da consulta SQL, que teve como objetivo identificar as áreas, caracterizadas por funções e subfunções, que foram mais beneficiadas pelas transferências de recursos, obtivemos um total de 79 linhas como resultado. Isso significa que existem 79 combinações únicas de funções e subfunções que receberam recursos através de transferências (Figura 10).

Tabela

Descrição gerada automaticamente

Figura 10 - Funções e subfunções beneficiadas peça transferência de recursos

Este resultado é bastante significativo, pois nos fornece uma visão abrangente de como os recursos públicos federais foram distribuídos entre diferentes áreas. A presença de 79 entradas distintas revela uma diversidade de áreas beneficiadas, indicando que os recursos não foram concentrados em um número limitado de setores, mas sim distribuídos por uma variedade de funções e subfunções.

**Solução do Problema - Como os recursos são distribuídos entre os diferentes tipos de favorecidos?**

Após a execução da consulta SQL para investigar como os recursos são distribuídos entre os diferentes tipos de favorecidos, foi possível observar uma variação significativa nos volumes de recursos destinados a cada categoria de favorecido. A consulta nos permitiu identificar e quantificar o montante total transferido para cada tipo de favorecido, proporcionando insights valiosos sobre as prioridades e ênfases na alocação de recursos públicos (figura 11).

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Figura 11 - Tipo de favorecido

A análise dos resultados evidenciou quais categorias de favorecidos foram mais beneficiadas e quais receberam menos recursos, destacando as discrepâncias e desigualdades na distribuição dos fundos. Esta informação é essencial para entender como as políticas públicas são implementadas na prática e quais grupos ou entidades estão recebendo maior apoio financeiro do governo.

**Solução do Problema - Como os recursos são distribuídos entre os diferentes tipos de favorecidos?**

Para resolução desta questão vamos utilizar os recursos do dataproc.

**O que é o Dataproc?**

O Dataproc é um serviço totalmente gerenciado e altamente escalonável para executar o Apache Hadoop, o Apache Spark, o Apache Flink, o Presto e mais de 30 ferramentas e frameworks de código aberto. Use o Dataproc para modernização do data lake, ETL e ciência de dados segura, em escala global, totalmente integrada ao Google Cloud e com custos bem menores.

**Configurando o Dataproc:**

Interface gráfica do usuário, Aplicativo, Word

Descrição gerada automaticamente

Figura 12- Dataproc

Interface gráfica do usuário, Texto, Aplicativo, Word

Descrição gerada automaticamente

Figura 13 - Dataproc

**Acesso Remoto aos Dados:**

Adicionamos o recurso do Jupyter Notebook para facilitar o acesso remoto aos dados. Abaixo, estão os passos para criar um túnel e conectar-se via Web no cluster (figuras 14 a 18):

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Figura 14 - Passos para se conectar na inteface Web

Texto

Descrição gerada automaticamente

Figura 15 - Exibição do tunnel, via putty

Texto

Descrição gerada automaticamente

Figura 16 - Executando o Chrome e conectando no proxy

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Figura 17 - Tela inicial do Hadoop

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Figura 18 - Carregando o Jupyter Notebook para utilização da base de dados com python

Após a execução da consulta SQL e o subsequente direcionamento dos dados para um DataFrame pandas, procedemos à análise da correlação entre **CODIGO\_ORGAO\_SIAFI** e **total\_transferido**. O resultado dessa análise indicou um coeficiente de correlação de −0.06397650836202501, revelando uma correlação negativa extremamente fraca entre as duas variáveis (Figura 19).

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Figura 19 - Correlação

Este valor, próximo de zero, sugere a ausência de uma relação linear significativa entre o código identificador do órgão e o valor total transferido. Dessa forma, com base nessa análise, não é possível estabelecer uma associação direta entre a natureza ou localização do favorecido, representada pelo **CODIGO\_ORGAO\_SIAFI**, e o **valor\_transferido**.

É crucial destacar que a natureza do **CODIGO\_ORGAO\_SIAFI**, por ser um identificador numérico atribuído aos órgãos, impede que a correlação com os valores transferidos tenha uma interpretação prática substantiva. Embora a correlação linear seja insuficiente, isso não elimina a possibilidade de existirem outras formas de relações não lineares entre as variáveis, que o coeficiente de correlação de Pearson não seria capaz de identificar.

Diante desse cenário, concluímos que a análise de correlação entre **CODIGO\_ORGAO\_SIAFI** e **total\_transferido** não proporcionou insights significativos acerca da relação entre o valor transferido e a natureza ou localização do favorecido. Para obtenção de resultados mais reveladores e pertinentes, seria recomendável explorar correlações entre variáveis que possuam uma relação lógica ou prática mais evidente, tais como a correlação entre o tipo de favorecido e o valor transferido, ou entre a unidade federativa e o valor transferido. Desta forma, poderíamos vislumbrar associações mais claras e informativas para responder às questões propostas no início deste trabalho.

**Conclusão:**

Ao longo deste estudo, priorizamos a compreensão e aplicação eficaz das técnicas e ferramentas de gerenciamento e análise de dados. Usando o Portal da Transparência do Governo Federal como fonte primária, fomos confrontados com o desafio de trabalhar com um volume substancial de informações.

Inicialmente, decidimos explorar a AWS (Amazon Web Services) como nossa plataforma de banco de dados. No entanto, encontramos obstáculos significativos durante essa tentativa. Mesmo após o processo de carregamento dos dados, enfrentamos problemas na recuperação e visualização dessas informações nas consultas SQL. As Figuras 20 e 21 ilustram algumas partes do trabalho feito. Este revés ressaltou a importância da seleção cuidadosa de ferramentas e plataformas para tarefas de grande volume e complexidade.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Figura 20 - Tentativa de configuração do AWS

Tabela

Descrição gerada automaticamente

Figura 21 - Apesar de carregar não aparecem dados nos sqls

Dado o impasse na AWS, migramos para o serviço BigQuery da Google Cloud. A capacidade desta plataforma de lidar eficazmente com grandes datasets e proporcionar consultas em tempo real provou ser crucial para a continuidade do nosso projeto.

A estratégia de ETL (Extração, Transformação e Carga) foi meticulosamente implementada, garantindo que os dados fossem precisamente importados, tratados e carregados no BigQuery. Durante a fase de transformação, a remoção de acentos e a correção de inconsistências emergiram como etapas vitais.

As vantagens do BigQuery tornaram-se ainda mais aparentes durante a fase de análise, permitindo consultas SQL complexas e insights detalhados sobre os dados. A tentativa de correlacionar variáveis, como o valor transferido e o CODIGO\_ORGAO\_SIAFI, reafirmou a relevância de uma infraestrutura robusta.

Concluindo, a jornada deste trabalho destacou o intrincado equilíbrio entre a seleção da plataforma, a preparação dos dados e a própria análise. Enquanto nossa experiência inicial com a AWS foi desafiadora, o BigQuery da Google Cloud emergiu como uma solução confiável. Esta experiência serviu como um testemunho do impacto das decisões tecnológicas na eficácia da análise de dados, sublinhando a necessidade de adaptação, experimentação e escolha cuidadosa das ferramentas.