Análise do Conjunto de Dados do Seguro Defeso

Analysis of the Defeso Insurance Dataset

Onofre André Dall'Oglio[1], Francisco Sanches Banhos Filho[1]

[1]Faculdade de Ciências Exatas – Universidade do Estado de Mato Grosso (UNEMAT)
Sinop – MT – Brasil

onofre.oglio@unemat.br, fsanches@unemat.br

Abstract. *This article highlights the importance of data processing and analysis, highlighting the challenge*
of dealing with large volumes of data. The study focuses on the application of data extraction, processing and loading techniques using Pentaho Data Integration and Microsoft Power BI software for data analysis. The article emphasizes the importance of combining data organization and analysis tools to obtain detailed insights to dataset the Seguro Defeso.
Keywords: *Seguro Defeso. Pentaho. ETL. Power BI. dataset*

Resumo. *Este artigo aborda a importância do tratamento e análise de dados, destacando o desafio de lidar com grandes volumes de dados. O estudo se concentra na aplicação de técnicas de extração, tratamento e carga de dados utilizando os softwares Pentaho Data Integration e Microsoft Power BI para análise de dados. O artigo enfatiza a importância da combinação de ferramentas de organização e análise de dados para obter insights detalhados do dataset do Seguro Defeso.*
Palavras Chave: *Seguro Defeso. Pentaho. ETL. Power BI. dataset*

1. Introdução

O crescimento exponencial na geração de dados, seu consumo e armazenamento tem criado desafios substanciais para o tratamento e análise conforme descreve Abreu (2014), apud Mayer-Schönberger, (2013), e como consequência, a necessidade que tais dados sejam processados para embasar a tomada de decisão em tempo hábil.

As plataformas e sistemas governamentais são parte integrante deste ecossistema e geram diariamente números astronômicos de dados. Uma parcela destes dados são públicos e podem ser consultados nas plataformas abertas como o Portal Transparência (https://portaltransparencia.gov.br, acesso em 21/10/2023) e o Portal Dados Abertos (https://dados.gov.br, acesso em 21/10/2023) que possui atualmente 12.296 conjuntos de dados ou *dataset* que segundo Junior (2015) é o resultado da coleta de dados de diversas fontes, agrupados em arquivos de grandes dimensões.

Como exemplo, o tamanho do *dataset* contendo as informações cadastrais dos CNPJ´s existentes no Brasil é superior a 5 Gigabytes (GB), o que nos dá uma dimensão, por exemplo, do volume do(s) *dataset(s)* contendo os registros de recolhimento de impostos, geração de multas de trânsito, folhas de pagamento, entre outros.

Os dados relativos aos pagamentos de benefícios sociais pelos governos, seja federal, estadual ou municipal, seguem a mesma lógica em volumetria e diversidade.

Ainda que existam mecanismos próprios e constantemente aprimorados para controlar e fiscalizar os pagamentos destes benefícios, conforme exige a legislação a exemplo da Lei de Responsabilidade Fiscal (Brasil, Lei 101/2000), é comum a ocorrência de notícias na mídia em geral sobre desvios ocorridos no pagamento destes benefícios.

Estas ocorrências não são diferentes quanto ao Seguro Defeso, que prevê o pagamento de um benefício social para pescadores durante o período em que a pesca artesanal é proibida, conhecido "defeso" (Brasil, Lei 10.779/2003).

Considerando os desafios para identificar possíveis irregularidades nestes pagamentos, objetivou-se neste trabalho, realizar um estudo de caso que caracteriza e explica o uso das ferramentas de organização e integração de dados *Pentaho Data Integrations (PDI)* e de análise de dados *Microsoft Power BI*, de forma a responder a pergunta problema do presente estudo, se a utilização destas ferramentas pode oferecer uma visão geral, objetiva e pormenorizada do conjunto de dados do seguro defeso?

Primeiramente, realizamos uma pesquisa bibliográfica onde buscamos os trabalhos e autores que tratam do tema. Também realizamos uma pesquisa quantitativa para obtermos informações quanto ao processamento dos dados e uma pesquisa qualitativa quanto a interpretação dos resultados.

Para solução da questão chave do trabalho, norteamo-nos pelo objetivo geral em utilizar as ferramentas de *Business Intelligence (BI): Pentaho Data Integration (PDI*) e *Microsoft Power BI (Power BI*) para demonstrar, como o tratamento e análise dos dados do seguro defeso pode gerar percepções relevantes na tomada de decisão.

Como objetivos específicos, pesquisou-se as fontes contendo os dados de pagamento do seguro defeso para o período 2012 à 2022, definiu-se os requisitos para criação de um banco de dados, realizou-se o processo de extração, tratamento e carga dos dados (ETL) no banco de dados (BD) e aplicou-se as ferramentas de BI sobre o BD para fazermos uma análise interpretativa das informações obtidas.

A estrutura deste artigo foi dividida, além da presente introdução, mais três seções. A próxima seção detalha a metodologia e as ferramentas utilizadas. Em seguida, apresentamos os dados utilizados e os resultados obtidos no estudo. Por último, na seção de conclusão, destacamos a importância dos pontos abordados neste artigo e também levantamos a possibilidade de expansão do tema para estudos futuros.

2 O Processo ETL

As etapas observadas na Figura 1, são definidos por Kimball (2004), como o processo de extração, tratamento e carga de dados (do inglês *ETL – Extrat, Transform and Load*), que se caracteriza por uma sequência de passos que permite criar fluxos de trabalho de forma que os dados originais de uma determinada fonte, sejam selecionados, limpos e reorganizados para que atendam aos requisitos de uma aplicação final. Ainda na Figura 1, é possível observar cada etapa deste processo, quando são realizadas operações de agendamento de processos, tratamento de erros, recuperação de informações e controle de qualidade.

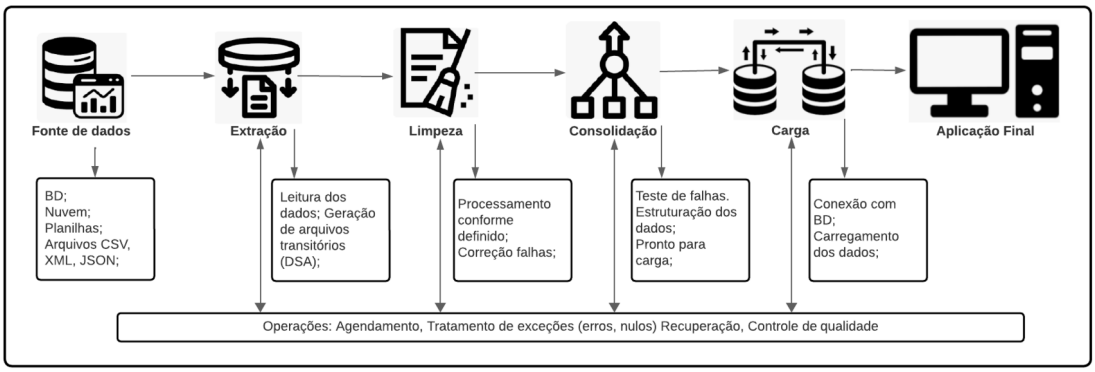

Figura 1: Fluxo de dados em processo de extração, transformação e carga. Fonte: Kimball (2004), adaptado pelo autor.

Geralmente, estas técnicas são aplicadas sobre grandes volumes de dados, desta forma, exige a definição de estratégias para divisão dos processos e subprocessos. Neste caso, o *software* reserva espaços em disco chamados *Data Staging Area (DSA)* onde são criados arquivos temporários para releitura, comparações e reprocessamento, conforme a necessidade da aplicação. Cada fase pode ser configurada para gerar arquivos de saída, que servirão de entrada para a próxima etapa, arquivos estes que também servem de *backup* em caso de falha, permitindo a recuperação do processo do ponto de parada.

2.1 Pentaho Data Integration (PDI)

Existem diversas ferramentas que realizam o processo ETL, como alguns exemplos podemos citar os descritos na Tabela 1.

Tabela 1 – Softwares que realizam o processo ETL

Software	**Endereço eletrônico**
Azure Data Factory	https://azure.microsoft.com/pt-br/products/data-factory/
Databricks	https://www.databricks.com/
Amazon Glue	https://aws.amazon.com/pt/glue/
Amazon Lambda	https://docs.aws.amazon.com/lambda/latest/dg/welcome.html
Informática Power Center	https://www.informatica.com/
Talend	https://www.talend.com/
Pentaho Data Integration	https://www.hitachivantara.com/

O software escolhido para realização do presente estudo foi o *Pentaho Data Integration,* e sua documentação (2023), descreve que seu processo ETL é executado em um servidor *Pentaho* ou em um servidor de aplicativos Java, onde pode ser programado, agendado e monitorado e é composto por três partes principais:

- O *Spoon* (colher) que é uma interface gráfica de usuário que permite a criação visual de fluxos de trabalho de ETL e abriga as transformações (*transformation*) que são composta por passos (*steps*) onde se definem os parâmetros da execução.
- O *Pan* (panela) é o seu mecanismo ETL que executa os fluxos de trabalho criados no *Spoon*.
- A *Kitchen* (cozinha) é um utilitário de linha de comando que permite a criação de códigos específicos quando não atendidos pelos modelos pré existentes do *Spoon*.

2.2 Sistema de Gerenciamento de Banco de Dados

Para receber os dados após o processo ETL, utilizamos o *MySQL* como SGBD, que utiliza linguagem de consulta estruturada *SQL (Strutured Query Language)* e foi escolhido por ser compatível com o PDI, o *Power BI,* o computador e seu sistema operacional *Windows*.

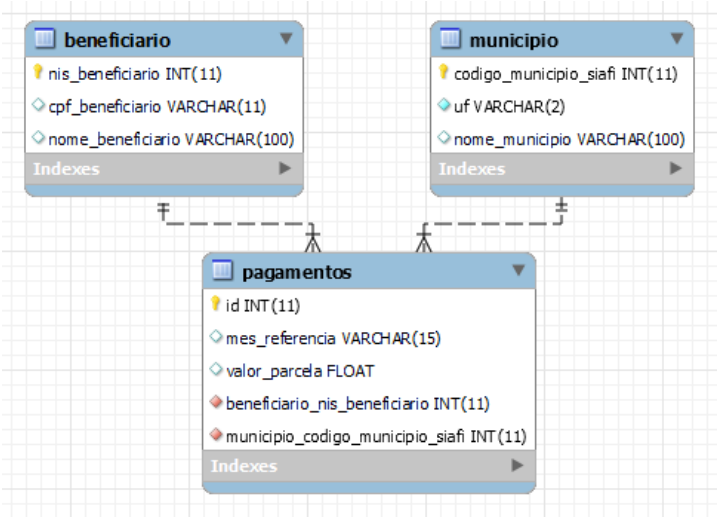

Figura 2: SGBDI, modelo entidade-relacionamento *MySQL*.

O modelo utilizado para armazenar dados no banco, representado na Figura 2, foi elaborado de forma que, a consulta ao banco de dados tanto pelo *PDI* como *Power BI* fosse minimamente eficiente e, para tal, dividimos os dados em três tabelas menores e referenciadas, conforme orienta Volaco (2004).

Tabelas menores possuem maior facilidade de indexação, e os índices criados em banco de dados SQL para realizar consultas, aceleram a recuperação de dados, e permitem ao *Microsoft Power BI* otimizar as consultas, realizar o carregamento seletivo de dados e a gravação em *cache* do resultado de consultas para reutilização.

Desta forma, para receber os dados de cadastro dos beneficiários e municípios do *dataset* original, criamos as respectivas tabelas 'beneficiario' e 'municipio' e para os dados relativos aos pagamentos dos benefícios criamos a tabela 'pagamentos'. Tanto o NIS como o Código do Município SIAFI, respectivamente chaves primárias nas tabelas 'beneficiario' e 'município', são chaves estrangeiras na tabela 'pagamentos'.

Tabela 2 – Descrição dos campos de origem e destino.

***Dataset* de Origem**	**Bando de dados**	**Tipo de dado**	**Tamanho do campo**
"Nis_Favorecido"	'beneficiario' = "nis_favorecido", Chave Primária	*Int*	15
"Nome_Favorecido"	'beneficiario' = "nome_favorecido"	*String*	100
"Codigo_Municipio_Siafi"	'municipio' = "codigo_municipio_siafi", Chave Prim.	*Int*	15
"Nome_Municipio"	'municipio' = "nome_municipio"	*String*	100
"Mes_Referencia"	'pagamentos' = "mes_referencia"	*String*	15
"Codigo_Municipio_Siafi"	'pagamentos' = "codigo_municipio_siafi", Chave Est.	*Int*	15
"Nis_Favorecido"	'pagamentos' = "nis_favorecido", Chave Est.	*Int*	15
"Valor_Parcela"	'pagamentos' = "valor_parcela"	*Float*	15

Os tamanhos e tipos de dados descritos na Tabela 2, foram definidos após realizar uma pré consulta aos dados originais, através de uma funcionalidade do PDI que permite realizar a leitura de uma amostra dos dados. Neste processo, são carregados as colunas, tipos e tamanho de dados e uma lista com os primeiros registros do *dataset*.

2.3 Microsoft Power BI

A documentação do *Microsoft Power BI* (2023), o descreve como uma coleção de serviços de *software*, aplicativos e conectores que trabalham juntos para transformar dados de diversas fontes, que através de sua interface gráfica, permite a criação de relatórios e visualizações interativas de forma rápida e eficiente para utilização inclusive por usuários com conhecimento básico em computação.

A escolha pelo *Microsoft Power BI* ocorreu em função de sua versatilidade em conectar-se à diferentes bases de dados, a facilidade para gerar apresentações e aplicações gráficas e velocidade com que produz percepções importantes e análises refinadas em tempo real.

3. Dados e Resultados

Para execução do processo ETL, utilizamos um *dataset* contendo os registros de pagamento do Seguro Defeso referentes aos anos de 2012 a 2022 que foram copiados do Portal Transparência (https://portaltransparencia.gov.br) no formato CSV (valores separados por vírgula, do inglês *Comma-Separated Values*).

O pagamento do Seguro Defeso é efetuado pela Caixa Econômica Federal, levando em consideração o mês de referência no qual o pescador tem direito ao recebimento do benefício. Esses pagamentos são registrados pelo Número de Inscrição Social (NIS) do beneficiário. O NIS é único para cada pessoa natural, conforme estabelecido pela Portaria nº 177, de 16 de junho de 2011, do Ministério do Desenvolvimento Social e Combate à Fome. Dessa forma, uma mesma pessoa não pode receber mais de um pagamento para um mesmo período.

Os registros de pagamentos do seguro defeso, no período de 2012 à 2022, totalizam 26 milhões de linhas, distribuídos em 9 colunas totalizando 118 arquivos com tamanho de 2,82 GB conforme a Tabela 03.

Tabela 3 – Dicionário de dados do Seguro Defeso

COLUNA	DESCRIÇÃO
Ano/Mês Referência	Ano/Mês da folha de pagamento do benefício.
UF	Sigla da Unidade Federativa do beneficiário do Seguro Defeso.
Código SIAFI Município	Código, no SIAFI (Sistema Integrado de Administração Financeira), do município do beneficiário do Seguro Defeso.
Nome Município	Nome do município do beneficiário do Seguro Defeso.
CPF Favorecido	Número do CPF do beneficiário do Seguro Defeso.
NIS Favorecido	NIS do favorecido do Seguro Defeso Criado pela Caixa Econômica Federal o NIS significa Número de Identificação Social.
RGP Favorecido	Número de registro geral do pescador.
Nome Favorecido	Nome do favorecido do Seguro Defeso.
Valor Parcela	Valor da parcela do benefício disponibilizada.

Para realizar o processamento dos dados, utilizamos um computador pessoal com a configuração:

- Processador Intel(R) Core(TM) i5-3317U CPU @ 1.70GHz, com 2 núcleos e 4 *threads*;
- 12 GB de memória RAM;
- Placa aceleradora gráfica com 2 GB de memória;
- HD de estado sólido 480 GB;
- Sistema operacional Windows;

Neste computador, foram instalados o *Pentaho Data Integratioin* (PDI) para realização do processo ETL, o *MySQL* como sistema de gerenciamento de banco de dados (SGBD) e o *Microsoft Power BI* como ferramenta de análise dos dados.

3.1 Resultados Obtidos

O processo de tratamento dos dados foi dividido em três transformações (etapas 1, 2 e 3), composta pelos "passos" (*steps*), visualizados na Figura 3.

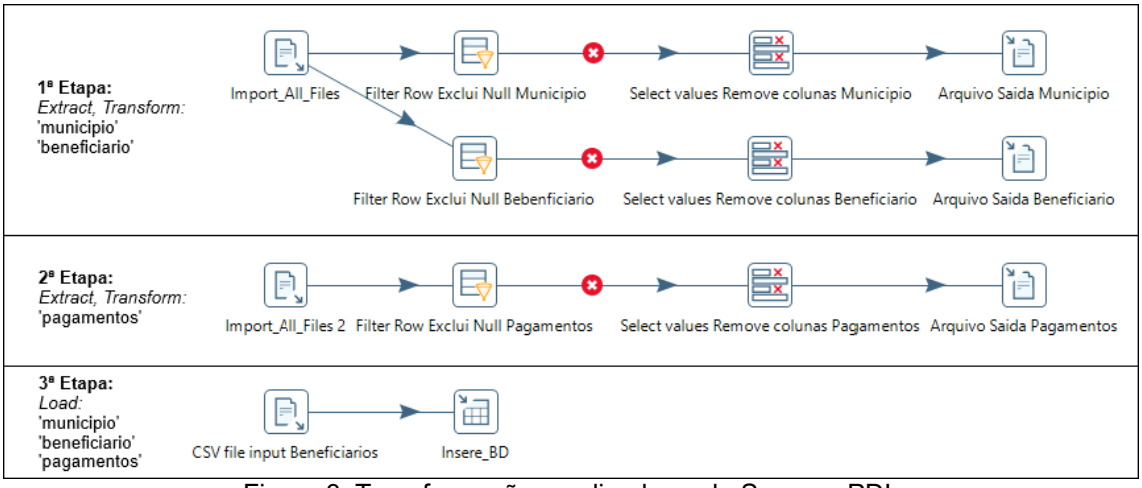

Figura 3: Transformações realizadas pelo Spoon – PDI.

A primeira e segunda etapas, demonstradas na Figura 3, refere-se ao tratamento dos dados que tem como destino as tabelas "municipio", "beneficiario" e "pagamentos" no BD denominado 'sd_big' e foram configuradas com quatro passos (*steps*): importar o *dataset* → excluir registros com dados vazios → selecionar colunas de interesse (Tabela 2) → Gerar arquivo CSV de saída. Pelo maior volume de dados da tabela "pagamentos" criamos uma transformação exclusiva, considerando o tempo de processamento, capacidade computacional e maior suscetibilidade a falhas, dada a quantidade de registros.

A terceira etapa visualizada na Figura 3, refere-se a última fase do processo ETL, que é a carga dos dados transformados no BD. Esta transformação possui dois passos (*steps*): importar os arquivos CSV (saída das etapas anteriores) → conectar-se e carregar os dados no BD 'sd_big'. A conexão com o banco de dados ocorre através de um conector JDBC (Conector de banco de dados Java, do inglês *Java Database Connectivity*), com os seguintes parâmetros: URL: *localhost*, banco de dados: 'sd_big', login e senha.

Tabela 4 – Tempo de execução das fases ETL.

Fase	Organização dos dados	N° Arquivos	Tamanho	Tempo de Processamento
Download	Arquivos CSV	118	2,82 Gb	
Extract, Transform	Tabela 1 ‘municipio’	1	372 Kb	30 minutos
Extract, Transform	Tabela 2 ‘beneficiario’	1	154,30Mb	
Extract, Transform	Tabela 3 ‘pagamentos’	1	1,39 Gb	60 minutos
Load MySQL	Carga das tabelas 1, 2 e 3	3	2,00 Gb	90 minutos

A Tabela 4 apresenta o comparativo do tamanho dos arquivos antes e após a execução do ETL e o tempo de processamento para executar todas as suas fases. Houve redução de 29% no tamanho dos arquivos salvos no banco de dados, e um tempo total de processamento de 180 minutos na máquina descrita na seção 2 deste artigo.

O processo de análise de dados utilizando o *Microsoft Power BI*, se iniciou através da conexão com a fonte de dados, através do conector JDBC com os mesmos parâmetros utilizados no processo ETL. Após o acesso aos dados, todas as funcionalidades ficam disponíveis na tela principal, inclusive os modelos de painéis que se deseja utilizar.

A Figura 4, mostra o volume de recursos do Seguro Defeso alocado por unidade da federação (UF). O painel *treemap* apresenta os quadros legendados para cada UF, sendo seu tamanho proporcional à soma dos valores do referido benefício social ao longo de todo período analisado. O estado do Pará recebeu R$ 6,8 bilhões, seguido pelo Maranhão (R$ 5 bilhões) e Bahia (R$ 2,9 bilhões), de um total de R$ 35 bilhões distribuídos ao longo de todo o período.

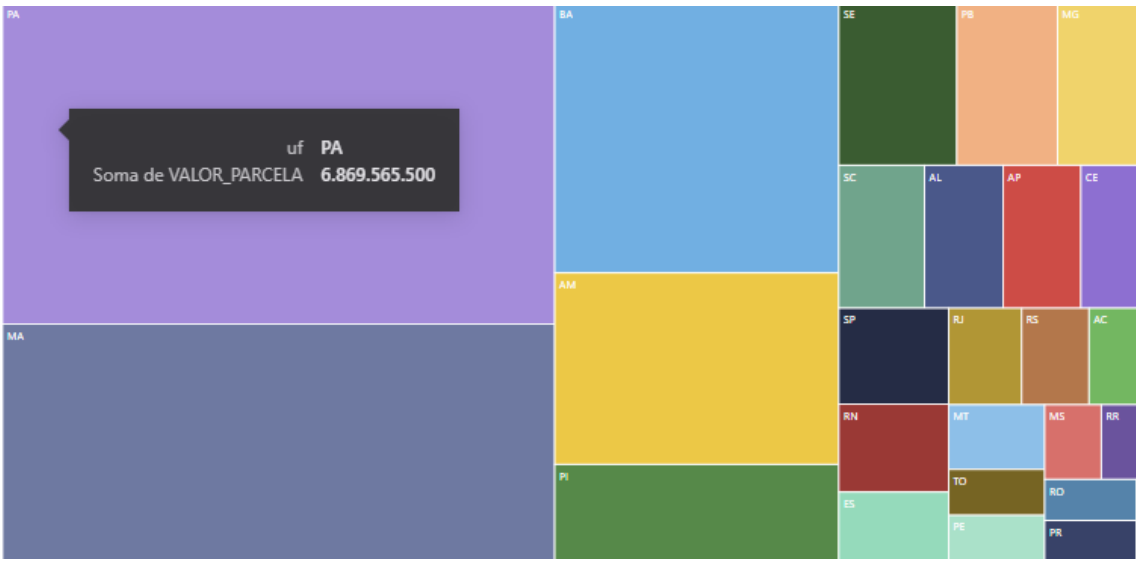

Figura 4: Distribuição do pagamento do Seguro Defeso por UF.

A figura 5 ilustra o município de Cameta – PA como a cidade com o maior volume recebido de Seguro Defeso do país (R$ 834.651.034,00).

Figura 5: Municípios com o maior volume recebido de Seguro Defeso.

Refinando a granularidade dos dados, evidenciamos os indivíduos que receberam o maior valor acumulado (2012 à 2022), conforme a Figura 6, a média foi de R$ 59 mil para cada pescador que recebeu o benefício entre 59 e 70 vezes ao longo da série histórica, porém, o NIS 21014835412 recebeu R$ 262.404.

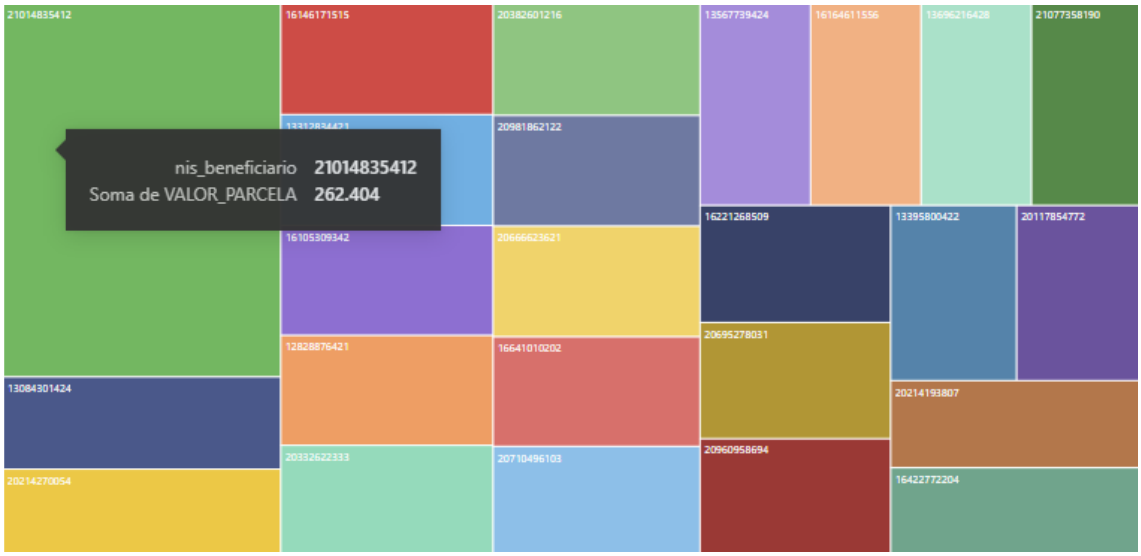

Figura 6: Valor recebido do seguro defeso por NIS.

O mesmo NIS, é identificado em pagamentos de 07 nomes diferentes vinculados ao registro de cinco números diferentes de CPF. Sabendo que o NIS é único por pessoa natural, é possível inferir que há erro de registro ou possível fraude.

Como contra-teste realizamos um processo ETL de forma que todo o *dataset* fosse novamente tratado e gerasse um arquivo de saída exclusivamente com o NIS observado na Figura 6, e a soma dos valores coincidiu com os dados apontados pelo *Power BI*.

4 Conclusão

Na presente pesquisa, realizamos uma investigação sobre os desafios enfrentados na análise de grandes volumes de dados, com foco nos registros de pagamento do Seguro Defeso. O crescente aumento na geração e armazenamento de dados tem exigido estratégias inovadoras para o tratamento e análise dessas informações. Utilizando ferramentas como o *Pentaho Data Integration* (PDI) e o *Microsoft Power BI*, buscamos responder à questão fundamental deste estudo: seria possível obter uma visão detalhada e significativa do conjunto de dados do Seguro Defeso através do uso dessas ferramentas?

Nossa pesquisa foi guiada pelo objetivo de utilizar as ferramentas de *Business Intelligence* (BI) para demonstrar como o tratamento e a análise dos dados do Seguro Defeso podem gerar *insights* necessários para a tomada de decisões. Nesse sentido, aplicamos o processo ETL para organizar os dados brutos, criando um banco de dados estruturado e otimizado. A escolha do *Pentaho Data Integration* (PDI) se mostrou fundamental, permitindo a execução do ETL de maneira eficiente, com destaque para a divisão dos processos e na manipulação dos dados, aspectos importantes para lidar com um *dataset* tão extenso.

Além disso, o *Microsoft Power BI* revelou-se uma ferramenta indispensável na análise interpretativa dos dados. Com sua interface intuitiva, conseguimos criar visualizações complexas de forma rápida e eficaz. Através de painéis, conseguimos ilustrar de forma clara o volume de recursos alocados por unidade federativa e municípios, identificar padrões e destacar possíveis discrepâncias nos dados.

Ao longo do processo, deparamo-nos com descobertas intrigantes. Identificamos casos onde um mesmo Número de Identificação Social (NIS) estava vinculado a diferentes nomes, números de CPFs e diversos pagamentos do benefício social, indicando possíveis erros de registro ou até mesmo fraudes. Essas revelações enfatizam a importância de métodos avançados de análise de dados para detectar irregularidades nestes pagamentos, portanto, respondendo a questão original do nosso trabalho, sim, a utilização do *Pentaho Data Integration*, em conjunto ao *Microsoft Power BI*, permitem obter uma visão detalhada e significativa do conjunto de dados do Seguro Defeso.

Por fim, como trabalhos futuros, recomenda-se uma investigação mais aprofundada desses casos suspeitos, aprimoramento das técnicas e métodos de tratamento e análise de dados, envolvendo inclusive estudos relacionados com outras bases de dados de outros programas sociais. Além disso, a exploração de técnicas avançadas de aprendizado de máquina pode ser uma abordagem promissora para identificar padrões complexos nos dados. Os estudos futuros podem contribuir significativamente para ampliar o conhecimento sobre o tema além de contribuir para o aprimoramento da fiscalização da aplicação dos recursos dos programas sociais e a democratização do acesso à informações sobre o pagamento destes benefícios.

5 Referências

Abreu. G. O. L. (2014) BIG DATA: Como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana, Instituto Federal de Educação, Ciência e Tecnologia da Paraíba. Temática, v. 16, n. 2, p. 155-178, 2019, https://periodicos.ufpb.br/ojs2/index.php/tematica/article/view/21510, Março 2023.

Brasil. Lei 10.779, de 25 de novembro de 2003. Estabelece normas de prevenção, controle e fiscalização da pesca em águas sob jurisdição nacional, https://www.planalto.gov.br/ccivil_03/Leis/2003/L10.779.htm, Março 2023.

Brasil. Lei 101, de 04 de maio de 2000.Estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências, https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp101.htm, Outubro 2023.

Kimball, Ralph. The Data Warehouse ETL Toolkit. 2004, https://archive.org/details/2004TheDataWarehouseETLToolkitRalphKimball, Setembro 2023.

Microsoft. Power BI documentation, https://docs.microsoft.com/en-us/ power-bi/, Março 2023.

Pentaho. Pentaho Data Integration, https://help.hitachivantara.com/Documentation/ Pentaho/9.4/Products/Pentaho_Data_Integration, Março 2023.

Volaco, E. B. Otimização de comandos SQL. Revista SQL Magazine. Out 2007. Ano 01. Edição 01. São Paulo: DevMedia. 2006, https://www.devmedia.com.br/artigo-sql-magazine-1-otimizacao-de-comandos-sql/6926, Outubro 2023.