Universidade Federal de Pernambuco

Centro de Informática

Graduação em Sistemas de Informação

Rastreabilidade de Fluxos para Ingestão de Dados em um Data Lake

Gabriel Estevam Longuinhos

Recife

2021

Gabriel Estevam Longuinhos

Rastreabilidade de Fluxos para Ingestão de Dados em um Data Lake

Trabalho apresentado ao curso de Sistemas de Informação da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação

Orientadora: Bernadette Farias Lóscio

Recife

2021

Resumo

Com o crescimento exponencial de dados sendo produzidos e processados no mundo, Data Lakes surgiram como uma solução para as deficiências percebidas nos Data Marts, podendo exercer a função de um sistema ou repositório essencial para análise de dados. Recentemente, entrou em vigor a Lei Geral de Proteção de Dados (LGPD), que dispõe sobre o tratamento de dados pessoais com o objetivo de proteger os direitos fundamentais de liberdade e privacidade da pessoa natural. A fim de atender à LGPD, organizações que possuem Data Lakes precisam fornecer transparência sobre o tratamento de dados pessoais. Neste contexto, a coleta e o uso de metadados, capazes de descrever os tratamentos realizados sobre os dados, tornam-se essenciais. Este trabalho aborda o problema da coleta de metadados em Data Lakes e apresenta o desenvolvimento de um fluxo de ingestão de dados, o qual possibilita a extração de metadados de proveniência e seu armazenamento em um repositório de dados. Dessa forma, torna-se possível a execução de consultas e análises para fins de monitoramento do cumprimento da LGPD, por exemplo.

Palavras-chave: Metadados de proveniência, Data Lake, LGPD.

Abstract

With the exponential growth of data being produced and processed in the world, Data Lakes emerged as a solution for the deficiencies perceived in Data Marts, being able to play the role of an essential system or repository for data analysis. Recently, the General Data Protection Law (LGPD) came into force and provides a new legal framework for privacy and data protection of the natural person in Brazil. In order to comply with the LGPD, organizations that have Data Lakes need to provide transparency about the handling of personal data. In this context, the collection and use of metadata, capable of describing the treatments performed on the data, becomes essential. This work addresses the problem of collecting metadata in Data Lakes and presents the development of a data ingestion flow, which allows the extraction of provenance metadata and its storage in a data repository. In this way, it becomes possible to carry out queries and analyzes for the purpose of monitoring compliance with the LGPD, for example.

Keywords: Metadatas provenance, Data Lake, GDPL.

Lista de Figuras

Sumário

1. Introdução

1.1 Contextualização

A Lei 13.709/2018, conhecida como Lei Geral de Proteção de Dados Pessoais (LGPD), aprovada em setembro de 2018, dispõe sobre o tratamento de dados pessoais com o objetivo de proteger os direitos fundamentais de liberdade e privacidade da pessoa natural. Dois anos após sua aprovação, a lei entrou em vigor, deliberando mudanças estruturais significativas nas organizações que utilizam dados para fins comerciais, a fim de que os titulares dos dados obtenham mais transparência sobre a utilização de seus dados por essas instituições.

Com o crescimento exponencial de dados sendo produzidos e processados no mundo, Data Lakes surgiram como uma solução para as deficiências percebidas nos Data Marts, que são subdivisões específicas de negócios de data warehouses que permitem apenas subconjuntos de perguntas a serem respondidas (Dixon, 2010). Atualmente, Data Lakes vem sendo utilizados como um sistema ou repositório essencial para análise de dados (Isuru Suriarachchi, Beth Plale 2016). Data Lakes são soluções de Big Data que armazenam dados heterogêneos em seu formato mais bruto, dentre esses dados, podem haver muitos dados pessoais, os quais podem receber diversos tratamentos, como: coleta, retenção, processamento, eliminação, modificação, entre outros.

Com o surgimento da LGPD, as organizações que possuem Data Lakes precisaram adotar soluções a fim de fornecer transparência sobre os tratamentos realizados nos dados, uma vez que, de acordo com a lei, é um direito do titular dos dados ter conhecimento sobre os tratamentos realizados nos seus dados pessoais. Como não há muitas soluções claras para resolução deste problema, uma possível solução consiste em, inicialmente, coletar metadados que descrevem os tratamentos realizados sobre os dados, persistir tais metadados em um repositório e, posteriormente, disponibilizá-los para fins de monitoramento, consultas e análises. Os metadados são informações de valor agregado geradas para organizar, descrever, rastrear e melhorar o acesso a objetos de informação, itens físicos e coleções, relacionados a esses objetos (Gilliland, 2016). A coleta de metadados pode ser realizada em diversos momentos do fluxo de dados em um Data Lake, incluindo

a ingestão de dados, edição de campos, clone de arquivos, remoção de alguns registros e tratamentos de dados em geral.

Atualmente, existem diversas ferramentas, como Dremio[1], SnowPlow[2] e DBT[3], que conseguem ter acesso a metadados de proveniência. Entretanto, nem sempre essas ferramentas são simples de utilizar ou disponibilizam uma versão gratuita que seja completa o suficiente para permitir a coleta e o uso de metadados necessários para as auditorias de cumprimento da LGPD. A literatura neste contexto é escassa, já que a LGPD é algo recente. Com isso, a fim de contribuir para o cumprimento da LGPD, neste trabalho apresentamos uma alternativa para a coleta e persistência de metadados. Acreditamos que a nossa proposta poderá ser utilizada por organizações que necessitem adaptar seu ambiente de Data Lake com o intuito de atender à LGPD. Diferentemente das ferramentas disponíveis, nossa proposta permite que, de forma simples e gratuita, e a partir de tecnologias existentes, organizações possam se adequar à LGPD.

1.2 Objetivo

O objetivo principal deste trabalho consiste em criar e disponibilizar um repositório com metadados de proveniência acerca do fluxo de ingestão de dados em um Data Lake, que permita a execução de consultas e análises para fins de monitoramento e cumprimento das leis de proteção de dados.

1.3 Estrutura do Trabalho

Este trabalho apresenta mais 4 capítulos. O Capítulo 2 apresenta de maneira mais aprofundada os conceitos básicos relacionados à Data Lake, fluxo de dados, metadados, dados de proveniência e rastreabilidade de dados. O Capítulo 3 introduz as ferramentas utilizadas e o desenvolvimento feito, enquanto o capítulo 4 mostra os resultados deste trabalho. Por fim, o Capítulo 5 apresenta as considerações finais e possíveis trabalhos futuros.

---

[1] https://www.dremio.com/
[2] https://snowplowanalytics.com/
[3] https://www.getdbt.com/

2. Fundamentação teórica

Neste capítulo serão abordados os principais conceitos do escopo deste trabalho, como Data Lakes e metadados de proveniência.

2.1 Data Lakes

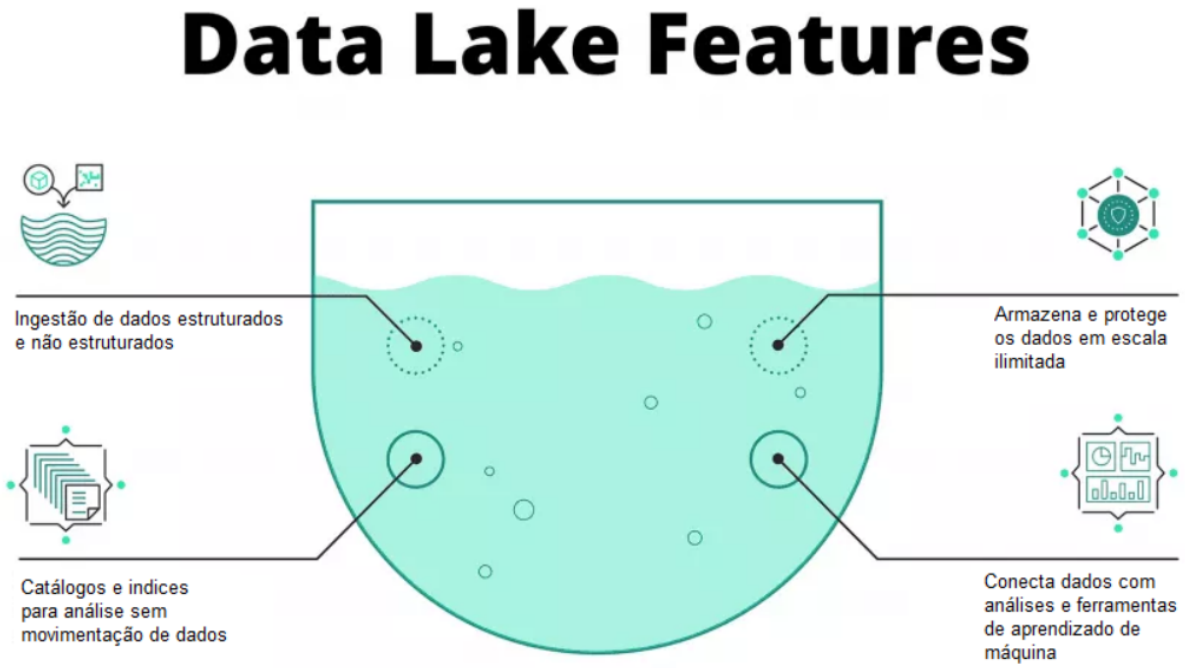
O conceito de Data Lake foi introduzido por Dixon como uma solução para as deficiências percebidas nos Data Marts, que são subdivisões específicas de negócios de data warehouses que permitem apenas subconjuntos de perguntas a serem respondidas (Dixon, 2010). Na literatura, Data Lakes também são referenciados como repositórios de dados (Chessell et al, 2014) e hubs de dados (Ganore, 2015; Laskowski, 2016), embora o termo Data Lake seja o mais frequente. Dixon define um Data Lake como um grande sistema de armazenamento para dados brutos e heterogêneos, alimentados por várias fontes de dados, para que os usuários explorem, extraiam e analisem os dados.

De acordo com Pegdwendé Sawadogo e Jérôme Darmont (2021), Data Lake também pode ser compreendido como um sistema de armazenamento e análise escalonável para dados de qualquer tipo, retidos em seu formato nativo e usados principalmente por especialistas em dados (estatísticos, cientistas e engenheiros de dados ou analistas) para extração de conhecimento.

A Figura 1 apresenta as principais características de um Data Lake, as quais são descritas a seguir:

- Ingestão de dados estruturados e não estruturados;
- Conexão com ferramentas de aprendizagem de máquina e análise de dados;
- Catalogação e indexação de dados para análise sem movimentação de dados;
- Armazenamento e segurança dos dados em escala ilimitada;

Figura 1: Data Lakes Features



Fonte: https://lakefs.io/data-lakes/

2.1.1 Componentes de um Data Lake

Um Data Lake possui 5 componentes principais em sua arquitetura (Figura 2), são eles:

- *Ingestão de dados* - altamente escalonável para extração de dados de várias fontes.
- *Armazenamento de dados* - escalável para armazenamento e processamento de dados brutos.
- *Segurança de dados* - Data lakes devem ser seguros com o uso de autenticação multifator, autorização, proteção de dados, entre outros.
- *Análise de dados* - podem ser analisadas de forma rápida e eficiente quando utilizadas ferramentas de aprendizado de máquina e feito as análises dos dados.
- *Gestão de dados* - ocorre durante todo o processo de ingestão, preparação, catalogação, integração e consulta de dados. Importante também para conseguir rastrear as mudanças que estiverem ocorrendo nos dados.

Figura 2: Componentes de um Data Lake

fonte: https://www.bbvaopenmind.com/en/technology/digital-world/data-lake-an-opportunity-or-a-dream-for-big-data/

2.1.2 Vantagens e desvantagens de um Data Lake

Um Data Lake possui várias vantagens sobre outras soluções de big data, assim como também possui desvantagens. Inicialmente, serão apresentadas as vantagens.

- O Data Lake oferece aos usuários acesso imediato a todos os dados que foram inseridos.
- O Data Lake não se limita a banco de dados relacionais ou transacionais, uma vez que aceitam qualquer tipo de estrutura e qualquer tipo de dados.
- Ajuda totalmente com a produção e análises mais avançadas sobre os dados.
- Oferece escalabilidade e flexibilidade para o sistema.
- Uma de suas principais vantagens é a centralização de diferentes fontes.

Assim como todo sistema, Data Lake também possui pontos contras que podem levar à equipe a não utilizá-lo. Com isso, vão ser apresentadas as desvantagens de um Data Lake.

- Área desconhecida de processamento de dados.
- Como são vários dados, de várias fontes, pode acabar sendo um caos.
- Ter que lidar com a questão da privacidade.
- Questão de integração.
- O maior risco de se utilizar um Data Lake é a segurança e o controle de acesso, pois alguns dados podem ser inseridos sem qualquer supervisão.

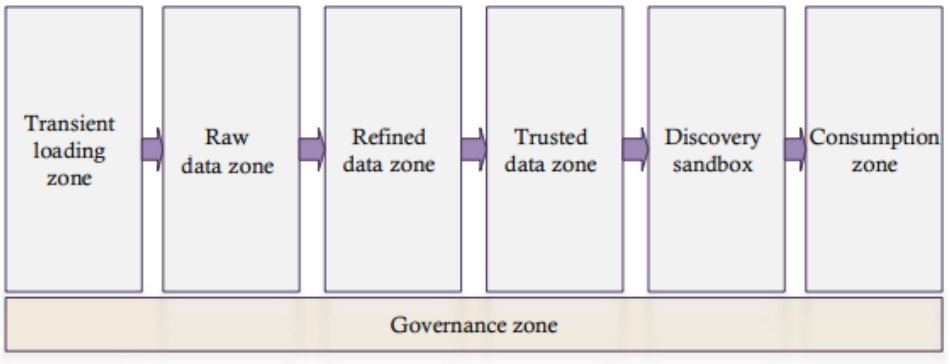
2.1.3 Arquiteturas de Data Lake

A literatura comumente aborda dois tipos de arquiteturas de Data Lake, são as arquiteturas de Zonas e arquiteturas de Lago (Giebler et al. 2019; Ravat and Zhao 2019a). Neste trabalho, utilizamos a arquitetura de Zonas. A seguir serão apresentados os principais componentes dessa arquitetura e seus respectivos papéis.

A arquitetura de Zonas tem como objetivo principal atribuir os dados a uma zona específica de acordo com seu grau de refinamento (Giebler et al. 2019). Para melhor compreensão dessa arquitetura, abordaremos o Data Lake de Zaloni (LaPlante e Sharma 2016) que adota uma arquitetura dividida em seis zonas (Figura 3):

1. A *Transient loading zone* lida com a ingestão de dados.
2. A *Raw data zone* lida com dados no formato bruto oriundos da zona de carregamento.
3. *Refined data zone* é onde os dados são tratados.
4. Na *Trusted data zone*, os dados são movidos para a *Discovery sandbox*, onde podem ser acessados por cientistas.
5. *Discovery sandbox* é a zona onde os usuários de negócios acessam.
6. Por fim, a *Consumption zone* possibilita gerenciar, monitorar e governar metadados, dados de qualidade e um catálogo de dados e segurança.

Figura 3: Zaloni's zone architecture



Fonte: LaPlante and Sharma, 2016

Nem todas as arquiteturas de zonas precisam seguir à risca a proposta sugerida por (LaPlante and Sharma 2016). A zona de descoberta (*Discovery sandbox*), por exemplo, não é utilizada neste trabalho, pois não há a necessidade no presente escopo. Sendo assim, foi desenvolvida uma arquitetura funcional que permite o recebimento de qualquer tipo de dado ou conjunto de dados para a execução de um tratamento e ingestão em um Data Lake. Durante esse processo, os metadados de proveniência poderão ser coletados.

2.2 Metadados de proveniência

Como dito anteriormente, metadados são informações de valor agregado para organizar, descrever, rastrear e melhorar o acesso a objetos de informação e itens físicos e coleções, relacionados a esses objetos (Gilliland, 2016).

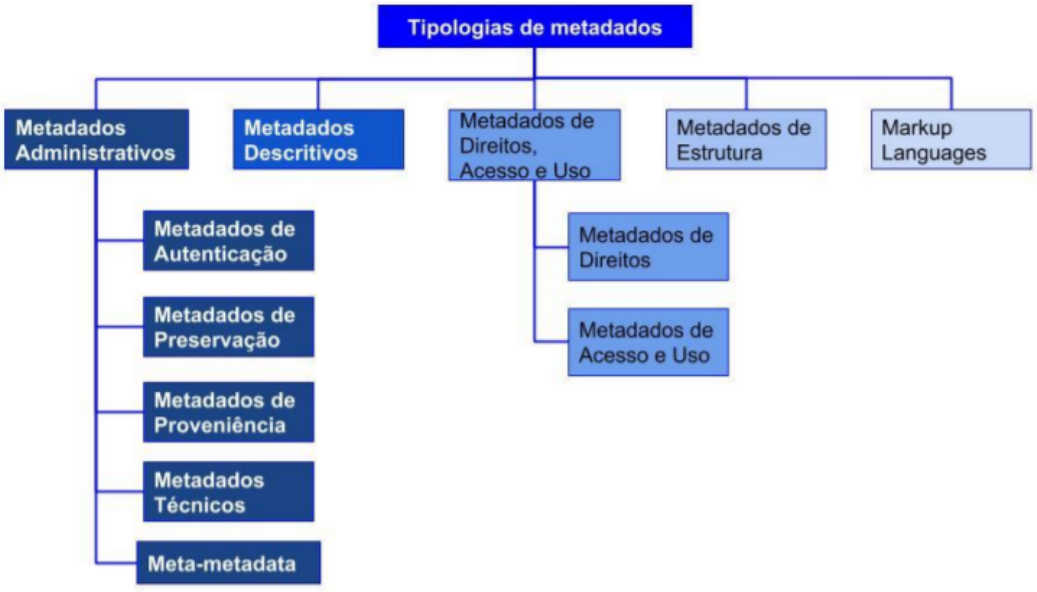
Porém, de acordo com o contexto, os metadados podem ter uma função específica, a qual está diretamente ligada a sua tipologia (Arakaki, 2019). A Figura 4 apresenta alguns exemplos de tipologias de metadados, dentre os quais destacam-se as tipologias administrativa, autenticação, preservação, proveniência, descritivos, entre outros.

Tanto os dados quanto os metadados são essenciais para a capacidade de interpretar um determinado item de dados. Mesmo quando o mesmo indivíduo está coletando os dados e os interpretando, os metadados e a proveniência são importantes. A proveniência de metadados é uma área importante de aplicação de

metadados. Os metadados de proveniência informam os usuários sobre as fontes e origens dos metadados, contribuindo assim para sua credibilidade e veracidade.

Este trabalho foca apenas na tipologia de proveniência, pois é por meio da proveniência que o usuário é capaz de determinar de onde veio um objeto de informação específico.

Figura 4: Tipologias de metadados



Fonte: Arakaki, 2019

2.3 Considerações finais

Este capítulo apresentou os conceitos fundamentais sobre Data Lakes e metadados de proveniência.

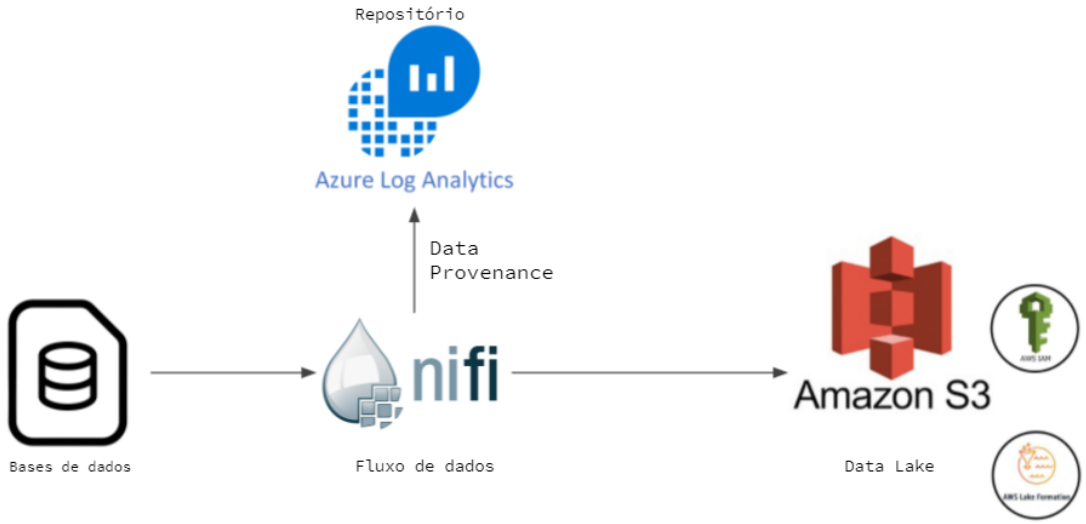
3. Especificação da Arquitetura para Coleta de Metadados

Neste capítulo, serão apresentados detalhes da arquitetura projetada para desenvolvimento do estudo, assim como as ferramentas e fluxos utilizados. O objetivo dessa arquitetura é apresentar uma forma pela qual é possível a criação de um repositório com metadados de proveniência acerca do fluxo de ingestão de dados em um Data Lake, que permita a execução de consultas e análises para fins de monitoramento.

3.1 Arquitetura proposta

A Figura 5 apresenta a arquitetura proposta para a criação de um repositório com metadados de proveniência acerca do fluxo de ingestão de dados em um Data Lake.

Figura 5: Arquitetura



Fonte: o autor, 2021

Nesta arquitetura, um dos principais componentes é o Apache Nifi. Por meio dele, são desenvolvidos os fluxos de ingestão de dados e também é realizada a coleta dos metadados de proveniência. Além dele, o Azure Analytics desempenha um papel essencial no armazenamento dos metadados e disponibilização destes para consulta e análise. No capítulo 4, estes metadados serão detalhados. A seguir, serão apresentadas as tecnologias e serviços utilizados no escopo do presente trabalho.

3.2 Apache Nifi

Apache Nifi é um software de código aberto que possibilita automatizar e gerenciar fluxos de dados entre sistemas. Fluxo de dados é o caminho pelo qual os dados percorrem para chegar em algum repositório final. Durante o fluxo, um arquivo pode passar por vários procedimentos, tratamentos e refinamentos.

Além disso, o Apache Nifi fornece uma interface de fácil entendimento para o desenvolvimento dos *pipelines* de dados. O diferencial dessa ferramenta comparada às outras é a geração de metadados de proveniência detalhados sobre eventos que ocorrem nos fluxos de dados. Para cada transformação executada dentro de um fluxo de dados, são gerados metadados que evidenciam o que foi realizado naquele procedimento.

Além disso, ela fornece processos e configurações de controle que facilitam a criação dos fluxos de dados para os usuários. Sabemos que a instalação de qualquer ambiente pode ser trabalhoso, mas isso não ocorre com o Apache Nifi, pois a documentação é de fácil acesso e leitura, sendo necessário apenas alguns passos para uma instalação bem sucedida. A Figura 6 apresenta os principais componentes do Apache Nifi.

- *JVM (Java Virtual Machine)* é onde o NIFI é executado.
- *Web Server* é para hospedar a API de comando e controle baseada em HTTP do Nifi.
- *Processador* fica responsável pela busca de dados do sistema ou armazenar no sistema de destino.
- *FlowFile Repository* é onde controla o arquivo durante o fluxo.
- *Content Repository* é onde estão os bytes de conteúdo vigentes de um arquivo.
- *Provenance Repository* é onde ficam armazenados os dados de origem.

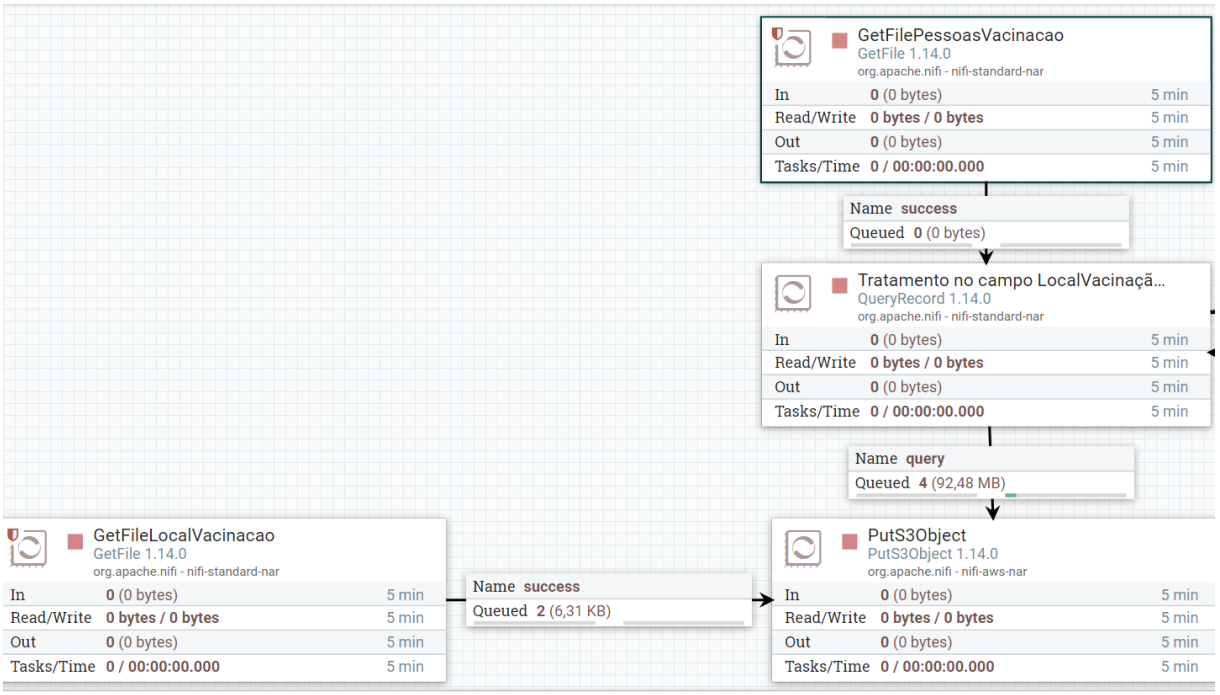
Figura 6: Nifi Architecture

Fonte: https://blog.dsacademy.com.br/ingestao-de-dados-em-tempo-real-com-apache-nifi/

O fluxo desenvolvido para a validação da proposta, apresentado na Figura 7, extrai duas bases de dados localmente, aplica um refinamento em um atributo comum às duas bases, e por fim realiza a inserção dos dados em um Data Lake. Simultaneamente, são gerados metadados de proveniência sobre todos os procedimentos realizados no fluxo, desde a extração até a ingestão dos dados no lake.

Para a execução do fluxo, foram utilizados alguns recursos que a ferramenta disponibiliza. Um deles é o *GetFile*, com ele conseguimos coletar os arquivos contendo os dados que estão armazenados localmente no computador. Outro recurso utilizado é o *QueryRecord*, esse recurso avalia uma ou mais consultas SQL em relação ao conteúdo de um arquivo. O resultado da consulta SQL então se torna o conteúdo do arquivo de saída. Por último, foi utilizado o *PutS3Object* para a inserção do resultado do fluxo no banco de dados.

Figura 7: Flow / Pipeline (Ingestion)



Fonte: o autor, 2021

3.3 AWS

Amazon Web Services, também conhecida como AWS, é uma plataforma de serviços para computação em nuvem oferecida pela empresa Amazon. Os serviços são oferecidos de forma distribuída, sendo esse um dos principais benefícios da plataforma. Amplamente utilizada, a AWS disponibiliza diversos recursos que permitem escalabilidade e segurança para os Data Lakes.

Para armazenar os dados do fluxo apresentado anteriormente, foram utilizados alguns serviços para integração entre o Apache Nifi e o Data Lake. Estes serviços foram selecionados com o objetivo de facilitar o processo de validação da arquitetura. Na sequência, introduziremos os principais aspectos destes serviços.

3.3.1 Amazon S3

A Amazon Simple Storage Service (Amazon S3) é um serviço de armazenamento de qualquer tipo de objeto que oferece escalabilidade, segurança e performance. O S3 fornece recursos de gerenciamento fáceis de usar, de maneira que é possível organizar os dados e configurar os controles de acesso para atender a requisitos específicos comerciais, organizacionais e de conformidade.

Existem outras ferramentas de armazenamento que podem ser utilizadas para receber o resultado do fluxo de ingestão apresentado na contextualização, no entanto, a Amazon S3 dispõe de uma configuração mais prática e amigável, além de uma documentação completa de fácil acesso e leitura.

3.3.2 Amazon IAM

A AWS *Identity and Access Management* (IAM) é um serviço que a Amazon disponibiliza para gerenciamento de acesso aos recursos da AWS de forma segura. Resumidamente, o IAM é utilizado para controlar quem é autenticado e autorizado a usar os recursos.

Por meio dele, foi criado o grupo de políticas responsável por gerenciar o nível de acesso de um usuário do Data Lake.

3.3.3 Data Lake Formation

O AWS Lake Formation, descrito na Figura 8, se trata de um serviço oferecido pela Amazon que facilita a configuração de um Data Lake. O Lake Formation ajuda a criar, proteger e gerenciar Data Lakes. Com ele também é possível gerenciar fluxos de ETL, catálogo de dados, configurações de segurança e controle de acesso. Além disso, o Lake Formation é integrado a outros sistemas e serviços que a própria Amazon disponibiliza, como o Amazon S3.

Utilizamos o Lake Formation para criação do Data Lake em si. Na prática, ele se integra com o S3 para listar e gerenciar os dados inseridos.

- *Source Crawlers* é onde se indica as fontes de dados para o Lake Formation para que ele examine essas fontes e mova os dados para um novo data lake no Amazon S3 ou bancos relacionais.
- *ETL and data prep* são feitos fluxos de dados que ingerem, limpam, transformam e organizam os dados brutos.
- *Data Catalog* pode ser criada e gerenciada contendo metadados sobre fontes de dados e dados no Data Lake.

- *Security settings* e *Access control* é onde são definidas políticas de acesso a dados granulares para os metadados e dados, por meio de um modelo de permissões de concessão ou revogação.

Figura 8: AWS Lake Formation

Fonte: https://aws.amazon.com/pt/lake-formation/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc

3.4 Log Analytics Workspace

O Log Analytics Workspace permite criar, por meio do portal do Azure, um ambiente exclusivo para armazenar os logs de metadados coletados dos seus dados. Cada espaço tem seu próprio repositório e suas configurações de dados, além de ter sua própria configuração de ambiente.

Existem outros repositórios que podem ser utilizados para receber os metadados de proveniência, mas essa ferramenta do Azure foi escolhida pela facilidade de configuração e por disponibilizar, na própria ferramenta, meios para consultar os metadados coletados.

3.5 Integração entre as tecnologias

Para realização da coleta de todos os metadados necessários para validação da arquitetura proposta no estudo, foi necessária a integração entre os sistemas e

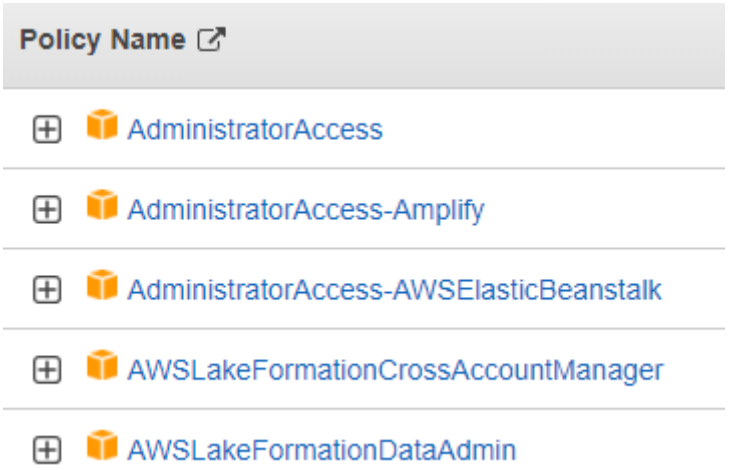
serviços descritos anteriormente. Nesta seção, introduziremos as configurações e permissões utilizadas para integração e funcionamento da arquitetura.

3.5.1 Apache Nifi e Amazon S3

Para conexão entre o Nifi e a Amazon S3, utilizamos um processo que o Nifi disponibiliza, denominado de *PutS3Object*. Para execução, é necessário configurar os seguintes campos: *Bucket*, *Access Key Id* e *Secret Access Key*.

Para obtenção dos campos *Access Key Id e Secret Access Key*, é preciso criar um usuário no Amazon IAM que possua as permissões de um usuário administrador. Ao criar um usuário, serão fornecidos os campos necessários para a configuração. Para que o usuário possua permissões de um administrador, é preciso criar um grupo e adicionar as permissões sugeridas na Figura 9.

Figura 9: Permissões IAM

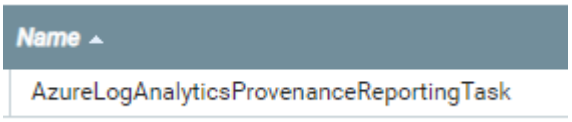


Fonte: o autor

3.5.2 Coleta dos metadados de proveniência

Para armazenamento dos metadados de proveniência, foi utilizada uma configuração de controle denominada *AzureLogAnalyticsProvenanceReportingTask* (Figura 10). Esta configuração é responsável por publicar os eventos de proveniência do Nifi em um espaço de trabalho do Azure Log Analytics.

Figura 10: Nifi Settings

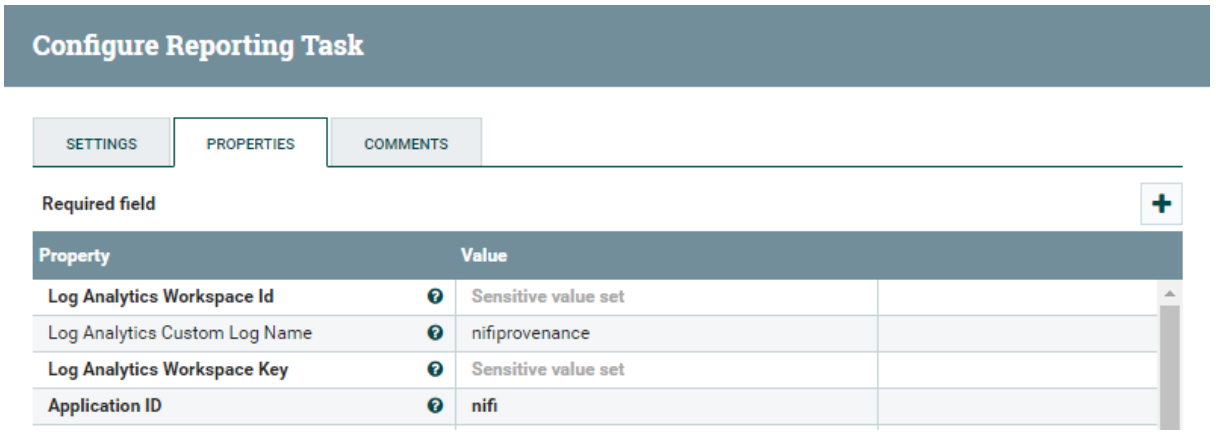


Fonte: o autor.

Para realizar a coleta, é necessário preencher os campos Log Analytics Workspace Id e Log Analytics Workspace Key (Figura 11). Para preenchimento correto desses campos, é preciso criar uma conta no Azure e obter tais informações. O passo a passo seguinte, fornece um guia de como realizar a abertura da conta no Azure:

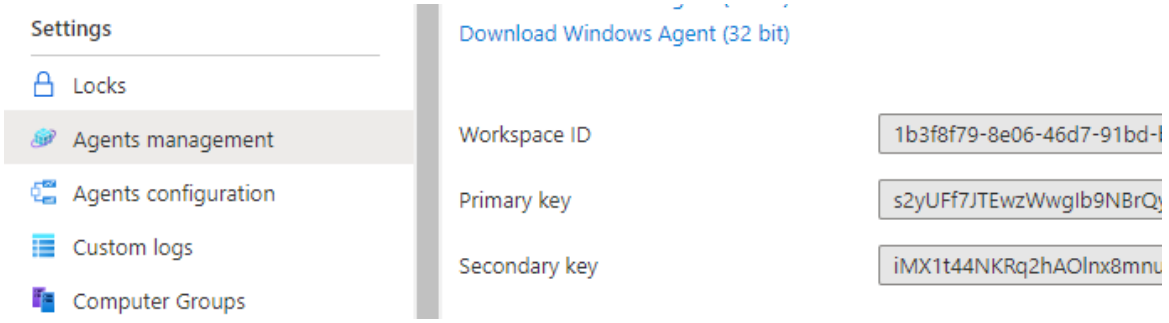
1. Criar uma conta no Microsoft Azure
2. Criar um Log Analytics Workspace
3. Ir no Menu Agents management
4. Copiar o Id e a Key (Figura 12)

Figura 11: Configure Reporting Task



Fonte: o autor

Figura 12: Workspace ID e Key



Fonte: o autor

3.6 Considerações Finais

Neste capítulo, foram descritas as tecnologias utilizadas para a elaboração da arquitetura que coleta e armazena metadados de proveniência para fluxos de ingestão de dados em Data Lakes. No capítulo a seguir, serão apresentados os resultados obtidos.

4. Aplicação da Arquitetura Proposta

Este capítulo apresenta um exemplo de uso da arquitetura proposta e discute a coleta dos metadados de proveniência durante o fluxo de ingestão de dados em um Data Lake.

4.1 Fontes de dados

Para exemplificar o uso da arquitetura proposta, foram utilizados dados do Portal de Dados Abertos da Prefeitura de Recife. O Portal de Dados Abertos da Prefeitura de Recife disponibiliza dados gerados pelas secretarias e órgãos da gestão municipal. Neste portal são disponibilizados dados e recursos nos formatos de CSV, JSON, PDF, GeoJSON, XLSX, ZIP e SHP. Para validação da arquitetura proposta, utilizamos arquivos no formato CSV.

A escolha para utilização do Portal de Dados Abertos de Recife justifica-se pelo fato de ser um repositório público de recursos, além da variedade de temas e formatos de bases. Foram extraídas duas bases de dados que estabelecem um relacionamento por meio de um atributo em comum. A primeira base (Figura 13) dispõe de 10 atributos e cerca de 100 mil registros que listam as pessoas vacinadas contra a Covid-19 na cidade do Recife. Já a segunda base (Figura 14), detém 9 atributos e 26 registros que dizem respeito aos locais de vacinação disponíveis para a população. O relacionamento entre as bases ocorre por meio do atributo local_vacinacao.

Figura 13: Relação de pessoas vacinadas

Grade | Gráfico | Mapa 1487809 records « 1 – 100 » Sea

_id	cpf	nome	sexo	grupo	vacina	lote	dose	data_va...	local_va...
1	***.574.5...	AABAN ...	MASCU...	TRABAL...	2 - CHA...	216VCD...	1	2021-07-...	DRIVE T...
2	***.886.9...	AAMAN...	FEMININO	TRABAL...	1 - COR...	210043	2	2021-04-...	CENTR...
3	***.886.9...	AAMAN...	FEMININO	TRABAL...	1 - COR...	210016	1	2021-11-...	CENTR...
4	***.695.2...	AANTO...	MASCU...	IDOSOS	1 - COR...	210093	2	2021-03-...	CENTR...
5	***.695.2...	AANTO...	MASCU...	IDOSOS	1 - COR...	202009014	1	2021-10-...	CENTR...
6	***.091.0...	AARÃO ...	MASCU...	PÚBLIC...	2 - CHA...	215VCD...	1	2021-07-...	DRIVE T...
7	***.284.1...	AARÃO ...	MASCU...	PÚBLIC...	2 - CHA...	215VCD...	1	2021-06-...	DRIVE T...
8	***.599.4...	AARAO ...	MASCU...	PÚBLIC...	2 - CHA...	214VCD...	1	2021-01-...	DRIVE T...
9	***.599.4...	AARAO ...	MASCU...	PÚBLIC...	2 - CHA...	210197	2	2021-01-...	DRIVE T...

Fonte: http://dados.recife.pe.gov.br/dataset/relacao-de-pessoas-vacinadas-covid-19

Figura 14: Locais de vacinação

Grade | Gráfico | Mapa | 26 records | « | 1 – 26 | »

_id	Local	lograd...	bairro	fone	como_u...	horario	latitude	longitude
1	Centro d...	Av. Gen....	Cordeiro	32259400	Os usuár...	Domingo...	-8.0596332	-34.9260...
2	Centro d...	R. Cône...	Tamarin...		Os usuár...	Domingo...	-8.02759...	-34.8986...
3	Centro d...	Av. Mal. ...	Imbiribeira	34977455	Os usuár...	Domingo...	-8.1169725	-34.9132...
4	Centro d...	Av. Nort...	Macaxeira		Os usuár...	Domingo...	-8.01571...	-34.9312...
5	Centro d...	Av. Caxa...	Cordeiro		Os usuár...	Domingo...	-8.04682...	-34.9271...
6	Centro d...	Rua Carl...	Hipódromo	33556143	Os usuár...	Domingo...	-8.0335653	-34.8879...
7	Centro d...	Rua Do...	Dois Irm...		Os usuár...	Domingo...	-8.01815...	-34.9524...
8	Centro d...	Av. Dois ...	Ibura		Os usuár...	Domingo...	-8.109992	-34.936492
25	Centro d...	R. do Ap...	Recife		Recome...	Domingo...	-8.06028...	-34.8720...
9	Centro d...	Rua do ...	Boa Vista		Os usuár...	Domingo...	-8.0556667	-34.8881...

Fonte:http://dados.recife.pe.gov.br/dataset/campanha-de-vacinacao-covid-19/resource/dbf660d2-1ee5-451a-94ca-7c316e50d0ad

4.2 Eventos

Durante o fluxo de ingestão de dados ocorrem eventos. Esses eventos referem-se a tratamentos realizados nos dados, como, inserções, envio, modificações e exclusão. No fluxo desenvolvido para validação da arquitetura apresentada no capítulo anterior, um subconjunto de eventos que a ferramenta é capaz de executar, foram processados sequencialmente indicando quais tratamentos ocorreram naquele *pipeline*. São eles:

- ATTRIBUTES_MODIFIED: Indica que os atributos de um arquivo foram modificados de alguma forma.
- RECEIVE: Indica um evento de proveniência para receber dados de um processo externo.
- FORK: Indica que um ou mais arquivos foram derivados de um arquivo pai.
- ROUTE: Indica que um arquivo foi roteado para um relacionamento especificado e fornece informações sobre por que o arquivo foi roteado para este relacionamento.
- DROP: Indica um evento de proveniência para a conclusão da vida de um objeto por algum motivo diferente da expiração do objeto.

- SEND: Indica um evento de proveniência para enviar dados para um processo externo

A tabela a seguir, exemplifica como os eventos descritos anteriormente são exibidos na ferramenta (Figura 15).

Figura 15: Tipos de eventos

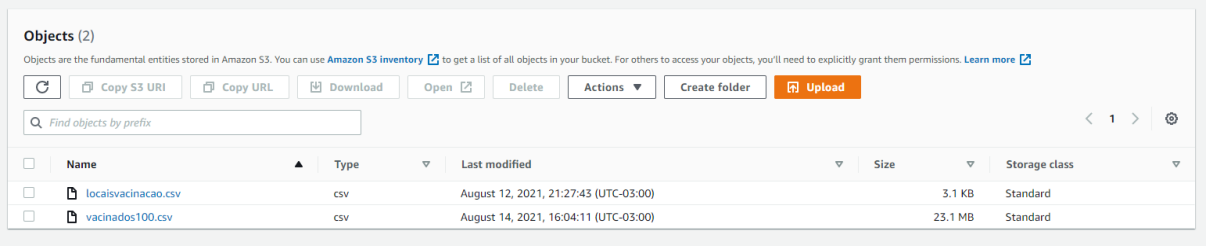
Type	Component Name
DROP	PutS3Object
ATTRIBUTES_MODIFIED	PutS3Object
SEND	PutS3Object
ROUTE	Tratamento no campo LocalVacinação
FORK	Tratamento no campo LocalVacinação
RECEIVE	GetFilePessoasVacinacao

Fonte: o autor

4.3 Armazenamento

No exemplo desenvolvido, a persistência dos dados ocorre em dois momentos e cada um deles utiliza uma ferramenta específica. Primeiro, são inseridos no Amazon S3 os dados refinados, como resultado do fluxo de ingestão (Figura 16). A partir de então, esses dados passam a fazer parte do Data Lake e estão disponíveis para uso.

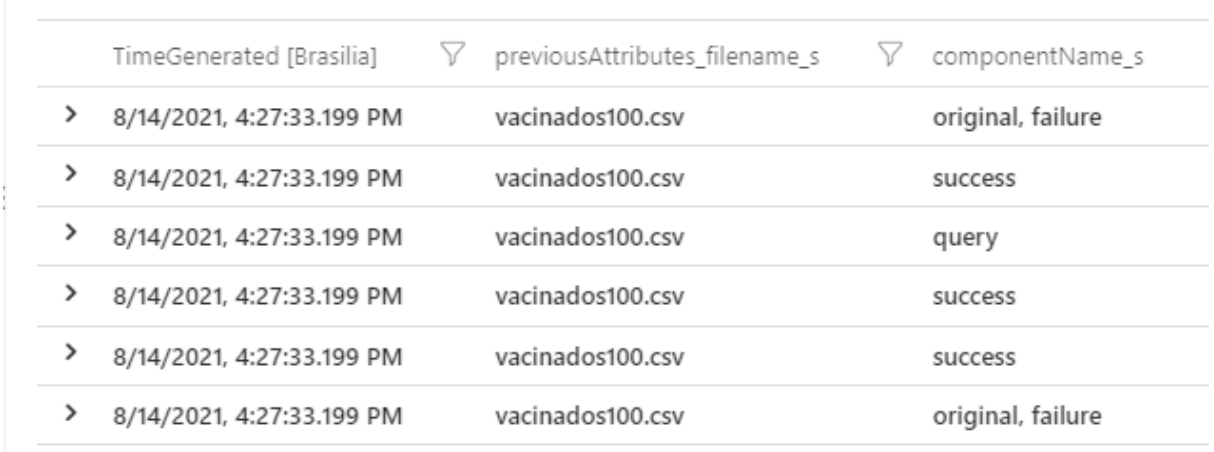
Figura 16: Armazenamento AWS S3



Fonte: o autor

Logo em seguida, são armazenados os metadados de proveniência no Azure Log Analytics, que são capturados pelo Nifi a partir dos eventos gerados para cada tratamento que ocorre nos dados durante o fluxo criado (Figura 17). Na seção a seguir, estes eventos e metadados associados serão detalhados.

Figura 17: Repositório metadados proveniência



	TimeGenerated [Brasilia]	previousAttributes_filename_s	componentName_s
>	8/14/2021, 4:27:33.199 PM	vacinados100.csv	original, failure
>	8/14/2021, 4:27:33.199 PM	vacinados100.csv	success
>	8/14/2021, 4:27:33.199 PM	vacinados100.csv	query
>	8/14/2021, 4:27:33.199 PM	vacinados100.csv	success
>	8/14/2021, 4:27:33.199 PM	vacinados100.csv	success
>	8/14/2021, 4:27:33.199 PM	vacinados100.csv	original, failure

Fonte: o autor

4.4 Metadados

Durante a execução do fluxo dos dados, ocorrem vários tipos de eventos, como inserção, alteração de atributo, clonagem, entre outros, e sobre cada evento são gerados vários metadados de proveniência. A própria ferramenta Nifi mostra alguns dados coletados durante a execução do fluxo, porém no repositório é possível ter acesso a dados mais detalhados sobre o que ocorreu durante o evento.

O Nifi disponibiliza uma tabela (figuras 18 e 19) onde é possível visualizar, de modo simples, os metadados de proveniência. Essa tabela, disponibiliza algumas informações em colunas, sendo elas:

- *Date/Time:* informa o momento em que ocorreu um evento em um determinado dado.
- *Type:* informa qual foi o tipo desse evento.
- *FlowFile Uuid:* identificador universalmente único (UUID) atribuído a um arquivo.
- *Size:* informa o tamanho do arquivo.
- *Component Name:* informa qual o nome do arquivo.
- *Component Type:* informa qual componente fez o procedimento sobre o dado.

Figura 18: Nifi Data Provenance (parte 1)

NiFi Data Provenance

Displaying 15 of 15
Oldest event available: 08/06/2021 20:48:48 GFT

Filter | by component name

Date/Time	Type	FlowFile Uuid
08/18/2021 18:25:25.222 GFT	ROUTE	4702912c-cf2a-4037-b88e-342b4c5af874
08/18/2021 18:25:25.222 GFT	FORK	684ff7ae-5cd1-48c5-b0bf-2aeb598e0e7e
08/18/2021 18:25:25.075 GFT	DROP	aebc6a63-2497-4318-b75b-b08a562c7882
08/18/2021 18:25:25.075 GFT	ATTRIBUTES_MODIFIED	aebc6a63-2497-4318-b75b-b08a562c7882
08/18/2021 18:25:25.075 GFT	SEND	aebc6a63-2497-4318-b75b-b08a562c7882
08/18/2021 18:25:17.773 GFT	RECEIVE	684ff7ae-5cd1-48c5-b0bf-2aeb598e0e7e
08/18/2021 18:25:15.657 GFT	RECEIVE	aebc6a63-2497-4318-b75b-b08a562c7882
08/18/2021 18:25:10.214 GFT	DROP	5571dd01-047c-4c66-8906-126ab4d6b5c0
08/18/2021 18:25:10.214 GFT	DROP	184cf02d-29de-4f32-94c2-8a758bbf638e
08/18/2021 18:25:10.214 GFT	DROP	afe8a9dd-67c9-4a16-922b-f7a4469429d7
08/18/2021 18:25:04.459 GFT	DROP	f8652846-1354-4960-9041-7b48a7a79767
08/18/2021 18:25:04.459 GFT	DROP	d9ff2381-e212-4a0c-913f-8eb802cd70ee
08/18/2021 18:24:59.568 GFT	DROP	d1442102-f200-42d2-b92b-233c45c797df
08/18/2021 18:24:59.567 GFT	DROP	5603463e-9105-455f-ad18-0553a2c5b22d
08/18/2021 18:24:59.566 GFT	DROP	96302e6b-b05d-4997-aa69-d820c4b1bc36

Fonte: o autor

Figura 19: Nifi Data Provenance (parte 2)

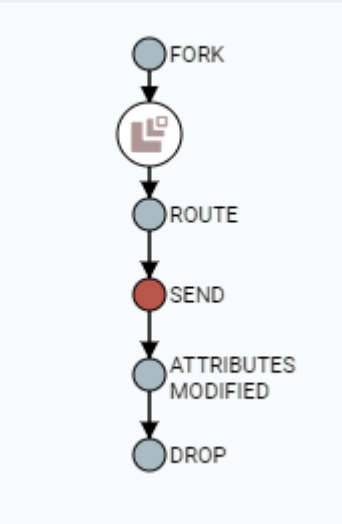
Size	Component Name	Component Type	
23,12 MB	Tratamento no campo LocalVacinação	QueryRecord	
24,1 MB	Tratamento no campo LocalVacinação	QueryRecord	
3,16 KB	PutS3Object	PutS3Object	
3,16 KB	PutS3Object	PutS3Object	
3,16 KB	PutS3Object	PutS3Object	
24,1 MB	GetFilePessoasVacinacao	GetFile	
3,16 KB	GetFileLocalVacinacao	GetFile	
0 bytes	original, failure	Connection	
0 bytes	original, failure	Connection	
0 bytes	original, failure	Connection	
0 bytes	success	Connection	
0 bytes	success	Connection	
0 bytes	query	Connection	
0 bytes	query	Connection	
0 bytes	query	Connection	

Fonte: o autor

Também nessa tabela de proveniência dos metadados (Figura 19), conseguimos ter acesso ao fluxo que o arquivo percorreu clicando no ícone de "três bolinhas" que está na última coluna, e os eventos que ocorreram durante o fluxo de forma gráfica. (Figura 20).

Na figura 20 podemos observar que tem um ponto em vermelho, isso se trata de qual tipo de evento o usuário quis olhar com mais detalhes. No caso, desse exemplo, foi clicado no evento do tipo “SEND”.

Figura 20: Fluxo de eventos

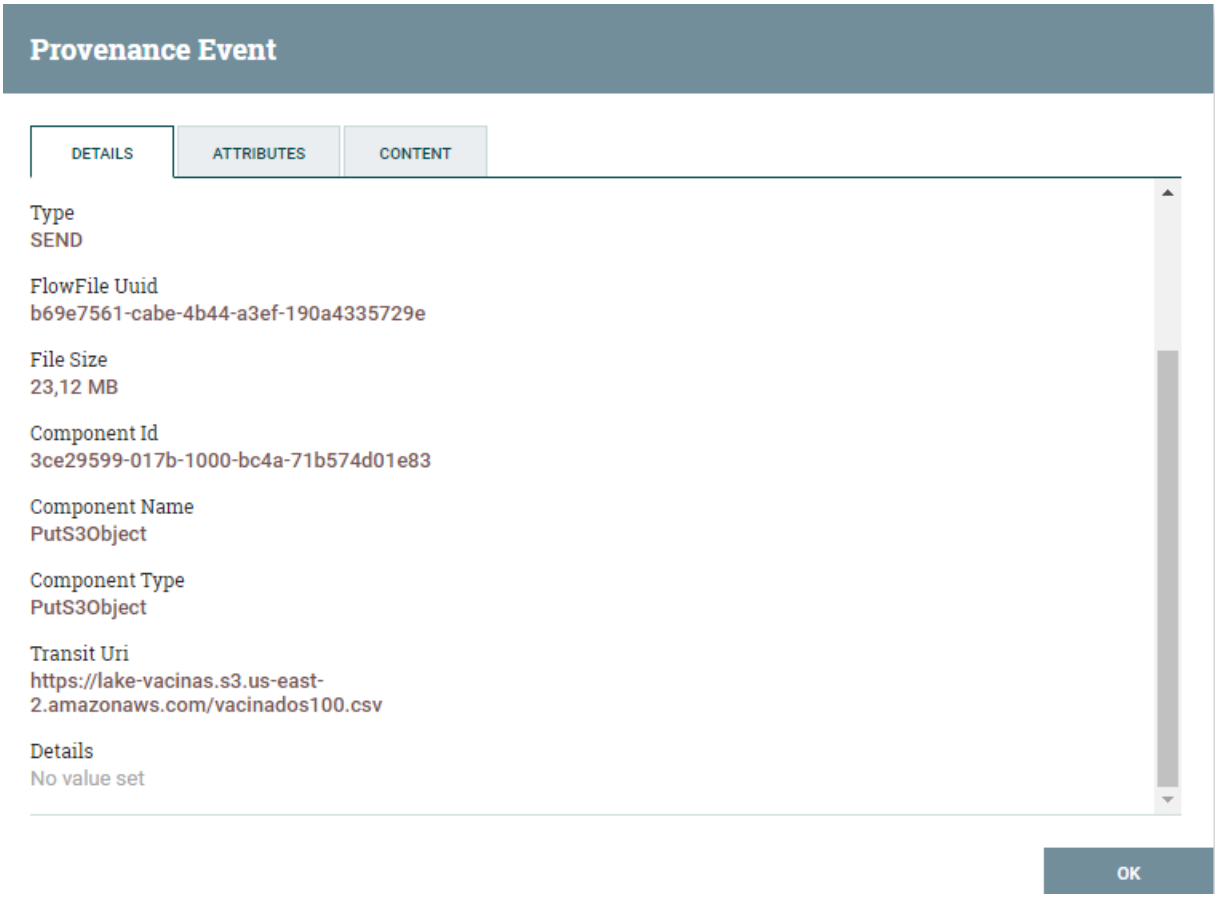


Fonte: o autor

Em cada evento nesse fluxo, ao clicar duas vezes em uma dessas bolinhas, conseguimos acessar com mais detalhes o que ocorreu (Figura 21). Com isso, conseguimos ter acesso aos detalhes do evento, atributos e ao conteúdo.

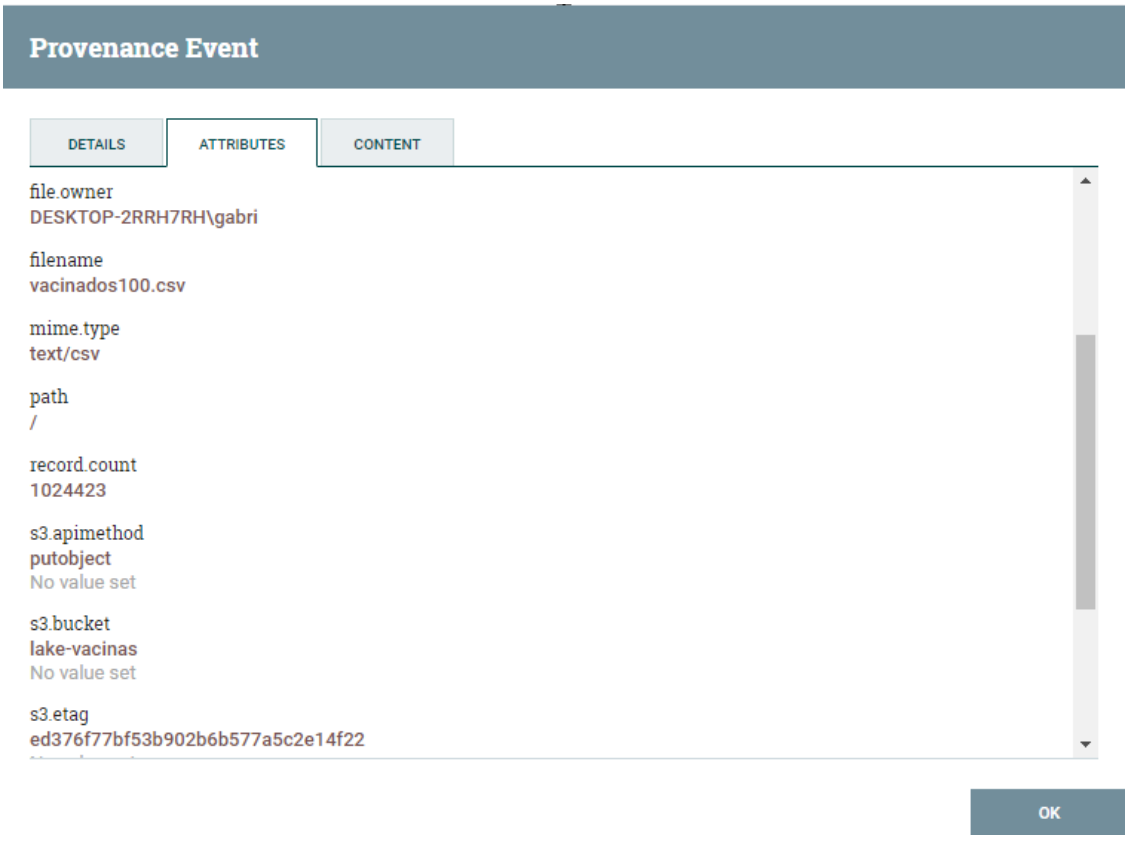
No detalhe (Figura 21) conseguimos visualizar em qual componente ocorreu aquele evento, o tamanho do arquivo, qual o tipo do evento, o identificador único do arquivo no fluxo, entre outros. Nos atributos (Figura 22) conseguimos visualizar o nome do arquivo, quantos registros o arquivo possui, o tipo do arquivo, a fonte do arquivo, entre outros metadados. No conteúdo (Figura 23) conseguimos visualizar e fazer download do arquivo antes e após passar pelo evento.

Figura 21: Detalhe de um evento



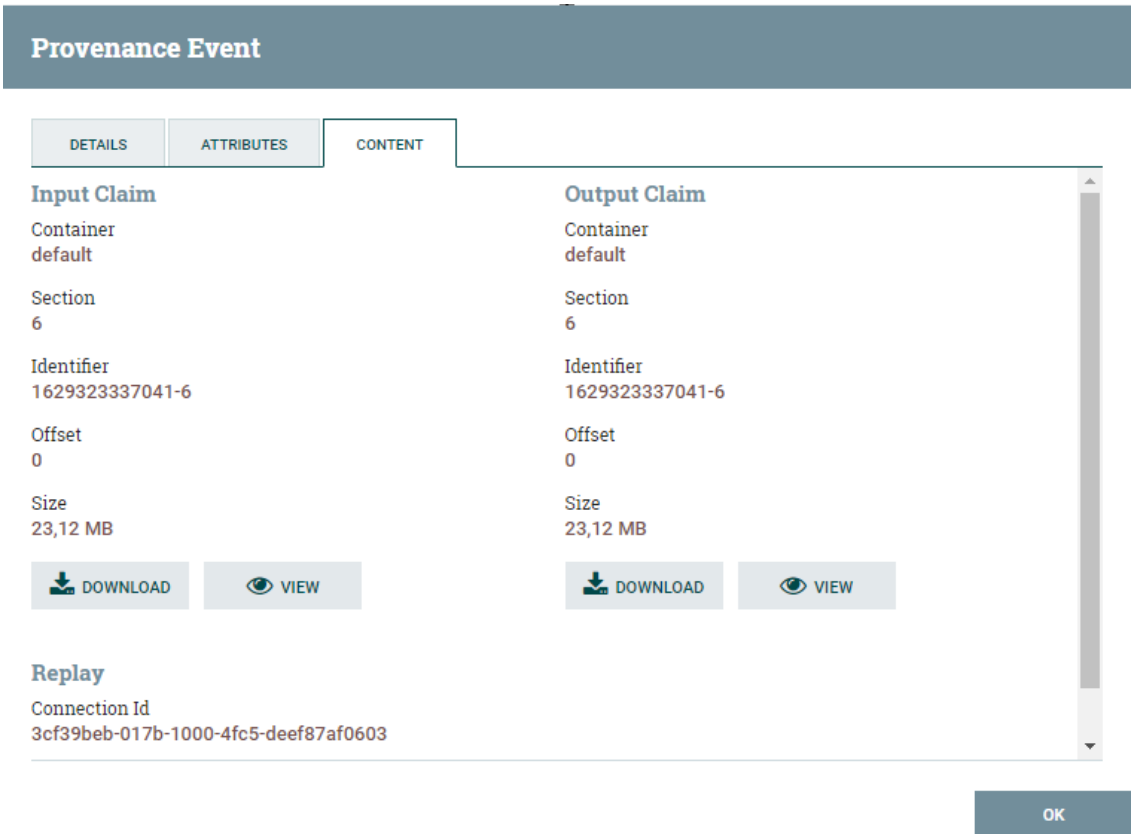
Fonte: o autor

Figura 22: Atributos de um evento



Fonte: o autor

Figura 23: Conteúdo de um evento



Fonte: o autor

Os metadados de proveniência também podem ser analisados a partir do Analytics Workspace logs (Figura 24). O acesso a partir do repositório, oferece maiores detalhes sobre os metadados de proveniência, assim podemos utilizar as colunas disponíveis para monitorar um data lake. Existem outras colunas, mas as principais para esse trabalho são:

- *TenantId:* contém a ID do registro no repositório.
- *TimeGenerated [UTC]:* mostra a hora em que o incidente ocorreu em UTC.
- *eventType_s:* indica qual evento ocorreu naquele momento do fluxo.
- *details_s:* mostra uma mensagem detalhando o incidente e quem causou.
- *actorHostname_s:* indica o nome do usuário na plataforma.
- *componentName_s:* indica em qual processo, ou qual momento do fluxo, ocorreu o registro.

- *previousAttributes_filename_s* indica qual o nome do arquivo foi alterado.
- *updatedAttributes_s3_bucket_s:* indica qual armazenamento foi alterado.
- *platform_s:* indica de qual plataforma foram coletados os metadados.

Figura 24: Analytics Workspace logs

TenantId	1b3f8f79-8e06-46d7-91bd-b2b70c423007
SourceSystem	RestAPI
TimeGenerated [UTC]	2021-08-23T19:35:51.336Z
updatedAttributes_s3_key_s	vacinados100.csv
previousAttributes_filename_s	vacinados100.csv
componentName_s	PutS3Object
processGroupId_g	185a4544-017b-1000-4342-44376619331f
processGroupName_s	NiFi Flow
previousAttributes_absolute_path_s	C:\Users\gabri\Desktop\TCC-csv/
previousAttributes_file_lastModifiedTime_t [UTC]	2021-08-12T20:29:08Z
previousAttributes_file_creationTime_t [UTC]	2021-08-12T20:29:07Z
previousAttributes_file_lastAccessTime_t [UTC]	2021-08-18T21:41:58Z
previousAttributes_file_owner_s	DESKTOP-2RRH7RH\gabri
previousAttributes_record_count_s	1024423
updatedAttributes_s3_etag_g	ed376f77-bf53-b902-b6b5-77a5c2e14f22
updatedAttributes_s3_storeClass_s	STANDARD
updatedAttributes_s3_apimethod_s	putobject
updatedAttributes_s3_bucket_s	lake-vacinas
transitUri_s	https://lake-vacinas.s3.us-east-2.amazonaws.com/vacinados100.csv
previousEntitySize_d	24242018
previousAttributes_path_s	/
previousAttributes_mime_type_s	text/csv
previousAttributes_uuid_g	838d11bf-d9ed-4ad3-a41b-6b1f0b8c6f4a
eventId_g	83b8ba9c-51bc-445c-9c61-96ffff19a65e
eventOrdinal_d	80456
eventType_s	SEND
timestampMillis_d	1629747305394
timestamp_t [UTC]	2021-08-23T19:35:05.394Z
durationMillis_d	9384
lineageStart_d	1629322917980
componentId_g	3ce29599-017b-1000-bc4a-71b574d01e83
componentType_s	PutS3Object
entityId_g	838d11bf-d9ed-4ad3-a41b-6b1f0b8c6f4a
entityType_s	org.apache.nifi.flowfile.FlowFile
entitySize_d	24242018
actorHostname_s	DESKTOP-2RRH7RH
contentURI_s	http://DESKTOP-2RRH7RH:8443/nifi-api/provenance-events/80456/content/output
previousContentURI_s	http://DESKTOP-2RRH7RH:8443/nifi-api/provenance-events/80456/content/input
parentIds_s	[]
childIds_s	[]
platform_s	nifi
application_s	NiFi Flow

Fonte: o autor

4.5 Considerações Finais

Este capítulo apresentou os resultados obtidos a partir da execução do fluxo desenvolvido na ferramenta Apache Nifi, em conjunto com ferramentas da AWS e da Azure. Esta demonstração teve como objetivo mostrar os metadados de proveniência coletados, após passar por todo o fluxo. Com essa demonstração, as empresas podem usufruir dessa solução para conseguir mostrar que podem estar em conformidade com a LGPD.

Apesar de não conseguir coletar algumas outras informações capazes de identificar melhor o usuário que realizou tratamentos sobre os dados disponíveis, foi possível coletar informações sobre quando houve alterações no arquivo, quais eventos ocorreram, o nome do arquivo, o hostname do usuário, entre outros dados.

5. Conclusão

O objetivo principal desse trabalho foi disponibilizar um repositório com metadados de proveniência acerca do fluxo de ingestão de dados em um Data Lake, que permita a execução de consultas e análises para fins de monitoramento, visando mostrar um meio para que empresas que utilizam Data Lake possam estar em conformidade com a Lei Geral de Proteção de Dados.

Inicialmente, foram realizadas pesquisas para identificar ferramentas adequadas para alcançar o objetivo determinado. Em seguida, a arquitetura para a coleta e armazenamento de metadados de proveniência foi elaborada. Finalmente desenvolvemos um exemplo, usando dois conjuntos de dados abertos, para ilustrar o uso da arquitetura proposta.

A partir dos resultados obtidos, foi possível concluir que, com o uso das ferramentas escolhidas, é possível coletar e armazenar os metadados de proveniência para um fluxo de ingestão de dados, sendo possível, posteriormente, analisar e consultar os metadados de proveniência.

Como a obrigatoriedade de adoção à LGPD ainda é recente, muitas empresas ainda estão se adaptando à lei proposta. Este trabalho propõe uma possível solução, baseada em um cenário específico de combinação de tecnologias existentes, para a coleta e inserção de metadados de proveniência em um repositório, facilitando a adequação de organizações à LGPD. Logo, outros cenários devem ser elaborados e outras tecnologias devem ser investigadas. Dessa forma, é necessário continuar avaliando outros meios que permitam alcançar o resultado esperado, mas que ofereçam melhorias, como uma variedade maior de metadados de proveniência.

Referências

Apache NiFi Overview. **What is Apache NiFi?** Disponível em: <https://nifi.apache.org/docs.html>. Acesso em: 11 de Agosto de 2021.

ARAKAKI, F. A. **Metadados administrativos e a proveniência dos dados: modelo baseado na família.** PROV. 2019. 139 f. Tese (Doutorado) - Doutorado em Ciência da Informação, Universidade Estadual 2019. Disponível em: https://repositorio.unesp.br/handle/11449/180490.

BOVO, A. B.; BALANCIERI, R.; FERRARI, S. **Diagrama de Fluxo de Dados ao Use Case**. Universidade Federal de Santa Catarina, Florianópolis, 2004. Disponível em <http://inf.unisul.br/~ines/workcomp/cd/pdfs/3007.pdf>.

BRASIL, 2018. **Lei nº 13.709, de 30 de agosto de 2018**. Disponível em: https://bit.ly/2VfiMWX. Acesso em: 15 de Agosto de 2021.

Chessell, M., Scheepers, F., Nguyen, N., van Kessel, R., van der Starre, R. **Governing and managing big data for analytics and decision makers.** IBM, 2014.

Data Science Academy. **Ingestão de Dados em Tempo Real com Apache NiFi**. Disponível em: <https://blog.dsacademy.com.br/ingestao-de-dados-em-tempo-real-com-apache-nifi/> . Acesso em: 20 de Agosto de 2021.

Dixon, J.. **Pentaho, Hadoop, and data lakes.** https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. 2010.

Ganore, P. **Introduction To The Concept Of Data Lake And Its Benefits**. https://www.esds.co.in/blog/introduction-to-the-concept-of-data-lake-and-its-benefits. 2015.

Giebler, C., Groger, C., Hoos, E., Schwarz, H., Mitschang, B. **Leveraging the data lake - current state and challenges.** DaWaK, 2019.

GILLILAND, A. J. Setting the Stage. In: BACA, Murtha (Org.). **Introduction Metadata 3.** Getty Research Institute, 2016.

Isuru Suriarachchi, Beth Plale. **Provenance as Essential Infrastructure for Data Lakes [Preprint , forthcoming in IPAW 2016].** 2016.

LaPlante, A., & Sharma, B. **Architecting data lakes data management architectures for advanced business use cases.** O'Reilly Media Inc, 2016.

Laskowski, N. **Data lake governance: A big data do or die.** TechTarget, 2016.

POMERANTZ, J. **Metadata**. The MIT Press, 2015.

RAVAT, F.; ZHAO, Y. **Data lakes: Trends and perspectives.** In 30Th international conference on database and expert systems applications (DEXA (2019).

Sawadogo, P., Darmont, J. **On data lake architectures and metadata management.** Journal of Intelligent Information Systems 56, 97–120 (2021). https://doi.org/10.1007/s10844-020-00608-7

Techonology Digital World. **What is a Data Lake?** Disponível em: <https://www.bbvaopenmind.com/en/technology/digital-world/data-lake-an-opportunity-or-a-dream-for-big-data/>. Acesso em: 20 de Agosto de 2021.