UNIVERSIDADE FEDERAL DE UBERLÂNDIA BACHARELADO EM GESTÃO DA INFORMAÇÃO

RAQUEL SILVA ALVES BUSTAMANTE

ADOÇÃO DE SOLUÇÃO DE *CLOUD COMPUTING* NA ÁREA DE NEGÓCIOS DE UMA FÁBRICA DE FÁRMACOS E COSMÉTICOS

UBERLÂNDIA – MG

2022

RAQUEL SILVA ALVES BUSTAMANTE

ADOÇÃO DE SOLUÇÃO DE *CLOUD COMPUTING* NA ÁREA DE NEGÓCIOS DE UMA FÁBRICA DE FÁRMACOS E COSMÉTICOS

Monografia apresentada ao Curso de Graduação em Gestão da Informação, da Universidade Federal de Uberlândia, como exigência parcial para a obtenção do título de Bacharel.

Orientador: José Eduardo Ferreira Lopes

UBERLÂNDIA – MG

2022

ADOÇÃO DE SOLUÇÃO DE *CLOUD COMPUTING* NA ÁREA DE NEGÓCIOS DE UMA FÁBRICA DE FÁRMACOS E COSMÉTICOS

Monografia apresentada ao Curso de Graduação em Gestão da Informação, da Universidade Federal de Uberlândia, como exigência parcial para a obtenção do título de Bacharel.

Orientador: José Eduardo Ferreira Lopes

_____

Prof.

_____

Prof.

_____

Prof.

Uberlândia, 21 de março de 2022

AGRADECIMENTOS

Toda a minha graduação, em especial este trabalho, é dedicado à minha família. Faço um agradecimento especial aos meus avós por terem me criado e me formado em caráter e educação. Sem vocês, Marilene e Osmar, eu não seria quem eu sou hoje, agradeço a vocês dois pela paciência, incentivo e dedicação. Por todas as vezes que em minhas crises de ansiedade e dúvidas sobre meu curso, vocês estiveram ali para segurar minha mão, secar minhas lágrimas e me darem forças para continuar.

Agradeço também aos meus pais e padrasto (Ledimar, Renato e Flávio), por terem sempre me guiado ao caminho dos estudos e terem me dado toda a base para chegar até aqui. Além disso, sou grata às minhas tias (Leidiane e Letícia) e ao meu irmão (Lucca) por sempre terem me admirado e me apoiado durante a graduação e a vida. Meu muito obrigada também ao meu namorado Felipe, por ter me acompanhado nessa etapa tão desafiadora que foi o final da faculdade.

RESUMO

O objetivo deste relato técnico foi de descrever a implementação de um projeto de *Cloud Computing*, mais especificamente, uma estrutura *Big Data* para uma indústria de fármacos e cosméticos, valendo-se das melhores práticas de mercado, que suporta e agiliza o processo de tomada de decisão. Para tanto, foi realizada a migração de determinados dados para a nuvem, onde o armazenamento pudesse ser acessado somente através da rede interna ou via VPN. Com a finalidade de realizar essa entrega, foi adotada a metodologia *Scrum* para gerenciamento e execução do projeto. O tipo de intervenção adotada foi a geração de uma infraestrutura provisionada em nuvem de um *Data Lake* que disponibilizasse os dados em um *dashboard*, substituindo a página inicial do antigo *dashboard* da empresa. Os resultados obtidos mostraram que foi possível visualizar os dados na ferramenta de visualização de dados Microsoft Power BI por meio da conexão com o *Redshift*, um dos produtos da *Amazon Web Services*, proporcionando mais agilidade e contribuindo para a qualidade e segurança no processo de tomada de decisão baseada em dados.

Palavras-chave: Big Data. *Cloud Computing. Dashboard.*

ABSTRACT

The objective of this technical report was to describe the implementation of a *Cloud Computing* project, more specifically, a *Big Data* structure for a pharmaceutical and cosmetics industry, using the best market practices that support and streamline the decision-making process. Therefore, certain data were migrated to the cloud, where the storage could be accessed only through the internal network or via VPN. In order to carry out this delivery, the *Scrum* methodology was adopted for project management and execution. The type of intervention adopted was the generation of a cloud-provisioned infrastructure of a *Data Lake* that would make the data available in a dashboard, replacing the home page of the company's old dashboard. The results obtained showed that it was possible to visualize the data in the Microsoft Power BI data visualization tool through the connection with Redshift, one of the Amazon Web Services products, providing more agility, quality and security in the decision-making process based on data.

Keywords: Big Data. *Cloud Computing. Dashboard.*

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, a *Cloud Computing* – ou computação em nuvem - se tornou uma solução alternativa inevitável em diversas organizações e setores. Independentemente do tamanho, segundo a empresa de pesquisas Everest Group, a maioria das companhias hoje adotam a nuvem de alguma forma (softline, 2017). Embora a estratégia de implementação seja usada principalmente para melhorar a escalabilidade de seus recursos de banco de dados baseados na Internet e, ao mesmo tempo, reduzir o risco e o custo, há vários outros benefícios dessa tecnologia. Como a redução dos riscos comparado ao *on-primise*, custos mais baixos (dado o uso por demanda) e segurança dos dados (SUNYAEV, 2020). Atuando como um catalisador para o crescimento, a abordagem da *Cloud* revolucionou a maneira como as empresas operam seus negócios.

A *Cloud Computing* consiste em um mecanismo que possibilita o acesso *on-demand* abrangente e conveniente a uma rede de campos que compartilham recursos de computação configuráveis, tais como servidores, redes, *data warehouses*, aplicativos, serviços, entre outros. Estes podem ser rapidamente provisionados e feitos disponíveis com o mínimo de envolvimento do provedor de serviços (TAURION, 2009; SUNYAEV, 2020). Ou seja, é um serviço digital que permite aos usuários acessar programas e dados importantes armazenados em um servidor remoto em qualquer lugar que tenham uma conexão com a Internet.

Com isso, a *Cloud Computing* é uma maneira eficiente para as empresas melhorarem e aumentarem sua produtividade geral, além de poderem navegar com eficácia no ambiente de negócios em rápida evolução. Os serviços baseados em computação em nuvem oferecem uma infraestrutura de Tecnologia da Informação muito mais escalonável e confiável, projetada especificamente para otimizar o desempenho dos negócios e suportar seu crescimento e desenvolvimento (SUNYAEV, 2020).

Ante a isso, ao considerar a adoção de solução de *Cloud Computing* na área de negócios, tem-se que ela permite que muitas das necessidades de negócios atuais sejam tratadas com mais eficiência em um servidor distante, em

vez de no equipamento interno da empresa. Assim, objetiva-se com este relato técnico descrever a estruturação de um projeto de *Cloud Computing*, mais especificamente, uma estrutura *Big Data* para uma indústria de fármacos e cosméticos que suporta e agiliza o processo de tomada de decisão, valendo-se das melhores práticas de mercado.

A empresa contratante (indústria de fármacos e cosméticos) contratou uma consultoria para fazer a migração de determinados dados para a nuvem, onde o armazenamento pudesse ser acessado somente através da rede interna ou via *Virtual Private Network* (VPN). Como resultado, pode-se visualizar os dados na ferramenta de visualização de dados da Microsoft (Power BI) através da conexão com o Redshift, um dos produtos da Amazon Web Services.

Cabe destacar que, por ocasião do desenvolvimento deste projeto, esta pesquisadora compunha a equipe da consultoria contratada e atuou diretamente no desenvolvimento da solução apresentada.

2 REFERENCIAL TEÓRICO

Para embasar este relato, foi necessário estruturar e discutir alguns conceitos de *Cloud Computing* e das ferramentas AirFlow, Spark com PySpark, Redshift, Gitlab CI/CD, Terraform, Amazon S3 e Microsft PowerBI, utilizadas no projeto.

2.1 CLOUD COMPUTING

De acordo com especialistas em tecnologia, grandes mudanças na tecnologia aparecem a cada 20 anos, uma delas é a *Cloud Computing* (ou Computação em Nuvem). Cada vez mais dominante no mundo da tecnologia da informação (TI), essa nova tecnologia veio para desafiar o setor de telecomunicações e surge no contexto de um cenário concebido principalmente por organizações no setor com o objetivo de reduzir custos (TAURION, 2009; SUNYAEV, 2020).

Esta arquitetura de nuvem é conhecida como um modelo onde a computação (processamento, armazenamento e software) é localizada em algum lugar da rede e pode ser acessada remotamente (JOSHI et al., 2019).

Assim, já não há necessidade de instalar um software em uma máquina local, já que o acesso a todos os recursos será por meio da Internet (mais precisamente, pela nuvem), dentro de um sistema operacional a partir de qualquer computador e de qualquer lugar em que seja possível acessar informações, arquivos e programas processados em um único sistema, independentemente da plataforma (TAURION, 2009).

O termo nuvem é empregado historicamente como uma representação para a Internet. A sua utilização, deu-se por meio de sua representação em um diagrama de rede, com o esquema de uma nuvem, a qual era adotada para representar o transporte de dados através de plataformas em nuvem (RITTINGHOUSE; RANSOME, 2009).

O conceito surgiu em 1961, quando o professor John McCarthy propôs que a tecnologia teria no futuro potencial para promover um processo no qual o poder de computação e até mesmo de aplicativos específicos poderiam ser vendidos por meio do modelo de utilidade. Essa ideia se difundiu na década de 1960, mas na década de 1970, a mesma se esvaiu. Apesar disso, desde a virada do milênio, o conceito foi revivido, de forma que em 2007, Eric Schmidt, então CEO do Google, usou o termo *Cloud Computing* para se referir à forma como a empresa gerenciava seus *Data Centers*. Foi durante esse contexto que o termo *Cloud Computing* começou a aparecer no cenário da tecnologia (RITTINGHOUSE; RANSOME, 2009).

Assim, a *Cloud Computing*, também conhecida como Computação em Nuvem é tida como “um conjunto de serviços de rede, que proporciona escalabilidade, qualidade de serviço, infraestrutura barata de computação sob demanda, que pode ser acessado de uma forma simples” (SUNYAEV, 2020, p. 198).

Na visão de Furht e Escalante (2010) a Cloud Computing e definida como:

> Um novo estilo de computação em que os recursos dinamicamente escaláveis e muitas vezes virtualizados são fornecidos como serviços através da Internet. Computação em nuvem se tornou uma tendência tecnológica significativa, e muitos especialistas esperam que a computação em nuvem irá reformular a tecnologia da informação (TI) os processos e o mercado de TI. Com a tecnologia de computação em nuvem, os usuários usam uma variedade de dispositivos, incluindo PCs, laptops, smartphones e PDAs para acessar programas, armazenamento e aplicação de desenvolvimento de plataformas pela

> Internet, através de serviços oferecidos por provedores de computação em nuvem (FURHT; ESCALANTE, 2010, p. 3).

Portanto, a *Cloud Computing* passou a ser compreendida como um meio decorrente da evolução natural da computação, de forma que a descoberta dessa nova infraestrutura resultou em um cenário permeado pelo desenvolvimento de aplicativos. Assim, as empresas passaram a poder optar por usar um provedor de serviços em nuvem ou até mesmo alavancar seu próprio *data center* e prestar esse serviço (FURTH; ESCANLANTE, 2010; SUNYAEV, 2020). Para isso, diversas ferramentas têm se mostrado valiosas nesses processos.

2.2 FERRAMENTAS

É inevitável que diferentes partes da empresa pesquisem e usem mais de uma solução de software, geralmente desenvolvida por vários fornecedores. Entre elas, as ferramentas escolhidas para realização deste caso foram as distribuídas pelo Apache: AirFlow e Spark. Além do serviço da Amazon Web Service: Redshift, como também as ferramentas Gitlab, Terraform, Amazon S3 e Microsoft PowerBI.

Acerca do AirFlow, esta é uma plataforma desenvolvida para criar, programar e monitorar fluxos de trabalho de forma programática**.** Ela foi criada pelo Airbnb, escrita na linguagem de programação Python, que se tornou *open-source* em 2015 e logo depois foi cedida para o Apache Foundation (BEAUCHEMIN, 2015). A relevância do Apache Airflow vem do fato de que um dos principais desafios que as empresas que trabalham com dados enfrentam hoje é garantir que os fluxos de dados sejam executados da maneira mais suave possível (BABUJI et al., 2018).

Com isso, é fundamental que as empresas fiquem por dentro do cronograma e que todos os usuários responsáveis sejam alertados e tenham visibilidade se algo der errado. E foi exatamente tais problemas que o Apache Airflow resolveu, tendo em vista que ele é um coordenador de fluxo de trabalho. Com ele pode-se programar, agendar e acompanhar consultas de diferentes fontes de dados e realizar tratamentos de forma simples (BABUJI et al., 2018).

Quanto ao Spark, ele é um mecanismo multilíngue para executar engenharia de dados, ciência de dados e aprendizado de máquina em máquinas de nó único ou *clusters*. O Apache Spark se integra às suas estruturas favoritas, ajudando a escaloná-las para milhares de máquinas (MENG et al., 2016).

Dessa forma, ele é um mecanismo de Big Data que visa processar grandes conjuntos de dados de maneira paralela e distribuída, uma vez que foi construído com objetivo de executar com maior velocidade, simplicidade de uso e análises sofisticadas. A fim de oferecer uma estrutura voltada para gerenciar e processar Big Data com vários conjuntos de dados de diferentes naturezas e de diferentes fontes (MENG et al., 2016).

Sobre o Redshift, ele é um produto de armazenamento de dados rápido, confiável e totalmente gerenciado que faz parte da plataforma de computação em nuvem da Amazon Web Services (AWS). Seus produtos são desenvolvidos com base na tecnologia de data warehouse MPP (Massive Parallel Processing). O produto é uma maneira simples e econômica de analisar todos os dados de um negócio usando as ferramentas de inteligência de negócios existentes (GUPTA et al., 2015).

O Redshift é um serviço baseado em nuvem e é hospedado diretamente na AWS (Amazon Web Service). Uma de suas principais vantagens é sua arquitetura versátil, que pode ser dimensionada em segundos para atender às suas necessidades crescentes de armazenamento, de forma que um grande problema que as organizações enfrentam com os requisitos de dados que mudam rapidamente é que o dimensionamento pode ser caro e complexo (GUPTA et al., 2015).

Dessa maneira, por meio do Amazon Web Services, o Redshift pode ser ampliado ou reduzido ativando rapidamente nós individuais de tamanhos diferentes. Essa escalabilidade também significa economia de custos, já que as empresas não precisam gastar dinheiro mantendo servidores não utilizados ou comprando rapidamente espaço no servidor quando necessário. Isso é especialmente útil para pequenas empresas que estão passando por um crescimento significativo e precisam dimensionar suas soluções existentes (GUPTA et al., 2015).

Acerca do GitLab, ele é responsável por armazenar os repositórios com base em git, gerenciar atividades e realizar CI/CD. Esta ferramenta é similar ao

GitHub, com a diferença que os desenvolvedores têm a capacidade de reunir seus scripts em seus próprios servidores, e não em um servidor terceiro (como o GitHub). O Gitlab CI/CD é uma ferramenta para desenvolvimento de software que utiliza o método de integração contínua e entrega contínua para detectar *bugs* e erros no início do ciclo de desenvolvimento (SINGH et al., 2019).

Sobre o Terraform, a referida ferramenta é uma infraestrutura de código aberto como ferramenta de software de código que fornece um fluxo de trabalho consistente para gerenciar centenas de serviços em nuvem (NAIK, 2021).

Já o Amazon Simple Storage Service (Amazon S3), é um serviço de armazenamento de objetos que oferta tanto escalabilidade, como disponibilidade de dados, segurança e performance líderes do setor, sendo assim, uma ferramenta que visa fornecer armazenamento como um serviço de baixo custo e alta disponibilidade, com um modelo de cobrança simples (PALANKAR et al., 2008).

Em relação ao PowerBI, ele é um serviço de análise de negócios da Microsoft, que fornecer visualizações interativas e recursos de *business intelligence* com uma interface simples, de forma que seus usuários finais possam criar seus próprios relatórios e *dashboards* (POWER BI, 2018).

Com base nas referidas ferramentas depreende-se que estas se mostram úteis para estruturação deste projeto de *Big Data*. Todas essas ferramentas se conectam em uma arquitetura de Data Lake, como na do projeto e que pode ser visualizada abaixo.

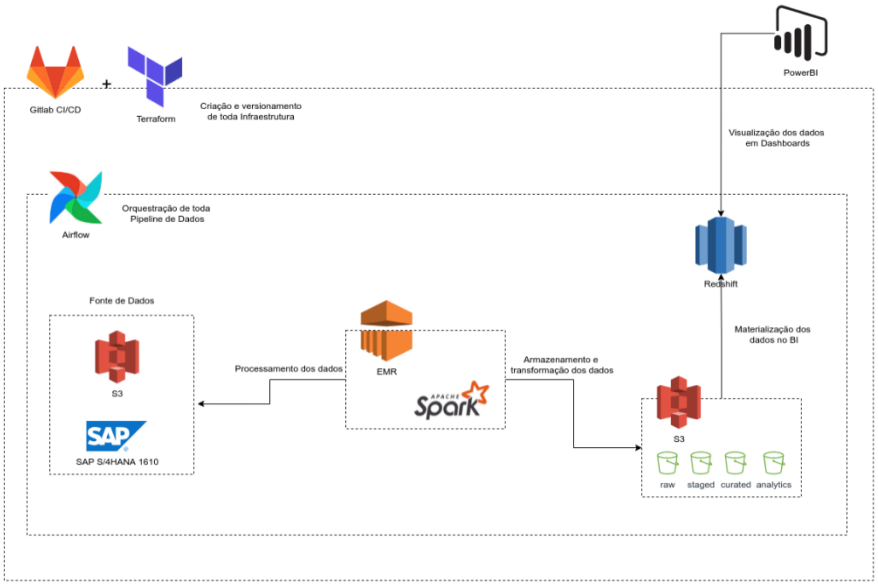

Figura 1. Arquitetura Data Lake.
Fonte: própria.

3 SITUAÇÃO PROBLEMA

A empresa em questão é uma indústria que atua na produção de itens relacionados as linhas cosmética, farmacêutica, hospitalar e suplemento alimentar, situada no estado de Minas Gerais. A empresa se encontra no mercado há mais de 40 anos, estando presente em todo o território brasileiro.

Tendo em vista a finalidade da empresa que se concentra em crescer com sustentabilidade e excelência operacional, a mesma demandava que suas necessidades de informações relacionadas aos negócios fossem tratadas com mais eficiência, velocidade, confiabilidade dos dados, disponibilidade dos dados e segurança. Contudo, tal processo não acontecia, de forma que a empresa apresentava um cenário que estava dificultando e até mesmo, em algumas situações, inviabilizando o processo de tomada de decisão baseada em dados. Tal problema era decorrente da expressiva dificuldade para organizar as regras de negócio das áreas, o que acabava por resultar em complicações para criação de consultas e relatórios a partir do banco de dados existente.

Além dos fatores apontados, a empresa não dispunha de uma centralização dos dados que utilizavam, o que acabava por gerar uma dificuldade para conseguir acesso às informações fidedignas para as áreas de negócios como marketing, financeiro, etc.

Vale dizer ainda que a empresa apresentava problemas em relação à periodicidade de disponibilização dos dados, visto que a mesma só conseguia acesso aos dados de 1 a 2 dias após a ocorrência de eventos, como por exemplo, gestão de estoques. Essa falta de periodicidade atrasava as áreas no processo de tomada de decisão e por isso, com o projeto, queriam que os dados fossem entregues em até 1 hora ou o mais próximo disso.

Resumindo, os principais problemas a serem resolvidos eram a falta de dados fidedignos, dados desatualizados, dados não íntegros, dados não disponíveis em tempo hábil, dentre outros aspectos.

Em vista disso, o trabalho aqui apresentado tem como situação problema a necessidade de implementação de uma infraestrutura de dados na nuvem para

controlar, gerenciar e permitir o processo de tomada de decisão em uma indústria de cosméticos e produtos farmacêuticos.

4 INTERVENÇÃO ADOTADA

Mediante a situação problema levantada, a qual refere-se ao fato de os dados da empresa não apresentarem a confiabilidade necessária para tomadas de decisão e acompanhamento, tem-se que a intervenção proposta para a situação problema foi a contratação de uma empresa de consultoria, como forma de promover a migração e o armazenamento dos dados para a nuvem.

Com a finalidade de realizar essa entrega, a empresa adotou o modelo de referência em métodos ágeis *Scrum* para realizar o projeto. A utilização dessa metodologia foi importante pois no decorrer da execução do projeto, os *stakeholders* estiveram também interligados ao processo. Ajudando na criação e validação das regras de negócios, além de participarem das reuniões diárias para acompanhar as entregas de cada *Sprint*. O projeto teve um roadmap inicial que predizia o término em cinco meses, contudo, para melhor qualidade da entrega, foi realizado em seis meses.

Para tanto, o tipo de intervenção adotada foi a geração de uma infraestrutura provisionada em nuvem de um Data Lake que disponibilizasse os dados em um *dashboard*, substituindo a página inicial do antigo *dashboard* da empresa.

A arquitetura proposta para solução do referido problema, é apresentada na Figura 1. Em seguida, são descrita as partes da arquitetura e as funcionalidades de cada ferramenta.

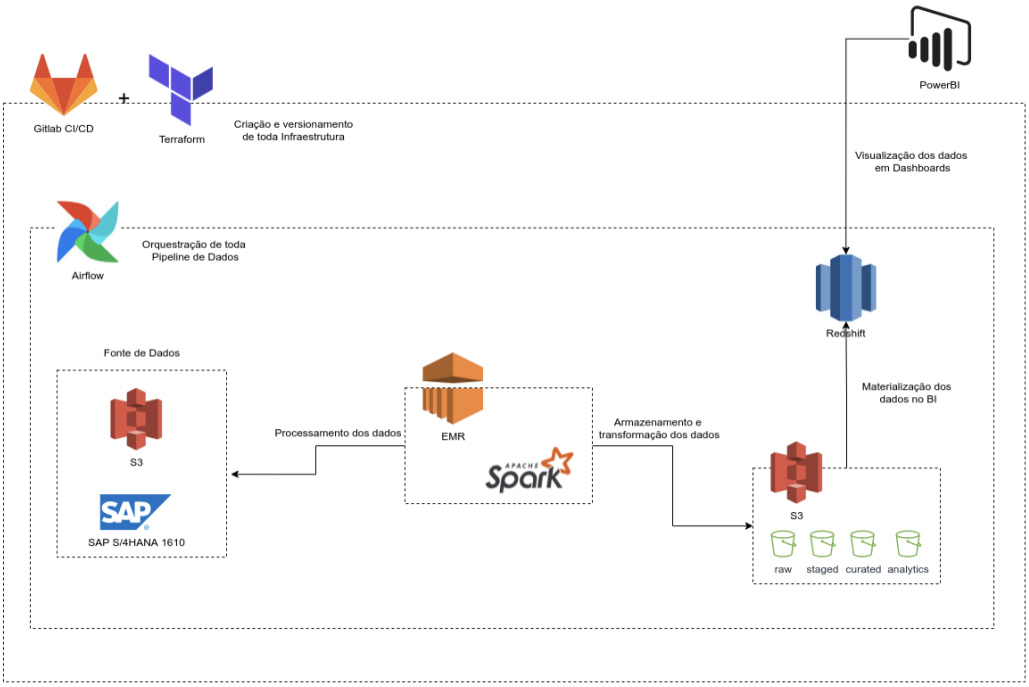

Figura 2. Arquitetura Data Lake.

Fonte: própria.

Visando seguir um método de mercado para análise, construção e validação da arquitetura proposta, foi seguido o AWS Well-Architected Framework (WAF), arcabouço utilizado por grandes corporações do mundo inteiro que adotam as melhores práticas existentes de projetos na nuvem (AWS, s.d). Para tal procedimento, é possível ver pela Figura 1 que os dados que se encontravam armazenados no banco de dados Oracle do SAP S/4HANA foram extraídos e acondicionados em *buckets* S3 (serviço de armazenamento da AWS) na camada raw.

Posteriormente, representado na parte central da Figura 1, deu-se o processamento e ingestão dos dados, os quais foram realizados mediante utilização do Apache Spark, rodando no serviço da AWS EMR. Em sequência, ocorreu a orquestração utilizando o Apache Airflow, instalado em uma máquina EC2 (serviço da AWS), a fim de haver controle sobre a pipeline de dados e gerar as camadas do Data Lake.

Um exemplo de processo de execução na ferramenta Apache Airflow é apresentando na Figura 2, para ilustrar o procedimento descrito.

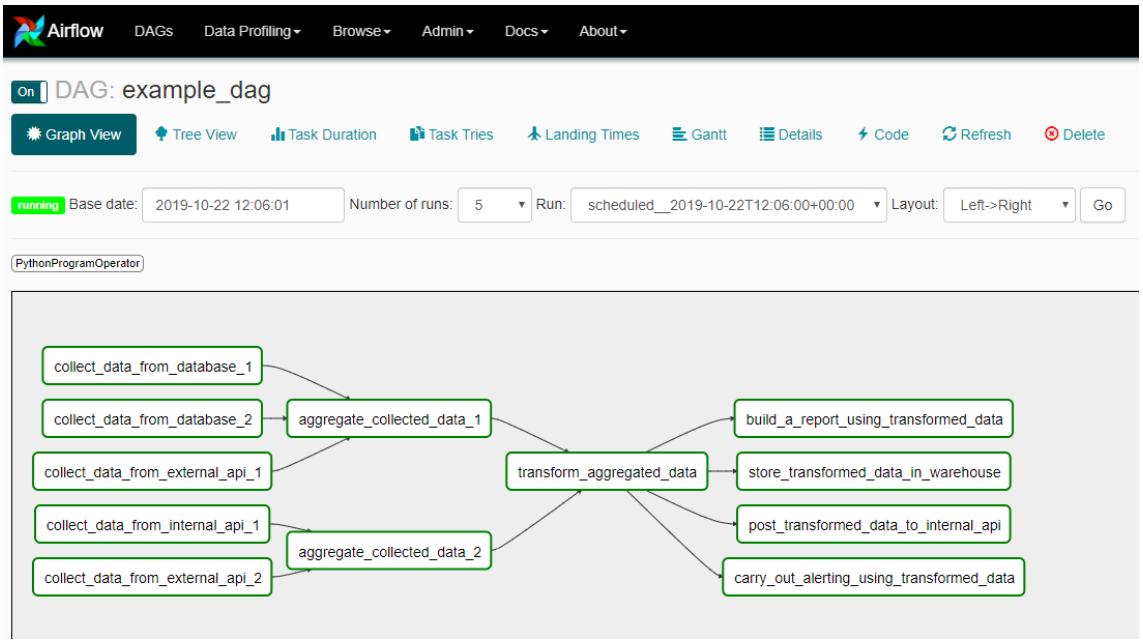

Figura 3. Demonstração da execução na ferramenta Apache Airflow.
Fonte: Pidy, 2019.

Para maior compreensão, a Figura 2 traz um fluxo de dados gerados a partir da ferramenta Airflow, onde os retângulos da primeira coluna representam a extração de dados vindo de bancos de dados e APIs (Interface de Programação de Aplicações). Na segunda etapa, os retângulos do fluxo que possuem o prefixo de “aggregate” têm o objetivo de fazer alguns processamentos de agregação nos dados vindos da fonte. Em seguida, há a transformação dos dados e por último são realizadas publicações, construção de relatórios, armazenamento em um *data warehouse*, etc.

Além do entendimento das ferramentas, é importante destacar o que são as camadas, como também mencionar sobre o que cada uma é responsável, visto que os dados são inseridos dentro de cada camada a partir de *scripts* que serão detalhados posteriormente.

Acerca dessas camadas, conforme a Figura 1, no canto inferior direito, tem-se a camada *raw*, que contém os dados brutos e puros, sem transformações e que vêm a partir dos dados recebidos da ingestão. A camada *staged* recebe dados em processamento a serem tratados, de forma que nessa etapa tem-se as padronizações dos tipos dos dados, como também modificações mais simples e quais os dados mais recentes após extração na camada *raw*.

A camada *curated* já possui um processamento de dados prontos para consumo sem muitas agregações, mas que já podem ser visualizadas pelas

áreas de negócio, especialmente para analistas de dados realizarem consultas. Já a camada *analytics* possui o dado mais pronto para análises, sem precisar realizar grandes agregações. Os dados dessas camadas foram particionados e armazenados no serviço de armazenamento de dados da Amazon, o S3 - Simple Storage Service.

Retornando à arquitetura, tem-se que todos os códigos responsáveis para geração de cada camada do *Data Lake* foram armazenados no GitLab (que também é executado dentro do serviço EC2 da AWS) a fim de se ter a gestão de versionamento dos códigos e centralização, para que estes possam ser utilizados na esteira de CI/CD.

Assim sendo, o provisionamento e criação da infraestrutura foi baseada no framework de Infraestrutura como Código (IaC) Terraform, no qual toda a arquitetura é provisionada com o framework, onde o código será versionado pelo Gitlab e o *deploy* automatizável dos códigos pelo CI/CD também do GitLab.

A infraestrutura como código (IaC) é o processo de gerenciamento e provisionamento de centros de processamentos de dados usando arquivos de configuração ao invés de configurações físicas de hardware ou ferramentas de configuração interativas. Ou seja, o objetivo principal é automatizar o provisionamento da infraestrutura de TI.

As ferramentas de IaC podem ser aplicadas tanto para orquestrar uma infraestrutura, quanto para gerenciar configurações que tem o objetivo de administrar em geral um software em execução. No contexto em questão, foi utilizada uma ferramenta com o objetivo de fazer o provisionamento e gerenciamento dos componentes do ambiente, sendo escolhida para tanto o Terraform.

Sequencialmente ao processamento dos dados, deu-se a materialização dos dados no RedShift, que é uma ferramenta de menor complexidade comparado aos scripts das camadas do Data Lake. O Amazon Redshift emprega o SQL para analisar dados estruturados e semiestruturados em data warehouses, bancos de dados operacionais e Data Lakes. Assim, os dados foram dispostos em uma grande tabela (conceito de One Big Table) a fim de construir um Data Lake.

Uma etapa que deve ser destacada em nível de análise de dados (para além da engenharia de dados), visando garantir a fidelização dos dados, foi

utilização do método CDM (Common Data Model). Este foi um processo de validação dos analistas de dados sobre os metadados gerados. Foi realizada uma perícia sobre a quantidade e os tipos de dados, além de trazer a média dos atributos de cada esquema da tabela, entre outros valores técnicos cabíveis aos analistas de dados.

Por fim, foi realizada uma conexão no software Power BI através do Redshift, o qual viabilizou que os analistas de dados pudessem realizar consultas no Data Lake e construir *dashboards* a partir desses dados.

5 RESULTADOS OBTIDOS

A arquitetura de nuvem proposta para a indústria farmacêutica possuía alguns resultados esperados e que foram alcançados. A partir do projeto, foi possível criar relatórios mais dinâmicos em ferramentas de BI, gerando confiança àqueles que utilizavam os dados.

Um dos resultados alcançados foi o intervalo temporal dos dados para as áreas. Antes do projeto, os dados possuíam um *delay* de cerca de dois dias, e após a implementação do projeto os dados passaram a ser disponibilizados entre uma e duas horas (em média oitenta e seis minutos). Este era um dos objetivos iniciais do projeto e foi alcançado com êxito.

Também foi possível realizar consultas sobre os dados de forma mais simples utilizando menos tabelas, já que o objetivo era gerar uma grande tabela (one big table). A *Big Table* é um sistema de armazenamento distribuído que gerencia dados estruturados, possui grande oferta e demanda de dados e grande escala. Sua estrutura parece um banco de dados, ele compartilha muitas estratégias de implementação com bancos de dados, utilizando conceitos de bancos de dados paralelos [1] a bancos de dados de memória principal [2], portanto alcançando escalabilidade e alto desempenho (SILVA, s.d).

Outro ponto de melhoria foi a consistência dos dados, já que por não ser mais estruturado por dimensões e sim em *One Big Table*, passou a não ocorrer casos de uma dimensão não ser atualizada enquanto outras são.

Alguns pontos de atenção que foram levantadas no projeto foram relacionadas ao fato de que a atualização da tabela poder demorar mais para One Big Table comparado com várias tabelas menores. *A priori*, isso não foi um

problema ocorrido e possivelmente não trará grandes problemas visto a periodicidade de *backups* e quantidade de dados que a indústria possuía. Contudo, é um ponto de atenção para outros projetos com o mesmo objetivo e para o próprio a longo prazo.

6 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi de relatar a implementação e desenvolvimento de um Data Lake em nuvem a fim de fornecer dados para as áreas de negócio de uma indústria de fármaco e cosméticos.

É possível dizer que isto foi alcançado, já que houve a implantação da arquitetura sugerida e a partir disso, os *stakeholders* do projeto puderam ter os dados de forma rápida e com integridade, validada através da a aplicação do CDM (Common Data Model).

Como dito na introdução deste trabalho, esta pesquisadora compunha a equipe da consultoria contratada e teve atuação no desenvolvimento do projeto. Isso trouxe grande valor à mesma, visto que foi possível aprofundar seu conhecimento prático sobre a área de dados visto em teoria durante o curso de Gestão da Informação.

Para além, os resultados obtidos trouxeram satisfação à empresa contratante e foram alertados sobre os possíveis problemas que podem ocorrer no futuro, dando a oportunidade de que sejam realizadas ações antecipadas. Uma sugestão foi da incorporação de um time de engenheiros e analistas de dados para sustentação do ambiente criado, visto que os dados são valiosos para o crescimento da empresa.

REFERÊNCIAS BIBLIOGRÁFICAS

BABUJI, Y. N., CHARD, K., FOSTER, I. T., KATZ, D. S., WILDE, M., WOODARD, A., WOZNIAK, J. M. Parsl: Scalable Parallel Scripting in Python. In: **IWSG**. 2018.

FURHT, B., ESCALANTE, A. **Handbook Of Cloud Computing**. Springer, 2010.

GUPTA, A., AGARWAL, D., TAN, D., KULESZA, J., PATHAK, R., STEFANI, S., SRINIVASAN, V. Amazon Redshift and the case for simpler data warehouses. In: **Proceedings of the 2015 ACM SIGMOD international conference on management of data**. p. 1917-1923, 2015.

MENG, X., BRADLEY, J., YAVUZ, B., SPARKS, E., VENKATARAMAN, S., LIU, D., TALWALKAR, A. Mllib: Machine learning in apache spark. **The Journal of Machine Learning Research**, v. 17, n. 1, p. 1235-1241, 2016.

NAIK, Nitin. Plataforma de processamento de big data leve e independente da nuvem em várias nuvens usando docker swarm e terraform. In: **Workshop do Reino Unido sobre Inteligência Computacional**. Springer, Cham, 2021. p. 519-531.

PALANKAR, M. R, IAMNITCHI, A., RIPEANU, M., GARFINKEL, S. Amazon S3 para grades científicas: uma solução viável?. In: **Anais do Workshop Internacional de 2008 sobre Computação Distribuída Consciente de Dados**, p. 55-64, 2008.

PIDY, Aakash. **Building a Production-Level ETL Pipeline Platform Using Apache Airflow**. 2019. Disponível em: https://towardsdatascience.com/building-a-production-level-etl-pipeline-platform-using-apache-airflow-a4cf34203fbd. Acesso em fevereiro 2022.

POWER BI. **Power BI**. 2018.Disponível em: https://excelprince.com/pdf/BI.pdf. Acesso em fevereiro 2022.

RITTINGHOUSE, J. W., RANSOME, F. J. **Cloud Computing: Implementation, Management and Security.** CRC PRESS, 2009.

SILVA, Valter Henrique. BigTable: **Um sistema de armazenamento distribuído para dados estruturados (Maio 2011)**. Disponível em: https://dcomp.ufscar.br/verdi/topicosCloud/BigTable.pdf. Acesso em fevereiro 2022.

SINGH, C., GABA, N. S., KAUR, M., KAUR, B. Comparison of different CI/CD tools integrated with cloud platform. In: **2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)**. IEEE, p. 7-12, 2019.

SUNYAEV, Ali. Cloud computing. In: **Internet computing**. Springer, Cham, 2020. p. 195-236.

TAURION, Cezar. **Cloud computing: computação em nuvem: transformando o mundo da tecnologia da informação**. Rio de Janeiro: Brasport, 2009.

SOFTLINE. **IaaS, PaaS e SaaS: entenda os modelos de nuvem e suas finalidades.** 2017. Disponível em: https://brasil.softlinegroup.com/sobre-a-empresa/blog/iaas-paas-saas-nuvem. Acesso em março 2022.

JOSHI, Y., RANJAN, M., CHIRAYIL, Z. **Business Transformation through Multi cloud.** 2019. Disponível em: https://www.accenture.com/_acnmedia/PDF-111/Accenture-Business-Transformation-through-Multi-cloud.pdf. Acesso em março 2022.

AWS. **AWS Well-Architected**. s.d. Disponível em: https://aws.amazon.com/pt/architecture/well-architected/?wa-lens-whitepapers.sort-by=item.additionalFields.sortDate&wa-lens-whitepapers.sort-order=desc. Acesso em março 2022.

BEAUCHEMIN, Maxime. **Airflow: a workflow management platform.** 2015. Disponível em: https://medium.com/airbnb-engineering/airflow-a-workflow-management-platform-46318b977fd8. Acesso em março 2022.