UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO

CONSTRUÇÃO DE UM PIPELINE DE DADOS UTILIZANDO SERVIÇOS DA NUVEM

KAROLAYNE FERNANDES ARRAIS

ORIENTADOR (A): Profª. Dra. Marcela Xavier Ribeiro

SÃO CARLOS - SP

2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

CONSTRUÇÃO DE UM PIPELINE DE DADOS UTILIZANDO SERVIÇOS DA NUVEM

KAROLAYNE FERNANDES ARRAIS

SÃO CARLOS
2022

AGRADECIMENTOS

Agradeço aos meus pais Karine e Marcos pelo esforço e apoio para que eu pudesse realizar o sonho da formação acadêmica, sem eles eu não cresceria compreendendo que a educação é um dos bens mais importantes que temos nesta vida e que por meio da mesma, conseguirei realizar meus sonhos.

Agradeço especialmente à minha mãe Karine e irmã Luna, por serem meu apoio, inspiração de luta e meus grandes amores.

Sou grata por todo o convívio e situações que vivenciei no ambiente acadêmico, todos os professores, amigos e demais funcionários que cruzaram o meu caminho e fizeram parte de uma das épocas mais importantes da minha vida, onde tive contato com a pluralidade e pude crescer como pessoa. Construí com meus professores o meu conhecimento técnico, tão importante para a minha vida profissional, mas também aprendi e reforcei princípios humanos.

Também tenho agradecimentos aos meus colegas de trabalho, que me ensinaram muito por meio de suas experiências e que me incentivam a compartilhar o conhecimento que venho adquirindo como forma de apoiar outras pessoas.

Por fim, agradeço imensamente à minha orientadora Marcela, pela paciência, disponibilidade e apoio no trabalho que finaliza um ciclo ao qual sentirei tanta saudade.

.

“Feliz aquele que transfere o que sabe e aprende o que ensina. ”

CORA CORALINA

RESUMO

A evolução da rede de computadores e o crescente acesso e interação da população mundial à internet vem proporcionando uma mudança no cenário de dados. A todo instante são gerados dados em quantidades exorbitantes e das mais variadas estruturas, rompendo com os sistemas convencionais de dados que eram voltados para operações transacionais, iniciando um processo onde os sistemas evoluíram para atender as demandas analíticas que crescem com o conceito da orientação de processos e decisões por meio de informações (Data Driven). Na era do Big Data, além da evolução dos modelos de dados e infraestrutura para processamento e armazenamento, também houve a especialização dos profissionais da área para que cada um tivesse domínio sobre processos específicos do ciclo de vida dos dados. Seguindo este contexto, o objetivo do presente estudo é construir um entendimento sobre o cenário do Big Data e sua influência nas evoluções de processos e conceitos atuais da área, realizando um desenvolvimento prático da criação de uma solução de pipeline de dados utilizando serviços da computação em nuvem para integrar, coletar, modelar, processar e analisar dados da COVID-19 e indicadores de desenvolvimento mundial.

Palavras-chave: Big Data; Data Driven; Pipeline de dados; Computação em nuvem; COVID-19, Indicadores de desenvolvimento mundial.

ABSTRACT

The evolution of the computer network and the increasing access and interaction of the world population to the internet has provided a change in the data scenario. At all times, data is generated in exorbitant amounts and of the most varied structures, breaking with conventional data systems that were focused on transactional operations, initiating a process where systems have evolved to meet the analytical demands that grow with the concept of process orientation and decisions through information (Data Driven). In the era of Big Data, in addition to the evolution of data models and infrastructure for processing and storage, there was also the specialization of professionals in the area so that each one had mastery over specific processes of the data life cycle. Following this context, the objective of the present study is to build an understanding of the Big Data scenario and its influence on the evolution of current processes and concepts in the area, carrying out a practical development of the creation of a data pipeline solution using computing services in cloud to integrate, collect, model, process and analyze COVID-19 data and world development indicators.

Keywords: Big Data; Data Driven; Data pipeline; Cloud computing; COVID-19; World development indicators.

LISTA DE FIGURAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

Desde o surgimento da primeira rede de computadores interconectados em 1969, a *Arpanet,* projeto iniciado pelo cientista Larry Roberts no MIT (STRAWN, 2014); o desenvolvimento da Rede Mundial de Computadores (www) ao longo dos anos possibilitou a proliferação existente hoje de browsers, sites e redes sociais; houveram evoluções na forma e qualidade de acesso à internet que acompanhavam a demanda da mesma e contribuíram para o processo de difusão e inclusão tecnológica da sociedade. Segundo o relatório Digital 2021 da We Are Social e Hootsuite, duas agências de marketing digital com atuação global, foi analisada uma porcentagem de 59,5% da população global como usuários da internet em 2020, com um crescimento anual de 7,3% e uma média de interação de 6 horas e 54 minutos, dando destaque para Brasil no ranking de tempo diário gasto, com 10 horas e 8 minutos por dia (usuários entre 16 e 64 anos) (WE ARE SOCIAL AND HOOTSUITE, 2021).

O acesso à internet, além de modificar a estrutura de comunicação, permitindo uma comunicação de todos para todos, como discutido por Lemos e Lévy (2010), trouxe inovações de produtos e serviços conectados à mesma, gerando dados que analisados e monitorados podem ser transformados em informações importantes juntamente com o fluxo de comunicação e interações dos usuários conectados à rede.

Portanto, essa crescente conexão e interação vem proporcionando a geração de uma quantidade exorbitante de dados com grande variedade. A infraestrutura para trabalhar com essas informações teve que evoluir para atender os volumes e variedade, assim como a especialização dos profissionais responsáveis pelo entendimento e manipulação dessas informações teve que ocorrer.

1.2 Objetivo geral

O objetivo deste trabalho é desenvolver e implementar uma solução por meio de um pipeline de dados, usando ferramentas de big data e nuvem para integrar, coletar, modelar, processar e analisar dados de Covid e de desenvolvimento humano representados por indicadores globais de saúde, educação e renda, variáveis.

1.3 Objetivos específicos

De forma específica, os objetivos são:

- Pesquisar o estado da arte referente à Big Data, a evolução de modelos de dados, conceitos de armazenamento e processamento e infraestrutura para manipulação de Big Data;
- Obter fontes de dados da COVID-19 e indicadores de desenvolvimento global disponíveis na web sem necessidade de preocupação com o formato destes dados;
- Construir um pipeline de ELT (Extract, load, transform) de dados utilizando serviços de computação em nuvem (plataforma Microsoft Azure);
- Consultar e analisar os resultados obtidos pela execução do pipeline construído do item anterior e compartilhar as conclusões obtidas ao final deste trabalho.

1.4 Organização do trabalho

O trabalho segue a seguinte organização:

- No Capítulo 1, é apresentada a contextualização da mudança proporcionada pela evolução e crescimento da conexão à internet e a sua relação com o aumento exponencial da produção de dados. Também são apresentados os objetivos gerais e específicos deste trabalho.
- No Capítulo 2, é realizada a fundamentação teórica do estado da arte sobre Big Data e suas principais características relacionadas ao volume, velocidade, variedade, veracidade e valor; a evolução dos modelos de dados e sistemas de gerenciamento para lidar com a variedade de dados estruturados e não estruturados; pipelines como fluxos de atividades de manipulação de dados e a diferença entre as abordagens de ETL e ELT; armazenamento de dados para lidar com processamento de transações

e processamento analítico, abordando superficialmente sobre a arquitetura de Data Warehouses e Data Lakes; e breve apresentação sobre ambientes on-premise e em nuvem no contexto de dados. Também são apresentadas informações sobre a Microsoft Azure e os serviços desta plataforma que são utilizados no Capítulo 4, juntamente com uma breve abordagem sobre o tema de análise de dados e a ferramenta Microsoft Power BI.

- No Capítulo 3, é discorrido sobre a metodologia utilizada para este trabalho, desde o levantamento do estado da arte sobre os assuntos descritos no item anterior; a identificação dos dados utilizados no desenvolvimento, até a construção do pipeline de dados como trabalho prático e a análise realizada nos dados resultantes.
- No Capítulo 4, é discutido sobre os passos que envolveram a criação do pipeline de processamento de dados utilizando os serviços da Microsoft Azure, e a conexão do Microsoft Power BI aos dados resultantes no Data Lake para a análise dos mesmos por meio de visualizações.
- Por fim, o Capítulo 5 apresenta as conclusões e contribuições deste trabalho para o estado da arte.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem o objetivo de contextualizar sobre o termo Big Data, seu surgimento e o impacto das mudanças proporcionadas pela produção crescente de informações. Também são abordados conceitos importantes do mundo de dados além de serviços para manipulação dos mesmos, em que a compreensão por meio da leitura auxiliará no entendimento do projeto descrito no capítulo 4.

2.1. Big data

Observações e medições de eventos, sejam elas eventos físicos, como um processo de controle de temperatura; comportamental, como o comportamento de um grupo de pessoas diante de um acontecimento específico; e sensorial, como a avaliação de um prato culinário; documentadas, são consideradas dados (IME UNICAMP, s.d.).

É fácil observar a transformação digital que estamos vivenciando. Como relatado anteriormente, 62.5% da população global foi identificada como usuários da internet, com média de utilização diária de 7 horas, desconsiderando ainda o tempo gasto diariamente com serviços de streaming em TVs, serviços de músicas, jogos e outras interações com fins comunicativos, educacionais, profissionais e de entretenimento (WE ARE SOCIAL AND HOOTSUITE, 2022).

Em (MARQUESONE, 2016), é comentado sobre a era de dados que vivenciamos. Praticamente todas as nossas atividades diárias envolvem de alguma forma tecnologia conectada à rede, que está em constante desenvolvimento. Em nossos trabalhos, temos videoconferências, agendas online compartilhadas e vagas totalmente remotas, que acabam resultando na relação online durante todo o período de trabalho; no âmbito do lazer, temos serviços de streaming de filmes, séries, músicas, além de jogos online e leituras eletrônicas; compras podem ser realizadas de forma online, com facilidade em procura e comparação em diferentes sites. Resumidamente, existem serviços diversos para cada uma de nossas necessidades, resultante das evoluções de hardware, software e infraestrutura de redes.

Dados produzidos por diferentes serviços, em 1996 eram armazenados quase que em totalidade em meios físicos. Em 2007, 94% dos dados já eram armazenados em meios digitais, com dependência da internet (MARQUESONE, 2016).

Temos sensores em todos os lugares, crescimento da utilização de dispositivos conectados à rede, aumento do poder de armazenamento com um arquivamento quase infinito, nuvens de processadores com evoluções em todos os recursos computacionais. Todas essas habilidades em constante evolução resultam em uma geração massiva de dados, que vem mudando a ciência, a medicina, os negócios e a tecnologia (ANDERSON, 2009). Todas essas informações, com grande volume e variedade, necessitam de formas inovadoras para ingestão, transformação, armazenamento e análise, pensando na riqueza que esses dados podem conter para a tomada de decisões.

Big Data é um conjunto de dados que devido a quantidade e a variedade não podem ser manipulados com as ferramentas computacionais tradicionais (MANYIKA *et al*., 2011). Temos, portanto, um crescimento exponencial de dados, heterogêneos, oriundos de diferentes fontes, de forma distribuída e descentralizada (FAGUNDES, MACEDO E FREUND, 2018).

Além da evolução das ferramentas para a manipulação destes conjuntos de dados, os profissionais atuantes na área também receberam o desafio de se adaptarem para atenderem as necessidades diante da maior complexidade em cada trabalho desenvolvido. Os profissionais que trabalham com Big Data devem possuir conhecimentos específicos de maneira que cada equipe complete as demais. De acordo com (ANDERSON, 2020), existem três equipes de profissionais diferentes que trabalham na área de dados:

- Ciência de Dados: esta equipe consome dados resultantes de pipelines de dados iniciais e os manipulam para criação de dados derivados, por meio de análises avançadas, como o aprendizado de máquina. Estes profissionais possuem experiências com disciplinas matemáticas, principalmente a estatística. Estas experiências combinadas à programação são utilizadas normalmente para a criação de modelos computacionais pelos quais é possível obter informações valiosas de previsões ou detecção de anomalias para o seu negócio. Normalmente esses

profissionais possuem um conhecimento mais raso sobre ferramentas de Big Data e programação.

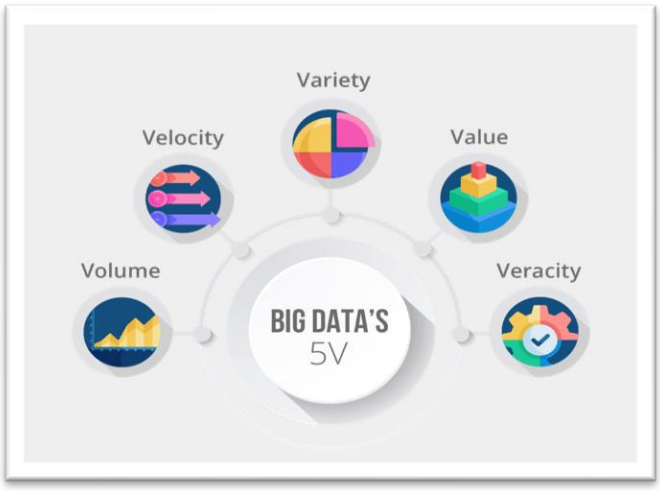
- Engenharia de Dados: para a aplicação da ciência de dados em escala para os negócios, é necessário ter profissionais que possuem o conhecimento para a criação de projetos e sistemas de dados sustentáveis. Os engenheiros de dados constroem pipelines de dados que alimentam os setores da organização, incluindo o time de ciência de dados. Estes profissionais devem possuir habilidades de limpeza, validação, manutenção, escalabilidade e sustentabilidade para adições e melhorias nos dados. Engenheiros de dados também são responsáveis pela escolha da infraestrutura de dados de acordo com o projeto e as especificidades do ambiente disponível para desenvolvimento das soluções. Normalmente, são profissionais especializados em Big Data, com no mínimo habilidade intermediária em programação e conhecimento em engenharia de software.
- Operações: este time é responsável por manter o desenvolvimento rodando e funcionando no ambiente de produção. Atualmente as organizações trabalham de duas maneiras com este time. Na primeira, a equipe de operações é responsável por auxiliar engenheiros de dados na automação, sem escrita de códigos para pipelines. A segunda maneira, menos tradicional, mistura a engenharia de dados com as funções operacionais para evitar o problema de um desenvolvimento com qualidade questionável por parte de engenheiros e a responsabilidade forçada da equipe de operações para resolvê-los. A equipe de operações, portanto é responsável pela operação dos softwares em produção, bom funcionamento dos clusters e tecnologias de big data operacionalizadas, otimização da rede para lidar com a grande quantidade de dados trafegando pela mesma, correção de problemas no hardware, instalação e configuração de softwares e sistemas operacionais para funcionar e otimizar as aplicações, entendimento dos dados enviados e acessados pelos sistemas.

Com a contextualização da mudança do cenário de dados, o surgimento do termo Big Data e a evolução que o mesmo trouxe com relação às ferramentas e especializações profissionais, os próximos tópicos abordam características e conceitos importantes sobre os dados que foram utilizados durante o desenvolvimento deste trabalho.

2.1.1. Os 5 Vs

O Big Data é referenciado por algumas características que definem seu significado. Inicialmente era referenciado por 3 Vs (conceitos de volume, velocidade e variedade), mas ao decorrer do tempo esses conceitos foram sendo modificados e hoje é possível encontrar referências que abordam e definem conceitualmente mais Vs.

Figura 1 - 5 V's da Big Data



Fonte:<https://www.analyticsvidhya.com/blog/2021/05/what-is-big-data-introduction-uses-and-applications/>. Acesso em: 24 mar. 22.

Neste trabalho são utilizados os conceitos apresentados na Figura 1 descritos por Bahga e Madisetti (2019):

- Volume: Big Data é formado por um conjunto de dados tão grande que é necessário ferramentas e estruturas especializadas para seu armazenamento, processamento e análise. Esse grande volume é gerado pelas indústrias, saúde, Internet das Coisas e demais sistemas de forma exponencial. Não existe um limite fixo para que uma quantidade de dados seja considerada como big data, mas normalmente são dados em grande escala que oferecem desafios para armazenamento, gerenciamento e processamento, necessitando de ferramentas não tradicionais.
- Velocidade: este conceito está relacionado com a rapidez em que os dados são gerados, o que influencia diretamente no crescimento exponencial dos dados e no volume alto para

armazenamento. Algumas aplicações precisam analisar esses dados em tempo real, e, portanto, é necessário ter ferramentas especializadas para essa ingestão e análise com alta velocidade para decisões mais assertivas.

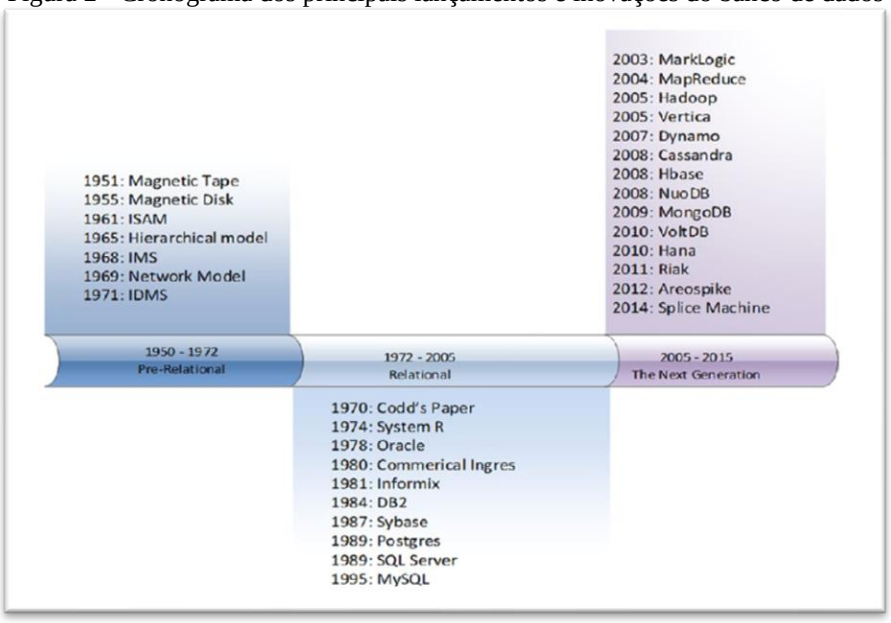
- Variedade: refere-se à variedade de dados que são gerados (dados estruturados, não estruturados e semiestruturados). Temos dados em formato de textos, imagens, áudios, vídeos, oriundos de sensores, de operações; e os sistemas precisam ser flexíveis para atender todos esses formatos e suas especificidades para que possa ser possível o tratamento e aquisição de informações a partir dos mesmos.
- Veracidade: este conceito se refere ao aspecto de confiabilidade dos dados. Para extrair valor, gerando informações de qualidade, é necessário que os dados sejam limpos e que ruídos (dados incorretos e/ou faltantes sejam identificados e/ou eliminados).
- Valor: refere-se à utilidade que os dados possuem para a finalidade pretendida, ou seja, para atender alguma necessidade ou resolver um problema. Esse valor está associado diretamente à veracidade e precisão dessas informações, podendo também em algumas aplicações depender da velocidade de processamento para tomada de decisões.

2.1.2 Evolução dos modelos de dados

Bancos de dados são conjuntos de dados organizados que tem como objetivo atender uma comunidade de usuários e sistemas com necessidades específicas. Para gerenciar esta necessidade de atendimento a diferentes sistemas, existem sistemas de gerenciamento de bancos de dados (SGBDs), que são softwares que incorporam funções de definição, recuperação e alteração de dados em um banco de dados (HEUSER, 1998).

As principais tecnologias de Big Data na linha do tempo entre 1960 e 2015 segundo Harrison (2015) são apresentadas na Figura 2.

Figura 2 - Cronograma dos principais lançamentos e inovações do banco de dados



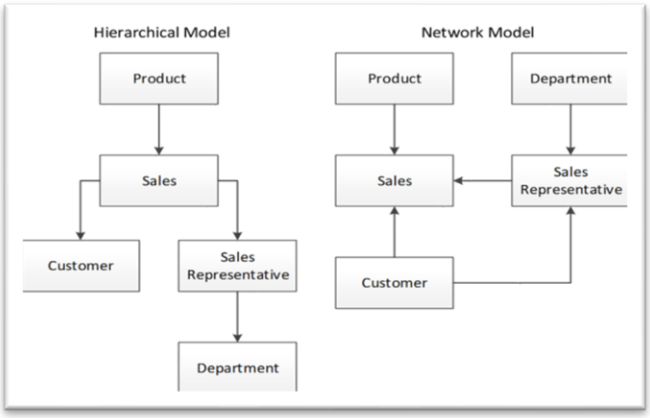
Fonte: (HARRISON, 2015, p. 4)

Após a 2ª Guerra Mundial, computadores eletrônicos surgiram e representaram a primeira revolução nos bancos de dados, onde tais computadores realizavam manipulação de dados, como processamento de comunicações militares criptografadas. Com o surgimento do disco magnético em 1950, o acesso a registros direto e com velocidade se tornou possível. Também surgiram métodos de indexação que tornaram o acesso orientado e rápido a registros, dando origem aos primeiros sistemas de transação online (OLTP) e primeiros bancos de dados eletrônicos (HARRISON, 2015).

Sem a existência ainda de sistemas de gerenciamento de banco de dados, notou-se problemas de produtividade causados pela necessidade de cada aplicação escrever seu próprio código para manipulação dos dados, aumentando os riscos de erros de código levando à dados corrompidos; não existia o controle de acesso aos dados, portando usuários poderiam acessar e alterar dados simultaneamente, os corrompendo fisicamente; além do processo de otimização de acessos por técnicas que tinham que ser implementadas com algoritmos complicados e especializados que dificilmente podiam ser duplicados para outras aplicações. A partir destes conflitos, surgiu a camada responsável por essa lógica de manipulação de banco de dados para garantir desempenho e integridade: os SGBDs (HARRISON, 2015).

Os primeiros SGBDs necessitavam da definição da estrutura dos dados que formariam o BD, além do caminho de acesso para navegar de um registro para outro. No início dos anos 1970, dois modelos de SGBDs ficaram conhecidos: o modelo de rede e o modelo hierárquico, como mostra a figura 3.

Figura 3 - Modelos de banco de dados hierárquicos e de rede



Fonte: (HARRISON, 2015, p. 6)

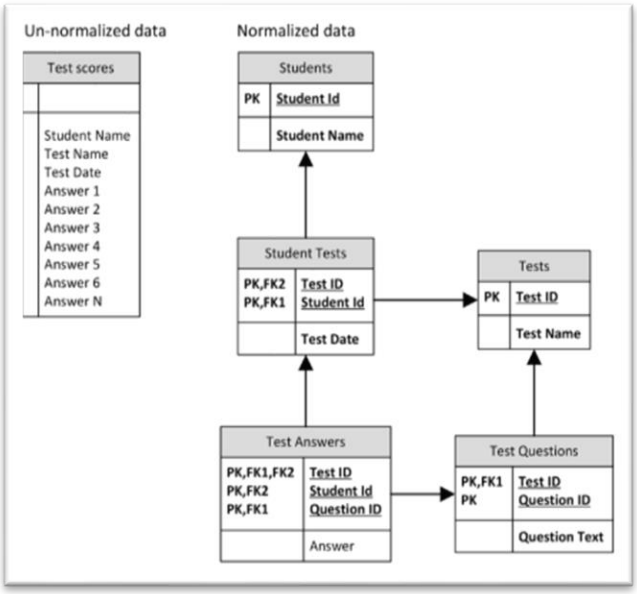
Esses dois modelos são conhecidos como modelos baseados em estruturas de navegação, onde é possível navegar de um objetivo para o outro por meio de ponteiros/links. Eram modelos inflexíveis com relação à estrutura dos dados e consultas, sendo difícil adicionar novos elementos de dados depois do sistema já existir (adicionar funcionalidades); e eram destinados para processamento de transações - CRUD (Criar, Ler, Atualizar, Excluir), operações de consulta com análises mais complexas exigiam também complexidade na codificação.

No final da década de 1980, um “matemático de programação”, chamado Edgar Codd, trabalhava para a IBM e realizou alguns questionamentos sobre os bancos de dados da época: a dificuldade de utilização por pessoas sem habilidades de programação avançadas, a falta de consistência lógica e as confusões entre implementações lógicas e físicas; com essas inquietações, Codd publicou um artigo interno descrevendo formalmente um modelo relacional para sistemas de banco de dados grandes e compartilhados, com ideias centrais do modelo relacional (HARRISON, 2015). Temos então a segunda revolução dos bancos de dados, marcada pelos modelos relacionais. Esses modelos focam em como os dados de um conjunto devem ser apresentados aos usuários, com a finalidade de manter integridade e diminuir redundâncias de dados, incluindo os conceitos-

chave de tuplas, relações, restrições e operações (HARRISON, 2015).

Neste modelo é possível identificar cada linha em uma tabela, de forma eficiente por meio de um valor de chave exclusivo (conceito de chave primária). As redundâncias e inflexibilidade desses bancos são tratadas por meio das várias formas normais (uma sequência de regras), que correspondem à normalização dos dados visualizada na figura 4, um processo onde os mesmos são organizados, com a criação de tabelas e relacionamentos entre as mesmas.

Figura 4 - Dados normalizados e não normalizados



Fonte: (HARRISON, 2015, p. 9)

No começo dos anos 2000, bancos de dados relacionais estavam bem estabelecidos, sem visão de mudanças bruscas no modelo para atender novas demandas. Porém, os problemas com big data começaram a ser encontrados por empresas como o Google, que se viu obrigada a desenvolver novas arquiteturas de hardware e software para lidar com o grande volume e velocidade de dados que necessitava armazenar e processar.

Em 2003 o Google revelou informações sobre o seu sistema de arquivos distribuídos desenvolvido (Google File System), além de detalhes sobre o algoritmo de processamento distribuído MapReduce, seu banco de dados estruturado distribuído BigTable. Neste momento temos o início da terceira revolução dos bancos de dados, com essas tecnologias que foram base

para o projeto do ecossistema Hadoop, um conjunto de aplicações facilitadoras para o trabalho com Big Data (HARRISON, 2015).

Michael Stonebraker, um dos pioneiros do banco de dados Postgres, com uma equipe de pesquisa publicou em 2007 o artigo “The End of an Architectural Era (It’s Time for a Complete Rewrite)” onde foi apresentada a necessidade de flexibilidade da arquitetura de SGBDRs para bancos de dados modernos para atender a variedade de cargas de trabalho. Dois projetos se tornaram muito relevantes como primeiros sistemas NewSQL (sistemas que possuem as características principais de SGBDRs tradicionais, mas com variações para atender as demandas modernas). Um deles foi o H-Store, um banco distribuído em memória; e o outro, o C-Store, possuía o design para BDs colunares (STONEBRAKER et al., 2007).

A partir dos anos 2000, uma série de sistemas de bancos de dados surgiram, como o *MongoDB* e Cassandra, sistemas que possuem participação no mercado até hoje. Em 2009, o termo *NoSQL* se familiarizou e todos os sistemas dessas novas gerações rompiam com o modelo relacional. Segundo Fatima e Wasnik (2016, p. 1), o modelo *NoSQL* apresenta um esquema flexível para trabalhar com a variedade de dados, apresentando também escalabilidade e disponibilidade, favorecendo o trabalho com *big data*.

O cenário moderno de dados, como comentado até o momento, envolve uma característica importante: a variedade. Os diferentes formatos de dados são abordados por Eberendu (2016):

- Dados estruturados: correspondem aos dados que possuem uma estrutura definida, com fácil organização, armazenamento e análise, já que essas estruturas são previamente estabelecidas antes da ingestão de dados nas mesmas; essa estrutura fixa contínua permite consultas para recuperação de informações de forma prática com linguagem de consulta estruturada (*SQL*). São dados tipicamente armazenados em BDs relacionais, organizados em tabelas com linhas e colunas e aceitam dados formados por números e caracteres textuais (*strings*); lembrando que, por possuírem uma estrutura pré-estabelecida, o tipo de dado que cada coluna será formada é definido inicialmente e se torna inflexível.
- Dados semi-estruturados: estes dados não possuem uma estrutura definida e inflexível. São dados irregulares, que podem estar incompletos, possuem uma estrutura, mas a mesma é passível de mudanças imprevisíveis, ou seja, é uma estrutura flexível. Esses tipos de dados

possibilitam o agrupamento de informações de diferentes fontes por meio de propriedades que os relacionam. São exemplos de dados semi-estruturados arquivos XML, que possui tags de identificação de estruturas, podendo representar informações complexas, com objetos compostos, relações de hierarquia, arrays, etc.
- Dados não estruturados: são dados que não possuem uma estrutura definida, com padrões, e podem ser compostos por uma variedade de elementos. Exemplos de dados não estruturados são imagens, vídeos, áudios e corpos de e-mails. Segundo Dijcks (2013), os dados não estruturados representam cerca de 80% dos dados das organizações, informações que não eram utilizadas antes do desenvolvimento das tecnologias de Big Data.

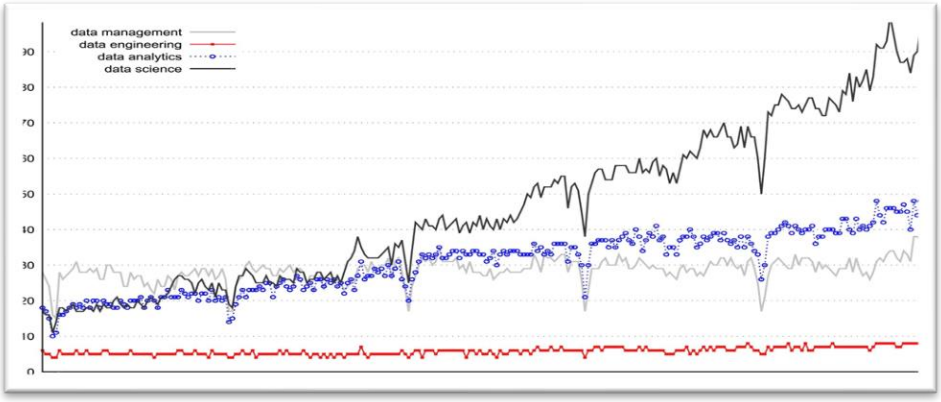
2.1.3 Pipelines de dados (ETL e ELT)

Como discutido anteriormente, a evolução dos modelos de dados e a complexidade que o Big Data trouxe para a manipulação dos mesmos, também teve influências sobre a divisão das equipes e áreas de atuação. Transformar os dados em informações de valor é o grande objetivo, e para atingir este propósito, tarefas distintas são exigidas, envolvendo diferentes disciplinas.

A ciência de dados tem a finalidade de extrair conhecimento e insights de dados em suas várias formas (estruturados, semiestruturados e não estruturados) por meio de algoritmos e ferramentas de análise (DHAR, 2013). Engenheiros de dados dão apoio a essas atividades, sendo responsáveis pela construção de ecossistemas que tornam o trabalho dos cientistas viável e mais facilitado, gerenciando o ciclo de vida dos dados, desde a coleta, integração e persistência, fornecendo-os com qualidade, eficiência e segurança.

Apesar da importância dos engenheiros de dados discutida até o momento, o artigo de Romero, Wrembel e Song (2020) identifica a baixa popularidade relacionada a problemas e tecnologias de engenharia de dados entre os usuários da internet, como visualizado na figura 5.

Figura 5 - Popularidade dos tópicos de gerenciamento de dados



Fonte: (ROMERO, WREMBEL e SONG, 2020)

A prática desenvolvida neste trabalho, evidencia uma parte do trabalho realizado por engenheiros de dados: a construção de pipelines de processamento de dados (DPPs).

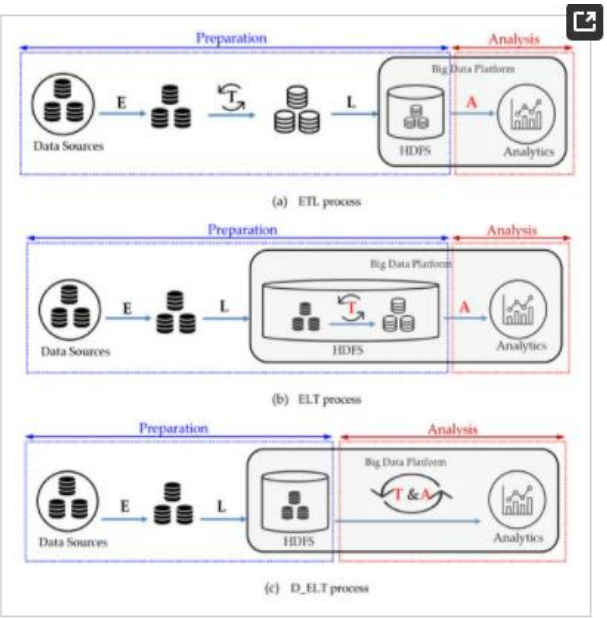
Para automatizar um fluxo de dados desde sua origem até seu destino final, pipelines de dados são criados. Os mesmos funcionam como uma cadeia com uma sequência de atividades simples e complexas, que realizam a manipulação desde a obtenção dos mesmos em diversas origens, até a transformações desses dados diante da finalidade desejada para o destino que envolve o armazenamento e aplicações em ferramentas de visualização e modelos de Machine Learning, como exemplo (RAJ, BOSCH, OLSSON e WANG, 2020).

DPPs incluem tarefas de preparação de dados e tarefas de análise. Tarefas de preparação normalmente incluem integração de dados heterogêneos oriundos de distintas fontes e transformação para uma representação comum; limpeza e padronização; eliminação de dados redundantes; e armazenamento dos dados limpos em um repositório centralizado. Já as tarefas de análise incluem a extração de dados de um repositório centralizado para armazenamento de visualizações específicas; pré-processamento para um algoritmo analítico; criação de conjuntos para testes e validações; e análises estatísticas.

O termo ETL é utilizado para se referir a essas tarefas de extração, transformação e carga de dados, onde os dados são extraídos de fontes variadas, passam por transformações com a finalidade de carregamento em banco final centralizado como discutido no parágrafo anterior. Para a aplicação dessas tarefas na quantidade e variedade de dados derivados do big data, este processo

clássico teve tentativas de evolução utilizando o conceito de processamento paralelizado e distribuído, adicionando etapas de particionamento antes das transformações e redução antes do carregamento (JO e LEE, 2019).

Figura 6 - Processo de preparação e análise de dados[1]



Fonte: (JO e LEE, 2019, p.6)

Na figura 6 temos o processo de ETL tradicional, com dados sendo extraídos de diversas fontes e em seguida transformados por um servidor ETL, e carregados em um sistema de arquivos distribuídos (HDFS), possibilitando então a análise. Esse processo não tem o melhor desempenho diante de grandes quantidades e variedades de dados, e por esse motivo, os processos de ELT (Extração, carga e transformação) foram criados para acelerar a preparação de dados.

A ideia principal do ELT observado na figura 6 é realizar o carregamento logo após a extração, para só então realizar as transformações; temos então o armazenamento dos dados brutos de origem e aproveitamento do sistema de destino para realizar as transformações, eliminando o servidor ETL. Ainda pode-se acelerar as transformações usando distribuição com suportes de arquiteturas como o Ecossistema Hadoop. A figura supracitada mostra ainda uma variação do ELT

[1] Retirado de um caso de big data geoespacial. (a) representa o processo de ETL, (b) ELT e (c) D_ELT. E significa extrair, T significa transformar, L significa carga e A significa análise.

aplicado no processamento de dados geoespaciais, para a execução da transformação de maneira distribuída e paralela com a análise.

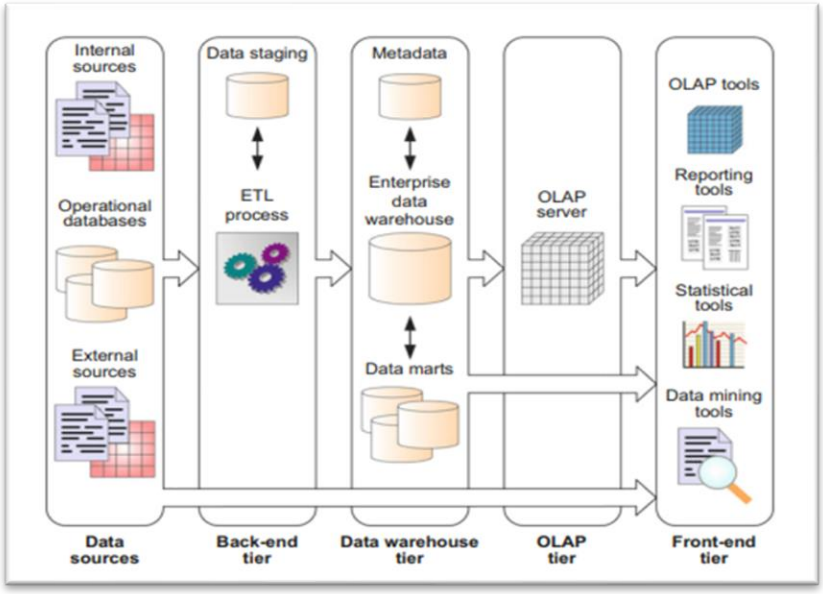
2.1.4 Data Warehouse e Data Lake

Sistemas de processamento de transações online (OLTP) são bancos de dados tradicionais projetados para processamento de transações diárias de uma organização, com controle sobre o acesso simultâneo e técnicas de recuperação para garantir a consistência dos dados. Esses sistemas além de não conterem dados históricos, são altamente normalizados para evitar inconsistências durante atualizações, e por esse motivo não são caracterizados como sistemas indicados para análise de dados, pelo baixo desempenho na execução de consultas complexas que necessitam agregar dados oriundos de muitas tabelas relacionais criadas pelo particionamento resultante da normalização (VAISMAN e ZIMANYI, 2016).

Os bancos de dados orientados a processamento analítico online (OLAP), surgiram para atender as necessidades que as organizações passaram a ter sobre a análise de dados instantâneos e históricos para a tomada de decisões, com foco no atendimento de cargas pesadas de consultas. Para atender esse paradigma, um novo modelo de banco de dados surgiu: o data warehouse (DW), uma consolidação centralizada de dados vindos do ambiente interno e externo a uma organização, com um modelo multidimensional, onde os dados são representados como hipercubos em que cada dimensão se refere à uma perspectiva do negócio e cada célula contém as medidas que são analisadas (VAISMAN e ZIMANYI, 2016).

Os DWs, como mencionado, surgiram com o objetivo de centralizar dados para análise de uma organização inteira. Porém, cada departamento da organização pode ter o interesse em uma parte específica do mesmo que atende especificamente às necessidades desta divisão. Assim, temos o conceito de data mart, uma espécie de DW departamental (VAISMAN e ZIMANYI, 2016). A figura 7 mostra a arquitetura de um DW clássico com suas diferentes camadas/níveis:

Figura 7 - Arquitetura de um DW



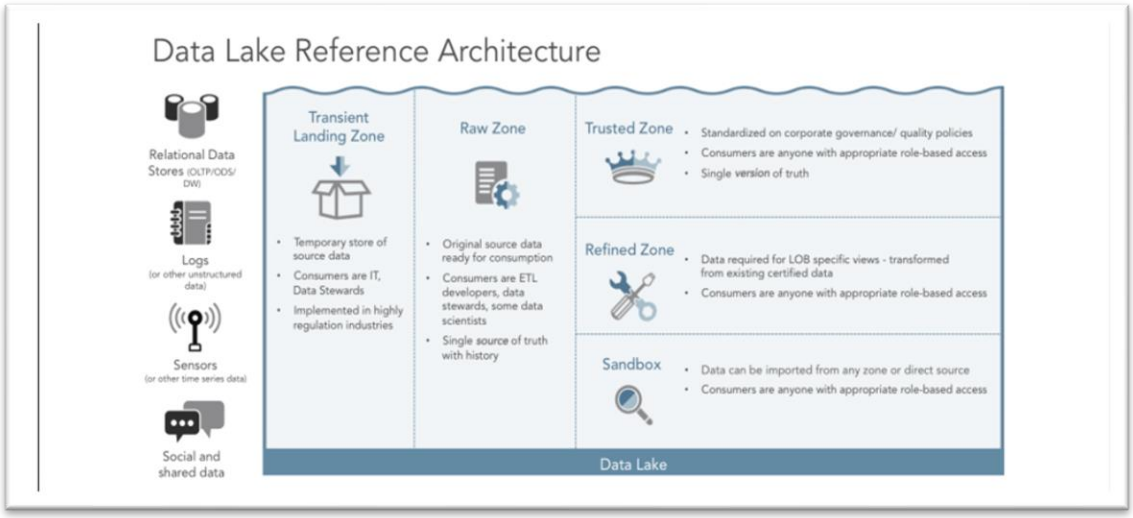
Fonte: (VAISMAN e ZIMANYI, 2016, p. 77)

- Camada back-end: formada por todos os processos de ETL, utilizados para sustentar a ingestão de dados no DW a partir das diferentes fontes de dados internas e externas da organização. Nesta camada também existe um banco de dados intermediários para auxílio aos processos de transformação.
- Camada data warehouse: composta pelos vários data marts e o DW corporativo, juntamente com um banco que armazena informações sobre estrutura e conteúdo do DW (metadados).
- Camada OLAP: responsável pelo fornecimento da visão multidimensional dos dados através de um servidor OLAP, de forma independente ao armazenamento realizado no sistema subjacente.
- Camada front-end: utilizada para análise e visualização dos dados, contendo ferramentas OLAP, de relatórios, estatísticas e de mineração de dados.

Para complementar esses designers de bancos de dados que atendem as análises de uma organização, temos a definição de Data Lake (DL), destinado ao armazenamento e processamento de big data. DL pode ser visualizado assim como DW, como sendo um repositório que centraliza dados de uma organização, mas com a diferença de armazenar dados de todos os formatos

(estruturados, semi-estruturados e não estruturados), escalonável e sem a necessidade de definição de estruturas e esquemas para esses dados antes do armazenamento. Temos, portanto, uma visão lógica de todos os conjuntos de dados em seus formatos brutos, que é acessível para compreensão e extração de conhecimento por todos os profissionais. Em DLs temos a consideração da variedade de dados e uma abordagem schema-on-read, onde os requisitos de esquema e os dados só são corrigidos (transformados) e definidos durante a consulta dos dados, de acordo com as partes do sistema corporativo (SAWADOGO e DARMONT, 2021).

Figura 8 - Arquitetura de Referência de um Lago de Dados



Fonte: (SHARMA, 2018, p.16)

Com relação à arquitetura, um DL pode ser dividido em zonas/camadas de acordo com o grau de refinamento dos dados. A figura 8, apresentada a seguir contém a arquitetura de referência adotada por Zaloni (SHARMA, 2018), que pode ser utilizada pelas organizações para aplicação de melhores práticas, visualizando e entendendo os componentes e tecnologias envolvidos em cada processo e etapa rastreada, para assim derivar um modelo que atenda cada solução da melhor maneira possível.

- Zona de carregamento transitória: é uma camada temporária, sendo destino para dados onde verificações e transformações básicas de qualidade e conformidade relacionadas a medidas de segurança podem ser aplicadas antes do armazenamento e acesso aos dados.

- Zona bruta: destino de armazenamento dos dados após a aplicação das transformações de segurança e verificações de qualidade aplicadas na camada transitória, quando existente; caso contrário, essa camada armazena permanentemente os dados em seu formato bruto/original.
- Zona confiável: baseada nos dados da camada bruta, onde os mesmos são limpos e padronizados de acordo com as políticas estabelecidas para atender as demandas do negócio.
- Zona refinada: nesta camada, os dados são preparados para finalmente serem utilizados para derivação de insights de acordo com as necessidades específicas do negócio. Os dados são integrados em um formato comum para facilitar criação de modelos e relatórios, e mais processos de qualidade e manutenção são aplicados.
- Zona de descoberta (Sandbox): essa camada permite criações de casos de uso e exploração de variáveis que afetam o negócio, testes dos dados e desenvolvimento de aplicações sem necessidades especiais de envolvimento de departamentos, TI e dedicação de fundos para criação de ambientes para testes. Dados de qualquer uma das zonas pode ser importada, inclusive das fontes iniciais.

DW e DL são repositórios de dados com conceitos, estruturas e implementações diferentes. Duas dessas diferenças diz respeito ao esquema de dados e a ordem de processamento dos mesmos; enquanto em um DW ocorre o processo de ETL e existe uma abordagem “Schema-on-Write”, onde o esquema dos dados precisa ser definido antes do carregamento dos dados, em um DL tem-se a ocorrência do processo de ELT, com abordagem “Schema-on-Read”, sem a necessidade de definição de esquema prévia, pois só no momento de solicitação dos dados que os mesmos serão transformados para um formato apropriado utilizando os metadados (KHINE e WANG, 2017).

2.1.5 Armazenamento On-Premise e Cloud Computing

Quando é discutido sobre um ambiente on-premise, temos a referência de um armazenamento local do servidor trabalhado, significando que o mesmo se encontra hospedado na infraestrutura da própria organização, que fica responsável por assegurar toda a administração, controle e manutenção do mesmo (DIAMOND, 2020). Esse controle sobre a infraestrutura exige tempo e mão de obra qualificada para tamanha responsabilidade com complexidade operacional,

ponto que supera o controle mais rígido sobre segurança e privacidade de dados que esse ambiente proporciona (LAPLANTE e SHARMA, 2016).

Com relação às tecnologias que foram desenvolvidas para a manipulação de big data no ambiente on-premise, tem-se destaque para o ecossistema Hadoop. Essas tecnologias utilizam do conceito da computação distribuída, onde o trabalho de processamento e armazenamento é escalável e otimizado pela adição de nós ao cluster, que funcionam como novas unidades computacionais trabalhando em conjunto e paralelamente (TANENBAUM e STEEN, 2006).

O Apache Hadoop foi desenvolvido para promover soluções para problemas com integridade de dados, disponibilidade de nós, escalabilidade e recuperação de falhas para computação distribuída e foi utilizado por grandes empresas como Facebook, Twitter, LinkedIn e The New York Times (GOLDMAN, KON, JUNIOR, POLATO e PEREIRA, 2012).

Em contrapartida ao ambiente on-premise, a computação em nuvem popularizada a partir de 2007, tem como base servidores virtuais com hospedagem distribuída em grandes data centers, onde os recursos são compartilhados e disponibilizados pela internet (BIBI, KATSAROS e BOZANIS, 2012).

Três tipos de sistemas são oferecidos pelos provedores de computação em nuvem, como mostrado na Figura 9.

Figura 9 – SaaS, PaaS e IaaS

Fonte: <https://azure.microsoft.com/en-us/overview/what-is-paas/>. Acesso em: 16 mar. 22.

- Plataforma como infraestrutura (IaaS): oferecimento sobre demanda de recursos essenciais de computação, armazenamento e rede, pagando em conformidade com o uso. Oferece maior flexibilidade com relação ao dimensionamento de recursos de acordo com demanda, reduzindo gastos com hardware e manutenções locais e tirando a responsabilidade sobre o gerenciamento de infraestrutura. Com esse sistema, uma infraestrutura necessária para uma nova iniciativa fica pronta em minutos ou horas, ao invés de levar dias (AZURE MICROSOFT, s.d.).
- Plataforma como serviço (PaaS): oferecimento de infraestrutura (servidores, armazenamento e rede) e serviços que dão suporte à construção, teste, implantação, gerenciamento e atualização de aplicativos. Neste sistema, o contratante se concentra no gerenciamento dos serviços que desenvolve e o provedor se encarrega do gerenciamento do restante (AZURE MICROSOFT, s.d.).
- Software como serviço (SaaS): onde uma solução de software completa é oferecida de acordo com a contratação e o uso da mesma é compartilhado para sua organização e usuários por meio da conexão pela internet. Hardware e software é gerenciado pelo provedor, garantindo disponibilidade e segurança ao contratante. Um exemplo são os serviços de e-mail como o Outlook (AZURE MICROSOFT, s.d.).

A computação em nuvem é um modelo com recursos adquiridos sob demanda, de forma rápida, sendo provisionados e liberados com mesma velocidade e sem grandes esforços de gerenciamento. Ela permite que o contratante se concentre na solução, no core business, ao invés de dedicar uma grande quantidade de tempo com questões de disponibilidade de recursos, infraestrutura e flexibilidade (HASHEM, YAQOOB, ANUAR, MOKHTAR, GANI e KHAN,

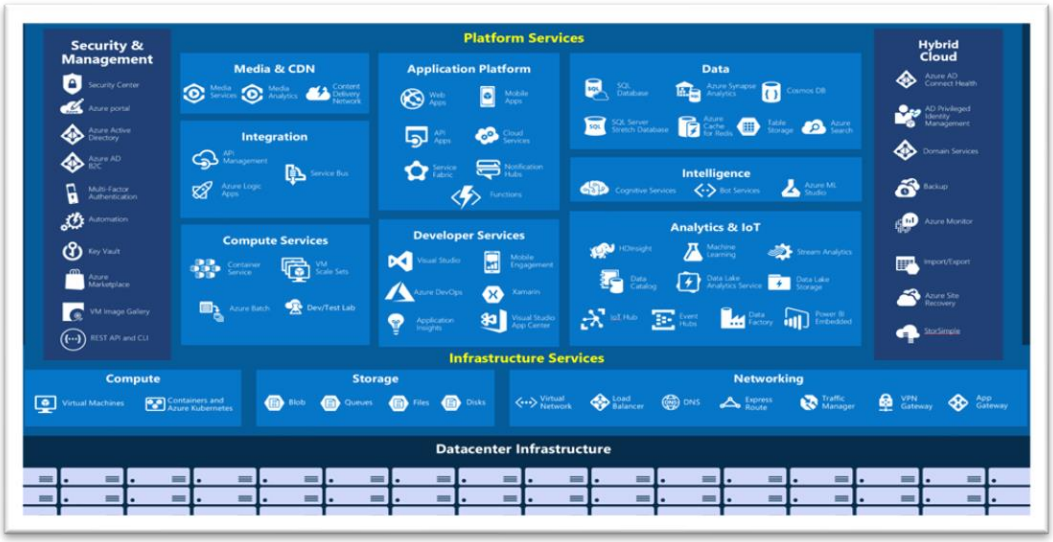
2.2 Microsoft Azure

Microsoft Azure, comumente mencionada apenas como Azure, é a plataforma de nuvem da Microsoft que reúne mais de 200 produtos e serviços, em inovação contínua e possibilitando abertura para construções de seus termos, migração de aplicação locais para a nuvem e operação híbrida, com suporte de especialistas para garantir confiabilidade e segurança (MICROSOFT, s.d.)

A Azure possui um console web por onde é possível gerenciar sua assinatura por meio de uma interface gráfica, ao invés de linhas de comando. Os serviços estão divididos em várias

categorias, dando destaque para serviços de computação, armazenamento, rede, bancos de dados, Internet das Coisas, Big Data, inteligência artificial e DevOps (MICROSOFT, s.d.). Na figura 10 é possível visualizar de forma geral alguns dos principais recursos disponíveis.

Figura 10 - Plataforma de serviços

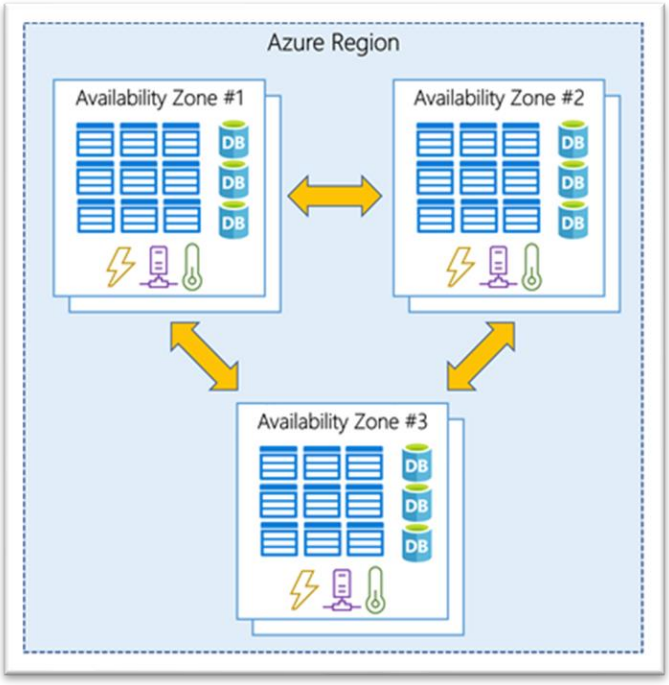


Fonte: <https://bit.ly/3LXPjcy>. Acesso em: 24 mar. 22.

Para acesso autenticado e autorizado aos recursos, é necessário possuir uma assinatura do Azure, que funciona como uma unidade lógica de serviços vinculados a uma conta, que por sua vez pode ter uma ou várias assinaturas com modelos de cobrança distintos que adicionam limites aos usos dos recursos. Para a realização da prática no capítulo 4, uma conta de estudante gratuita foi criada, onde por 12 meses é liberado o uso de um crédito disponibilizado e acesso gratuito a alguns serviços (MICROSOFT, s.d.)

O Azure possui datacenters distribuídos por todo o mundo, os quais não são expostos diretamente aos usuários e sim organizados em regiões geográficas do planeta. Essas regiões oferecem escalabilidade e redundância, além de possuir zonas de disponibilidade em algumas delas, para garantir a redundância de dados e proteção a falhas.

Figura 11 - Região Azure



Fonte: <https://bit.ly/37bGa1s> Acesso em: 24 mar. 22.

Como mostra a figura 11, essas zonas são datacenters localizados em uma mesma região, mas que estão fisicamente separados, conectados por meio de redes privadas de fibra óptica. Elas podem ser usadas para replicação de aplicações para criar alta disponibilidade em sua arquitetura. Se uma zona cai, outra continua funcionando normalmente (MICROSOFT, s.d.)

A disponibilidade mencionada anteriormente é um dos benefícios da computação em nuvem, que pode oferecer aos usuários de suas aplicações uma experiência contínua, sem aparentar tempo de inatividade. Outros benefícios são a escalabilidade, com a facilidade de dimensionamento vertical e horizontal; elasticidade e agilidade em implantar e configurar recursos novos que se adaptam às mudanças de requisitos da sua aplicação; distribuição geográfica, com implantações em regiões de todo o mundo, garantindo o melhor desempenho para o cliente de cada região; e a recuperação de desastres, garantida pelos serviços de backup e a replicação de dados (MICROSOFT, s.d)

As próximas seções tratam sobre a explicação breve dos recursos que são utilizados neste trabalho.

2.2.1 Resource groups

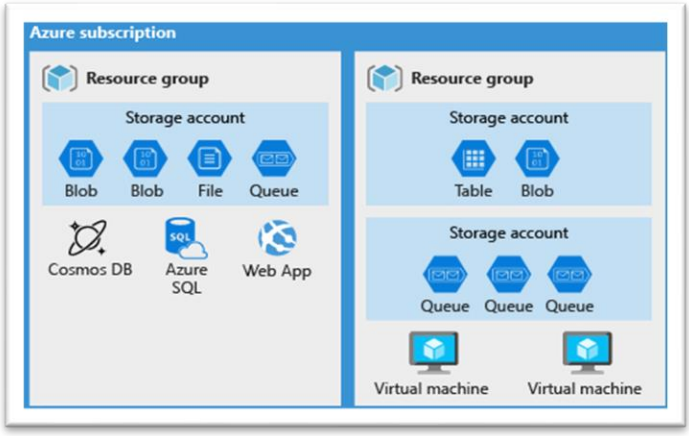
Os resource groups são uma estrutura fundamental na plataforma Azure. Eles funcionam como um contêiner lógico para recursos, e estes só podem ser provisionados com a premissa da existência de um grupo de recursos para ser associado.

Estes grupos auxiliam no gerenciamento dos recursos do Azure de forma lógica, possibilitando a organização dos mesmos por tipo, uso e/ou local. Além do benefício de agrupamento lógico de recursos, ao excluir um resource group, todos os recursos associados a ele também serão excluídos, facilitando assim a remoção de recursos de uma única vez. Outro benefício está relacionado ao poder de aplicação de controle de acesso baseado em função (RBAC), limitando o acesso para permitir apenas o necessário aos usuários ou grupos.

2.2.2 Storage Account - Data Lake Storage Gen2

O Azure fornece diversos serviços de armazenamento de dados, considerando toda a variedade de estruturas. Assim como os resource groups, existe uma estrutura que funciona como um contêiner para agrupar serviços de armazenamento, as storage accounts.

Figura 12 - Inscrição Azure



Fonte: <https://bit.ly/3Jy9AEf>. Acesso em: 22 mar. 22.

Uma storage account faz parte de um resource group como visualizado na figura 12, e possibilita o gerenciamento sobre os serviços de armazenamento associadas a ela. Todas as configurações especificadas para uma storage account são aplicadas aos serviços que ela agrupa.

Essas configurações dizem respeito a identificação da assinatura na qual os serviços serão cobrados; a localização do datacenter que irá armazenar os serviços da conta; o desempenho, que determina quais tipos de serviços de dados podem ser associados à sua conta e o tipo de disco de hardware que será usado para armazenamento; e estratégia de replicação usada para fazer cópias dos dados para proteção contra falhas de hardware e desastres naturais; camada de acesso que controla a rapidez que os dados podem ser acessados na conta; transferência segura que diz respeito os protocolos suportados para acesso; e redes virtuais que permite acessos apenas de redes virtuais especificadas por segurança (MICROSOFT, s.d.).

Uma solução de data lake para manipulação de big data criada pelo Azure é o Azure Data Lake Storage Gen2, que possui a combinação de um sistema de arquivos com uma plataforma de armazenamento. Esse serviço se baseia no armazenamento de Blobs do Azure, um armazenamento sem restrições sobre os tipos de dados, podendo conter, por exemplo, documentos PDF, imagens, arquivos JSON, CSV e vídeos.

Figura 13 - Armazenamento de Lago de Dados Azure Gen2

Fonte: <https://bit.ly/37BV94A>. Acesso em: 14 mar. 22.

O Azure Data Lake Storage funciona como o sistema distribuído de arquivos do Hadoop, podendo armazenar dados de diferentes formatos e fontes em um mesmo lugar e acessá-los com as tecnologias de computação como o Azure Databricks. Ele foi projetado para lidar com a variedade e volume de dados, com soluções em tempo real e em lote. Os dados são armazenados em uma hierarquia de diretórios e subdiretórios, semelhante a um sistema de arquivos, onde o controle de

acesso e permissões aos diretórios e arquivos podem ser configurados. A redundância de dados também é garantida pelo armazenamento localmente redundante (LRS) ou o armazenamento geograficamente redundante (GRS) (MICROSOFT, s.d.).

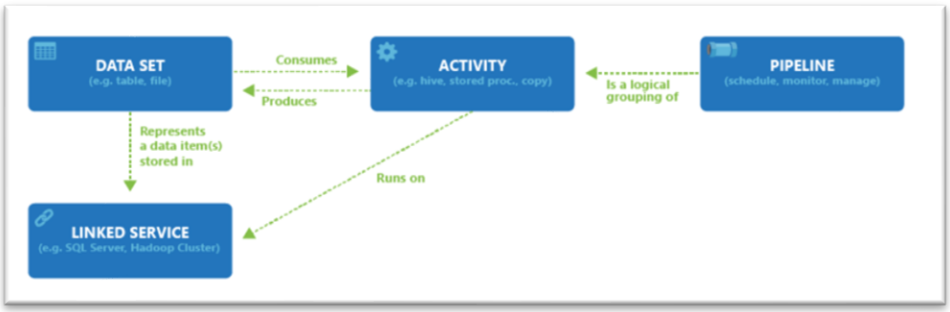
2.2.3 Azure Data Factory

O Azure Data Factory (ADF) é um serviço para integração de dados. Ele é composto por uma série de ferramentas de manipulação (ETL/ELT), dando suporte para acesso em fontes variadas, encontradas tanto na Azure, como localmente ou em plataformas de nuvem de terceiros.

A combinação das funções do ADF fornece um ambiente para análise de dados. A conexão com fontes de dados é realizada por meio da criação de um linked service, que funciona como uma cadeia de conexão entre o Data Factory e os recursos externos de armazenamento de dados e também recursos de computação (MICROSOFT, s.d.).

Uma vez conectado, é possível coletar dados estruturados, semiestruturados e não estruturados gravando-os no destino (coletor). Dados coletados podem ser processados e transformados de acordo com a demanda do negócio por data flows, que realizam execução no Spark (uma estrutura de processamento paralelo na memória que melhora o desempenho de aplicações que analisam Big Data); e por fim podem ser carregados em mecanismos de análise. Todo esse fluxo de atividades mencionadas acima, pode ser organizado em um agrupamento lógico chamado pipeline, ou uma série de pipelines específicos para a execução de um trabalho especializado, automatizando o processo com a possibilidade de adição de um gatilho (trigger) para a execução, além da melhoria de gerenciamento das atividades desenvolvidas (MICROSOFT, s.d.).

Figura 14 - Relacionamento entre pipelines, atividades, conjunto de dados e serviço vinculado

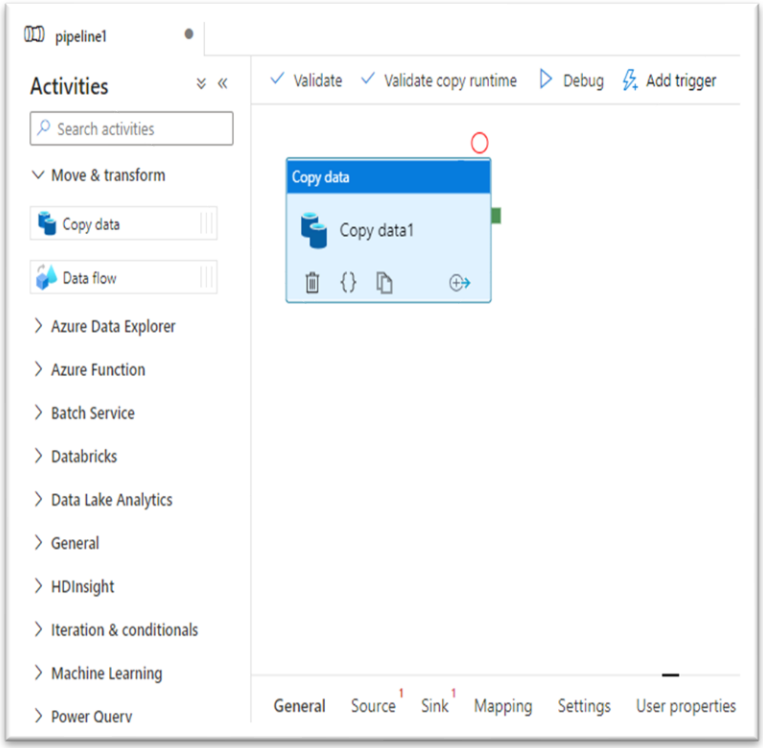


Fonte: <https://bit.ly/3KATGdn>. Acesso em: 24 mar. 22.

Os pipelines de dados do ADF podem ser formados por uma série de atividades que executam ações de movimentação, transformação e controle. Abaixo serão descritas as atividades que foram utilizadas para o desenvolvimento deste projeto. Como visualizado na figura 14, as atividades possuem uma organização em grupos de ações. Para realizar uma cópia de dados entre armazenamentos locais e na nuvem, é possível utilizar a atividade *Copy data* encontrada no grupo de movimentação e transformação.

A atividade de cópia ilustrada na figura 15 faz a leitura de dados de um armazenamento origem e então executa as operações de integração em tempo de execução (serialização/desserialização, compactação/descompactação, mapeamentos e assim por diante, de acordo com as configurações realizadas no conjunto de dados de entrada e saída, e na própria atividade de cópia). Por fim, os dados são gravados no armazenamento configurado como destino/coletor (MICROSOFT, 2021).

Figura 15 - Copy Data

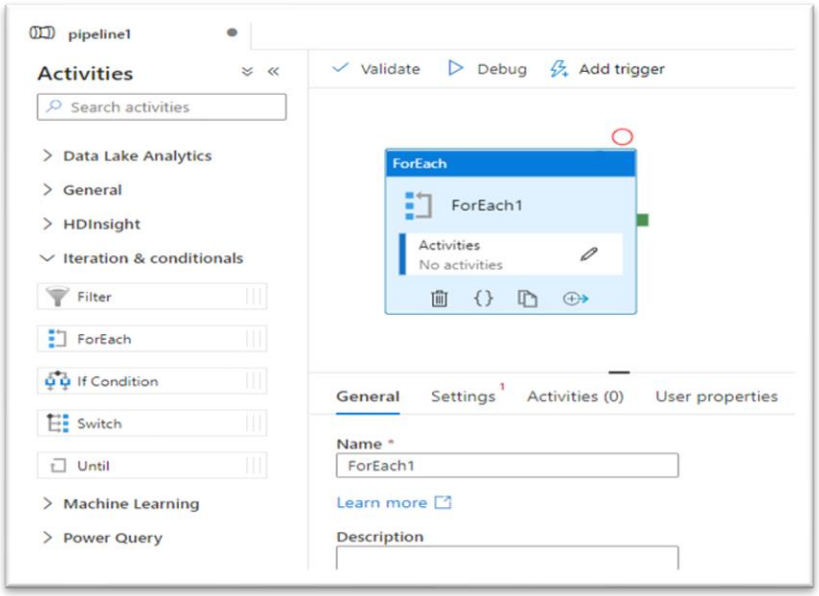


Fonte: <https://bit.ly/3ri6rC5>. Acesso em: 26 mar. 22.

A atividade de cópia foi utilizada neste trabalho para realizar a cópia de dados de uma fonte HTTP utilizando a interface do usuário e ingestão final em um Azure Data Lake Storage Gen2 (MICROSOFT, 2022).

Uma atividade interessante que também foi utilizada é *ForEach* presente no grupo de iteração e condicionais. Esta atividade exemplificada na figura 16 funciona como um fluxo de controle de repetição, utilizada para iteração e execução de uma atividade específica em um loop, semelhante a estrutura *for* utilizada nas linguagens de programação. Pode conter como entrada qualquer variável do tipo matriz ou saída de uma atividade executada anteriormente. Este loop de repetição pode ser executado de forma sequencial ou paralela.

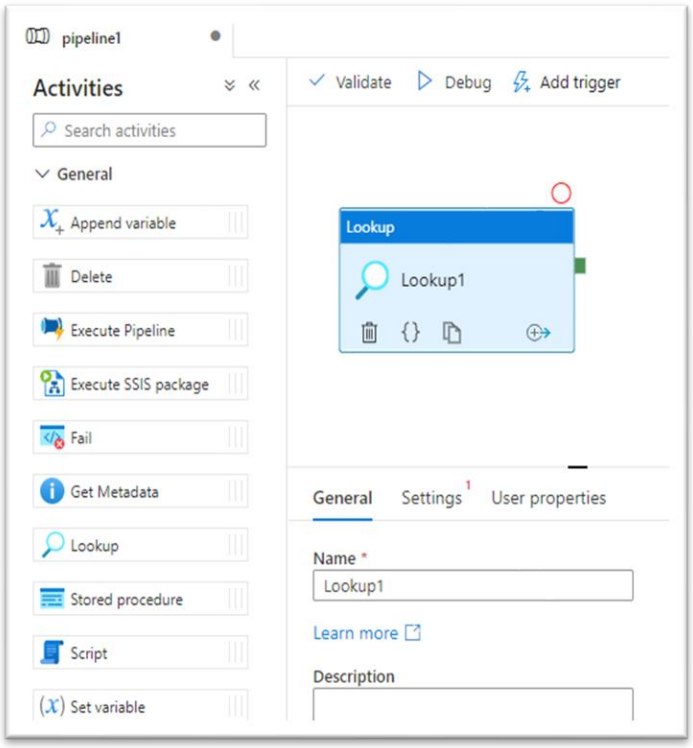
Figura 16 - ForEach



Fonte: <https://bit.ly/3LYgfco>. Acesso em: 18 mar. 22.

Para a leitura de arquivos com parâmetros, configurações ou tabelas, é possível utilizar a atividade *Lookup* do conjunto de atividades gerais. Esta atividade representada na figura 17 recupera um conjunto de dados de qualquer uma das fontes de dados suportadas pelos pipelines do ADF e então retorna como saída um valor único ou uma matriz que pode ser consumida pelas demais atividades do pipeline.

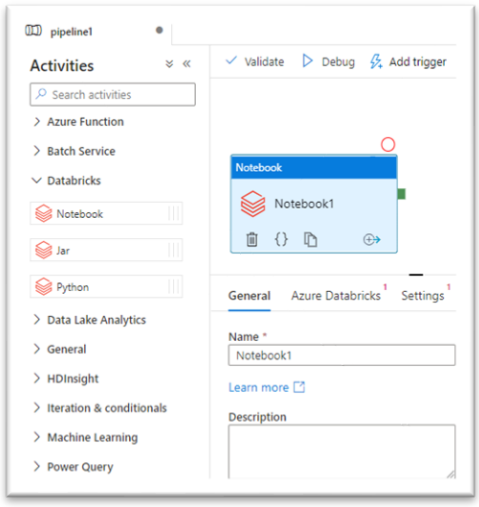
Figura 17 - Lookup1



Fonte: <https://docs.microsoft.com/pt-br/azure/data-factory/control-flow-lookup-activity>. Acesso em: 21 mar. 22.

Um notebook Databricks pode ser inserido no pipeline de dados do ADF, como retratado na figura 18. A atividade *Notebook* se encontra no conjunto de atividades Databricks. Para a utilização é preciso ter um workspace Azure Databricks e o notebook criado dentro do mesmo. É preciso criar um linked service para a conexão do Data Factory ao recurso Databricks criado (MICROSOFT, 2022).

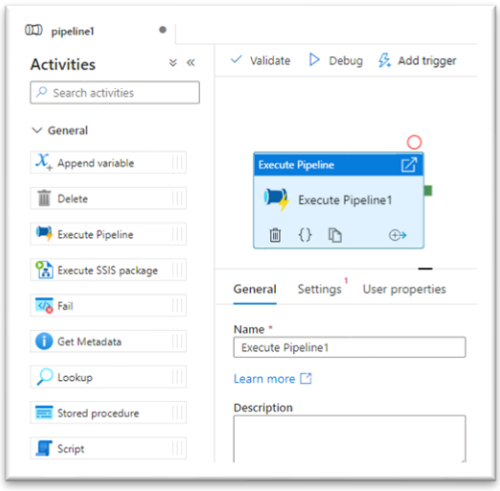
Figura 18 - Notebook1



Fonte: <https://bit.ly/3KI7bbh>. Acesso em: 20 mar. 22.

Uma forma de concentrar as execuções de mais de um pipeline criado, é a utilização da atividade *Execute Pipeline* visualizada na figura 19, pertencente ao grupo de atividades gerais. Esta atividade permite que um pipeline do ADF (pipeline mestre) invoque a execução de outro pipeline (pipeline invocado).

Figura 19 - Pipeline executado



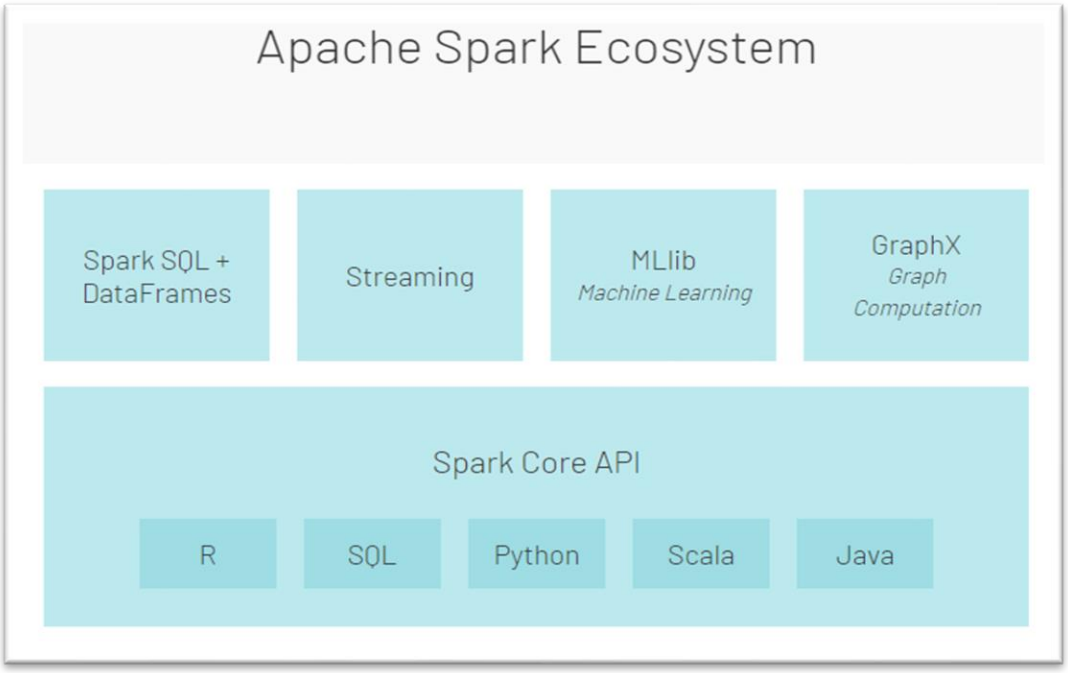
Fonte: <https://bit.ly/3xElQkj>. Acesso em: 13 mar. 2022.

2.2.4 Azure Databricks

O Azure Databricks é um serviço em nuvem para Big Data e Machine Learning, desenvolvido a partir do trabalho da equipe que iniciou o projeto Apache Spark e a equipe Microsoft, para proporcionar aos engenheiros e cientistas um ambiente otimizado, seguro e simplificado para desenvolvimento de aplicações em nível empresarial (MICROSOFT, s.d.).

Apache Spark é uma ferramenta para processamento de Big Data, que fornece suporte para APIs de alto nível em Java, Scala, Python e R, com ferramentas para processamento de dados estruturados como o Spark SQL, MLlib para aprendizado de máquina, GraphX para processamento de gráfico e Structured Streaming para computação incremental e processamento de fluxos em tempo real (APACHE SPARK, s.d.)

Figura 20 - Ecossistema Apache Spark



Fonte: <https://databricks.com/spark/about>. Acesso em: 13 mar. 2022.

Essa ferramenta aberta e seu ecossistema representado na figura 20 logo se tornou amplamente utilizada por potências da internet, como a Netflix, com uma grande comunidade de mais de 1000 colaboradores de mais de 250 organizações. O Spark arquiteturalmente foi projetado

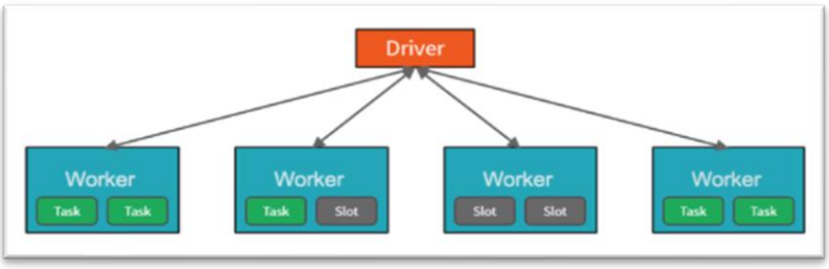
para exploração da computação em memória e otimizações que o tornaram 100x mais rápido que o Hadoop para o processamento de Big Data (DATABRICKS, s.d.).

A Databricks foi fundada em 2013 pela equipe que iniciou o projeto Apache Spark, Delta Lake e MLflow; sendo uma plataforma multi-cloud, fornecendo uma nova arquitetura que combina o melhor do data warehouses e do data lakes: data lakehouse. O Databricks fornece uma plataforma aberta, unificada e simples para o trabalho com dados, desde análises SQL até o aprendizado de máquina (DATABRICKS, s.d.).

Portanto, na plataforma da Microsoft Azure temos o Azure Databricks como um mecanismo de computação que une todas essas otimizações da plataforma Databricks, tendo integração com outros serviços da Azure, como o ADF.

Para finalizar o entendimento sobre esse serviço, temos a visão geral em alto nível da arquitetura envolvida no mesmo. O Azure Databricks inicia e gerencia clusters Apache Spark, que são grupos de computadores que trabalham como uma única unidade para lidar com as execuções dos comandos presentes nos notebooks executados. Os clusters trabalham com uma arquitetura do tipo mestre-trabalhador visualizada na figura 21, onde o nó mestre (driver) distribui o trabalho/processamento de dados para os workers, o paralelizando, para melhorar o desempenho (MICROSOFT, s.d.).

Figura 21 - Driver



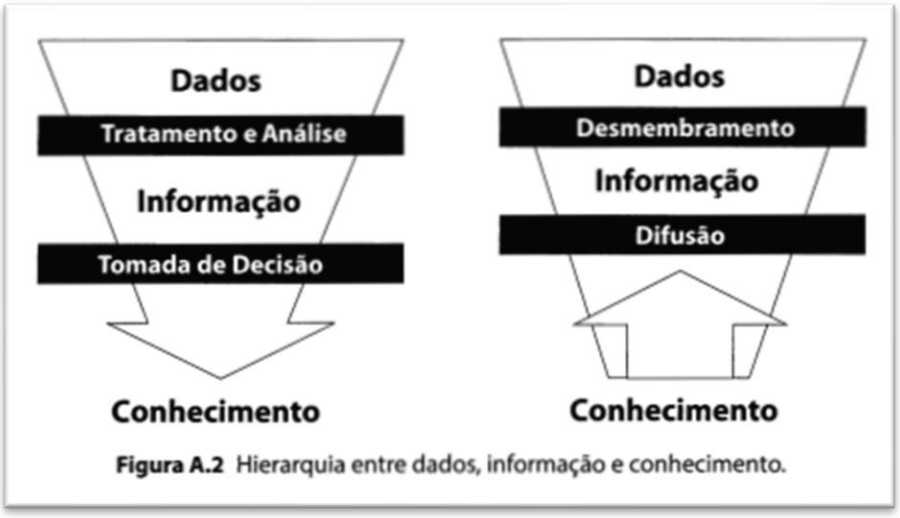
Fonte: <https://bit.ly/3xm8adp>. Acesso em: 20 mar. 2022.

2.3 Análise de Dados

Como comentado na seção Big Data, a humanidade presencia atualmente um volume exacerbado de dados, fruto do crescimento da capacidade computacional, monitoramento de fenômenos e surgimento das mídias sociais. Esses dados com variabilidade e complexidade, são produzidos em alta velocidade e disponibilizados para tratamento e análise, visando o auxílio na tomada de decisões em ambientes que se tornam cada vez mais competitivos.

Dados produzidos, ao serem tratados e analisados, tornam-se informações que podem ser reconhecidas e aplicadas para tomadas de decisão e planejamentos estratégicos. Esse é o conceito de data driven, processos organizacionais que além de baseados em experiências, são baseados na orientação a dados. Na figura 22 abaixo temos a hierarquia entre dados, informações e conhecimento (FÁVERO e BELFIORE, 2017):

Figura 22 - Dados



(FÁVERO e BELFIORE, 2017, p.14)

A área de análise de dados (Analytics) compreende estratégias e tecnologias aplicadas para processar um fluxo de dados de forma a oferecer insumos para que uma organização tenha a capacidade de tomar decisões e corrigir cursos com rapidez, acompanhando tendências e novas oportunidades de negócio. Com a análise de dados é possível ter uma percepção mais profunda e

precisa, analisando padrões e correlações, sustentando a vantagem competitiva (VIANNA; DUTRA, 2016).

As tarefas de análise são divididas nas seguintes categorias:

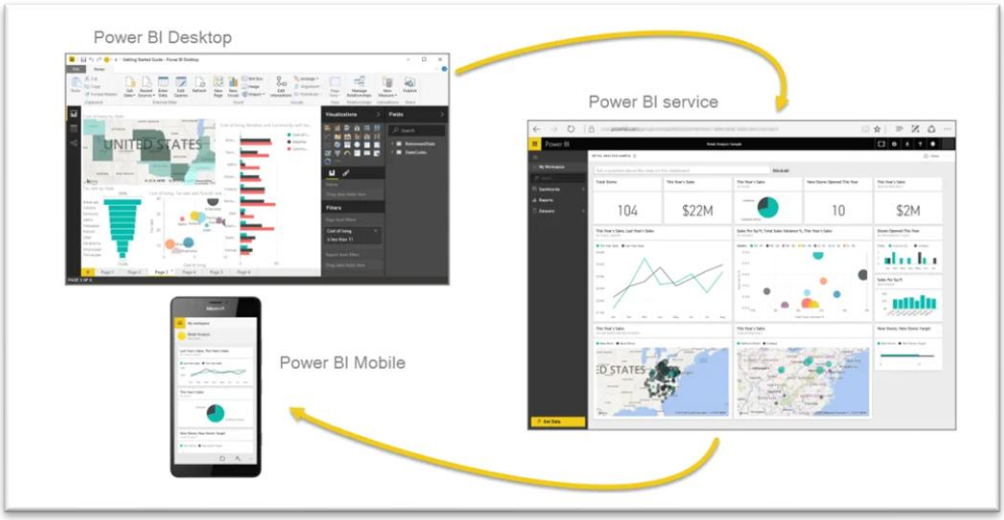
- Análise descritiva: utilizada para analisar acontecimentos com base em dados históricos. Suas técnicas resumem grandes conjuntos de dados para descrever resultados obtidos para as partes interessadas. Exemplos de análise descrita são indicadores chave de desempenho (KPIs), que ajudam a rastrear o sucesso ou fracasso de objetivos; e a geração de relatórios para fornecer a visão dos dados financeiros de uma organização.
- Análise de diagnóstico: essa análise complementa a análise descritiva para a descoberta da causa de determinados eventos. Normalmente é composta por três etapas: identificação de anomalias, que podem ser valores inesperados para uma métrica; coleta de dados relacionados a essas anomalias; e então aplicação de técnicas estatísticas para descobrir os relacionamentos e tendências que expliquem essas anomalias.
- Análise preditiva: análise voltada para auxílio no entendimento e respostas sobre acontecimentos futuros. Com base em dados históricos, identificam-se tendências e a possibilidade de acontecerem novamente. As técnicas de análise preditiva incluem estatísticas e aprendizado de máquina, como regressão, redes neurais e árvores de decisão.
- Análise prescritiva: auxiliam na indicação de ações que devem ser tomadas para atingir uma meta, permitindo que as organizações tomem decisões orientadas por dados. Essa análise depende das estratégias de aprendizado de máquina que encontram padrões nos conjuntos de dados e estimam a probabilidade de resultados diferentes diante dos eventos históricos.
- Análise cognitiva: análise para auxiliar no conhecimento dos cenários que podem ocorrer diante de diferentes circunstâncias, para que a organização saiba lidar com essas diferentes situações. Hipóteses não estruturadas (inferências) coletadas de diferentes fontes de conhecimento com diversos graus de confiança e identificação de padrões, são utilizadas para derivar conclusões que são adicionadas à base de conhecimento, que passa por futuras inferências (ciclo de autoaprendizagem). Normalmente é uma análise que envolve conceitos de processamento de linguagem (MICROSOFT, s.d.)

2.3.1 Microsoft Power BI

O Microsoft Power BI é uma coleção de serviços de software que permite a conexão a fontes de dados de diversas origens, possibilitando de forma simples a criação de insights de forma personalizada e análises em tempo real, sendo, portanto, uma ferramenta que pode servir como mecanismo auxiliar para as tomadas de decisões comentadas na seção anterior.

Como visualizado na figura 23, o Power BI possui uma versão de aplicativo local para o Windows, chamado Power BI Desktop; também possui a versão online, como um serviço SaaS; e por fim, existe a versão para aplicativos móveis. Essas opções permitem a criação e compartilhamento de insights de negócios de maneira eficiente e abrangente (MICROSOFT, s.d.).

Figura 23 - Power BI

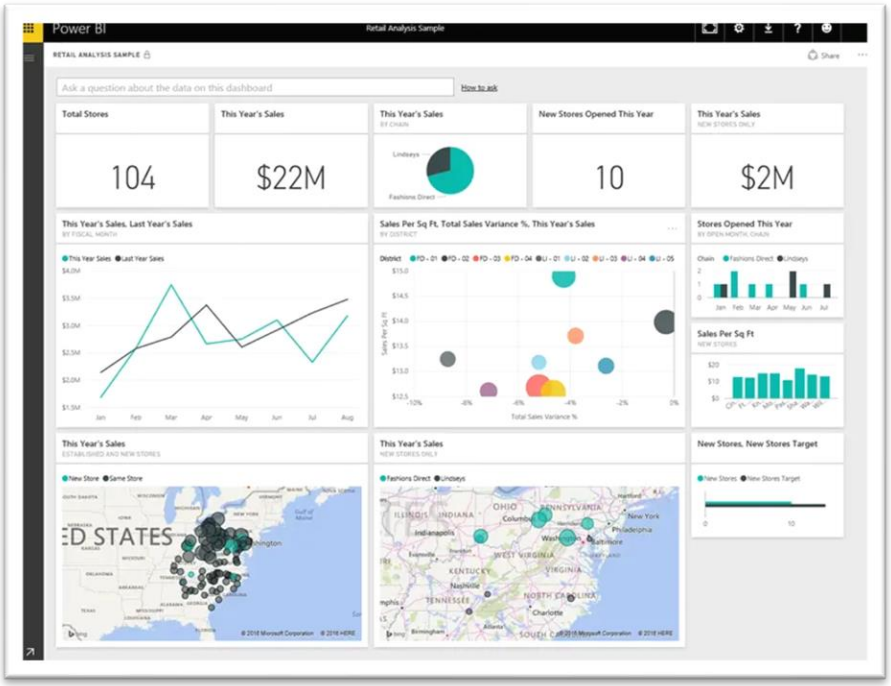


Fonte: <https://bit.ly/37ayPiE>. Acesso em: 24 mar. 2022.

Todo o desenvolvimento realizado no Microsoft Power BI é construído a partir de alguns blocos de construção básicos:

- Visualizations (visualizações): são representações visuais dos dados para que seja possível fornecer um contexto e insights que seriam difíceis de compreender e visualizar a partir de tabelas. O Power BI tem diversos tipos de representação como pode ser observado na figura 24, desde representações mais simples, como gráficos de barra ou um número significativo, até mapas codificados por cores, e novas representações continuam sendo adicionadas ao software (MICROSOFT,s.d.).

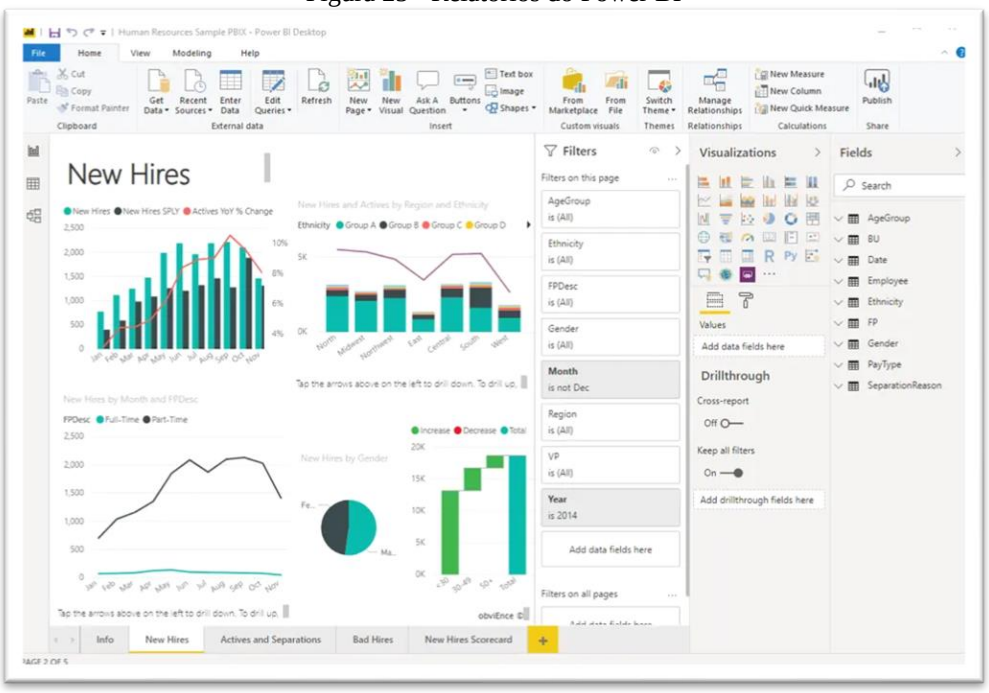
Figura 24 - Visualização do Power BI



Fonte: <https://bit.ly/3rgeKyc>. Acesso em: 22 mar. 2022.

- Datasets (conjuntos de dados): são conjuntos de dados utilizados pelo Power BI para criação das visualizações. Esses conjuntos de dados podem ser de uma única fonte simples ou da combinação de fontes, por exemplo uma tabela de um site, dados de um arquivo excel e resultados online de uma campanha (todos esses dados serão combinados e considerados um conjunto único). Filtragens podem ser aplicadas para criação de subconjuntos de acordo com os requisitos que serão atendidos pelas suas visualizações. O Power BI possui uma variedade de conectores de dados que permite conexão com diversos sistemas e serviços, como Microsoft SQL Server, Azure e Facebook.
- Reports (relatórios): relatórios são coleções de visualizações que estão relacionadas entre si, como apresentado na figura 25, organizadas em uma ou mais páginas, de forma a contar a sua história.

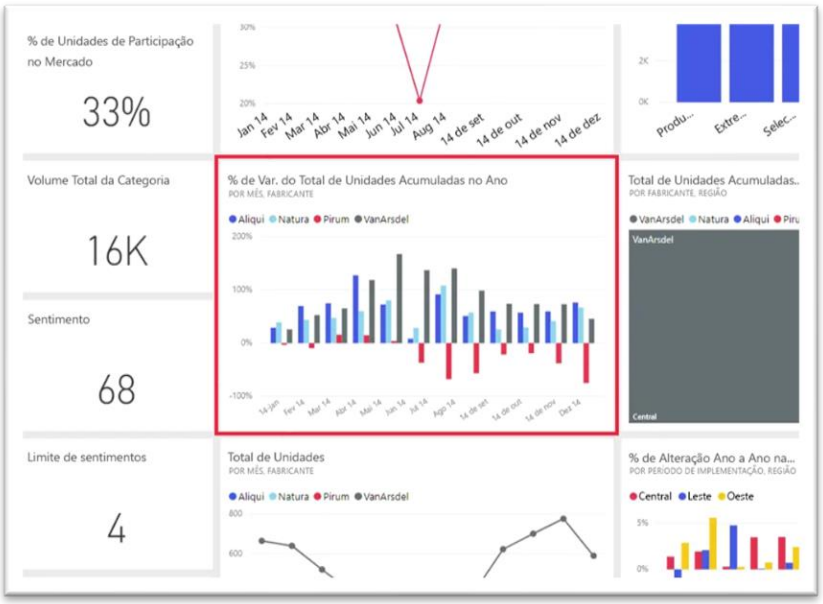
Figura 25 - Relatórios do Power BI



Fonte: <https://bit.ly/3rgeKyc>. Acesso em 26 mar. 2022.

- Dashboards (painéis): um dashboard é construído para o compartilhamento de um relatório. Um dashboard no Power BI é uma coleção de visualizações organizadas em uma única página/tela que pode ser compartilhada com usuários e grupos, possibilitando a interação dos mesmos para obter informações rápidas e relevantes sobre os dados do assunto apresentado.
- Tiles (blocos): um bloco corresponde à uma visualização individual presente em um dashboard. Esses blocos podem ser organizados da maneira que o criador acreditar que seja melhor para os usuários visualizarem o que é preciso. Eles podem ser movidos e sua altura e largura também podem ser modificadas. Na figura 26 abaixo, um dos blocos está identificado pela delimitação por um retângulo vermelho.

Figura 26 - Blocos Power BI



Fonte:<https://bit.ly/3E2ksJj>. Acesso em: 24 mar. 2022.

Conclusão

Ao final da leitura deste capítulo, é possível ter a visão sobre o termo Big Data, algumas características importantes e as mudanças que este novo cenário trouxe para o mundo de dados, desde a evolução dos modelos de dados e a evolução dos sistemas on-premise, até a popularização da computação em nuvem.

Também foram abordados conceitos importantes para entendimento no desenvolvimento de soluções com dados, desde a construção de pipelines, até questões de extração, transformação e armazenamento. Conceitos sobre a Microsoft Azure e seus serviços direcionados à dados também foram discutidos.

Portanto, este capítulo oferece a fundamentação teórica que visa retratar o trabalho realizado por engenheiros de dados com o desenvolvimento de um pipeline automatizado utilizando a computação em nuvem (Microsoft Azure). Neste trabalho é feito desde a extração de dados de fontes distintas sobre a covid-19 e indicadores de desenvolvimento globais, até a realização de transformações e integrações, a fim de disponibilizar dados que podem ser visualizados na ferramenta Power BI para a criação de insights e análise de possíveis relações.

Capítulo 3

METODOLOGIA

Este capítulo descreve com maiores detalhes o caminho utilizado para a realização da pesquisa e desenvolvimento deste trabalho, apresentando os processos utilizados para chegar ao resultado final.

3.1 Objetivo do trabalho

O presente trabalho teve como objetivo a pesquisa do estado da arte referenciando Big Data, a evolução dos modelos de dados e infraestrutura para processamento e armazenamento destes; com a finalidade de aprofundamento de conhecimento e aplicação dos conceitos na criação de um pipeline de dados utilizando serviços da nuvem para coletar, processar, armazenar, integrar e analisar dados da COVID-19 e indicadores de desenvolvimento global.

3.2 Aquisição dos dados

Para desenvolvimento prático do conceito de ELT, conjuntos de dados a respeito da COVID-19 e de indicadores de desenvolvimento foram extraídos de fontes HTTP pelo pipeline de dados criado no Capítulo 4. A seguir, apresenta-se uma descrição breve das fontes que foram utilizadas.

3.2.1 COVID-19 (WHO)

No site da Organização Mundial de Saúde (World Health Organization - WHO), é possível visualizar o painel do coronavírus da WHO com contagens diárias de casos, mortes e vacinações relatadas pelo mundo[2].

3.2.2 COVID-19 (OWD)

Our World in Data[3] contém dados cujo intuito é a análise de problemas mundiais. Um dataset sobre a COVID-19 com informações de vacinação, testes, hospitalização, mortes e outras

[2] É possível realizar o download dos dados exibidos no painel em formato de arquivo separado por vírgula (CSV) e verificar o nome de cada campo, seu tipo e descrição no endereço: <https://covid19.who.int/data.>.
[3] Site: <https://ourworldindata.org/about>. Acesso em: 27 mar. 2022.

variáveis pode ser extraído em formato CSV, XLSX e JSON acessando a página da organização no GitHub[4].

3.2.3 Indicadores (TWB)

The World Bank[5] (TWB) é uma parceria formada por cinco instituições voltadas ao trabalho de financiamento e desenvolvimento de soluções sustentáveis a fim de reduzir a pobreza e aumentar a prosperidade com desenvolvimento sustentável para países em desenvolvimento[6].

3.3 Definição da arquitetura e pipeline

Para aplicação dos conceitos de ELT, arquitetura e funcionamento de um Data Lake e processamento com finalidade de análise descritiva, um fluxo de atividades foi esquematizado após ser definida a localização origem dos dados a serem trabalhados.

Um pipeline para o processo de extração, armazenamento, transformação e análise foi então definido. Inicialmente, a arquitetura foi estabelecida, definindo a plataforma Microsoft Azure para o desenvolvimento. Na escolha da Azure foi levado em conta a popularização dos serviços em nuvem para o trabalho com dados; a interface da Azure que proporciona facilidade de interação; e a vasta documentação disponibilizada pela Microsoft de forma gratuita e online sobre os seus serviços.

Visualizando o fluxo de execução de atividades que a prática envolveria, foram definidos os componentes da solução. De forma resumida, foi escolhido o serviço Azure Data Factory para o processo de criação dos pipelines, já que o mesmo fornece uma série de atividades para manipulação dos dados, desde a extração até a integração; e também foi definido o serviço de armazenamento a ser utilizado para prática dos conceitos de armazenamento de dados que envolvem Big Data: o Data Lake Storage Gen2.

O pipeline então foi definido contendo o seguinte fluxo de atividades:

---

[4] Site: <https://github.com/owid/covid-19-data/tree/master/public/data.>. Acesso em: 28 mar. 2022.

[5] Site: <https://www.worldbank.org/en/who-we-are>. Acesso em: 22 mar. 2022.

[6] Dados abertos podem ser consultados no site do TWB, entre eles indicadores de desenvolvimento mundial acessíveis no endereço: <https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators.>. Acesso em: 26 mar. 2022.

(1) Extração dos dados da COVID-19 e indicadores de desenvolvimento disponibilizados de forma aberta na internet e carga sem aplicação de transformação na zona bruta do Data Lake;

(2) Leitura dos dados da zona bruta e padronização com seleções dos dados convenientes para a finalidade de integração da atividade posterior, realizando a carga na zona confiável do Data Lake;

(3) Leitura dos dados presentes na zona confiável e refinamento com integração dos dados da COVID-19 e indicadores de desenvolvimento de saúde, educação e renda. Dados consolidados integrados prontos para realização de análise e geração de insights são carregados na zona refinada do Data Lake;

(4) Dados da zona refinada são acessados pela ferramenta de análise, possibilitando explorações e criação de insights.

No Capítulo 4, figura 27, é possível verificar de forma visual a arquitetura e fluxo de atividades descrito anteriormente.

3.4 Avaliação final (resultados da integração)

Ao final da execução do pipeline de dados construído, que executa todo o processo ponta a ponta de extração de dados brutos, armazenamentos e transformações; os dados integrados e acessados na zona refinada indicam o sucesso da execução de todo o fluxo de atividades definido, resultando no alcance do objetivo final esperado.

Capítulo 4

PRÁTICA

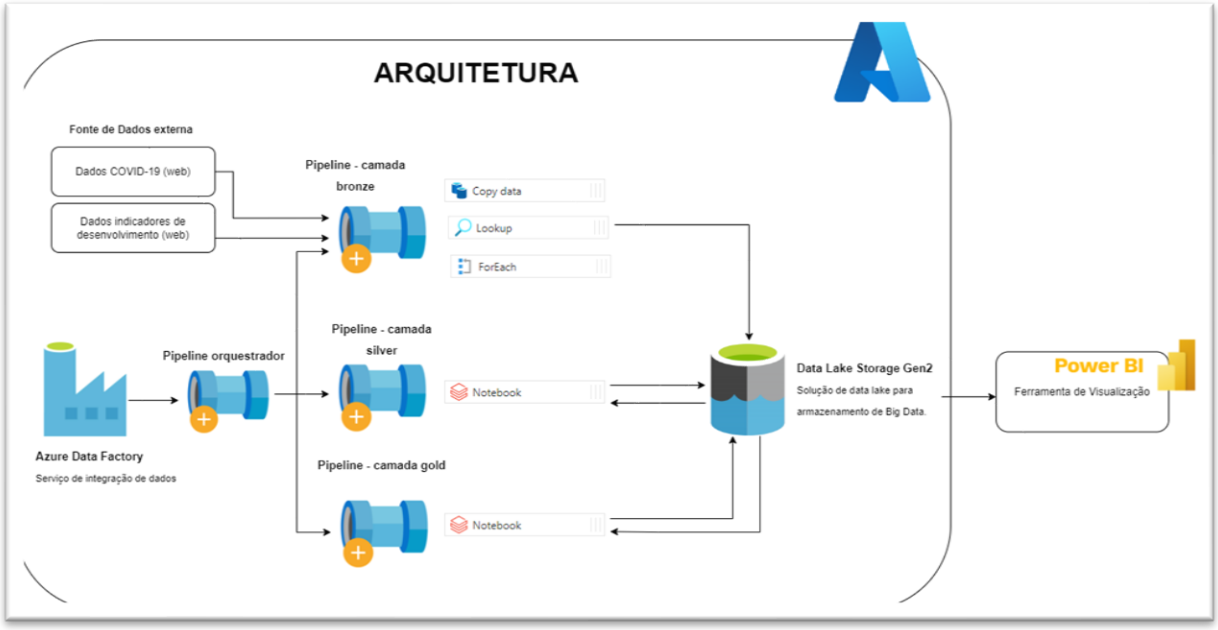
Este capítulo contém o detalhamento do trabalho prático desenvolvido na Microsoft Azure, envolvendo a construção de pipelines de dados para coleta, armazenamento, transformação e análise de dados da COVID-19 e indicadores de desenvolvimento disponibilizados de forma aberta na internet.

4.1 Fluxo de execução

A figura 27 contém todo o fluxo de trabalho que foi desenvolvido neste capítulo. É possível visualizar a utilização do serviço de integração de dados Azure Data Factory como base do desenvolvimento, onde um pipeline cross trabalha como orquestrador para execução de outros três pipelines:

- Pipeline - camada bronze: responsável pela extração dos dados externos disponibilizados em fonte HTTP e carga dos mesmos na camada landing (bronze) do Data Lake.
- Pipeline - camada silver: responsável pela leitura dos dados da camada bronze e transformação dos mesmos por meio de sintetizações e padronizações e carga na zona confiável do Data Lake (camada silver).
- Pipeline - camada gold: realiza a leitura dos dados da zona confiável e aplica transformações e integração para carga na camada gold, zona refinada do Data Lake, a qual é acessada por serviços de análise para geração de insights.

Figura 27 - Fluxo de execução da prática com os componentes utilizados



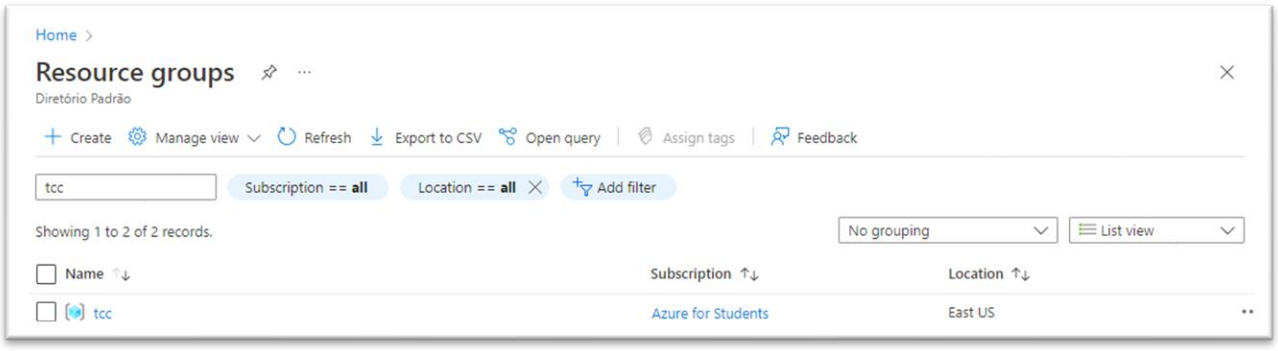
Fonte: autoria própria

4.2 Preparações iniciais do ambiente

Para dar início ao desenvolvimento deste trabalho que visa a criação de um pipeline para ingestão e tratamentos de dados com a finalidade de integração para análise, foram necessárias algumas configurações iniciais no portal da Azure para prosseguir com a ingestão dos dados.

Como discutido no capítulo anterior, um resource group na Azure se comporta como um contêiner que mantém recursos relacionados à uma solução específica. Portanto, foi criado um Resource Group **tcc** que pode ser visualizado na figura 28, que armazenará todos os recursos que são utilizados para execução deste projeto.

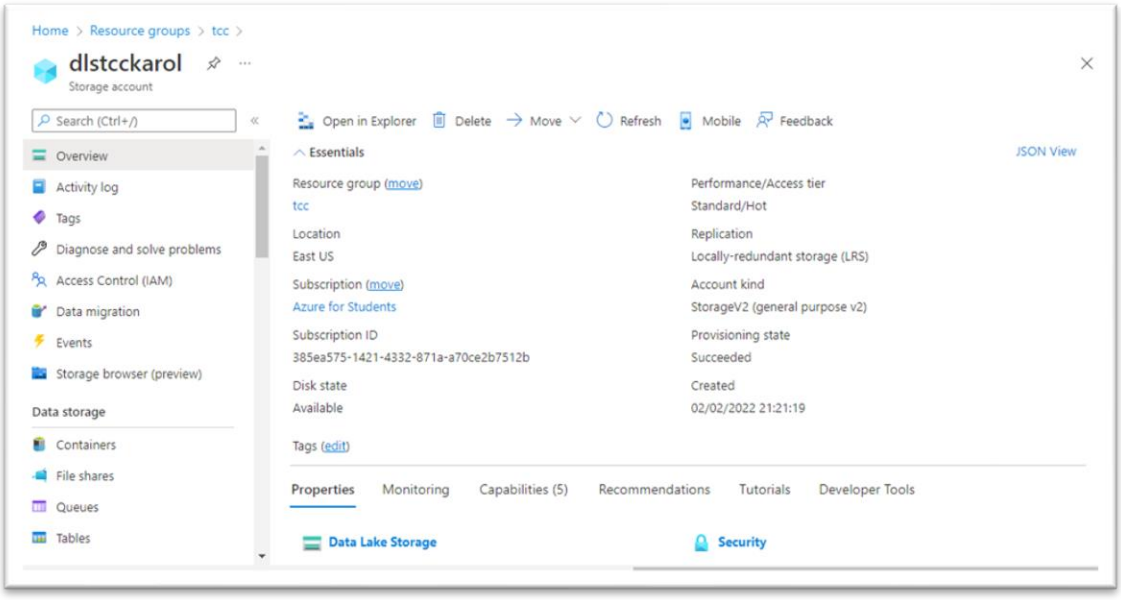
Figura 28 - Interface dos resource groups da Microsoft Azure



Fonte: autoria própria

Tendo em vista o Data Lake Storage Gen2 como recurso de armazenamento, criou-se dentro do resource group **tcc** uma storage account, serviço gerenciado pela Microsoft que fornece armazenamento em nuvem com alta disponibilidade, segurança, escalabilidade e redundância.

Figura 29 - Interface com a visão geral do Data Lake Storage

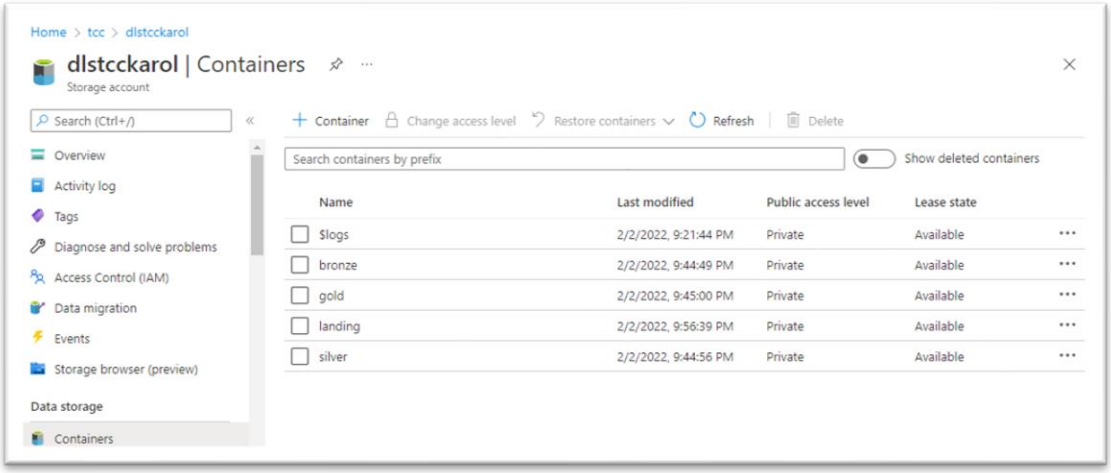


Fonte: autoria própria

Após a criação da Data Lake Storage da figura 29, gerou-se os containers, que funcionam como um diretório em um sistema de arquivos, fornecendo a possibilidade de organização de variados tipos de dados. Neste projeto, utilizou-se a criação de múltiplas camadas no Data Lake, que

conterão dados em diferentes estágios de manipulação/tratamento, referenciando-se na arquitetura exposta no Capítulo 2. Portanto, originou-se 4 containers como mostrado na Figura 30.

Figura 30 - Visualização dos containers criados no Data Lake



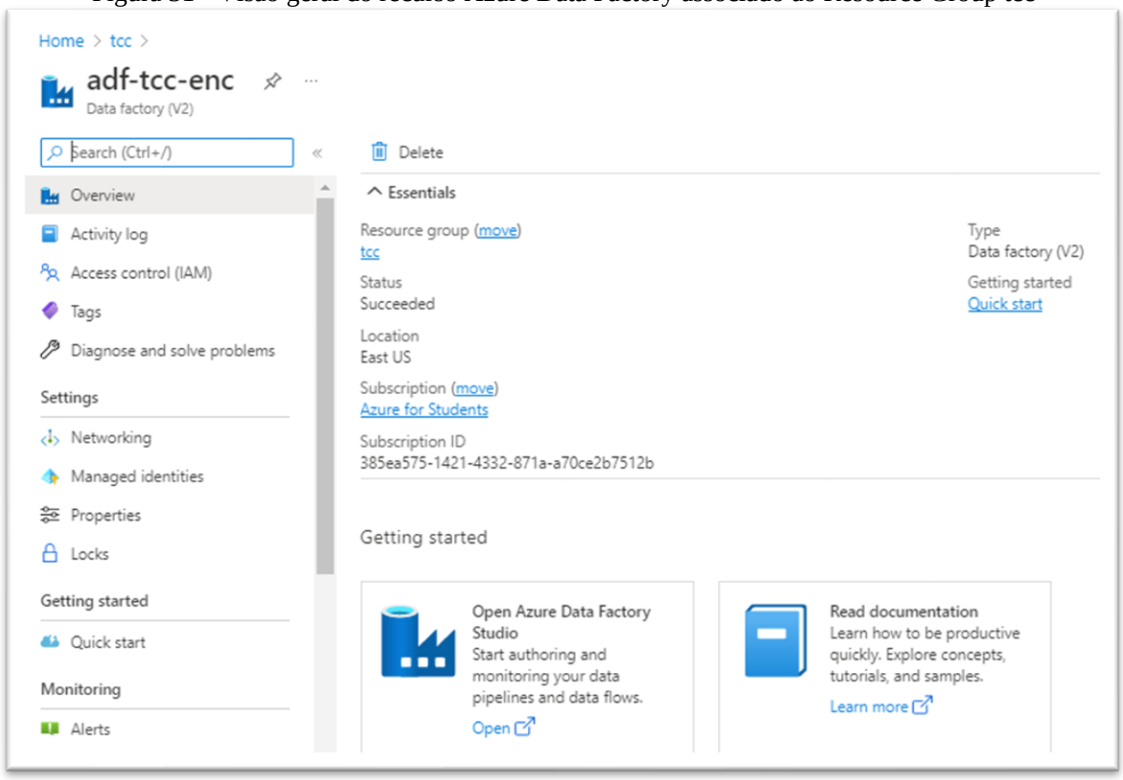
Fonte: autoria própria

O container bronze representa a camada onde os dados brutos foram armazenados, ou seja, nela são armazenados os dados do COVID e os indicadores de desenvolvimento da forma em que são encontrados na origem (web). O silver representa a camada intermediária que armazena os dados da camada bronze que passaram por um tratamento e transformações iniciais para atender a demanda específica posterior de integração das informações para análise. Por fim, a camada gold corresponde ao local de armazenamento da integração das informações com valor agregado, prontas para serem analisadas.

Também foi criado um contêiner landing para armazenamento de arquivos gerais, como arquivos de parâmetros.

Na figura 31, observa-se a instanciação do recurso ADF associado ao resource group tcc, já que todo o processo de ingestão até a integração de dados e a disponibilização na camada de valor gold é orquestrada por este recurso.

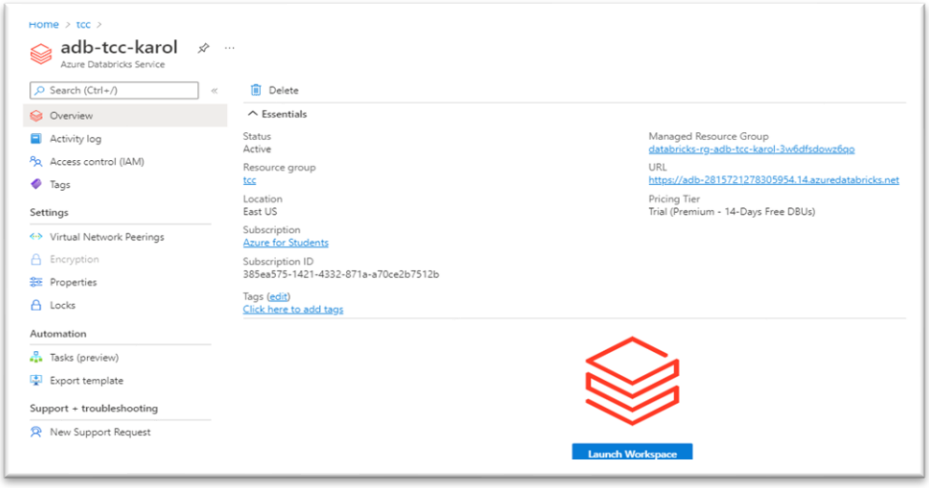
Figura 31 - Visão geral do recurso Azure Data Factory associado ao Resource Group tcc



Fonte: autoria própria

Criou-se também o serviço Azure Databricks visualizado na figura 32, onde iremos utilizar notebooks deste serviço para o processamento dos dados.

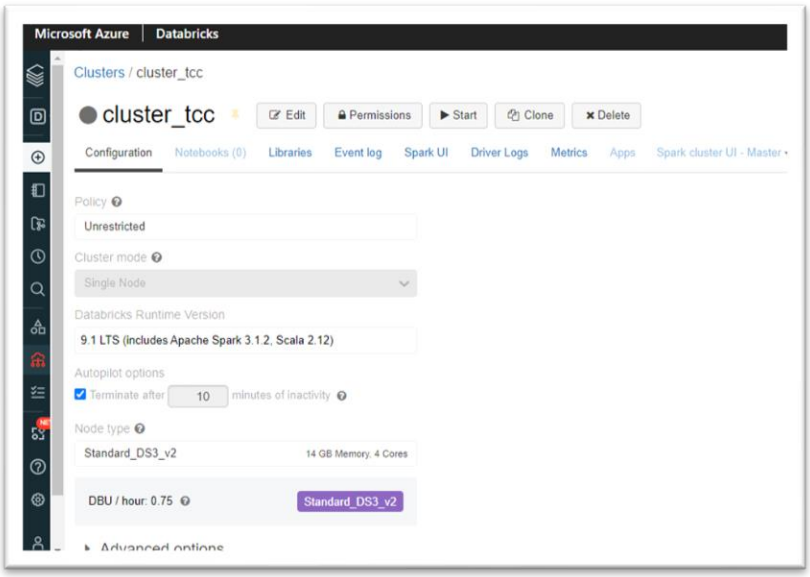
Figura 32 - Visão geral do recurso Azure Databricks associado ao Resource Group tcc



Fonte: autoria própria

É necessária a configuração deste serviço iniciando a área de trabalho do mesmo e realizando a criação do cluster na aba Compute. A figura 33 exibe o cluster criado para execução deste projeto.

Figura 33 - Área de trabalho do Azure Databricks com visualização do cluster criado.



Fonte: autoria própria

Após realizar todas essas criações e configurações iniciais, pôde-se iniciar o desenvolvimento do pipeline no ADF.

4.3 Ingestão dos dados primários

A primeira etapa do pipeline de dados consistiu no acesso às fontes dos dados que foram trabalhados e a extração dos mesmos, prosseguindo com a ingestão na camada bronze (zona bruta do data lake) da conta de armazenamento criada e configurada no início deste capítulo.

Abrindo o Azure Data Factory Studio, pôde-se iniciar a criação de 3 pipelines com a finalidade de executar a tarefa de acesso aos dados na fonte e exportação para ingestão na camada bronze do Data Lake. Cada pipeline de cópia teve sua especificidade devido ao formato dos dados encontrados na origem, então iremos discorrer sobre tais particularidades a seguir.

4.3.1 Cópia dos dados COVID (WHO)

No site da Organização Mundial de Saúde, como foi apresentado no capítulo 3, pode-se realizar o download de 4 bases de dados que estão disponíveis em um formato estruturado, arquivos CSV separados por ponto e vírgula. Tínhamos, portanto, uma URL completa para cada uma dessas bases de dados e um tipo de compactação para duas dessas bases:

- (1) Casos diários e óbitos por data relatados à WHO:
  URL para download: <https://covid19.who.int/WHO-COVID-19-global-data.csv>.
  Tipo de compactação: nenhuma

- (2) Últimas contagens relatadas de casos e mortes
  URL para download: <https://covid19.who.int/WHO-COVID-19-global-table-data.csv>.
  Tipo de compactação: nenhuma

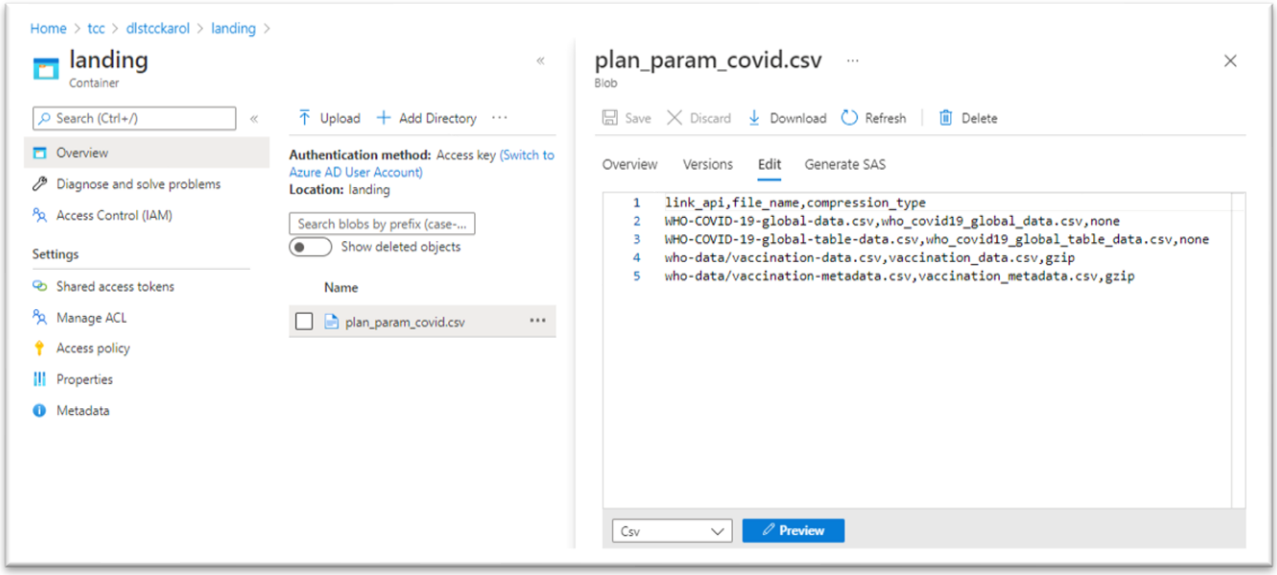
- (3) Dados de vacinação
  URL para download: <https://covid19.who.int/who-data/vaccination-data.csv>.
  Tipo de compactação: gzip

- (4) Metadados de vacinação
  URL para download:<https://covid19.who.int/who-data/vaccination-metadata.csv>.
  Tipo de compactação: gzip

Tínhamos, portanto, quatro URLs para fazer o acesso e o download dos dados, além de termos também um tipo de compactação ou não.

O pipeline para cópia dos dados dessas 4 URLs teve como processamento inicial a atividade Lookup do tipo General, que realizou a leitura de um arquivo .CSV de parâmetros demonstrado na figura 34 na camada landing do Data Lake.

Figura 34 - Arquivo CSV de parâmetros inserido na camada landing do Data Lake



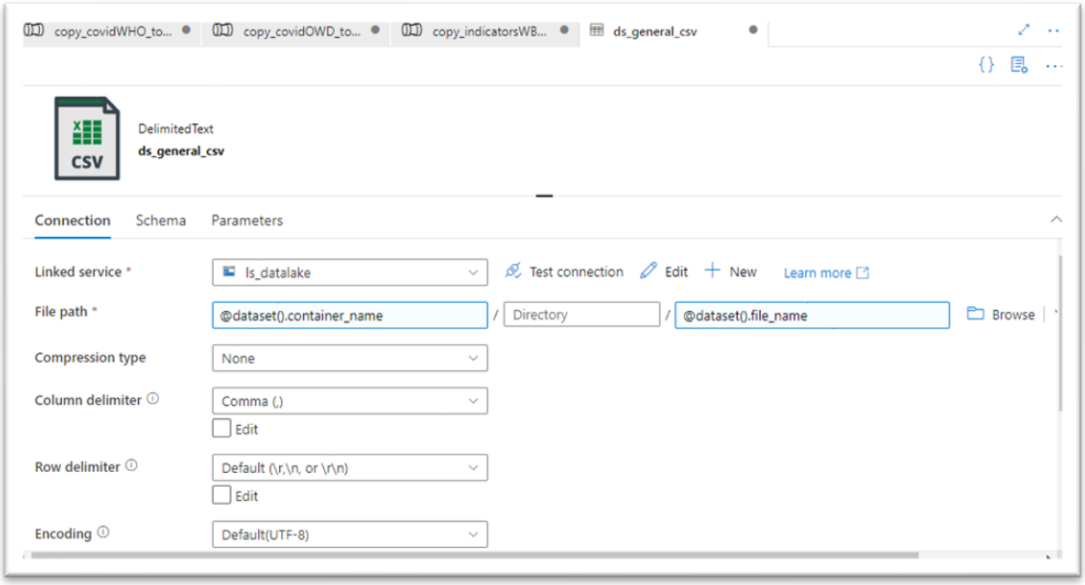
Fonte: autoria própria

Nele temos associado ao cabeçalho link_api, as URLs para download dos dados da OMS; em file_name, tem-se a nomeação padronizada com letras minúsculas que desejamos que os arquivos sejam salvos após a extração; e ao cabeçalho compression type temos associado o tipo de compressão do arquivo.

Criou-se um dataset genérico, configurando-o com armazenamento de dados Azure Data Lake Storage Gen2 e tipo de dado no formato DelimitedText. Foi associado a este dataset um linked service para a conta de armazenamento Azure Data Lake Storage Gen2, Data Lake criado anteriormente.

Neste dataset tivemos também a criação de dois parâmetros como exibido na figura 35, um para armazenar o nome do container e outro o nome do arquivo. Estes parâmetros são usados dinamicamente para identificar a localização do arquivo (File path).

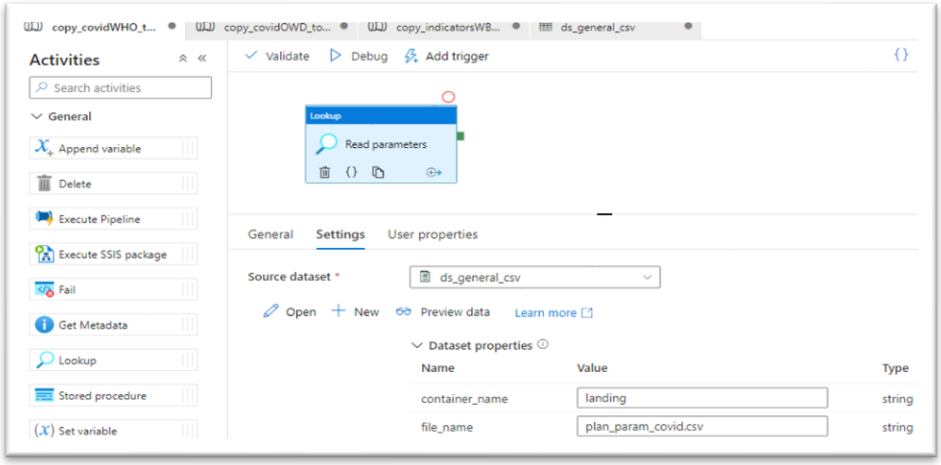
Figura 35 - Configurações do dataset genérico do tipo texto delimitado que se conecta ao Data Lake



Fonte: autoria própria

Então, a atividade de Lookup apresentada na figura 36 para leitura do CSV com os parâmetros, é configurada selecionando o dataset geral que foi criado acima como origem e passando as propriedades - nome do container (landing) e nome do arquivo de parâmetros (plan_param_covid.csv).

Figura 36 - Atividade Lookup para leitura de arquivo de parâmetros



Fonte: autoria própria

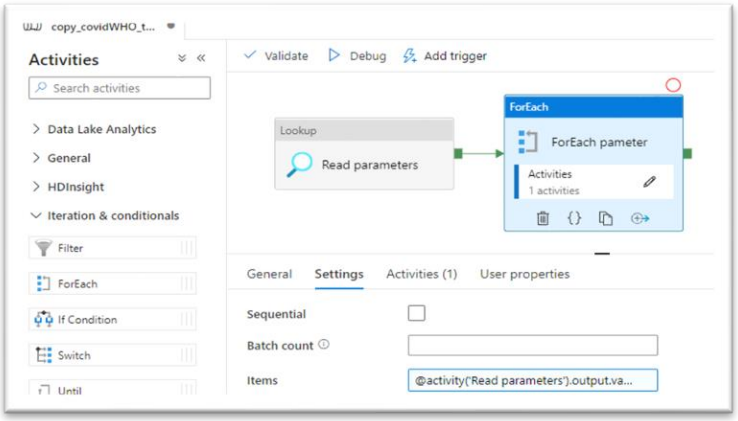
É possível realizar um preview dos dados como mostra a figura 37, para certificação de que a leitura está ocorrendo no arquivo correto de parâmetros.

Figura 37 - Visualização dos parâmetros lidos pela atividade Lookup

Fonte: autoria própria

Após a configuração do Lookup que terá como saída a leitura dos parâmetros, foi realizada a configuração da atividade ForEach do tipo Iteration & conditionals. Esta atividade visualizada na figura 38 recebeu dinamicamente a saída da atividade Lookup (parâmetros do CSV já descritos anteriormente) para execução da atividade Copy data do tipo Move & transform criada dentro do loop.

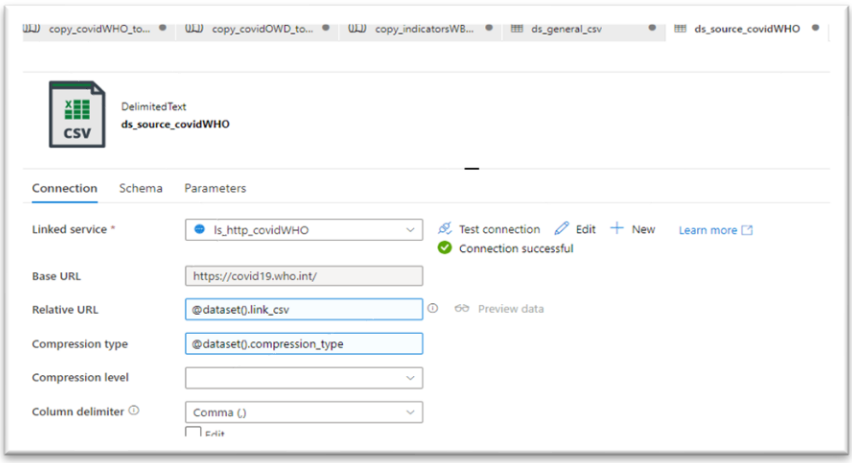
Figura 38 - Atividade de iteração ForEach que recebe como entrada os parâmetros lidos pela atividade Lookup.



Fonte: autoria própria

Para as configurações da atividade Copy data, um dataset de origem HTTP foi criado, com o tipo de dado no formato DelimitedText (já que os arquivos na origem são arquivos .CSV), associando um linked service HTTP que realiza a conexão com a URL base do site da OMS onde iremos extrair os dados.

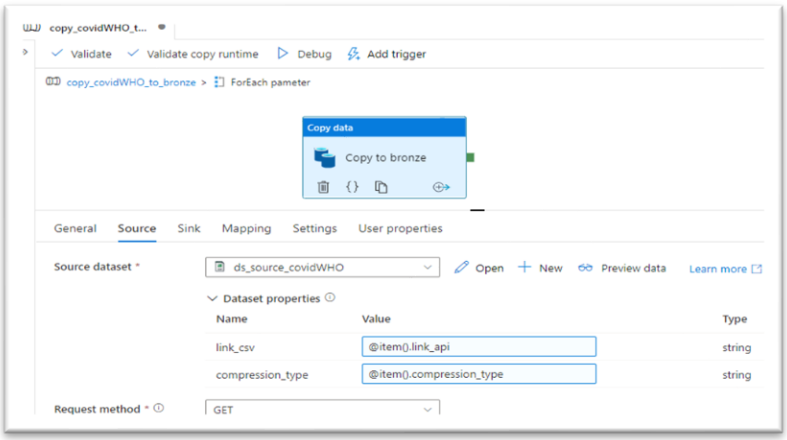
Figura 39 - Dataset de origem HTTP para carga de dados do tipo texto delimitado



Fonte: autoria própria

Com o dataset de origem criado e configurado, pôde-se realizar a configuração da aba origem da atividade de cópia exibida na figura 40, selecionando o dataset e passando dinamicamente os parâmetros link_csv e compression_type lidos na iteração do loop (parâmetros recebidos pelo ForEach como saída do Lookup)

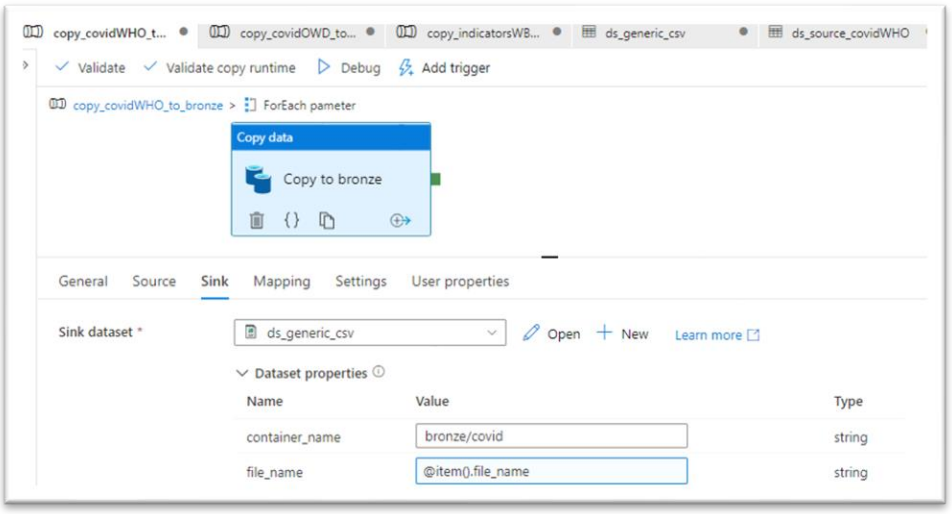
Figura 40 - Configuração da origem da atividade de cópia presente na iteração ForEach



Fonte: autoria própria

A figura 41 exibe por fim a configuração do destino dos arquivos lidos, ou seja, a camada bronze do Data Lake. Podemos utilizar o dataset genérico ds_generic_csv já criado com o linked service para o Data Lake, restando só passar o nome do container e o nome do arquivo (parâmetro dinâmico file_name - item corrente do loop).

Figura 41 - Configuração do destino da atividade de cópia presente na iteração ForEach



Fonte: autoria própria

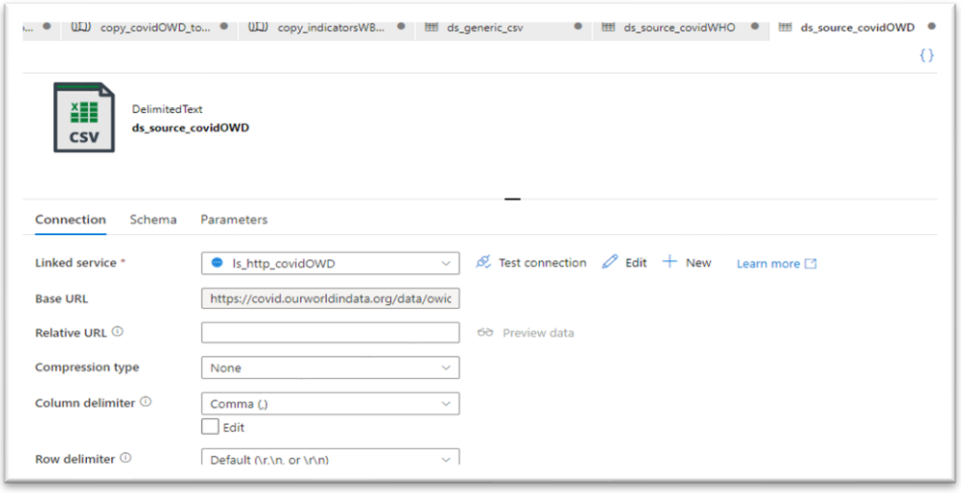
4.3.2 Cópia dos dados COVID (OWD)

No git do Our World in Data, como foi apresentado no capítulo 3, pode-se realizar o download de uma base de dados sobre COVID-19 disponível em um formato estruturado, arquivo CSV separado por vírgula. Temos, portanto, uma URL completa para um arquivo sem compressão:

- URL para download: <https://covid.ourworldindata.org/data/owid-covid-data.csv>.

Para a extração destes dados, precisamos apenas da criação da atividade Copy data do tipo move & transform. A figura 42 mostra o dataset de origem HTTP criado, com o tipo de dado no formato DelimitedText, associando com um linked service HTTP que realiza a conexão com a URL base do site da OWD onde iremos extrair os dados.

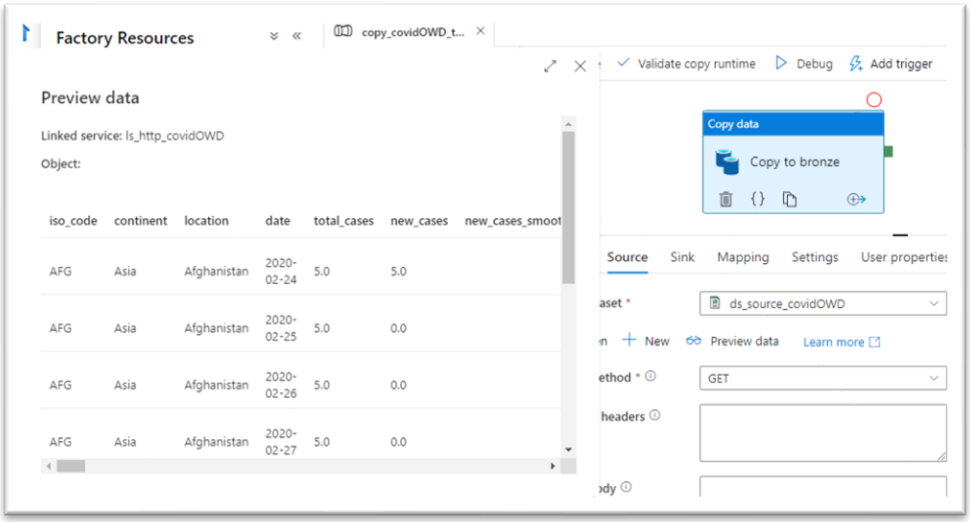
Figura 42 - Dataset de origem HTTP criado para carga dos dados de Covid do tipo CSV da Our World in Data



Fonte: autoria própria

Com o dataset de origem criado e configurado, pôde-se realizar a configuração da aba origem da atividade de cópia exibida na figura 43, selecionando o dataset ds_source_covidOWD . É possível fazer um teste para verificar a consistência da leitura dos dados na origem web, basta selecionar a opção Preview data na aba source:

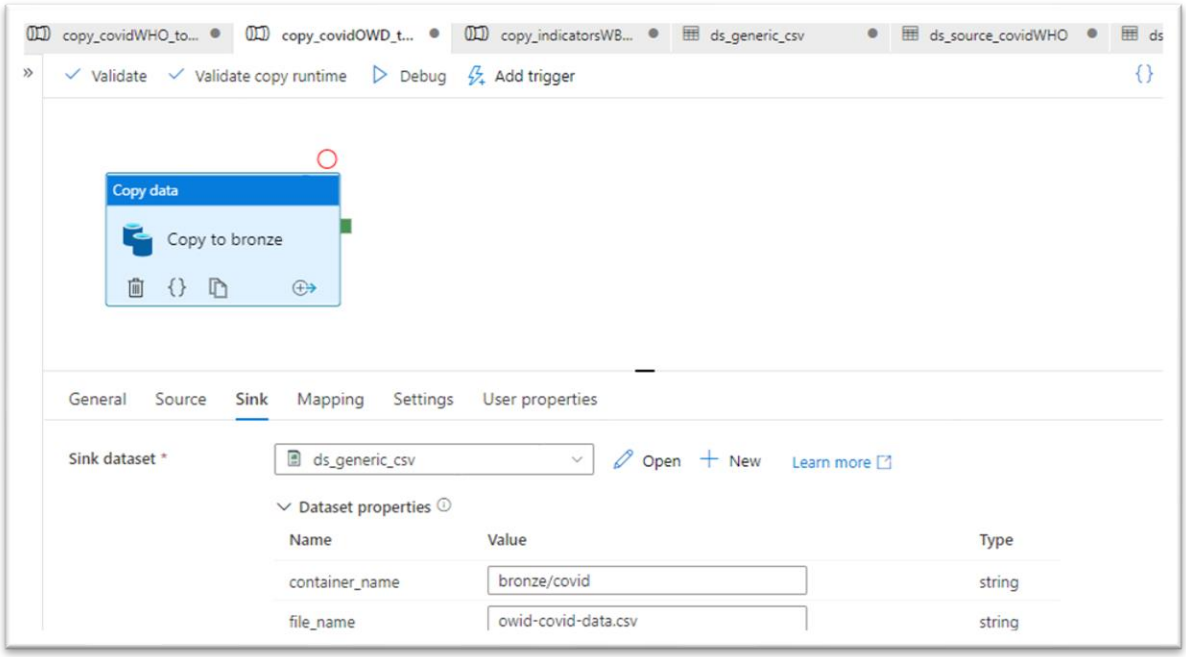
Figura 43 - Pré visualização da leitura da atividade de cópia



Fonte: autoria própria

Por fim, foi realizada a configuração do destino dos arquivos lidos, ou seja, a camada bronze do Data Lake. Podemos utilizar o dataset genérico ds_generic_csv já criado com o linked service para o Data Lake, já que o nosso dataset de origem também é do tipo DelimitedText e armazenaremos todos os dados deste projeto na mesma conta de armazenamento. Configurou-se ainda na aba de destino como visualizado na figura 44, o container_name (bronze/covid) e o file_name com o nome que desejamos que o arquivo seja salvo.

Figura 44 - Configuração do destino da atividade de cópia para carga dos dados



Fonte: autoria própria

4.3.3 Cópia dos dados Indicadores (TWB)

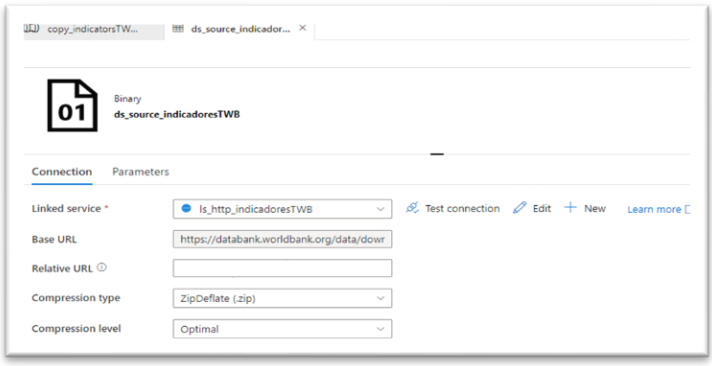
No site do The World Bank, como foi apresentado no capítulo 3, pode-se realizar o download de um arquivo compactado .ZIP que contém arquivos CSV com dados sobre indicadores de desenvolvimento globais. Temos, portanto, uma URL completa para um arquivo compactado:

- URL para download: <https://databank.worldbank.org/data/download/WDI_csv.zip>.
- Tipo de compactação: ZIP

Para a extração destes dados, precisamos apenas da criação da atividade Copy data do tipo Move & transform. Assim como realizado anteriormente, foi necessária a criação de um dataset de

origem HTTP, mas com a particularidade observada na figura 45, na seleção do tipo de dado por conta da compactação ZIP: Binary. Foi associado um linked service HTTP que realiza a conexão com a URL de download do TWD.

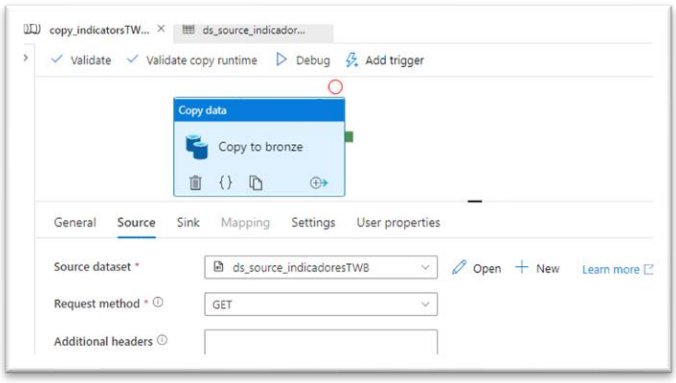
Figura 45 - Dataset de origem HTTP do tipo binário carga dos arquivos compactados dos indicadores



Fonte: autoria própria

Com o dataset de origem criado e configurado, pôde-se configurar a aba Source como a figura 46, selecionando o dataset de origem.

Figura 46 - Configuração da origem dos dados de indicadores para a atividade de cópia

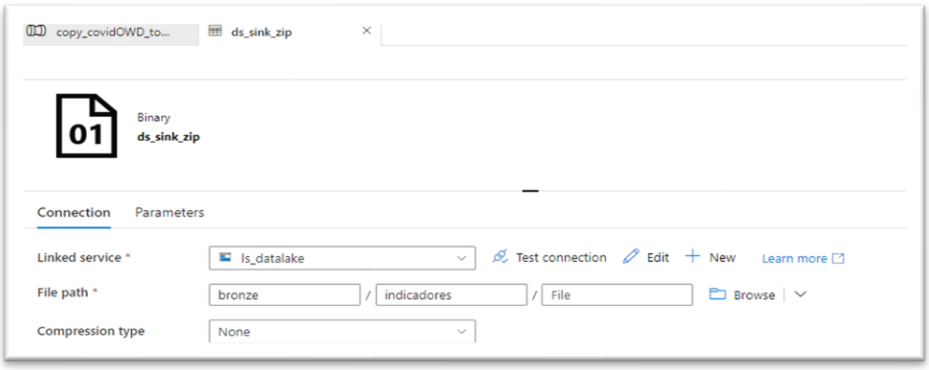


Fonte: autoria própria

Ao utilizar em uma atividade de cópia um dataset de origem definido como conjunto de dados binários, o destino também deve ser um dataset binário, essa informação pode ser encontrada na documentação do Data Factory (MICROSOFT, 2022).

Sendo assim, foi criado o dataset de destino ds_sink_zip mostrado na figura 47, configurado com o linked service já criado para o Data Lake e com a passagem do diretório onde os dados devem ser armazenados.

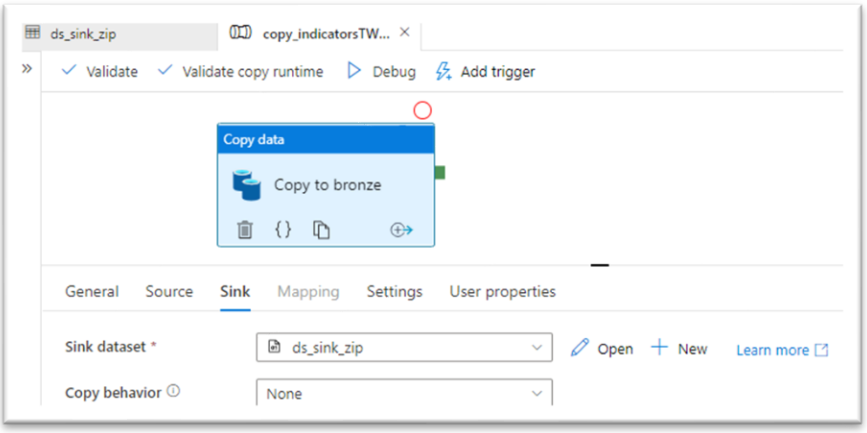
Figura 47 - Dataset de destino para carga dos dados dos indicadores já descompactados



Fonte: autoria própria

Configurou-se então a aba destino da atividade de cópia, como pode ser visualizado na figura 48, selecionando o dataset de destino criado anteriormente.

Figura 48 - Configuração do destino dos dados de indicadores para a atividade de cópia



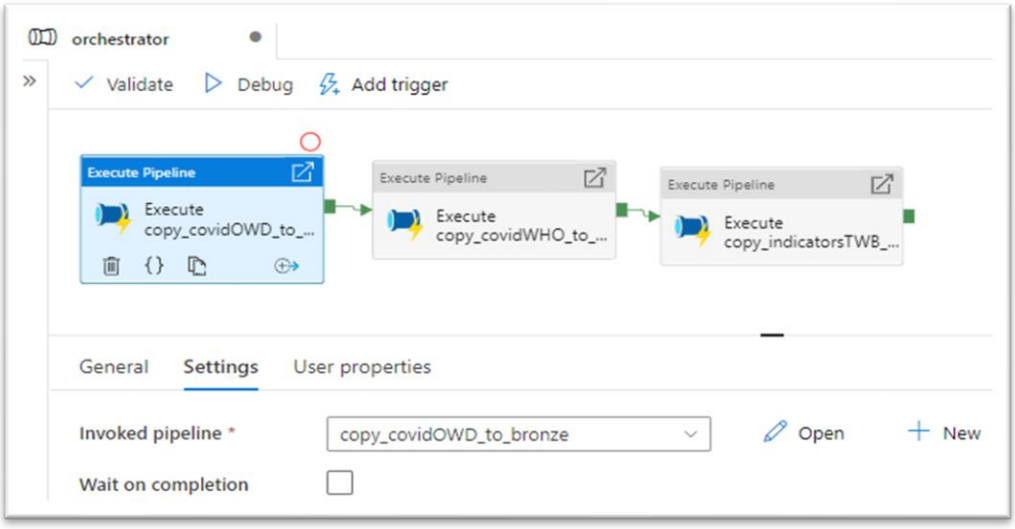
Fonte: autoria própria

4.3.3 Orquestrador e resultado da extração

Com as operações de extração de dados construídas nos pipelines de cópia, criou-se o pipeline orquestrador da figura 49, com a finalidade de automatizar esse fluxo de trabalho, sem a

necessidade de execução individual de cada um dos pipelines criados anteriormente. Este orquestrador foi formado por atividades Execute Pipeline do tipo General e em cada uma delas foi selecionado o pipeline invocado para execução e desmarcada a opção de espera de conclusão, já que as atividades de extração não possuem dependências entre si.

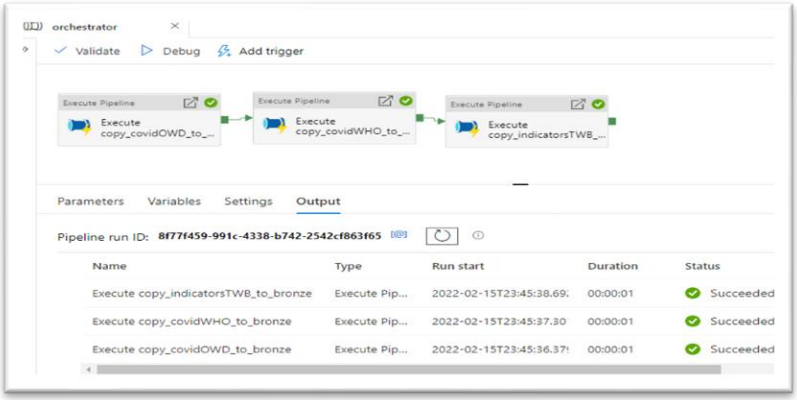
Figura 49 - Orquestrador de pipelines para automatização das execuções



Fonte: autoria própria

Ao executar o orquestrador na opção debug, conseguimos visualizar o sucesso da execução e a duração de cada extração realizada. Basta visualizar a aba output como mostra a figura 50.

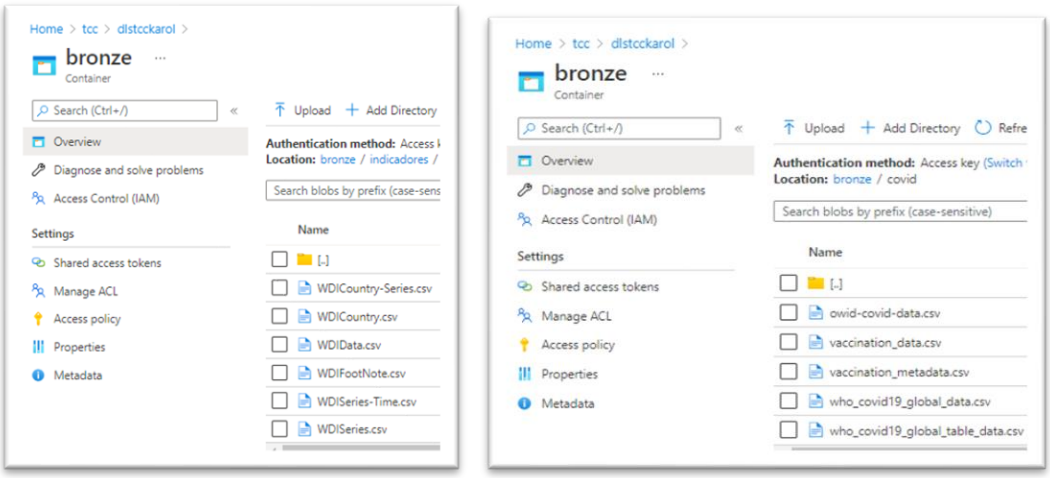
Figura 50 - Execução com sucesso do orquestrador de pipelines



Fonte: autoria própria

Consultando a camada bronze exibida na figura 51 podemos visualizar que a ingestão dos arquivos ocorreu como esperado.

Figura 51 - Visualização dos arquivos ingeridos na camada bronze do Data Lake



Fonte: autoria própria

4.3 Análise inicial e transformação

A segunda etapa da prática consistiu em uma breve visualização dos dados para identificar algumas transformações iniciais que formam a segunda parte do pipeline. Visando a simplicidade dessa prática que possui maior foco na construção e orquestração de um pipeline de dados, as transformações nesta etapa foram realizadas para consolidar tabelas que serão utilizadas nas etapas posteriores de integração e análise descritiva.

Essa etapa consiste em um processo intermediário de transformação dos dados que serão armazenados na zona confiável do Data Lake, abordada na arquitetura do mesmo no Capítulo 2. Esta camada que aqui nomeamos como camada silver guardará dados limpos e padronizados de acordo com a nossa demanda posterior de integração para análise.

No workspace do databricks criado e configurado como mencionado na seção 4.2 deste capítulo, foi criado um notebook para o processo de transformação dos dados de indicadores de desenvolvimento social. Todos os dados dos indicadores estão presentes no arquivo WDIData.csv que foi descompactado na atividade de cópia da seção anterior. Por meio do notebook foi realizado

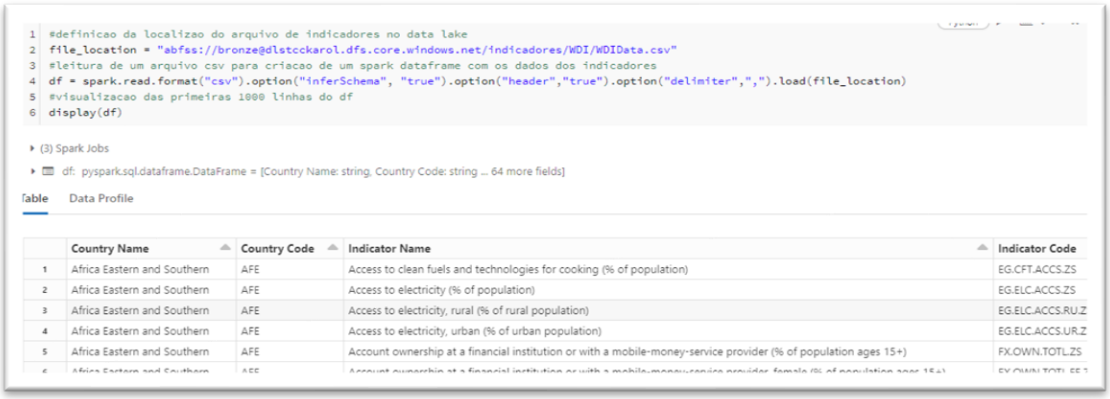
o acesso autenticado à conta do Data Lake onde o arquivo foi armazenado, como segue na figura 52.

Figura 52 - Conexão à conta do Azure Data Lake Storage Gen2 criada

Fonte: autoria própria

Na figura 53 temos a leitura do arquivo de indicadores presente no Data Lake utilizando a função de leitura do Spark para a criação do dataframe que pode ser visualizado.

Figura 53 - Visualização dos dados de indicadores em um DataFrame



Fonte: autoria própria

Para aqueles com maior familiaridade com a linguagem SQL, pode-se criar uma view temporária a partir do spark dataframe e então realizar as análises e transformações utilizando o SQL.Para visualizar informações básicas da tabela, utilizamos o DESCRIBE TABLE e verificamos que existem as colunas com nome do país, código do país, nome do indicador, código do indicador e colunas que identificam o ano de 1960 até 2020.

Análises elaboradas podem ser realizadas com Python, Scala, R ou SQL nos notebooks Databricks, mas este trabalho tem como intuito o foco na criação de um pipeline com serviços da

Azure, por isso não teremos muito aprofundamento em questões de análise e grandes transformações nos dados.

Na figura 52 abaixo podemos visualizar a quantidade de países representados nos dados e a quantidade de indicadores de desenvolvimento existentes nesse conjunto.

Figura 54 - Visualização da quantidade de países e indicadores na tabela tIndicators

Fonte: autoria própria

Nesta etapa, para os indicadores, realizou-se a consolidação de três tabelas focadas em setores distintos: educação, saúde e renda; e apenas os 5 últimos anos foram considerados (2016, 2017, 2018, 2019 e 2020). A tabela original não dispõe desta separação, tendo, portanto, 1443 indicadores que tratam de áreas diversas, todos reunidos.

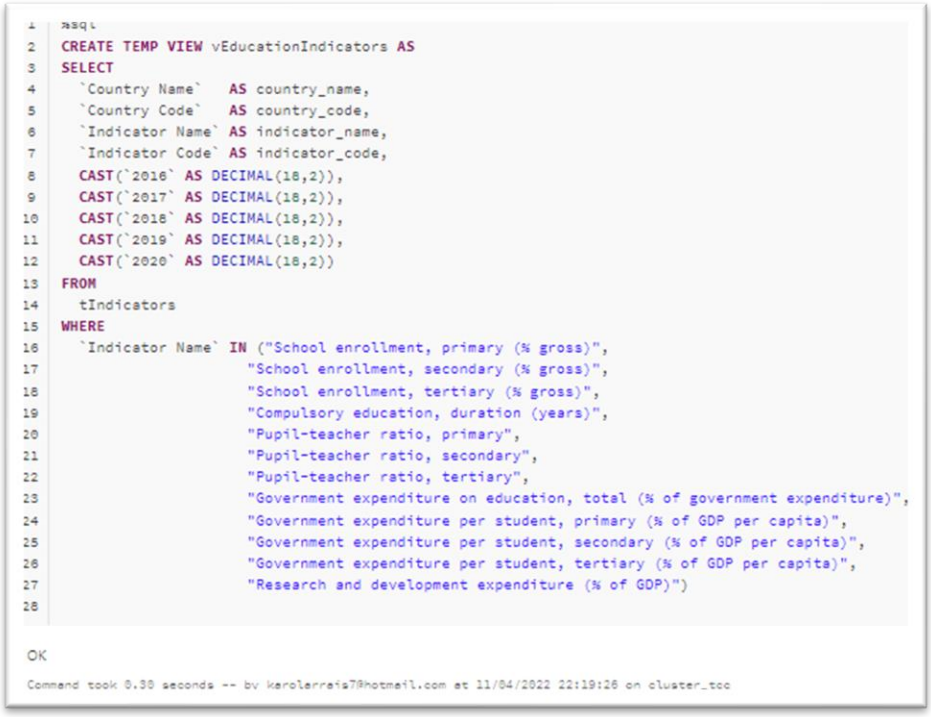
Para a tabela com indicadores de educação visualizada na figura 55, realizou-se a seleção dos indicadores listados abaixo (THE WORLD BANK, 2022.):

- School enrollment, primary (% gross) - Matrícula escolar, primária (% bruta): razão entre total de matrículas independente de idade, e a população da faixa etária oficial para o nível primário/fundamental. Ensino primário oferece habilidades básicas de leitura, escrita, matemática e compreensão elementar de outros assuntos como história, ciências naturais e sociais, para crianças.

- School enrollment, secondary (% gross) - Matrícula escolar, secundária (% bruta): razão entre total de matrículas independente de idade, e a população da faixa etária oficial para o nível secundário. O ensino secundário/médio complementa o ensino básico do nível primário, com um ensino mais orientado por disciplinas e habilidade com professores mais especializados.
- School enrollment, tertiary (% gross) - Matrícula escolar, ensino superior (% bruta): razão entre total de matrículas independente de idade, e a população da faixa etária oficial para o nível superior. Este ensino compreende ou não qualificação avançada, mas exige condições mínimas de admissão, como a conclusão com aproveitamento do ensino secundário.
- Compulsory education, duration (years) - Escolaridade obrigatória, duração (anos): número de anos legalmente exigidos como obrigatório para que crianças frequentem a escola.
- Pupil-teacher ratio, primary - Relação aluno-professor, primário: número médio de alunos por professor na escola primária.
- Pupil-teacher ratio, secondary - Relação aluno-professor, secundário: número médio de alunos por professor na escola secundária.
- Pupil-teacher ratio, tertiary - Relação aluno-professor, superior: número médio de alunos por professor no ensino superior.
- Government expenditure on education, total (% of government expenditure) - Despesas do governo em educação, total (% das despesas do governo): as despesas das administrações públicas em educação correspondem à percentagem da despesa total em todos os setores (saúde, serviços sociais, etc).
- Government expenditure per student, primary (% of GDP per capita) - Despesas do governo por aluno, primária (% do PIB per capita): despesa média do governo por aluno no ensino fundamental, expressa como porcentagem do PIB per capita.
- Government expenditure per student, secondary (% of GDP per capita) - Despesas do governo por aluno, secundário (% do PIB per capita): despesa média do governo por aluno no ensino médio, expressa como porcentagem do PIB per capita.
- Government expenditure per student, tertiary (% of GDP per capita) - Despesas do governo por aluno, ensino superior (% do PIB per capita): despesa média do governo por aluno no ensino superior, expressa como porcentagem do PIB per capita.

- Research and development expenditure (% of GDP) - Despesas com pesquisa e desenvolvimento (% do PIB): corresponde aos gastos internos brutos em pesquisa básica, aplicada e desenvolvimento experimental, incluindo as despesas de capital e correntes nos 4 setores: empresas, governo, ensino superior e privado sem fins lucrativos.

Figura 55 - Criação da view temporária para indicadores de educação



```
%sql
CREATE TEMP VIEW vEducationIndicators AS
SELECT
  `Country Name`   AS country_name,
  `Country Code`   AS country_code,
  `Indicator Name` AS indicator_name,
  `Indicator Code` AS indicator_code,
  CAST(`2016` AS DECIMAL(18,2)),
  CAST(`2017` AS DECIMAL(18,2)),
  CAST(`2018` AS DECIMAL(18,2)),
  CAST(`2019` AS DECIMAL(18,2)),
  CAST(`2020` AS DECIMAL(18,2))
FROM
  tIndicators
WHERE
  `Indicator Name` IN ("School enrollment, primary (% gross)",
                       "School enrollment, secondary (% gross)",
                       "School enrollment, tertiary (% gross)",
                       "Compulsory education, duration (years)",
                       "Pupil-teacher ratio, primary",
                       "Pupil-teacher ratio, secondary",
                       "Pupil-teacher ratio, tertiary",
                       "Government expenditure on education, total (% of government expenditure)",
                       "Government expenditure per student, primary (% of GDP per capita)",
                       "Government expenditure per student, secondary (% of GDP per capita)",
                       "Government expenditure per student, tertiary (% of GDP per capita)",
                       "Research and development expenditure (% of GDP)")

OK
Command took 0.30 seconds -- by karolarrais7@hotmail.com at 11/04/2022 22:19:26 on cluster_tcc
```

Fonte: autoria própria

Para a tabela com indicadores ligados à saúde da população, como exibido na figura 56, realizou-se a seleção dos indicadores listados abaixo (THE WORLD BANK, 2022):

- Community health workers (per 1,000 people) - Agentes comunitários de saúde (por 100 pessoas): Agentes comunitários de saúde englobam vários tipos de assistentes de saúde, como visitantes de saúde e agentes de saúde da família.
- Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total) - Causa de morte, por doenças transmissíveis e condições maternas, pré-natais e nutricionais (% do total): refere-se à parcela de todas as mortes independente da idade por causas básicas, incluindo doenças infecciosas e parasitárias, infecções respiratórias e deficiências nutricionais, como baixo peso e nanismo.

- Current health expenditure (% of GDP) - Despesas atuais com saúde (% do PIB): Nível da despesa corrente com saúde expresso em percentagem do PIB. As estimativas dos gastos atuais com saúde incluem bens e serviços de saúde consumidos durante cada ano. Não inclui gastos de capital com saúde, como prédios, maquinário, TI e estoques de vacinas para emergências ou surtos.
- External health expenditure (% of current health expenditure) - Despesas externas com saúde (% das despesas atuais com saúde): Parcela dos gastos atuais com saúde financiados por fontes externas. As fontes externas são compostas por transferências externas diretas e transferências externas distribuídas pelo governo abrangendo todos os fluxos financeiros para o sistema nacional de saúde de fora do país. As fontes externas fluem através do esquema governamental ou são canalizadas através de organizações não governamentais ou outros esquemas.
- Out-of-pocket expenditure (% of current health expenditure) - Despesas desembolsadas (% das despesas atuais com saúde): Parcela dos pagamentos diretos do total de gastos atuais com saúde. Os pagamentos diretos são gastos em saúde diretamente do próprio bolso das famílias.
- UHC service coverage index - Índice de cobertura do serviço UHC: Índice de cobertura para serviços essenciais de saúde (baseado em intervenções rastreadoras que incluem saúde reprodutiva, materna, neonatal e infantil, doenças infecciosas, doenças não transmissíveis e capacidade e acesso aos serviços). É apresentado em uma escala de 0 a 100.

Figura 56 - Criação da view temporária para indicadores de saúde

```
%sql
CREATE TEMP VIEW vHealthIndicators AS
SELECT
  `Country Name`   AS country_name,
  `Country Code`   AS country_code,
  `Indicator Name` AS indicator_name,
  `Indicator Code` AS indicator_code,
  CAST(`2016` AS DECIMAL(18,2)),
  CAST(`2017` AS DECIMAL(18,2)),
  CAST(`2018` AS DECIMAL(18,2)),
  CAST(`2019` AS DECIMAL(18,2)),
  CAST(`2020` AS DECIMAL(18,2))
FROM
  tIndicators
WHERE
   `Indicator Name` IN ("Community health workers (per 1,000 people)",
                        "Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)",
                        "Current health expenditure (% of GDP)",
                        "External health expenditure (% of current health expenditure)",
                        "Out-of-pocket expenditure (% of current health expenditure)",
                        "UHC service coverage index")

OK
```

Fonte: autoria própria

Ao final, na figura 57, realizou-se a seleção dos indicadores listados abaixo (THE WORLD BANK, 2022) para a consolidação da tabela com indicadores relacionados à renda:

- GDP per capita (current LCU) - PIB per capita (LCU atual): O PIB per capita é o produto interno bruto dividido pela população no meio do ano. O PIB é a soma do valor bruto agregado por todos os produtores residentes na economia mais quaisquer impostos sobre os produtos e menos quaisquer subsídios não incluídos no valor dos produtos. É calculado sem fazer deduções por depreciação de bens fabricados ou por esgotamento e degradação de recursos naturais. Os dados estão na moeda local atual.
- GNI per capita (current LCU) - RNB per capita (LCU atual): O RNB per capita é a renda nacional bruta dividida pela população no meio do ano. RNB (anteriormente PNB) é a soma do valor adicionado por todos os produtores residentes mais quaisquer impostos sobre produtos (menos subsídios) não incluídos na avaliação da produção mais as receitas líquidas de renda primária (compensação de empregados e renda de propriedade) do exterior. Os dados estão na moeda local atual.

- Gini index (World Bank estimate) - Índice de Gini (estimativa do Banco Mundial): O índice de Gini mede até que ponto a distribuição de renda (ou, em alguns casos, despesas de consumo) entre indivíduos ou famílias dentro de uma economia se desvia de uma distribuição perfeitamente igual. Uma curva de Lorenz traça as porcentagens acumuladas da renda total recebida em relação ao número acumulado de beneficiários, começando com o indivíduo ou família mais pobre. O índice de Gini mede a área entre a curva de Lorenz e uma linha hipotética de igualdade absoluta, expressa em porcentagem da área máxima sob a linha. Assim, um índice de Gini de 0 representa igualdade perfeita, enquanto um índice de 100 implica desigualdade perfeita.

- Poverty headcount ratio at national poverty lines (% of population) - Índice de pobreza nas linhas de pobreza nacionais (% da população): O índice nacional de pobreza é a porcentagem da população que vive abaixo da(s) linha(s) nacional(is) de pobreza. As estimativas nacionais são baseadas em estimativas de subgrupos ponderadas pela população de pesquisas domiciliares. Para as economias para as quais os dados são da EU-SILC, o ano relatado é o ano de referência da renda, que é o ano anterior ao ano da pesquisa.

Figura 57- Criação da view temporária para indicadores de renda

```
%sql
CREATE TEMP VIEW vIncomeIndicators AS
SELECT
  `Country Name`   AS country_name,
  `Country Code`   AS country_code,
  `Indicator Name` AS indicator_name,
  `Indicator Code` AS indicator_code,
  CAST(`2016` AS DECIMAL(18,2)),
  CAST(`2017` AS DECIMAL(18,2)),
  CAST(`2018` AS DECIMAL(18,2)),
  CAST(`2019` AS DECIMAL(18,2)),
  CAST(`2020` AS DECIMAL(18,2))
FROM
  tIndicators
WHERE
   `Indicator Name` IN ("GDP per capita (current LCU) ",
                        "GNI per capita (current LCU)",
                        "Gini index (World Bank estimate)",
                        "External health expenditure (% of current health expenditure)",
                        "Poverty headcount ratio at national poverty lines (% of population)")

OK

Command took 0.10 seconds -- by karolarrais7@hotmail.com at 11/04/2022 22:20:04 on cluster_tcc
```

Fonte: autoria própria

Após consolidação das views temporárias com indicadores de saúde, educação e renda, os dados foram carregados na camada silver. Na figura 58 é possível visualizar o processo de criação de um Spark DataFrame para utilização do PySpark para salvar os dados em formato CSV dos indicadores de educação, e também em formato parquet, um formato orientado a colunas altamente utilizado para armazenamento de big data pela alta performance em compactação e consultas (DATABRICKS, s.d.).

Figura 58 - Processo de carga na camada silver do Data Lake no formato CSV e parquet dos indicadores

```
dfEducation = spark.sql("SELECT * FROM vEducationIndicators")

▸ dfEducation:  pyspark.sql.dataframe.DataFrame = [country_name: string, country_code: string ... 7 more fields]

Command took 0.09 seconds -- by karolarrais7@hotmail.com at 12/04/2022 00:41:19 on cluster_tcc
```

Fonte: autoria própria

Um processo de consolidação dos dados de COVID também foi realizado para salvar informações pertinentes para a análise. De modo a resumir a prática para que a mesma não se estenda demasiadamente, foi utilizado apenas o conjunto de dados completo de COVID da Our World in Data, com uma tabela final contendo as seguintes colunas (GITHUB, s.d.):

- iso_code: código de país de três letras;
- location: localização geográfica;
- date: data de observação;
- population: número de habitantes da localização;
- median_age: idade média da população;
- aged_65_older: parcela da população com 65 anos ou mais;
- total_cases_per_million: total de casos confirmados de COVID-19 por 1.000.000 de pessoas. Pode incluir casos prováveis relatados;
- total_deaths_per_million: total de mortes registradas como COVID-19 por 1.000.000 de pessoas. Pode incluir mortes prováveis, quando relatadas;
- icu_patients_per_million: pacientes com COVID-19 em UTIs em um determinado dia por 1.000.000 de pessoas;
- hosp_patients_per_million: número de pacientes com COVID-19 no hospital em um determinado dia, por 1.000.000 de pessoas;
- stringency_index: Índice de Rigidez de Resposta do Governo, com valor de 0 a 100 (100 = resposta mais rigorosa). Inclui ações como fechamentos de escolas, locais de trabalho e proibição de viagens;
- total_tests_per_thousand: total de testes de COVID-19 por 1.000 pessoas;

- people_vaccinated_per_hundred: número total de pessoas que receberam pelo menos uma dose de vacina por 100 pessoas na população total;
- people_fully_vaccinated_per_hundred: número total de pessoas que receberam todas as doses prescritas pelo protocolo de vacinação inicial por 100 pessoas na população total;
- total_boosters_per_hundred: Número total de doses de reforço de vacinação COVID-19 administradas por 100 pessoas na população total.

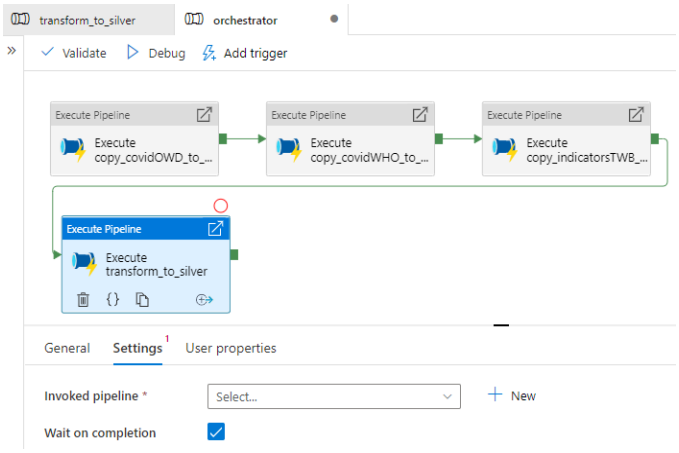
Ao final da consolidação tem-se 161.989 registros que podem ser carregados na camada silver assim como foi realizado com os dados dos indicadores, processo mostrado na figura 59.

Figura 59 - Processo de carga na camada silver do Data Lake no formato CSV e parquet dos dados COVID

Fonte: autoria própria

Ao final, é criado um pipeline do Data Factory com a atividade *Notebook* que se encontra no conjunto de atividades Databricks, onde os notebooks criados para a transformação dos dados e carga na camada silver são associados. Este pipeline é então adicionado no pipeline orquestrador como segue na figura 60, por meio da atividade Execute Pipeline do tipo General, onde a espera de conclusão das execuções de cópia para bronze é necessária e, portanto, habilitada. Ao final da execução do orquestrador obtém-se então os dados consolidados dos indicadores de saúde, educação e renda; e dados selecionados de covid, dados pré-selecionados e transformados para maior facilidade na integração e visualizações.

Figura 60 - Adição ao orquestrador do pipeline de transformação para a silver



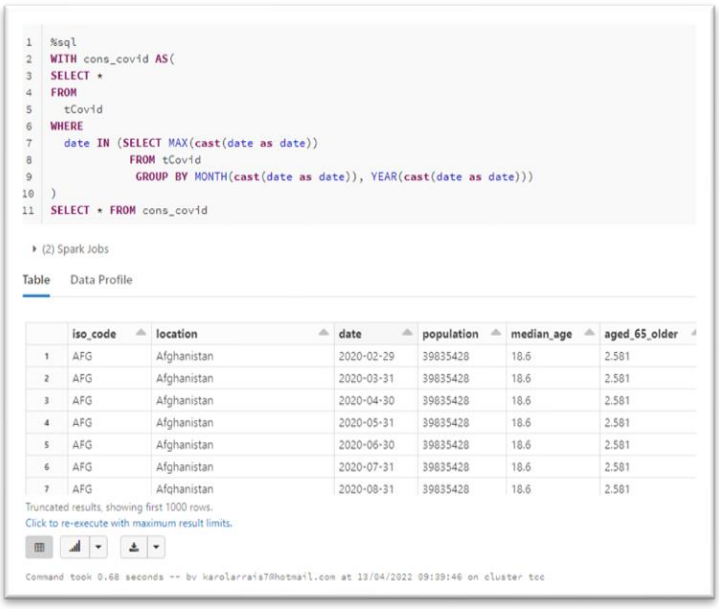
Fonte: autoria própria

4.4 Integração dos dados

A terceira etapa da prática consistiu no refinamento e integração dos dados consolidados na etapa anterior, como preparação para realizar a geração de insights pela plataforma Power BI. Neste processo de transformação os dados são armazenados na zona refinada do Data Lake, camada gold criada na preparação do ambiente. Assim como as transformações iniciais, o refinamento e integração dos dados consolidados pode ser realizada utilizando um notebook Databricks. Foram realizadas integração para geração de 3 novos arquivos: o primeiro contém a integração dos dados de COVID-19 com os indicadores de saúde; o segundo integra os indicadores de educação aos dados de COVID-19; e por último, ocorre a integração do consolidado de COVID-19 com indicadores de renda.

Os dados da COVID-19 transformados na etapa anterior possuíam uma estrutura de registros por dia, da data 01-01-2020 à 14-02-2022. Para a integração, uma nova consolidação dessas informações foi realizada, selecionando apenas o registro mais atual de cada mês e ano para o país correspondente, como segue a consulta da figura 61.

Figura 61 - Consulta SQL para seleção do registro mais atualizado de cada mês e ano dos dados COVID



```
%sql
WITH cons_covid AS(
SELECT *
FROM
  tCovid
WHERE
  date IN (SELECT MAX(cast(date as date))
            FROM tCovid
             GROUP BY MONTH(cast(date as date)), YEAR(cast(date as date)))
)
SELECT * FROM cons_covid
```

▸ (2) Spark Jobs

Table Data Profile

	iso_code	location	date	population	median_age	aged_65_older
1	AFG	Afghanistan	2020-02-29	39835428	18.6	2.581
2	AFG	Afghanistan	2020-03-31	39835428	18.6	2.581
3	AFG	Afghanistan	2020-04-30	39835428	18.6	2.581
4	AFG	Afghanistan	2020-05-31	39835428	18.6	2.581
5	AFG	Afghanistan	2020-06-30	39835428	18.6	2.581
6	AFG	Afghanistan	2020-07-31	39835428	18.6	2.581
7	AFG	Afghanistan	2020-08-31	39835428	18.6	2.581

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

Command took 0.68 seconds -- by karolarrais7@hotmail.com at 13/04/2022 09:39:46 on cluster tcc

Fonte: autoria própria

Com relação aos indicadores, na etapa anterior de transformação foram selecionados os valores para os anos de 2016,2017,2018,2019 e 2020. Uma consolidação destes dados também foi realizada antes da integração com os dados da COVID; para cada registro, foi calculada a média dos valores presentes nesses cinco anos do indicador da linha correspondente, realizando assim uma redução de dimensionalidade para uma medida de centralidade estatística que facilitará a análise. A código do cálculo foi feito com a linguagem Python que também é suportada pelo notebook Databricks, como pode ser visualizado na figura 62.

Figura 62 - Cálculo da média por registro dos dados de indicadores

```
#transformação para integração - realização do cálculo da média dos indicadores de 2016 a 2020
#transformação para pandas dataframe
pd_dtIncome = df_incomeInd.toPandas()
#array com as colunas dos anos que utilizaremos para calcular a media já foi declarada anteriormente
#cálculo da média
dfFinal_income = pd_dtIncome.assign(mean=pd_dtIncome[years].apply(np.mean, axis=1))
dfFinal_income
```

▸ (1) Spark Jobs

```
/databricks/spark/python/pyspark/sql/pandas/utils.py:79: UserWarning: The conversion of DecimalType columns is inefficient and may take a
2020] If those columns are not necessary, you may consider dropping them or converting to primitive types before the conversion.
  warnings.warn(
Out[10]:
```

	country_name	country_code	indicator_name	indicator_code	2016	2017	2018	2019	2020	mean
0	Africa Eastern and Southern	AFE	External health expenditure (% of current heal...	SH.XPD.EHEX.CH.ZS	11.92	12.18	12.90	None	None	12.333333
1	Africa Eastern and Southern	AFE	Gini index (World Bank estimate)	SI.POV.GINI	None	None	None	None	None	NaN
2	Africa Eastern and Southern	AFE	GNI per capita (current LCU)	NY.GNP.PCAP.CN	None	None	None	None	None	NaN
3	Africa Eastern and Southern	AFE	Poverty headcount ratio at national poverty li...	SI.POV.NAHC	None	None	None	None	None	NaN
4	Africa Western and Central	AFW	External health expenditure (% of current heal...	SH.XPD.EHEX.CH.ZS	13.23	11.61	10.96	None	None	11.933333
...	...	...	...	...	...	...	...	...	...	...
1059	Zambia	ZMB	Poverty headcount ratio at national poverty li...	SI.POV.NAHC	None	None	None	None	None	NaN
1060	Zimbabwe	ZWE	External health expenditure (% of current heal...	SH.XPD.EHEX.CH.ZS	27.66	21.04	20.01	None	None	22.903333
1061	Zimbabwe	ZWE	Gini index (World Bank estimate)	SI.POV.GINI	None	44.30	None	50.30	None	47.300000

Fonte: autoria própria

Com essas consolidações foi realizada a integração dos dados mencionada inicialmente, com a geração de 3 arquivos distintos que relacionam dados da COVID-19 com os 3 tipos de indicadores. A integração foi feita com SQL e teve como chave a coluna que possui código do país, representada pela coluna iso_code nos dados da COVID e como country_code nos dados dos indicadores. Segue visualização da criação da tabela de integração dos dados covid com os indicadores de renda:

```
%sql
-- CENÁRIO DE RENDA - integrar os dados de covid com os indicadores de renda
CREATE OR REPLACE TEMP VIEW tCovid_IncomeIndicator AS
WITH income_indicators AS(
SELECT
  country_name,
  country_code,
  indicator_name,
  cast(mean AS decimal(18,2)) AS mean_indicator
FROM
 tIncome
WHERE
  indicator_name IN ('GDP per capita (current LCU)',
                     'GNI per capita (current LCU)',
                     'Gini index (World Bank estimate)',
                     'Poverty headcount ratio at national poverty lines (% of population)')
)
```

```
-- consolidação dos dados covid pegando últimos registros de cada mês (dados covid tem informações do
dia 01-01-2020 até 14-02-2022 )
, cons_covid AS(
SELECT *
FROM
  tCovid
WHERE
  date IN (SELECT MAX(cast(date as date))
            FROM tCovid
             GROUP BY MONTH(cast(date as date)), YEAR(cast(date as date)))
)
, covid_join_income AS(
SELECT
  a.iso_code,
  b.country_name,
  a.date AS date_infoCovid,
  a.population,
  a.median_age,
  a.aged_65_older,
  a.total_cases_per_million,
  a.total_deaths_per_million,
  a.icu_patients_per_million,
  a.hosp_patients_per_million,
  a.stringency_index,
  a.total_tests_per_thousand,
  a.people_vaccinated_per_hundred,
  a.people_fully_vaccinated_per_hundred,
  a.total_boosters_per_hundred,
  b.indicator_name,
  b.mean_indicator
FROM
  cons_covid AS a
LEFT JOIN
  income_indicators AS b
ON a.iso_code = b.country_code)
SELECT * FROM covid_join_income
```

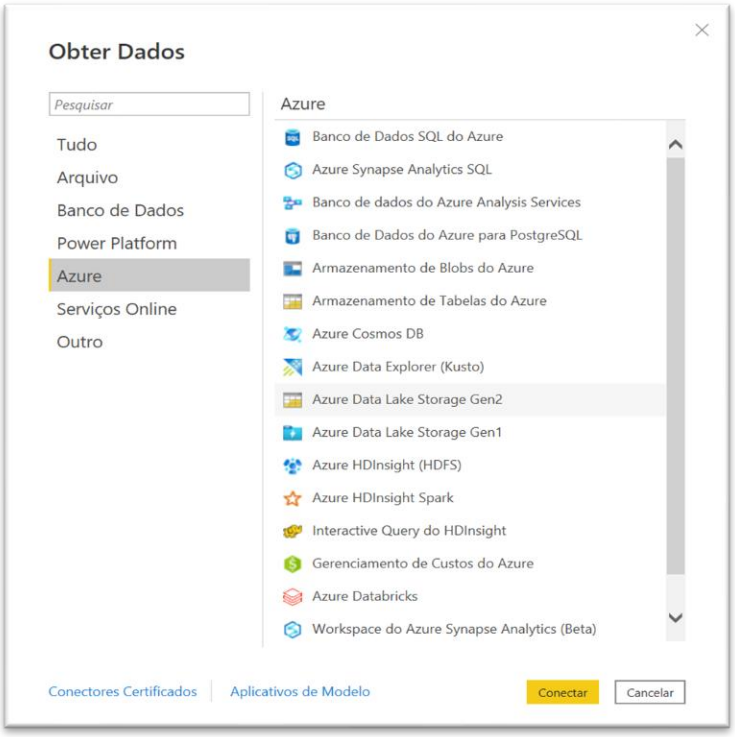
Com o processo de integração acima realizado também para os indicadores de educação e saúde, finaliza-se a construção da etapa de preparação dos dados para carga na camada gold do Data Lake. As views criadas podem ser transformadas em Spark Dataframes e exportadas para o Data Lake com processo semelhante ao que foi realizado no notebook de transformação para carga da camada silver.

4.5 Visualização dos resultados

Com os dados refinados e disponibilizados na camada gold do Data Lake, é possível iniciar todo o trabalho que envolve análise de dados para a geração de informações e conhecimento. É possível conectar na conta de armazenamento da Azure a partir do Power BI Desktop. Para isso, basta realizar a conexão por

meio da obtenção de dados escolhendo o serviço de armazenamento Azure Data Lake Storage Gen2 como mostrado da figura 63, e então informar a URL e chave de acesso do Data Lake.

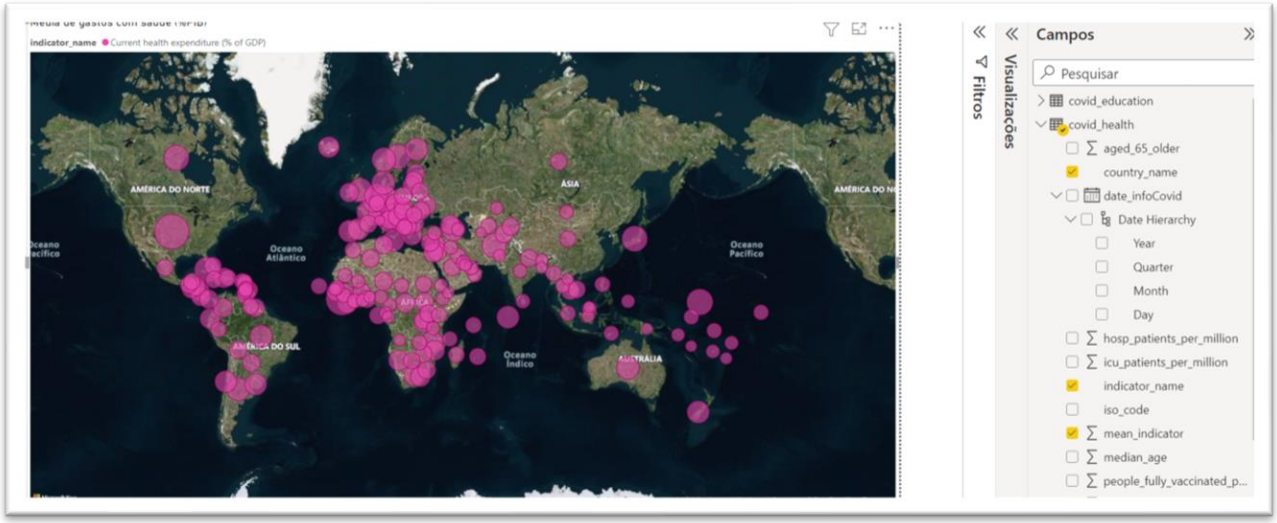
Figura 63 - Conexão ao Azure Data Lake Storage Gen2 pelo Power BI desktop



Fonte: autoria própria

Conectados, é possível visualizar toda a hierarquia de diretórios do Data Lake e acessar todos os arquivos para a realização das análises e geração de insights. Com a finalidade de ilustrar o processo de análise e criação de relatórios utilizando a ferramenta, na figura 64 mostra-se o print do início de um relatório conectado aos dados integrados de covid e indicadores de saúde presentes na camada gold.

Figura 64 - Acesso em dataset e construção de relatório no Power BI desktop



Fonte: autoria própria

Capítulo 5

CONCLUSÃO

Este trabalho se alinha ao contexto do Big Data e os impactos proporcionados pela crescente e variada produção de dados. Dando destaque aos conteúdos apresentados, o Capítulo 2 consistiu no levantamento das evoluções de conceitos (como Data Warehouse x Data Lake e ETL x ELT), ferramentas (Ecossistema Hadoop e Microsoft Azure) e profissões especializadas da área de dados, seguindo uma linha de discussão que aborda inicialmente conceitos essenciais até a finalização do capítulo com as discussões mais específicas sobre as funcionalidades de ferramentas que foram utilizadas no Capítulo 4 de forma prática.

O pipeline para aplicação destes conceitos de forma prática de ponta a ponta, foi construído e detalhado ao longo do Capítulo 4. Todo o fluxo de atividades envolvendo os ciclos de coleta, processamento, armazenamento e análise foi definido de forma prática e funcional por meio do pipeline e arquitetura que podem ser visualizados na figura 27 do Capítulo 4. O sucesso do acesso e extração dos dados abertos da COVID-19 e indicadores de desenvolvimento disponibilizados na internet foi alcançado de forma simples e executado com manipulação de recursos em uma interface visual, de forma descomplicada, sem a necessidade de grandes esforços com programação. Todo o processo de transformação e consolidação desses dados foi realizado com um recurso eficiente para processamento de Big Data e que possui compatibilidade com as linguagens comumente utilizadas para a manipulação de dados: SQL, Python e R. Junto com esse processo de extração, cargas e transformações, foram trabalhadas as zonas do Data Lake abordadas na arquitetura do mesmo, no Capítulo 2. Como resultado obtemos um fluxo de trabalho normalmente desenvolvido por engenheiros de dados até a disponibilização de dados refinados para que analistas e cientistas gerem informações e conhecimento.

5.1 Trabalhos futuros

Este trabalho contribuiu para o conhecimento de conceitos importante relacionados ao processamento de dados com foco nos serviços em nuvem e nos ciclos de extração, processamento e armazenamento. Como trabalhos futuros sugere-se um possível aprofundamento no ciclo de análise para demostrar o valor que pode ser gerado a partir dos dados. Também pode-se seguir uma

linha ainda dentro da engenharia de dados, com o desenvolvimento da mesma aplicação (pipeline) utilizando os serviços comumente empregados em ambiente On-Premise, como o ecossistema Hadoop e realizar comparações de usabilidade e performance com o desenvolvimento da mesma aplicação em serviços da nuvem.

Referências

ANDERSON, C. The Petabyte Age: because more isn't just more - more is different. **Wired**, 23 jun. 2009. Disponível em: <http://www.wired.com/science/discoveries/magazine/16-07/pb_intro>. Acesso em: 30 mar. 2022.

ANDERSON, J. **Data Teams.** Chapter one. 2020. Disponível em: < https://www.datateams.io/wp-content/uploads/2020/09/Anderson2020_Chapter_DataTeams.pdf>. Acesso em: 24 mar. 2022.

APACHE SPARK. **Spark Overview.** Disponível em: < https://spark.apache.org/docs/3.2.1/>. Acesso em: 20 mar. 2022.

AZURE MICROSOFT. **What is PaaS?** Disponível em: < https://azure.microsoft.com/en-us/overview/what-is-paas/>. Acesso em: 12 mar. 2022.

BAHGA, A.; MADISETTI, V. **Big Data Science & Analytics:** A Hands-On Approach. Published by Arshdeep Bahga & Vijay Madisetti. 2019.

BIBI, S., KATSAROS, D. e BOZANIS, P. "Aquisição de aplicativos de negócios: soluções baseadas em SaaS ou no local?", em **IEEE Software**, vol. 29, nº. 3, pp. 86-93, maio-junho 2012.

DATABRICKS. **Apache Spark™.** Disponível em: < https://databricks.com/spark/about>. Acesso em: 20 mar. 2022.

DATABRICKS. **Discover Lake House.** Disponível em: < https://databricks.com/discoverlakehouse>. Acesso em: 20 mar. 2022.

DATABRICKS. **Parquet.** Disponível em: < https://databricks.com/glossary/what-is-parquet>. Acesso em: 22 mar. 2022.

DHAR, Vasant. Data science and prediction. **Communications of the ACM**. Volume 56, Issue 12, December 2013, pp. 64–73. Disponível em: <https://doi.org/10.1145/2500499>. Acesso em: 31 mar. 2022.

DIAMOND, P. **Armazenamento em nuvem vs. servidores locais: nove considerações importantes.** 2020. Disponível em: < https://www.microsoft.com/pt-br/microsoft-365/business-insights-ideas/resources/cloud-storage-vs-on-premises-servers>. Acesso em: 13 mar. 2022.

DIJCKS, J.ÿP. Oracle: Big Data para a Empresa. **Um White Paper da Oracle.** Oracle Corporation. 2013.

FAGUNDES, P. B.; MACEDO, D. D. J.; FREUND, G. P. A Produção Científica sobre Qualidade de Dados em Big Data: um estudo na base de dados Web of Science. **Revista Digital de Biblioteconomia e Ciência da Informação.** Campinas, v. 16, n. 1, p. 194-210, jan-abr, 2018. Disponível em: < https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8650412 > Acessado em: 24 mar. 2022.

FATIMA, H.; WASNIK, K. Comparison of SQL, NoSQL and NewSQL databases for internet of things. In: **2016 IEEE Bombay Section Symposium (IBSS**). IEEE, 2016. p. 1-6, 2016.

FÁVERO, L.P.; BELFIORE, P. **Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel, SPSSe Stata.** Cerqueira César: Elsevier Editora Ltda., 2017.

GITHUB. **Data on COVID-19 (coronavirus) by Our World in Data.** Disponível em: < https://github.com/owid/covid-19-data/tree/master/public/data>. Acesso em: 12 mar. 2022.

GOLDMAN, Alfredo; KON, Fábio; PEREIRA JUNIOR, Francisco; POLATO, Ivanilto; PEREIRA, Rosangela de Fátima. Apache hadoop: conceitos teóricos e práticos, evolução e novas possibilidades. **Anais**. Porto Alegre: SBC, 2012.Disponível em:< https://repositorio.usp.br/item/002338037>. Acesso em: 12 mar. 2022.

HARRISON, G. **Next Generation Databases: NoSQLand Big Data.** New York: Apress, 2015.

HASHEM, I., YAQOOB, I.; ANUAR, N.; MOKHTAR, S.; GANI, A.; KHAN, S. The rise of "big data" on cloud computing: Review and open research issues, **Information Systems,** Volume 47, 2015, Pages 98-115. Disponível em: <https://doi.org/10.1016/j.is.2014.07.006>. Acesso em: 11 mar. 2022.

HEUSER, C.A. **Projeto de Banco de Dados – Número 4.** Instituto de Informática da UFRGS. Porto Alegre: Editora Sagra Luzzatto, 1998.

IME UNICAMP. **O Que São Dados?** Disponível em: <https://www.ime.unicamp.br/~hildete/dados.pdf>. Acesso em: 30 mar. 2022.

JO, J.; LEE, K. MapReduce-Based D_ELT Framework to Address the Challenges of Geospatial Big Data. **ISPRS Int. J. Geo-Inf**. 2019, 8, 475. Disponível em: <https://www.mdpi.com/2220-9964/8/11/475>. Acesso em: 11 abr. 2022.

KHINE, P.P.; WANG, Z.S. **Data Lake: A New Ideology in Big Data Era.** Conference: 2017 4th International Conference on Wireless Communication and Sensor Network. 2017. Disponível em: <https://www.researchgate.net/publication/321825490_Data_Lake_A_New_Ideology_in_Big_Data_Era>. Acesso em: 13 mar. 2022.

LEMOS, A. LÉVY, P. **O futuro da Internet:** Em direção a uma ciberdemocracia planetária. São Paulo: Paulus, 2010.

MANYIKA, J., et al. **Big Data:** The Next Frontier for Innovation, Competition, and Productivity. San Francisco, McKinsey Global Institute, CA, USA, 2010.

MARQUESONE, R. **Big Data:** Técnicas e tecnologias para extração de valor dos dados. Brasil: Casa do Código, 2016. Disponível em: < https://bit.ly/3NyB13U>. Acesso em: 30 mar. 2022.

MICROSOFT. **Azure regions, availability zones, and region pairs.** Disponível em: < https://docs.microsoft.com/en-us/learn/modules/intro-to-azure-fundamentals/what-is-microsoft-azure?ns-enrollment-type=LearningPath&ns-enrollment-id=learn.az-900-describe-cloud-concepts>. Acesso em: 16 mar. 2022.

MICROSOFT. **Building blocks of Power BI.** Disponível em: < https://docs.microsoft.com/en-us/learn/modules/get-started-with-power-bi/3-building-blocks-of-power-bi>. Acesso em: 24 mar. 2022.

MICROSOFT. **Describe cloud benefits and considerations.** Disponível em: < https://docs.microsoft.com/en-us/learn/modules/intro-to-azure-fundamentals/what-is-microsoft-azure?ns-enrollment-type=LearningPath&ns-enrollment-id=learn.az-900-describe-cloud-concepts>. Acesso em: 16 mar. 2022.

MICROSOFT. **Execute Pipeline activity in Azure Data Factory and Synapse Analytics.** 2021. Disponível em: <https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-pipeline-activity>. Acesso em: 16 mar. 2022.

MICROSOFT. **Explain Azure Databricks.** Disponível em: < https://docs.microsoft.com/en-us/learn/modules/describe-azure-databricks/2-explain?ns-enrollment-type=learningpath&ns-enrollment-id=learn.wwl.data-engineer-azure-databricks>. Acesso em: 20 mar. 2022.

MICROSOFT. **Formato Binary do Azure Data Factory e do Synapse Analytics.** 2020. Disponível em: <https://docs.microsoft.com/pt-br/azure/data-factory/format-binary>. Acesso em: 22 mar. 2022.

MICROSOFT. **Introduction – Microsoft Power BI**. Disponível em: < https://docs.microsoft.com/en-us/learn/modules/get-started-with-power-bi/1-introduction?ns-enrollment-type=learningpath&ns-enrollment-id=learn-bizapps.get-started-data-analytics>. Acesso em: 24 mar. 2022.

MICROSOFT. **Overview of data analysis.** Disponível em: < https://docs.microsoft.com/en-us/learn/modules/data-analytics-microsoft/2-data-analysis?ns-enrollment-type=learningpath&ns-enrollment-id=learn-bizapps.get-started-data-analytics>. Acesso em: 22 mar. 2022.

MICROSOFT. **Run a Databricks notebook with the Databricks Notebook Activity in Azure Data Factory.** 2022. Disponível em: <https://docs.microsoft.com/en-us/azure/data-factory/transform-data-using-databricks-notebook>. Acesso em: 13 mar. 2022.

MICROSOFT. **Understand the architecture of Azure Databricks spark cluster.** Disponível em: < https://docs.microsoft.com/en-us/learn/modules/spark-architecture-fundamentals/2-understand-architecture-of-azure-databricks-spark-cluster?ns-enrollment-type=learningpath&ns-enrollment-id=learn.wwl.data-engineer-azure-databricks>. Acesso em: 20 mar. 2022.

MICROSOFT. **What is Azure?** Disponível em: < https://docs.microsoft.com/en-us/learn/modules/intro-to-azure-fundamentals/what-is-microsoft-azure?ns-enrollment-type=LearningPath&ns-enrollment-id=learn.az-900-describe-cloud-concepts>. Acesso em: 13 mar. 2022.

RAJ, A., BOSCH, J, OLSSON, HH e WANG, TJ. "Modelling Data Pipelines,". 2020, **46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA),** 2020, pp. 13-20.

ROMERO, Oscar; WREMBEL, Robert, SONG, Il-Yeol. An Alternative View on Data Processing Pipelines from the DOLAP 2019 Perspective. **Science Direct**, Information Systems, Volume 92, 2020. Disponível em: <https://doi.org/10.1016/j.is.2019.101489>. Acesso em: 20 mar. 2022.

SAWADOGO, P.; DARMONT, J. On data lake architectures and metadata management. **Journal of Intelligent Information Systems**, Springer Verlag, 2021, 56 (1), pp.97-120. Disponível em: <ff10.1007/s10844-020-00608-7ff. ffhal-03114365f>. Acesso em: 01 abr. 2022.

SHARMA, B.: **Architecting Data Lakes Data Management Architectures for Advanced Business Use Cases**. O'Reilly Media, Inc., 2018. Disponível em: <https://github.com/ffisk/books/blob/master/architecting-data-lakes.pdf ->. Acesso em: 11 mar. 2022.

STONEBRAKER, M. et al. The End of an Architectural Era (It's Time for a Complete Rewrite). **VLDB Endowment:** 33rd International Conference, Vienna, 2007. Disponível em: < https://dl.acm.org/doi/10.5555/1325851.1325981>. Acesso em: 31 mar. 2022.

STRAWN, G. "Masterminds of the Arpanet", em **IT Professional**, vol. 16, não. 3, pp. 66-68, maio-junho 2014, doi: 10.1109/MITP.2014.32.

TANENBAUM. Andrew; STEEN VAN, Maarten. **Distributed System: Principles and Paradigms.** 2. Ed. Nova Jersey: Prentice Hall,2006.

THE WORLD BANK. **World Development Indicators.** 2022. Disponível em: < https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators >. Acesso em: 22 mar. 2022.

VAISMAN, A.; ZIMÁNYI, E. **Data Warehouse Systems:** Design and Implementation. USA: Springer, 2016.

VIANNA, W. B.; DUTRA, M. L. Big Data e gestão da informação: Modelagem do Contexto Decisional Apoiado pela Sistemografia. **Revista Informação e Informação**, Londrina, v. 21, n. 1, p. 185 - 212, jan./abr. 2016.

WE ARE SOCIAL AND HOOTSUITE. **Special Report Digital 2021.** Disponível em: <https://wearesocial.com/uk/blog/2021/01/digital-2021-uk/>. Acesso em: 30 mar. 2022.

WE ARE SOCIAL AND HOOTSUITE. **Special Report Digital 2022.** Disponível em: < https://wearesocial.com/uk/blog/2022/01/digital-2022/>. Acesso em: 27 mar. 2022.

WORLD HEALTH ORGANIZATION. WHO Coronavirus (COVID-19) Dashboard. Disponível em: < https://covid19.who.int/data>. Acesso em: 23 mar. 2022.