WASHINGTON LUIZ PERONI

MACHINE LEARNING ENGINEERING

BIG DATA CLOUD PLATFORMS

AULA 01

SUMÁRIO

O QUE VEM POR AÍ? 3
HANDS ON 5
SAIBA MAIS 6
O QUE VOCÊ VIU NESTA AULA? 15
REFERÊNCIAS.....16

O QUE VEM POR AÍ?

Olá, pessoal! Boas-vindas à aula de “AWS Glue”, um importante serviço que nos ajuda na extração, transformação e carga de dados em nossos datalakes e data-warehouses.

Desde os anos 90, as empresas vêm aumentando suas capacidades de geração, manipulação e tomada de decisão por meio de dados. No passado, conhecemos os termos “mineração de dados”, “business-intelligence” ou simplesmente BI. Nos dias atuais, unindo as tecnologias de Big Data e o aumento exponencial de volume, variedade e velocidade, essas atividades e ferramentas hoje são conhecidas como “Engenharia de Dados” e nos permitem processar Big Data em tempo real.

Nenhuma empresa consegue ter times de IA e Machine Learning competitivos sem um processo maduro de engenharia de dados; já repararam que a oferta de vagas cresceu tanto quanto a de ciência de dados? É um diferencial competitivo para qualquer empresa conseguir fazer ETL e BI de forma organizada e rápida para poder alimentar seus times de IA e Machine Learning.

Como vocês bem sabem, no mundo real não há conjuntos de dados prontos (limpos, classificados, ordenados, sumarizados) para serem consumidos por modelos de IA; assim, as atividades de ETL em Big Data são cruciais e estratégicas para times maduros de ML e IA.

Por essas e outras razões vocês aprenderão técnicas de Big Data que compreendem a disciplina de Engenharia de Dados para que, enquanto cientistas de dados, tenham a capacidade de avaliar a confiabilidade e a qualidade do dado e ajudar lideranças a montar um processo de fim-a-fim maduro o suficiente. Dado certo, modelo e previsões melhores!

Nessa aula conheceremos o “AWS Glue”, que segundo a definição oficial da AWS (2024, n.p.) é:

> O AWS Glue é um serviço de integração de dados com tecnologia sem servidor que facilita aos usuários de análise a descoberta, preparação, transferência e integração de dados de várias fontes. Você pode usá-lo para análise, machine learning e desenvolvimento de aplicações.

Entenderemos isso na prática, demonstraremos alguns casos de uso para que vocês possam usar no mundo real e entender arquiteturas de referência e boas práticas para endereçar as necessidades de Big Data de nossas empresas. Então vamos lá, mãos aos dados!

HANDS ON

Nosso objetivo será demonstrar, na prática, dentro do console da AWS, o uso do Glue, mostrando desde a tela inicial, os conceitos básicos, até um caso de uso real de integração com S3, que hoje é o Data Lake da maioria das empresas. Iremos dividi-lo nas seguintes etapas:

- AWS Glue via console (página da AWS).
- **Visão geral**: organização e como seus componentes se comunicam.
- **Conceitos básicos**: fluxo simplificado.
- **Job Glue**: como criar e executar um job glue.
- **Caso de uso real:** ETL de dados CSV, parquet e integração com S3.

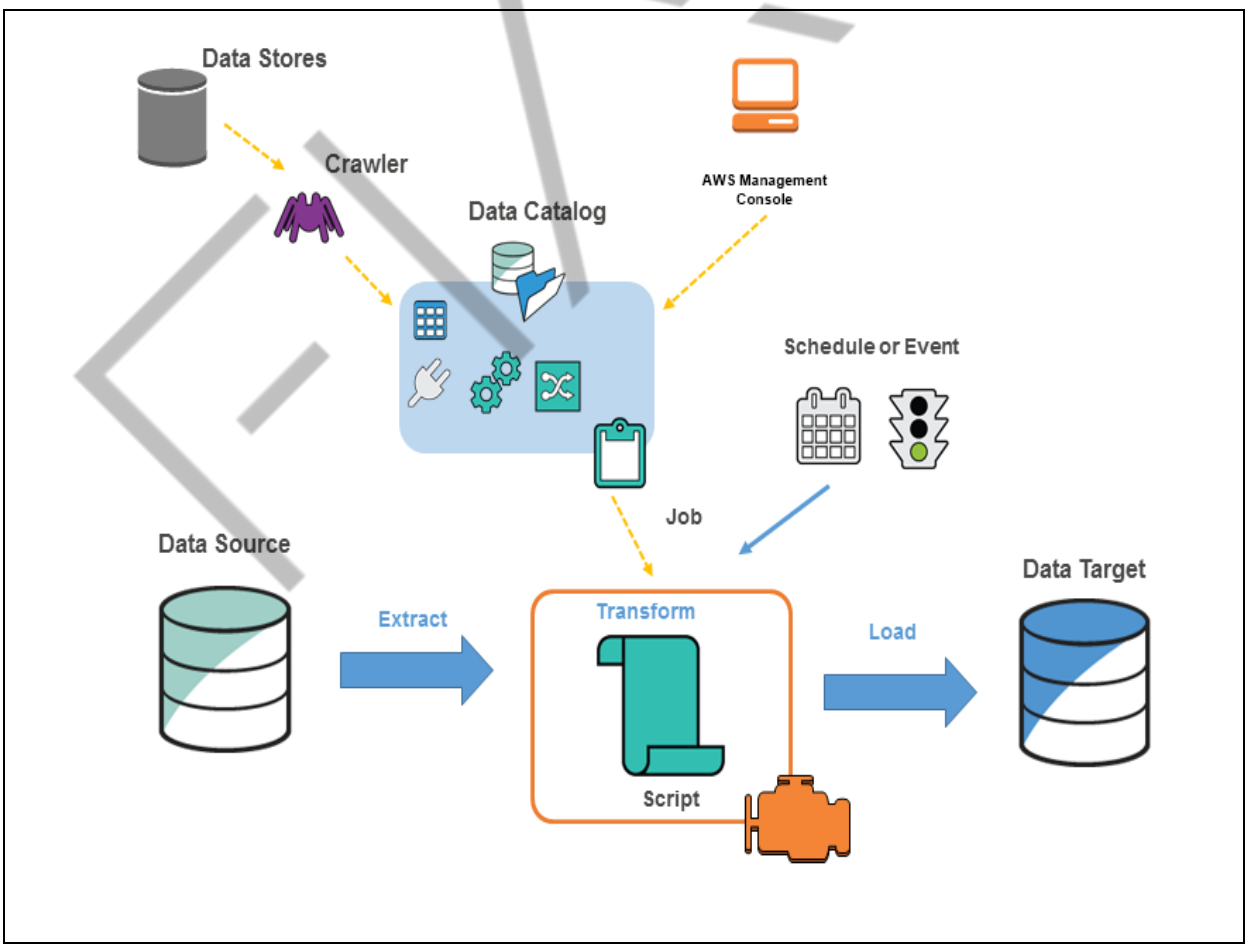

Figura 1 – Arquitetura do AWS Glue
Fonte: AWS (2024)

SAIBA MAIS

A seguir, temos alguns conceitos fundamentais para o total entendimento e alcance do que foi demonstrado nas videoaulas. Dessa forma, veja e reveja os vídeos e complemente o que foi aprendido com essa leitura.

Datalake

James Dixon, então líder de tecnologia da Pentaho, alegadamente inventou o termo em 2010 para contrastar com o Data Mart, que é um depósito menor de características interessantes derivadas de dados brutos. Ao promover o Data Lake, ele argumentou que os Data Marts enfrentam vários problemas inerentes, como a segmentação de informações.

A PricewaterhouseCoopers afirmou que os data lakes podem "eliminar os compartimentos de dados". Em uma pesquisa sobre reservatórios de dados, eles observaram que as empresas estavam iniciando a extração e o armazenamento de dados para análise em um único repositório que era baseado no Hadoop. Hortonworks, Google, Oracle, Microsoft, Zaloni, Teradata, Impetus Technologies, Cloudera e Amazon agora oferecem soluções de data lake.

Um Data Lake pode ser definido como um depósito centralizado que permite armazenar todos seus dados estruturados e não estruturados em qualquer escala. É possível manter seus dados como estão, sem a necessidade de organizá-los primeiro, e realizar vários tipos de análises, indo desde painéis e visualizações até processamento de dados volumosos, análises em tempo real e aprendizado de máquina para a orientação de decisões mais assertivas (AWS, 2024).

As companhias que conseguem obter valor comercial de seus dados têm desempenho superior em relação aos seus concorrentes. Uma pesquisa da Aberdeen mostrou que as organizações que implementaram um Data Lake superaram empresas semelhantes em 9% no crescimento orgânico da receita. Essas lideranças foram capazes de conduzir novos tipos de análises, como aprendizado de máquina, em novas fontes, como registros de eventos, dados de cliques, redes sociais e dispositivos conectados à Internet armazenados no Data Lake. Isso auxiliou a identificar e a agir diante das oportunidades de crescimento do negócio com mais

agilidade, atrair e reter clientes, aumentar a eficiência, manter dispositivos proativamente e tomar decisões fundamentadas.

À medida que as organizações constroem Data Lakes e uma plataforma de análise, elas precisam considerar uma série de recursos importantes. Os Data Lakes permitem importar qualquer volume de dados que possa chegar em tempo real e os dados são coletados de várias fontes e transferidos para o Data Lake em seu formato original. Dessa maneira, esse processo permite dimensionar dados de qualquer magnitude, economizando tempo na definição de estruturas de dados, esquemas e transformações (AWS, 2024).

Eles ainda permitem armazenar dados relacionais, como bancos de dados operacionais e dados de aplicativos de negócios, e dados não relacionais, como aplicativos móveis, dispositivos IoT e mídias sociais. Eles também oferecem a capacidade de compreender quais dados estão no lago por meio de rastreamento, catalogação e indexação de dados. Por fim, os dados devem ser protegidos para garantir a segurança de seus ativos.

Os Data Lakes permitem que várias funções em sua organização, como cientistas de dados, desenvolvedores(as) de dados e analistas de negócios, acessem dados com sua escolha de ferramentas e estruturas analíticas. Isso inclui estruturas de código aberto, como Apache Hadoop, Presto e Apache Spark, e ofertas comerciais de fornecedores de Data Warehouse e Business Intelligence.

Essas ferramentas também permitem que você execute análises sem a necessidade de mover seus dados para um sistema analítico separado. Elas permitem que as corporações gerem diferentes tipos de insights, incluindo relatórios sobre dados históricos e aprendizado de máquina, em que modelos são construídos para prever resultados prováveis e sugerir uma série de ações prescritas para alcançar o resultado ideal (AWS, 2024).

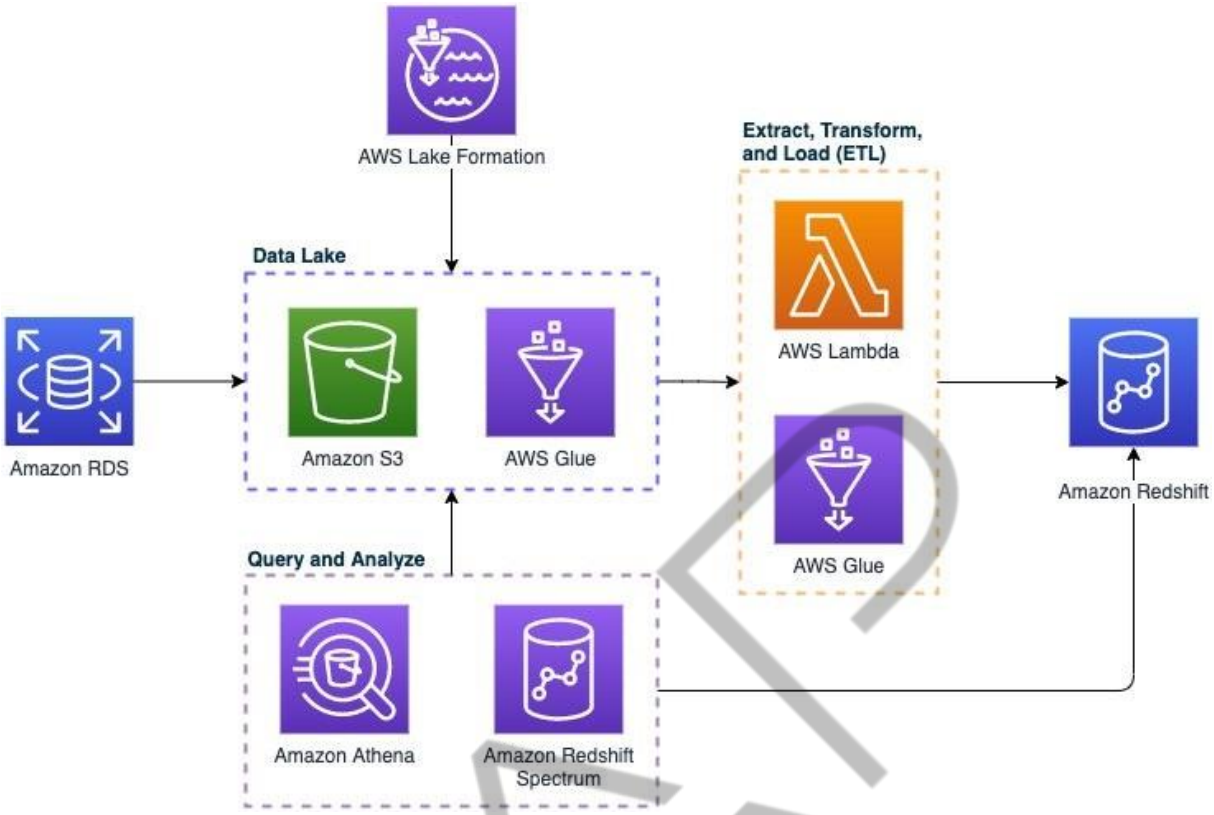

Figura 2 – Data Lake AWS
Fonte: Chen et al. (2021)

Data Warehouse

Os depósitos de dados surgiram como ideia acadêmica na década de 80. Com a evolução dos sistemas de informação empresarial, a necessidade de análise de dados cresceu concomitantemente. Dessa maneira, os sistemas OLTP não conseguiram atender às demandas de análise apenas com a geração de relatórios.

Nesse contexto, a implementação do Data Warehouse se tornou realidade nas grandes empresas. O mercado de ferramentas de Data Warehouse, que integra o mercado de Inteligência de Negócios, cresceu e ferramentas melhores e mais sofisticadas foram desenvolvidas para suportar tanto a estrutura do Data Warehouse quanto sua utilização.

OLTP

Sistemas OLTP (Processamento de Transações On-line) são sistemas encarregados de monitorar e processar as funções básicas e rotineiras de uma

empresa, como folha de pagamento, faturamento, controle de estoque etc. Os fatores críticos de sucesso para esse tipo de sistema incluem alto nível de precisão, integridade transacional e produção de documentos em tempo hábil. Os dados transacionais OLTP são utilizados por usuários no dia a dia em seus processos e transações, para escrita e leitura, como por exemplo consulta de estoque e registro de vendas.

O objetivo principal da modelagem relacional em um sistema OLTP é minimizar a redundância de maneira que uma transação que altere o estado do banco de dados atue da forma mais eficiente possível. Portanto, nas metodologias de projeto comuns, os dados são fragmentados em várias tabelas (normalizados), o que traz uma complexidade considerável à formulação de uma consulta por um usuário final.

Devido a isso, essa abordagem não parece ser a mais adequada para o projeto de um Data Warehouse, no qual estruturas mais simples, com menos normalização, devem ser buscadas.

OLAP

As ferramentas OLAP (Processamento Analítico On-line) geralmente são desenvolvidas para se trabalhar com bancos de dados normalizados, embora existam ferramentas que lidem com esquemas de armazenamento especiais, com dados normalizados. Elas conseguem navegar pelos dados de um Data Warehouse e contam com uma estrutura adequada para a realização de pesquisas e a apresentação de informações. Nas ferramentas de navegação OLAP, é possível navegar entre diferentes níveis de granularidade (detalhamento) de um cubo de dados.

Por meio de um processo chamado “Aprofundamento”, é possível aumentar ou diminuir o nível de detalhamento dos dados. Por exemplo: se um relatório estiver consolidado por países, ao aprofundar os dados passarão a ser apresentados por estados, cidades, bairros e assim por diante até o maior nível de detalhamento possível.

O processo contrário, a “Ampliação”, faz com que os dados sejam consolidados em níveis superiores de informação. Outra funcionalidade presente na maioria das ferramentas de navegação OLAP é um recurso nomeado como “Cortar” e “Fatiar”. Ele

é utilizado para criar visões dos dados por meio de sua reorganização, permitindo que sejam examinados sob variadas perspectivas.

O uso de recursos para manipular, formatar e apresentar os dados de maneira rápida e flexível é uma das principais características de um data warehouse. Isso faz com que a apresentação de relatórios na tela seja mais comum do que a impressão. Além disso, a pessoa usuária possui liberdade para explorar as informações de várias maneiras e, ao final, pode imprimir e até salvar as visões mais importantes para futuras consultas.

Mineração de Dados

Mineração de dados é o processo relativo à descoberta de padrões existentes em grandes volumes de dados. Apesar de haver ferramentas que ajudam na execução desse processo, a mineração de dados não é simplesmente automatizada (muitas pessoas discutem se é sequer viável) e precisa ser conduzida por uma pessoa, de preferência com formação em Estatística ou áreas correlatas.

Diferentemente do OLAP, a mineração de dados fornece informações sobre dados corporativos ocultos em grandes bancos de dados, o que possibilita prever comportamentos futuros, tornando-se uma ferramenta importante para a tomada de decisão de gestores e gestoras. Os tipos de informações obtidas com a mineração de dados incluem associações, sequências, classificações, aglomerações e previsões.

BI

Business Intelligence é uma forma de transformar dados em conhecimento. Informações de bases operacionais são coletadas, armazenando-as de forma modelada e, posteriormente, é possível realizar consultas por meio de ferramentas para fornecer informações que se traduzam em vantagem competitiva, empregando diversas ferramentas e metodologias.

Big Data

Com o surgimento da internet, o volume de dados gerados em todo o mundo aumentou de forma surpreendente ao longo dos anos. A adoção em larga escala de dispositivos móveis ampliou ainda mais a quantidade de dados gerados diariamente. Os métodos tradicionais de armazenamento e processamento de dados em grandes

empresas começaram a se tornar insuficientes, resultando em problemas e custos cada vez maiores para atender às suas necessidades.

Diante desses eventos, surgiu o conceito de Big Data, uma área de conhecimento destinada a estudar maneiras de lidar, analisar e extrair conhecimento de grandes conjuntos de dados que não podem ser processados em sistemas tradicionais. Para entender melhor o que é Big Data, podemos pensar na forma como o armazenamento e o processamento de dados são realizados no sistema tradicional.

Perceba que estamos falando no presente porque os processos de trabalho com Big Data não excluem a forma de trabalhar no sistema tradicional na maioria dos casos. Isso ocorre porque muitas empresas não precisam usar ferramentas de Big Data para manipular os dados e mesmo as grandes corporações utilizam um sistema híbrido. Assim, as duas maneiras de trabalhar com os dados coexistem. O sistema tradicional utiliza os conhecidos “SGBDs”, ou “Sistemas Gerenciadores de Banco de Dados“, que armazenam informações de forma estruturada em formato tabular, com linhas e colunas.

Eles usam máquinas com grande capacidade de armazenamento e processamento. Quando há necessidade de expandir a capacidade dessas máquinas, é preciso adicionar novos componentes de hardware para aumentar a memória e o processamento. Os problemas começam a surgir quando se atinge um grande volume de dados usando esse sistema tradicional e são relacionados à escalabilidade, disponibilidade e flexibilidade.

Como exemplos, podemos mencionar que é muito custoso aprimorar essas máquinas verticalmente sempre que for necessário realizar um upgrade; frequentemente nesse momento o sistema fica indisponível, pois a máquina está em processo de manutenção.

Para contornar os problemas grandes empresas pesquisaram um novo sistema que fosse escalável, surgindo então o Hadoop, uma forma de armazenamento e processamento distribuído. A ideia é utilizar clusters de máquinas ou agrupamentos de computadores. Individualmente, um único computador nesse cluster não tem um poder de processamento muito poderoso, mas, em conjunto, é possível fornecer poder de processamento e armazenamento capazes de atender às necessidades.

Nesse cluster, há uma máquina principal conhecida como “Nó Mestre” que é responsável por gerenciar o restante das máquinas, conhecidas como “Nós de Dados”. Os dados são replicados em diferentes nós de dados para que, caso uma máquina falhe, os dados não sejam perdidos e permaneçam sempre disponíveis. Esse conceito é conhecido em Big Data como disponibilidade.

O mais interessante é que, quando é necessário ampliar as capacidades, novas máquinas podem ser integradas ao cluster, crescendo de maneira indefinida. Essa é a escalabilidade horizontal, a solução encontrada para os problemas de Big Data. A partir do surgimento do Hadoop, muitas outras tecnologias foram desenvolvidas em paralelo, criando um ecossistema de ferramentas que se expande a cada dia.

Vale ressaltar o uso de bancos de dados NoSQL para lidar com dados não estruturados. Assim como na Ciência de Dados, são necessárias habilidades técnicas, bem como habilidades de comunicação e pensamento crítico. No aspecto técnico, as ferramentas de Big Data são variadas e podem gerar dúvidas sobre por onde começar a estudar. Confira a seguir alguns elementos essenciais.

O(a) profissional precisará aprender pelo menos uma linguagem de programação, como Python, R, Java ou Scala. Além disso, será necessário se familiarizar com frameworks como Apache Hadoop e Spark. Quanto aos Bancos de Dados, é necessário ter conhecimento tanto em bancos relacionais, quanto em NoSQL. Nesse caso, também é necessário conhecer sistemas de armazenamento distribuído.

O Hadoop é uma das principais estruturas para o processamento de Big Data. Portanto, é vantajoso conhecer o ecossistema do Hadoop com ferramentas como MapReduce, Hive, Pig e HBase. Por outro lado, existem várias plataformas em nuvem, como Google Cloud, Azure e AWS, que tendem a facilitar esse processo, além de permitir o armazenamento e processamento rápido de grandes volumes de dados.

Hadoop

O Apache Hadoop é uma estrutura de código aberto utilizada para armazenar e processar grandes conjuntos de dados com eficiência, com tamanho entre gigabytes e petabytes de dados. Em vez de usar um computador de grande porte para armazenar e processar os dados, o Hadoop permite o agrupamento de vários computadores em clusters para analisar em paralelo grandes conjuntos de dados.

O Hadoop consiste em quatro módulos principais (AWS, 2024):

- **Sistema de Arquivos Distribuído Hadoop (HDFS)**: um sistema de arquivos distribuído que opera em hardware padrão ou de baixo custo. Ele oferece melhor throughput de dados do que os sistemas de arquivos tradicionais, além de alta tolerância a falhas e suporte nativo a grandes conjuntos de dados.
- **Outro Negociador de Recursos (YARN)**: gerencia e monitora os nós do cluster e o uso de recursos, além de agendar trabalhos e tarefas.
- **MapReduce**: é uma estrutura que ajuda os programas a realizarem computação paralela em dados. A tarefa de mapeamento usa os dados de entrada e os converte em um conjunto de dados que pode ser calculado em pares de chaves/valores. A saída da tarefa de mapeamento é consumida por tarefas de redução para agregar a saída e fornecer o resultado desejado.
- **Comum do Hadoop**: fornece bibliotecas Java comuns que podem ser empregadas em todos os módulos.

O Hadoop facilita o uso de toda a capacidade de armazenamento e processamento em servidores de cluster e a execução de processos distribuídos em grandes volumes de dados. Assim, aplicações que coletam dados em vários formatos podem inserir dados no cluster do Hadoop usando uma operação de API para se conectar ao Nó Mestre, que rastreia a estrutura de diretórios de arquivos e o posicionamento dos "blocos" de cada arquivo, replicados em Nós de Dados.

Para executar um trabalho para consultar dados, forneça um trabalho MapReduce composto por vários mapeamentos e reduções de tarefas que são executadas com base nos dados do HDFS espalhados pelos Nós de Dados (AWS, 2024).

As tarefas de mapeamento são executadas em cada nó com base nos arquivos de entrada fornecidos e os redutores são executados para agregar e organizar a saída final. O ecossistema Hadoop cresceu significativamente ao longo dos anos devido à sua extensibilidade e atualmente inclui muitas ferramentas e aplicativos para ajudar a

coletar, armazenar, processar, analisar e gerenciar grandes conjuntos de dados. Algumas das aplicações mais populares são (AWS, 2024):

- **Spark**: um sistema de processamento distribuído de código aberto normalmente usado para cargas de trabalho de Big Data. Ele emprega o armazenamento em cache na memória e a execução otimizada para obter alta performance, além de oferecer suporte ao processamento geral em lotes, análises de streaming, aprendizado de máquina, bancos de dados de gráficos e consultas ad hoc.
- **Presto**: um mecanismo de consulta SQL distribuído de código aberto otimizado para baixa latência e análise de dados ad hoc. Ele suporta o padrão ANSI SQL, incluindo consultas complexas, agregações, junções e funções de janela. Ele pode processar dados de várias fontes, como o Sistema de Arquivos Distribuído Hadoop (HDFS) e o Amazon S3.
- **Hive**: permite que usuários utilizem o Hadoop MapReduce a partir de uma interface SQL, o que possibilita análises em grande escala, além de armazenamento de dados distribuído e tolerante a falhas.
- **HBase**: é um banco de dados de código aberto, não relacional e versionado que é executado no Amazon S3 (usando o EMRFS) ou no Sistema de Arquivos Distribuído Hadoop (HDFS). Ele pode ser definido como um armazenamento de Big Data distribuído e altamente escalável, criado para acesso aleatório, estritamente consistente e em tempo real a tabelas com bilhões de linhas e milhões de colunas.
- **Zeppelin**: é um notebook interativo que possibilita a exploração interativa de dados.

O QUE VOCÊ VIU NESTA AULA?

Nessa aula vimos como o AWS Glue nos ajuda a processar dados resolvendo parte dos problemas de ETL em nossos Data Lake e como isso é fundamental para que, no futuro, esses dados sejam consumidos pelos times de Machine Learning e IA.

Revise esse material quantas vezes forem necessárias e em caso de dúvidas entre em contato conosco! Obrigado e bons estudos.

REFERÊNCIAS

KIMBALL, R. **The Data Warehouse Toolkit**: The Complete Guide to Dimensional Modeling (Second Edition). [s.l.]: Wiley, 2002.

DIXON, J. **Pentaho, Hadoop, and Data Lakes**. 2010. Disponível em: <https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/>. Acesso em: 09 mai. 2024.

AWS. **O que é um Data Lake?** 2024. Disponível em: <https://aws.amazon.com/what-is/data-lake/>. Acesso em: 09 mai. 2024.

AWS. **O que é o AWS Glue?** 2024. Disponível em: <https://docs.aws.amazon.com/pt_br/glue/latest/dg/what-is-glue.html>. Acesso em: 09 mai. 2024.

AWS. **Conceitos do AWS Glue**. 2024. Disponível em: <https://docs.aws.amazon.com/pt_br/glue/latest/dg/components-key-concepts.html>. Acesso em: 09 mai. 2024.

CHEN, A.; GEYER, D.; MULLIS, J. **Reduce Operational Load using AWS Managed Services for your Data Solutions**. 2021. Disponível em: <https://aws.amazon.com/pt/blogs/architecture/reduce-operational-load-using-aws-managed-services-for-your-data-solutions/>. Acesso em: 09 mai. 2024.

AWS. **O que é o Hadoop?** 2024. Disponível em: <https://aws.amazon.com/pt/what-is/hadoop/>. Acesso em: 13 mai. 2024.

MIRANDA, J. **Uma breve introdução sobre o que é Big Data**. 2017. Disponível em: <https://site.alura.com.br/artigos/big-data>. Acesso em: 09 mai. 2024.

PALAVRAS-CHAVE

Palavras-chave: Big Data. ETL. ELT. Volumetria. Data Lake. Data Warehouse. Data Marts. Banco de Dados. BI. Tomada de Decisão. Engenharia de Dados. Ciência de Dados. Machine Learning. IA. Hadoop. Map Reduce. Data Mining. AWS Glue.