WASHINGTON LUIZ PERONI

MACHINE LEARNING ENGINEERING

BIG DATA CLOUD PLATFORMS

AULA 05

SUMÁRIO

O QUE VEM POR AÍ? 3
HANDS ON 4
SAIBA MAIS 5
O QUE VOCÊ VIU NESTA AULA? 11
REFERÊNCIAS 12

O QUE VEM POR AÍ?

Olá, pessoal! Sejam bem-vindos e bem-vindas à aula de AWS Data Firehose, um serviço de Streaming de Dados em tempo quase real que nos ajuda na ingestão e processamento de um grande volume de dados e em uma tomada mais rápida de decisão.

O Amazon Data Firehose era conhecido anteriormente como Amazon Kinesis Data Firehose. Este é um serviço totalmente gerenciado para fornecer dados de streaming em tempo real para destinos como Amazon Simple Storage Service (Amazon S3), Amazon Redshift, Amazon Service, Amazon Serverless, Splunk e qualquer endpoint HTTP personalizado ou endpoints HTTP de propriedade de provedores de serviços terceirizados compatíveis, incluindo Datadog LogicMonitor, Dynatrace, MongoDB, New Relic, Coralogix e Elastic OpenSearch (AWS, 2024).

Com o Amazon Data Firehose, você não precisa criar aplicativos ou gerenciar recursos: é possível configurar seus produtores de dados para enviar dados para o Amazon Data Firehose. Além disso, ele entrega automaticamente os dados para o destino que você especificou. Você também pode configurar o Amazon Data Firehose para transformar seus dados antes de entregá-los (AWS, 2024).

Vamos demostrar alguns casos de uso em conjunto com outros serviços como Athena e S3 para que vocês possam usar no mundo real. Então, sejam mais uma vez todas e todos bem-vindos à disciplina de Big Data e mãos aos dados!

HANDS ON

Demonstraremos na prática, dentro do console da AWS, o uso do AWS Data Firehose, mostrando desde a tela inicial e os conceitos básicos até um caso de uso real de integração com Athena e S3. Assim, evidenciaremos como o uso do AWS Data Firehose é importante para a ingestão em tempo quase real com um exemplo de minuto em minuto. As etapas serão as seguintes:

- AWS Data Firehose via console; configuração inicial.
- AWS Data Firehose: conexão com fonte de dados.
- AWS Data Firehose: ingestão de minuto em minuto.
- AWS Data Firehose: integração com S3 e Athena e consumo do streaming.

SAIBA MAIS

Aqui, veremos alguns conceitos fundamentais. Para total entendimento e alcance do que foi demonstrado nos vídeos, veja e reveja-os e complemente com essa leitura.

Principais conceitos

Batch (Lotes) e Streaming (Fluxo)

É o procedimento de reunir, processar e guardar dados em um sistema para análise posterior. A recepção de lotes e a recepção de fluxo são dois métodos essenciais para lidar com a entrada de dados em sistemas de processamento. Neste resumo, analisaremos as discrepâncias entre esses dois métodos, com ênfase nos serviços fornecidos pela AWS (Amazon Web Services).

Batch

na recepção de lotes, os dados são coletados e processados em blocos pré-determinados ou em lotes com intervalos regulares.

Exemplo de Serviços AWS:

- Amazon S3 (Simple Storage Service) para armazenamento de dados em lote.
- AWS Glue para alteração e preparação de dados.
- Amazon EMR (Elastic MapReduce) para processamento distribuído em grandes conjuntos de dados.

Vantagens:

- Eficiência no processamento de grandes volumes de dados em intervalos programados.
- Facilitar a administração de fluxos de dados. O custo operacional é menor em comparação com a recepção de fluxo para cargas de trabalho com baixa latência.

Desvantagens:

- Latência mais alta devido ao processamento em lotes.

- Capacidade inferior para lidar com dados em tempo real.
- Requer planejamento cuidadoso para determinar o tamanho ideal dos lotes e os intervalos de processamento.

Recepção de Fluxo

Na recepção de fluxo, os dados são processados à medida que são recebidos, permitindo processamento em tempo real e análise contínua. Alguns exemplos de Serviços AWS incluem Amazon Kinesis Data Streams para recepção e processamento de fluxos de dados em tempo real; AWS Lambda para execução de código sem servidor em resposta a eventos, como dados em fluxo; Amazon Data Firehose para entrega simplificada de dados quase em tempo real para destinos como Amazon S3 e Redshift.

Vantagens:

- Baixa latência, permitindo análises em tempo real e tomada de decisão imediata.
- Escalabilidade automática para lidar com picos de carga de dados.
- Ideal para casos de uso que requerem resposta em tempo real, como monitoramento de sistemas e análises de fluxo de cliques.

Desvantagens:

- Complexidade operacional maior devido à necessidade de lidar com eventos contínuos.
- Custos potencialmente mais altos devido ao processamento contínuo e à necessidade de recursos escaláveis.

AWS Data Firehose

Ao iniciar o uso do Amazon Data Firehose, você pode se beneficiar da compreensão de alguns conceitos (AWS, 2024). Um deles é o canal de fluxo, a base subjacente do Amazon Data Firehose. Você utiliza o Amazon Data Firehose criando um canal de fluxo e enviando dados para ele. Temos também os dados de interesse que seu produtor de dados envia para um canal de fluxo. Lembrando que um registro pode ter, no máximo, 1000 KB.

Já no caso do produtor de dados, os geradores enviam informações para os canais de fluxo. Por exemplo: um servidor web que envia registros de acesso para um canal de fluxo é um produtor de dados. Você também pode configurar seu canal de fluxo para ler automaticamente os dados de um fluxo de dados existente do Kinesis e carregá-los nos destinos.

Sobre tamanho e intervalo do Buffer, o Amazon Data Firehose armazena os dados de streaming recebidos em um tamanho específico ou por um determinado período antes de entregá-los aos destinos. O tamanho do buffer é medido em MBs e o intervalo do buffer é medido em segundos.

Por fim, há o fluxo de Dados. Para destinos no Amazon S3, os dados de streaming são entregues no bucket do S3. Se a transformação de dados estiver ativada, você também poderá fazer backup dos dados da fonte em outro bucket do Amazon S3 (AWS, 2024).

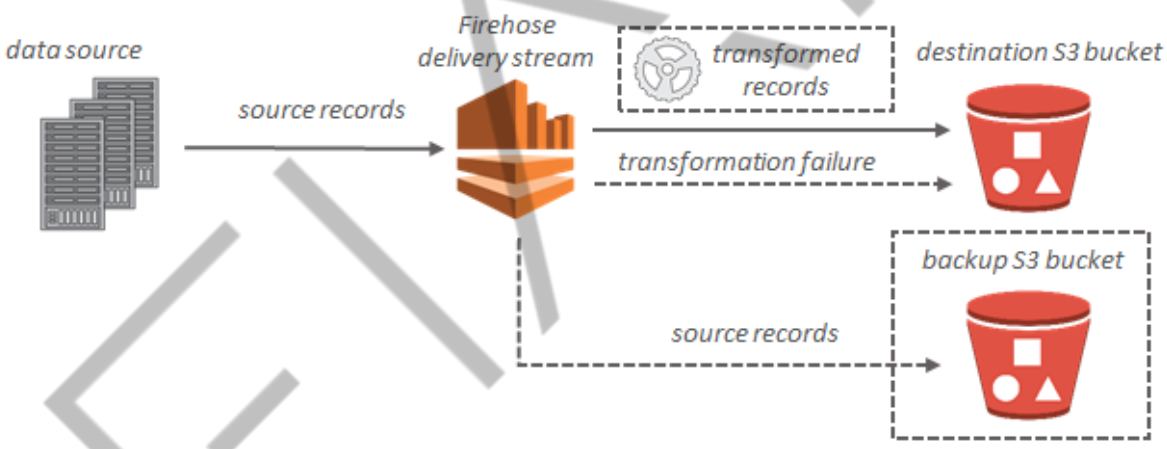

Figura 1 – Fluxo AWS Data Firehose até AWS S3
Fonte: AWS (2024)

Fluxo de transformação de Dados com Data Firehose

O Firehose armazena os dados recebidos em reserva. A sugestão de extensão da reserva varia entre 0,2 MB e 3 MB. A sugestão usual da extensão da reserva do Lambda é de 1 MB para todos os destinos, exceto o Splunk. Para o Splunk, a indicação de armazenamento em reserva usual é de 256 KB. A sugestão de espaço de tempo da reserva do Lambda varia entre 0 e 900 segundos.

Já a sugestão usual de espaço de tempo da reserva do Lambda é de sessenta segundos para todos os destinos. Para ajustar a extensão da reserva, defina o ProcessingConfiguration parâmetro da UpdateDestinationAPI CreateDeliveryStream

ou com o ProcessorParameter chamado e BufferSizeInMBs IntervalInSeconds Em seguida, o Firehose invoca a função Lambda especificada de maneira assíncrona com cada lote armazenado em reserva usando o modo de chamada síncrona. No AWS Lambda, os dados alterados são enviados do Lambda para o Firehose.

O Firehose então o encaminha para o destino quando a extensão do buffer de destino especificada ou o espaço de tempo da reserva é alcançado, o que acontecer primeiro. No formato de designação de coisa do Amazon S3, o Firehose acrescenta um prefixo de hora UTC no estilo YYYY/MM/dd/HH antes de gravar objetos no Amazon S3. Esse prefixo estabelece uma hierarquia lógica no bucket s3, no qual cada barra (/) cria um nível na hierarquia. É viável modificar essa estrutura especificando um prefixo adaptado.

Para obter informações sobre como especificar um prefixo adaptado, veja sobre prefixos adaptados para objetos do Amazon S3. A designação da coisa Amazon S3 segue o padrão DeliveryStreamName-DeliveryStreamVersion-YYYY-MM-dd-HH-MM-SS-RandomString, em que DeliveryStreamVersion começa com 1 e aumenta em 1 para cada mudança de configuração do fluxo de entrega do Firehose.

Você pode modificar as configurações do fluxo de entrega (por exemplo: o nome do bucket s3 do S3, as dicas de armazenamento em reserva, a compressão e a criptografia). Você pode fazer isso usando o console Firehose ou a operação da UpdateDestinationAPI.

O Amazon Data Firehose se integra às métricas CloudWatch da Amazon para que você possa recolher, ver e analisá-las para seus fluxos do Firehose. Por exemplo: você pode monitorar IncomingRecords, métricas IncomingBytes e acompanhar os dados ingeridos no Amazon Data Firehose pelos produtores de dados. O Amazon Data Firehose reúne e publica métricas CloudWatch a cada minuto.

Porém, se ocorrerem surtos de dados recebidos apenas por alguns instantes, elas podem não ser totalmente capturadas ou visíveis nas métricas de um minuto. Isso acontece porque no CloudWatch as métricas são agregadas do Amazon Data Firehose em intervalos de um minuto. Lembrando que as métricas reunidas para os fluxos do Firehose são livres.

Algumas outras vantagens do AWS Data Firehose incluem o fato de que este é um serviço manejado pela Amazon Web Services (AWS) que simplifica a entrega

de dados em tempo real para destinos como Amazon S3, Amazon Redshift e Amazon Elasticsearch.

Suas principais vantagens incluem (AWS, 2024):

- **Facilidade de Utilização**: o Data Firehose elimina a necessidade de escrever e manter um código complexo para a ingestão e entrega de dados em tempo real. Ele oferece configuração simples e integração direta com outros serviços da AWS.
- **Escalabilidade Automática**: o serviço é dimensionado automaticamente para lidar com picos de carga de dados, garantindo que a entrega seja confiável mesmo em situações de alto volume.
- **Integração com Destinos AWS**: o Data Firehose se integra perfeitamente com serviços populares da AWS, como S3, Redshift e Elasticsearch, simplificando o processo de armazenamento e a análise de dados em tempo real.
- **Gerenciamento de Dados Descomplicado**: com Dados Firehose, os usuários podem configurar facilmente pipelines de dados para limpar, transmutar e comprimir dados antes de entregá-los aos destinos desejados.
- **Exemplo de Caso de Utilização**: imagine uma empresa de comércio eletrônico que deseja analisar o comportamento dos clientes em tempo real para personalizar ofertas e melhorar a experiência do usuário. Utilizando o AWS Dados Firehose, a empresa pode capturar continuamente dados de interação do usuário em seu aplicativo ou site e entregá-los em tempo real ao Amazon Redshift para análise. Isso permite que a empresa identifique padrões de comportamento, tendências de compra e oportunidades de otimização instantaneamente, resultando em uma resposta mais ágil às necessidades de clientes e maior eficácia em campanhas de marketing.

Desvantagens

- **Limitações de Flexibilidade**: embora seja uma solução eficiente para casos de uso específicos, pode não ser adequado para cenários altamente personalizados ou complexos.

- **Custo Baseado em Utilização**: os custos associados ao uso do AWS Data Firehose podem aumentar com o volume de dados e a frequência de entrega, tornando-o menos econômico em comparação com soluções auto-gerenciadas em certos casos.

O QUE VOCÊ VIU NESTA AULA?

Nessa aula, vimos como o AWS Data Firehose nos ajuda na ingestão de dados em tempo quase real, possibilitando assim uma tomada de decisão mais rápida e confiável, aproveitando todo o potencial do Big Data.

Revejam a aula, pratiquem em casa e em caso de dúvidas entre em contato conosco! Obrigado e bons estudos.

REFERÊNCIAS

AWS. **AWS Big Data Blog**. 2024. Disponível em: <https://aws.amazon.com/pt/blogs/big-data/>. Acesso em: 13 mai. 2024.

AWS. **Amazon Data Firehose**. 2024. Disponível em: <https://aws.amazon.com/pt/firehose/>. Acesso em: 13 mai. 2024.

CHAKRABARTI, R.; GAGNE, D.; MAGUIRE, B.; MAKOTA, T. **Scalable Data Streaming with Amazon Kinesis**: Design and secure highly available, cost-effective data streaming applications with Amazon Kinesis. [s.l.]: Packt Publishing, 2021.

AWS. **O que é o Amazon Data Firehose?** 2024. Disponível em: <https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html>. Acesso em: 13 mai. 2024.

ZIKOPOULOS, P.; EATON, C. **Understanding Big Data**: Analytics for Enterprise Class Hadoop and Streaming Data. [s.l.]: McGraw-Hill Osborne Media, 2011.

PALAVRAS-CHAVE

Palavras-chave: Big Data. Streaming. Near-real-time. Batch processing. ingestão. ETL. AWS Data Firehose.