04.Velocidade: Processamento dos dados

segunda-feira, 4 de setembro de 2023 19:44

- Quando as empresas **precisam de informações rápidas** dos dados que estão coletando, mas os **sistemas** implantados simplesmente **não conseguem atender às necessidades**, há um problema de **velocidade**

- A coleta e o processamento dos dados são combinados em um único conceito conhecido como **processamento de dados**
- **Processamento de dados** = Coleta de dados + Manipulação dos dados

- **Tipos de processamento:**
  - **em lote ou batch** - processamento de dados em intervalos (Grande quantidade de dados, realiza-se em intervalos) - Tentativa de fraudes (Análises mais detalhadas)
    - Análises complexas
    - Ex: Logs de servidor, dados financeiros, relatórios de fraude, clickstreams
    - Todo o conjunto de dados é disponibilizado para consultas analíticas
    - Análises altamente complexas sejam executadas
    - Latência de minutos a horas
    - **Tipos de processamento em batch**
      - **Agendado -** Muitos volumes em rotinas regulares (Semanalmente)
      - **Periódico -** Aleatório (Imprevisíveis)
    - **Tipos de Armazenamento em batch**
      - Usado uma única aplicação, armazena os dados temporariamente em quanto está sendo processado, **etapa final** - Carregamento dos dados em um armazenamento para ser feito a análise.
      - Serviço: Amazon EMR
  - **processamento em streaming** - Processar dados em um fluxo contínuo
    - Análises iniciais - Em tempo real (Transações de cartão de crédito)
    - Feedback em tempo real
    - Insights contínuos
    - Dados do sensor de IoT
    - Compras de comércio eletrônico

        - Atividade do jogador no jogo
        - Clickstreams
        - Informações de redes sociais
    - **Tipos de processamento em Streaming**
        - **Tempo real -** Ocorre em milissegundos
        - **Próximo ao tempo real -** Ocorre em minutos
    - **Tipos de Armazenamento em Streaming**
        - Uso de vários serviços, 1 serviço para consumir o fluxo de dados constantes, 1 serviço para analisar o fluxo, 1 para carregar os dados em um data store se necessário.
        - Serviços: **Amazon Kinesis Data Firehose** e **Amazon Kinesis Data Streams** (Consumir e carregador)**, Amazon Kinesis Data Analytics**(Processar e Analisar o fluxo de dados).

OBS: Podem ser usados os 2 tipos no mesmo dataset

- **Amazon EMR** - Utiliza Apache Spark e Hive para processamento de dados complexas

Processamento em Batch

É a execução de uma séries de programas ou jobs em um ou mais computadores sem intervenção manual, os dados são coletados de forma assíncrona.

- **Cada batch** é enviado para um sistema de processamento quando condições específicas são atendidas (Ex: Quando um determinado horario do dia), os resultados são armazenados para serem consultados posteriormente quando necessário.
- Vantagens:
    - Pode ser executado quando há capacidade de baixo custo disponíveis
    - Com arquiteturas modernas, você pode otimizar sistemas de processamento em batch para a frequência e o tamanho dos lotes que você está processando
    - Permite priorização de trabalhos e o alinhamento da alocação de recursos com o objetivo de negócio.
    - Em um único batch pode haver milhões ou até bilhões de registros
    - Quando os requisitos dos sistemas de coleta e dos sistemas de processamento estão fora de equilíbrio demora horas para processar os dados.

- O Amazon EMR mitigou esse problema ao desacoplar o sistema de coleta, do sistema de processamento. Isso é feito implementando uma série de duas estruturas de trabalho em comum: **Hadoop ou Apache Spark**. Ambas as estruturas de trabalho processam os dados em alta velocidade, mas fazem isso de maneiras diferentes.
- Quando temos um Hadoop rodando no EMR, ele irá configurar um cluster de instâncias do EC2 para servir como uma única solução de armazenamento distribuído e processamento. Isso provê velocidade, tolerância a falhas e a habilidade de escalar separadamente as instâncias que coletam o dado, das instâncias que realizam o processamento.
- **O Apache Spark** é um framework diferente do **Hadoop**. A diferença é que o Spark usa o armazenamento em cache na memória e a execução é otimizada para uma performance mais rápida. As análises são primeiro realizadas filtrando os dados, depois agregando. O **Apache Spark** evita gravar dados no armazenamento, preferindo manter os dados na memória o tempo todo.
  Tanto o Hadoop quanto o Spark oferecem suporte ao processamento geral em batch, análise de streamings, machine learning, bancos de dados de grafos e consultas ad hoc.
- Usando o **AWS Lambda**, criaremos um programa que será executado uma vez a cada quatro horas, pra capturar quaisquer novos dados dentro do bucket do Amazon S3 e enviá-los para o **Amazon EMR** para processamento
- Assim que as análises e o processamento dos dados forem concluídos, os resultados serão enviados para um serviço chamado **Amazon Redshift**. **O Amazon Redshift** é um Data Warehouse rápido e escalável que torna simples e econômica a análise de todos os seus dados, de seu Data Warehouse e do seu Data lake.
- **O Amazon Glue** é um serviço de ETL totalmente gerenciado, que categoriza, limpa, enriquece e move dados de maneira confiável, entre vários armazenamento de dados. O Glue simplifica e automatiza tarefas difíceis e demoradas de descoberta, conversão, mapeamento e agendamento de jobs de dados. Em outras palavras, ele simplifica o processamento de dados.
- O Amazon EMR é um serviço gerenciado para a implantação de cargas de trabalho do Apache Hadoop. Além de executar o framework Apache Hadoop, você também pode executar outros frameworks distribuídos conhecidos, como Apache Spark, HBase, Presto e Flink no EMR. Você tem a vantagem adicional de poder interagir com dados em outros datastores da AWS, como o Amazon S3 e o Amazon DynamoDB.

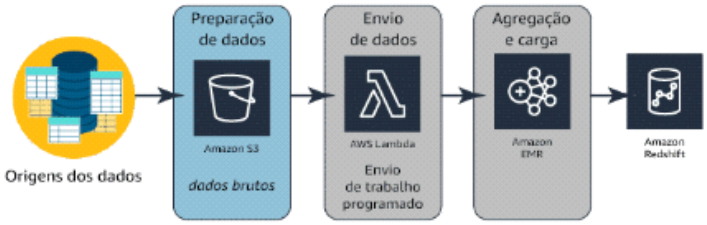
Apache Hadoop

- É um sistema escalável de armazenamento e processamento de dados em batch.
- Usa hardware de servidor de commodity
- Complementa sistemas de dados existentes, e processa simultaneamente grande volume de dados estruturados ou não, de qualquer quantidade de fonte
  - **Hadoop Common**
    - Biblioteca e utilitarios em Java
    - Suporte para outros modulos hadoop
    - Ajudam a abstrair o sistema de arquivos dos componentes de processamento
    - Arquivos e scripts Java são necessários para iniciar o Hadoop
  - **Hadoop Distributed File System (HDFS)**

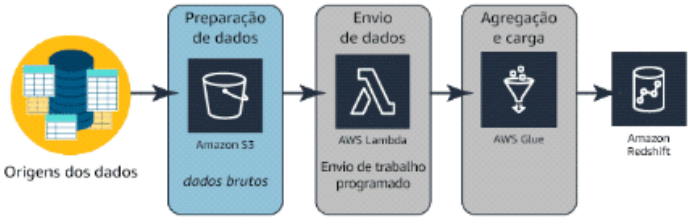
        - Armazena os dados em um ambiente de alta taxa de transferência
        - Garante acesso aos dados do app com largura de banda agregada alta
    - **Hadoop YARN**
        - Framework de gerenciamento de recursos responsável por programar e executar trabalhos de processamento
    - **Hadoop MapReduce**
        - Baseado no YARN permite processamento paralelo de grande conjunto de dados no cluster (Maquina)

Arquitetura em Batch

S3 - Serviço de armazenamento de objetos que oferece escalabilidade, disponibilidade de dados, segurança e desempenho lideres do setor
AWS Lambda - Consegue executar códigos sem servidores, paga apenas pelo tempo de computação consumido, não cobra quando o código não está em execução.
Amazon EMR - Fornece um framework que facilita o processamento rápido de grandes quantidades de dados em instancias EC2 dinamicamente escaláveis.
Amazon Glue - é um serviço de ETL que facilita o preparo e carregamento dos dados para análise.
Amazon Redshift - é um data warehouse rápido e escalável que torna simples e econômica a análise dos seus dados no DW(Data Warehouse) ou no DL(Data Lake).



Uso do **Amazon EMR**



Uso do **AWS Glue**

Processamento em Stream

processo de stream é a coleta e o processamento de um stream, um fluxo constante de dados, no processamento em batch a quantidade e a velocidade dos dados são relativamente estáveis, porém no Stream não, os dados não são consistentes.

o processamento de dados via streaming fornece às empresas a capacidade de obter dicas imediatas sobre os dados coletados.
Por exemplo, os dados de um sensor com porta de fechadura automática. Esse sensor pode enviar um sinal como um ping a cada 60 segundos para informar se a porta foi aberta nos últimos 60 segundos. Isso é importante para entender se alguém entrou na minha casa sem ter a permissão.

Ping, a porta está fechada? - Sim!
Está fechada? - Sim!
Está fechada? - Não! [grito] [fingindo som de alarme] Isso foi ótimo!

Benefícios do serviço e Streaming

- Os serviços de Stream tem como foco o desacoplamento onde difere a coleta, o produtor, do sistema que executa o processamento, soluções de streaming provêm um buffer persistente para os seus dados entrando, o dado pode ser processado e você manda pra onde for necessário.
- podemos gravar dados distintos em um único endpoint, a partir de vários produtores de stream. Isso permite fazer com que os dados de vários streams diferentes sejam combinados em um único stream de processamento. Por exemplo, uma solução de IoT que tem 1 milhão de dispositivos diferentes em todo o mundo, todos eles podem gravar os seus dados exclusivos no mesmo endpoint facilmente.
- Capacidade de preservar a ordem dos dados. Isso pode ser vital pra preservar a sequência de eventos em um stream. Por exemplo, o produtor vai, envia os dados na ordem: um, dois, três, quatro. O consumidor deve receber os dados na mesma ordem. Então, um, dois, três, quatro.

Como você processa os dados de streaming utilizando AWS?

- Usando um serviço chamado **Amazon Kinesis**. O Kinesis facilita a coleta, o processamento e a análise de dados de streaming em tempo real, permitindo que você obtenha insights oportunos e reaja rapidamente às novas informações. O Kinesis oferece recursos essenciais para processar dados de streaming em qualquer escala, de forma econômica. Além da flexibilidade de escolher a ferramenta mais adequada aos seus requisitos da aplicação.
- **Amazon Kinesis**
    - **Amazon Kinesis Data Firehose**
        - um serviço de extração, transformação e carregamento (ETL) que captura, transforma e entrega de forma confiável dados de transmissão para data lakes, armazenamento de dados e serviços analíticos, envia dados para o Kinesis Data Analytics.
        - O Amazon Kinesis Data Firehose é a maneira mais fácil de capturar, transformar e carregar streams de dados em datastores da AWS para análises quase em tempo real usando ferramentas existentes de business intelligence.

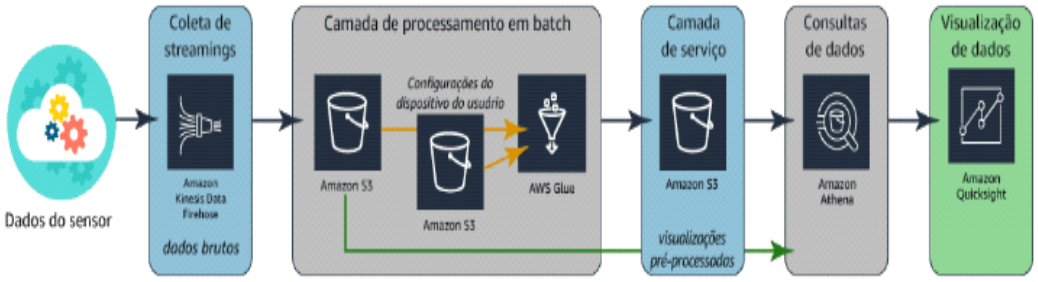
    - **Amazon Kinesis Data Stream** - coletar e processar grandescórregos de registros de dados em tempo real
    - **Amazon Kinesis Data Analytics**
        - é a maneira mais fácil de transformar e analisar dados de transmissão em tempo real com o Apache Flink
        - O Amazon Kinesis Data Analytics é a maneira mais fácil de processar streams de dados em tempo real com SQL ou Java sem

precisar aprender novas linguagens de programação ou frameworks de processamento.

    - **Amazon Kinesis Video Streams** - Facilita fazer stream de vídeos, seguro de dispositivos conectados a AWS, para análise de dados, machine learning ou outras análises.
- **Amazon Athena**
    - Serviço de consultas interativas que facilita a análise de dados no S3, utilizando SQL padrão.
    - É um Serveless (Computação sem servidor)
    - Pagamento apenas pelos dados escaneados pela consulta.
- **Amazon QuickSight**
    - Dashboards e relatórios detalhados
    - Em poucos minutos entrega soluções business em forma de insights

Podemos usar o **AWS Glue** para criar conjuntos de resultados combinados a partir dos dados coletados do stream do **Kinesis Data Firehose** e dos dados do dispositivo. Em seguida, consultar o **Amazon Athena**, visualizar com a **Amazon QuickSight**, como vimos na última arquitetura.

Uso do Kinesis Data analytics



Uso do AWS Glue

Agora que você viu as duas arquiteturas de forma independente, é hora de ver como elas trabalham juntas para formar uma solução completa.