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.


Os dados são coletados em batches
Os dados são processados em registros
Os registros são enviados para o armazenamento