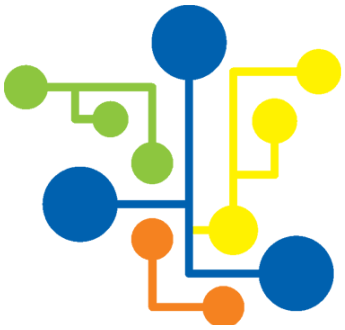

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Data Lake Processing

O processamento transforma os dados em um formato padronizado, útil para usuários corporativos e Cientistas de Dados, sendo trabalho do Engenheiro de Dados garantir que os dados sejam transformados e processados de acordo com os objetivos de análise. É necessário que, durante o processo de ingestão de dados em um Data Lake, o usuário não tome decisões sobre como transformar ou padronizar os dados. Em vez disso, tomamos esta decisão apenas quando o usuário fizer a leitura dos dados. Nesse ponto, os usuários do Data Lake possuem uma variedade de ferramentas para padronizar ou transformar os dados.

Um dos maiores benefícios dessa metodologia é que diferentes usuários de negócios podem realizar diferentes padronizações e transformações, dependendo de suas necessidades exclusivas. Ao contrário de um Data Warehouse tradicional, os usuários não estão limitados a apenas um conjunto de padronizações e transformações de dados que devem ser aplicadas na abordagem convencional de esquema por gravação. Nesse estágio, você também pode provisionar fluxos de trabalho para processamento repetitivos de dados.

Ferramentas apropriadas podem processar dados para casos de uso em lote e quase em tempo real. O processamento em lote é para cargas de trabalho tradicionais de extração, transformação e carregamento (ETL) - por exemplo, você pode querer processar informações de faturamento para gerar um relatório operacional diário. Streaming é para cenários em que o relatório precisa ser entregue em tempo real ou quase em tempo real e não pode esperar por uma atualização diária. Por exemplo, uma grande empresa de courier pode necessitar de dados de Streaming para identificar os locais atuais de todos os seus caminhões em um determinado momento.

Diferentes ferramentas são necessárias, dependendo se o seu caso de uso envolve lote ou streaming. Para casos de uso em lote, as organizações geralmente usam Pig, Hive, Spark e MapReduce. Para casos de uso de streaming, as empresas provavelmente usariam ferramentas diferentes, como Spark-Streaming, Kafka, Flume e Storm.

Podemos ainda usar outras ferramentas de integração como Apache Drill, Apache NiFi, Apache Beam e Apache Sqoop.