Resumo sobre material de Lakehouse – Grupo04

Bruno Sena  
Guilherme Patternezi  
Henrique Zigon  
Rhian Botura  
Thiago Yuiti  
Vinicius

O artigo nos diz sobre a arquitetura de dados chamada de “Warehouse” será substituída por um novo padrão de arquitetura, denominado por “Lakehouse”, se diferenciando nos seguintes aspectos:

* Suporte a “direct-acess data” como Apache Parquet e ORC, ambos formato de armazenamento de dados gratuito e de código aberto orientado a colunas do Hadoop.
* Suporte de primeira classe para Machine learning e bases para ciência de dados
* Performance no “estado da arte”

A história do Warehouse

O princípio de Warehouse começou para ajudar os líderes de negócio a conseguirem insights coletando dados de seus bancos de dados centralizados(warehouse). Este formato era utilizado para as tomadas de decisão e BI. Porém este modelo inicial era muito custoso, devido ao custo OnDemand e à gestão dos dados conforme o crescimento dos dados. Além disso, o crescimento de dados não estruturados não era englobado por tal modelo, limitando seu repertório de dados.

Como resposta, criaram-se os chamados “Datalakes”, tendo um sistema de armazenamento “low cost” e uma API de dados capaz de armazenar daods genéricos e comumente formatos como Apache Parquet e ORC. Tudo isso foi potencializado pelo movimento do Apache Hadoop com seu HDFS(Hadoop File System). Por ter uma arquitetura “Schema-on-read”, ela possibilitava um armazenamento de dados por baixo custo, mas agravou o problema de qualidade e governança dos dados.

Este segundo modelo funciona como uma fonte de dados única, onde um subconjunto era retirado para o tratamento(ETL) e após isso (dentro de um Warehouse), a tomada de decisão e BI poderiam ser feitos. Além disso, o dado passou a ser muito mais direcionado aos sistemas de machine learning.

Atualmente, o padrão HDFS foi fortemente substituído pelos modelos de Datalakes em núvem (S3, ADLS e GCS), pois sua durabilidade é maior, conseguem ter uma “geo-replicação” e um custo menor, tanto operacional ou para automações. Entretanto, a estrutura não mudou, e com isso, mantemos nossos atuais desafios:

* Manter um Warehouse confiável necessita de uma engenharia constante e cara. O processo de ETL corre risco de falhas ou bugs entre os sistemas usados.
* Dados desatualizados do Warehouse em comparação com o Datalake.(De acordo com uma pesquisa da Dimensional Research e Fivetran, 86% dos analistas usam dados desatualizados e 62% relatam espera em recursos de engenharia várias vezes por mês)
* Para realizar predições com os dados, é necessária uma base de dados grande, não usando o padrão SQL, onde o ODBC/JDBC passa a ser ineficiente para isso. Com isso, é recomendado que se estes dados sejam transformados em arquivos, o que acaba aumentando a complexidade (adiciona mais um ETL no processo) e acaba gerando mais custo para manter seus dados atualizados e confiáveis.
* Custo de se manter os dados em duplicidade e aos ETLs.