PONTÍFICA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Progama de graduação em Engenharia de Software - Resumo de Artigo

Victor Boaventura Goes Campos

Plataformas para análises de grandes dados: tendência para Era Híbrida

Belo Horizonte

2020

SWEBOK-O artigo pertence a área de Computing Foundations(Fundamentos da computação) por abordar temas como a estrutura e representação de dados (seção 6).

Vivemos em uma época que contém volumes grandes e complexos dados mantendo-se em plataformas e/ou sendo transionado,esse fato gera um papel importante nas mudanças das plataformas de gerenciamento de dados.Geralmente, essas plataformas são alteradas a fim de solucionar problemas que giram em torno da velocidade, volume , variedade e veracidade de dados.Atualmente é possível armazenar e utilizar esses conjuntos de dados com o assistência de sistemas distribuídos, onde elementos dos dados são mantidos em várias áreas e reunidos novamente por frameworks.

No Big Data, não há necessidade de colocar dados diretamente no banco de dados, pois essa tecnologia atual permite analisar os dados enquanto estão sendo gerados.Com essa tecnologia é possível aproveitar diversos tipos de dados e reuni-los novamente com os dados estruturados mais convencionais.Esse tipo de processamento é possível devido às características da tecnologia dos dados atualmente disponível.

O primeiro tópico se trata sobre o **escalamento de dados** que é a capacidade do sistema de lidar com o aumento da demanda de processamento de dados,de maneira geral as plataformas Big Data podem ser escaladas verticalmente e horizontalmente.

O escalamento horizontal inclui a distribuição de processamento em muitas máquinas combinadas,a escala horizontal é extremamente poderosa e pode ser expandida de forma ilimitada.alguns dos mais importantes pesquizadores dessas plataformas ponto a ponto(mais conhecida como Peer-to-peer) são o Apache hadoop e Spark.

Ao contrário da horizontal,a **escala vertical** consiste em só uma maquina com uma grande memória e hardware rápido,nessa escala o gerenciamento e a instalação de recursos é simples porém a capacidade de dimensionamento pode ser prejudicada a ponto de levar despesas gerais de invenstimentos financeiros.

O segundo tópico trata-se sobre a **nuvem**,uma inovação tecnológica muito poderosa para escalar dados complexos,isso ajuda na eleminação de gastos com hardware e software.O ponto crucial da computação em nuvem é a sua capacidade de prover qualquer tipo de infra-estrutura e variedade de ferramentas como por exemplo a sua capacidade de dados que pode ser ampliada e reduzida de acordo com a necessidade atual evitando disperdícios.

O proximo tópico aborda os gargalos de escalabilidade da extração de dados de antigas

gerações e plataformas abertas(open-source) atuais.Os algoritimos de extração de dados tem requerimento de que todos os dados devem estar disponíveis juntos e devem estar completos na memória principal para processamento.Mesmo se a memória for suficiente para carregar uma quantidade enorme de dados,o problema de escalabilidade não é resolvido pois a transferência de dados é cara e inviável.

A era atual é de plataformas abertas e elas desempenham um papel revolucionário no domínio do Big Data.Agora, há variedade de código aberto a nível empresarial distribuidos e amplamente disponíveis.

O ultimo tópico aborda as **integrações de plataformas**,um dos grandes desafios são incompatibilidade ao integrar plataformas heterogêneas juntas,este é o grande obstáculo em escalabilidade de DM.Nos últimos anos o aprendizado de máquina recebeu muito interesse para solucionar esse problema de escalabilidade.Essa soluções permitem uma integração consistente entre recursos e elimina o estágio de importação/exportação para ferramentas descutidas anteriormente como Hadoop,R,Python.A análise ded dados e muitas integrações são desenvolvidas usando: Rhadoop , Rhipe , Rhive , SparkR, Rpython , Psypark.Também emergiram variantes nesse comércio de integração de plataforma com hadoop como por exemplo: ORCH , ORAAH.

Justificativa da escolha do artigo

Quando eu olho para uma plataforma complexa paro e penso,pra onde vai todos essas dados? como o sistema faz para integrar esses dados? De longe parece simples,mas quando você pesquiza a fundo descobre-se uma evolução complexa e sem fim no gerenciamento de dados.

Importância para a Engenharia de Software

Os dados em geral são de extrema importância para nossa área poís tudo que nós progamamos e toda a informação que o usuário nos fornece é guardado em algum lugar e além da evolução tecnológica citada nesse artigo,temos questões de segurança de dados crucíais para a nossa área.

Bibliografia

https://ieeexplore.ieee.org/document/8390056

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8390056&tag=1