Curso 03 - Atividade Quinzenal
Quinzena 02 - Prof. Jose Fernando Rodrigues Junior

Leitura prévia:

The Five Key Differences of Apache Spark vs Hadoop MapReduce

Resilience and Vibrancy: The 2020 Data & AI Landscape

1) O Kylin é o projeto Apache cujo objetivo é a realização de consultas OLAP sobre uma infraestrutura Hadoop. Trata-se de um projeto capaz de lidar com Terabytes de informação, respondendo a consultas OLAP em segundos. Para alcançar tal desempenho, é correto afirmar que o Apache Kylin:

a) Precisa de processamento Spark sobre uma infraestrutura computacional com nós de processamento configurados com grande quantidade de memória RAM.

b) Realiza um pré-processamento de agregação segundo o algoritmo Layered Cubing. Este pré-processamento ocorre na construção do cubo de dados, podendo levar horas ou dias de processamento, dependendo do volume de dados e do poder computacional disponível.

c) Conta, necessariamente, com discos de estado sólido em todos os nós de processamento para acesso ultra-rápido aos dados.

d) Depende de processamento do tipo *cloud computing* advindo de *data centers* acessados via redes com altíssima largura de banda.

2) Com base no texto *The Five Key Differences of Apache Spark vs Hadoop MapReduce*, identifique a afirmação INcorreta:

(a) O Apache Spark tem desempenho potencialmente melhor do que o Apache Hadoop pois seu processamento é centrado em memória, ao passo que o Hadoop faz repetidas escritas e leituras em disco após suas operações de *map* e *reduce*.

(b) Por se basear em memória, o Spark suporta problemas que dependem de processamento iterativo, como em álgebra linear; ao passo que o Hadoop pode ter melhor desempenho em processamentos do tipo *batch* que passam pelos dados uma única vez, como em operações ETL.

(c) As APIs de processamento do Spark tornam seu uso mais simples do que o do Hadoop, fortemente atrelado ao paradigma map-reduce. Todavia, por ser uma tecnologia mais madura e central para diversos projetos que dependem da infraestrutura HDFS+map-reduce, o Hadoop possui uma quantidade maior de ferramentas que auxiliam seu uso.

(d) O Apache Spark pode substituir o Apache Hadoop em qualquer tipo de problema, provendo mais desempenho, segurança, e facilidade de uso. De fato, o uso de Spark permite que a infraestrutura Hadoop seja desnecessária ao fazer uso de soluções de processamento e armazenamento de outros fabricantes.

3) De acordo com o artigo Resilience and Vibrancy: The 2020 Data & AI Landscape, sobre o conceito de "*modern data stack*" é INcorreto afirmar que:

(a) A ideia de "*modern data stack*" admite uma nova compreensão e um novo *modus operandi* que advém do crescente uso de tecnologias de *cloud computing*; resumidamente, no novo *modern data stack*, os dados serão armazenados e processados em um *data center* cujos recursos se ajustarão dinamicamente à demanda.

(b) A abundância de recursos computacionais trazida pelo advento dos *data centers* tem permitido a prática de *Extraction Loading and Transformation* (ELT), em contraste à prática de *Extraction Transformation and Loading*. Segundo o ELT, os dados podem ser carregados sem maiores preocupações, sendo transformados posteriormente.

(c) Como o processo de *Transformation* tornou-se mais acessível graças aos *cloud data centers*, o mercado para Analistas de Dados versados em SQL e Python ampliou-se sobremaneira; estes analistas passaram a ocupar o espaço que antes era restrito aos Engenheiros de Dados.

(d) Mesmo com a crescente oferta de recursos computacionais - memória e processamento antes impensáveis, o *modern data stack* tende a entrar em um processo de completa estagnação como um protocolo computacional solidificado por décadas de prática bem estabelecida.