Big Data

DCA0133 - APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS (2018.1)
Mariana Beatriz Fonseca Alves
Felipe Ferreira Barbosa
Vanessa Dantas de Souto Costa

História

- O tema “Big Data", é um termo que descreve o grande volume de dados — tanto estruturados quanto não-estruturados — que impactam as empresas diariamente. Mas não é a quantidade de dados disponíveis que importa e sim, em que ou como podemos utilizá-los. Big data pode ser analisado para obter insights que levam a decisões melhores e ações estratégicas de negócio.

História

O conceito Big Data ganhou força no começo dos anos 2000, quando o analista Doug Laney articulou a definição de big data em três Vs:

• Volume: grande volume de dados é gerado na ordem de Zettabytes, 10^21.

• Velocidade: Os dados são transmitidos numa velocidade sem precedentes.

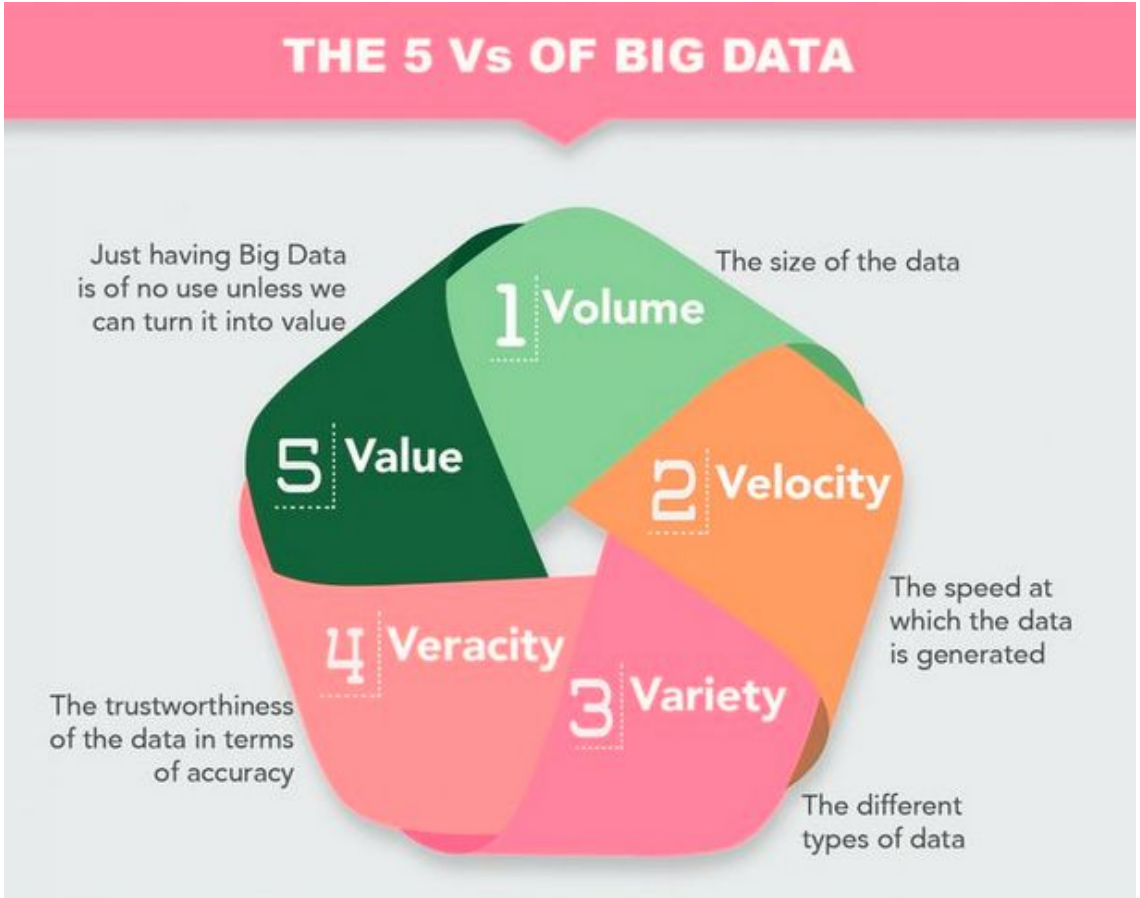
• Variedade: Dados são gerados em inúmeros formatos — desde estruturados (numéricos, em databases tradicionais) a não-estruturados (documentos de texto, e-mail, vídeo, áudio, cotações da bolsa e transações financeiras).

\quad No SAS, nós consideramos duas dimensões (mais 2 Vs) adicionais ao falar de big data:

• Veracidade: Quais os dados que realmente são importantes para reconhecimento de determinado padrão ou comportamento.

• Valor: O valor financeiro que os dados representam para uma determinada empresa.

História

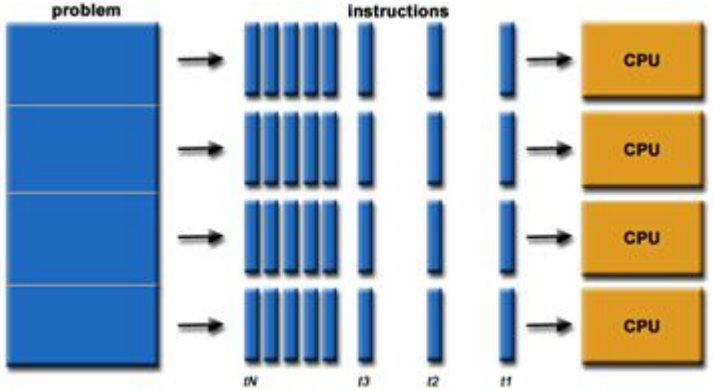

THE 5 Vs OF BIG DATA
Just having Big Data is of no use unless we can turn it into value
1 Volume
The size of the data
5 Value
2 Velocity
The speed at which the data is generated
4 Veracity
The trustworthiness of the data in terms of accuracy
3 Variety
The different types of data

Desafios do Big Data

- o volume de dados e a velocidade com que crescem, tornaram-se tão grandes que o processamento do Big Data se torna inviável em um único computador, por mais potente que seja.
- não é escalável
- consumo de energia consiste é um fator limitante
- limite físico de espaço para armazenamento de tanta informação

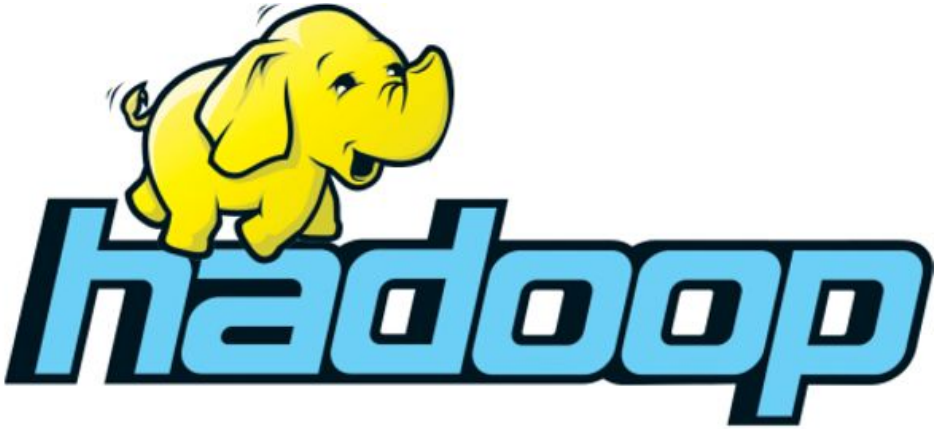
Computação Paralela é a solução?

- A utilização de computação paralela por meio de múltiplos computadores é uma “solução" inviável, pois é extremamente complexo, sendo necessário considerar a divisão e escalonamento de tarefas, balanceamento de carga, sincronismo entre tarefas, limitação na largura de banda, transferência de volumes de dados entre computadores. Em outras, palavras, essa solução seria eficiente apenas para pequenos volumes de dados.

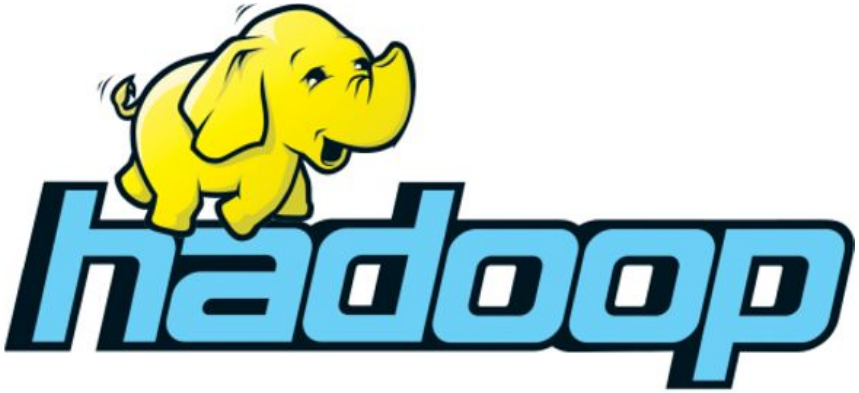


Ferramentas para trabalhar com Big Data

- No entanto, existem ferramentas pensadas para trabalhar exclusivamente com Big Data, exemplos que iremos abordar:
- Apache Hadoop
- Apache Spark
- MongoDB



Apache Hadoop



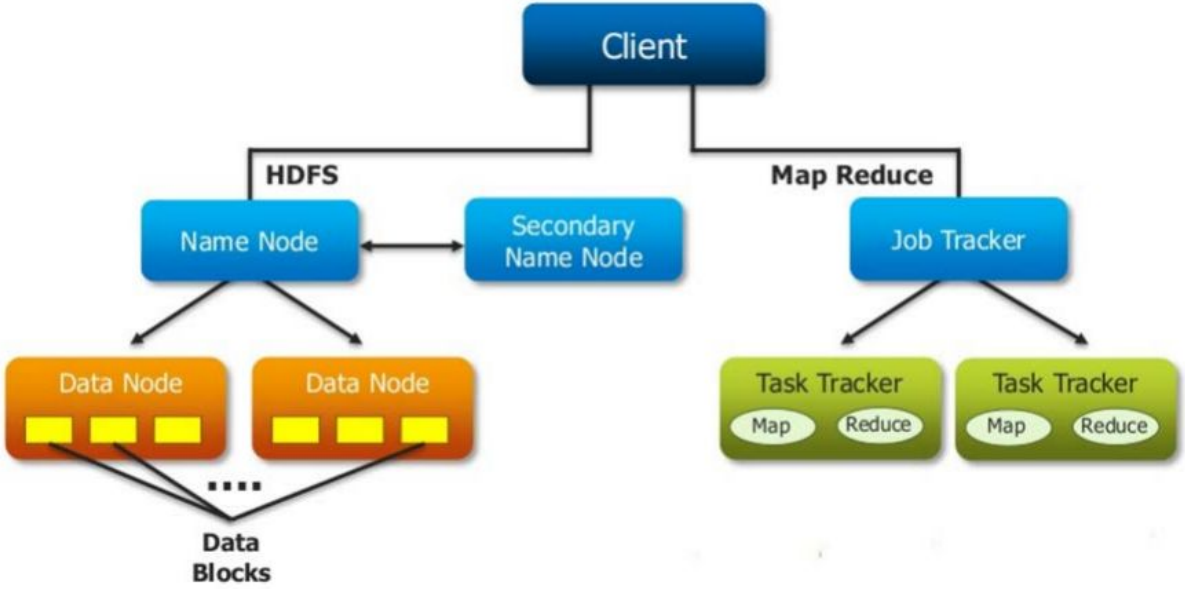
- O Hadoop é para problemas tão grandes que os sistemas tradicionais não são capazes de processar. Ele consiste em um Framework de código aberto criado em 2005 por Doug Cutting e Mike Carafella, desenvolvido em linguagem Java e projetado para armazenar e processar grandes volumes de dados em larga escala. Para tanto, ele utiliza computação distribuída.

- Algumas aplicações comuns do Hadoop: Processamento de texto em larga escala, aprendizado de máquina e mineração de dados, análise de dados em larga escala de redes sociais.

- As vantagens de sua utilização são: é gratuito, permite a utilização de computadores de baixo custo, não exige alterações na infraestrutura da rede (rede comum), é tolerante a falhas, de fácil uso e é escalável.

Apache Hadoop

- As componentes base do framework do Hadoop são HDFS (Hadoop Distributed File System), que serve para armazenamento distribuído, e o MapReduce (Computação distribuída). Podemos ver a arquitetura do Hadoop na imagem a seguir:

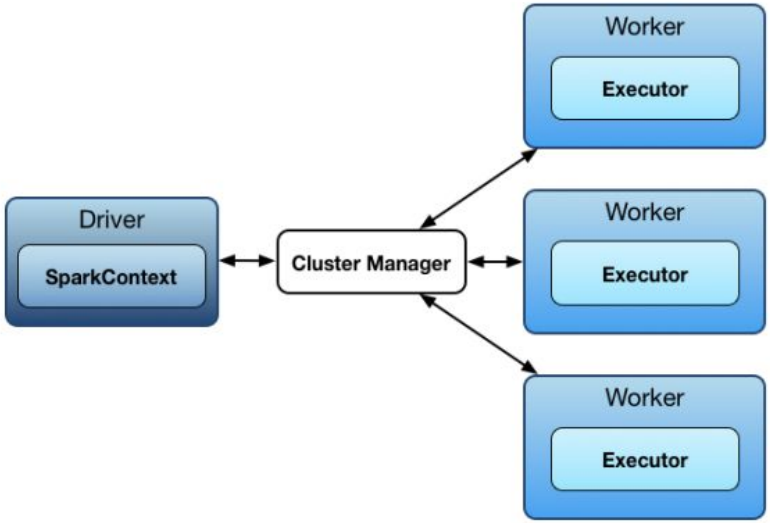


Apache Spark

- O Spark é um framework para análise de dados em tempo real em sistemas distribuídos (clusters). Ele realiza o processamento dos dados em memória, apresentando melhor desempenho que o Map Reduce. No entanto, apesar de ser mais rápido, requer mais poder de processamento (o Spark conta com uma série de bibliotecas de alto nível para processamento dos dados).

Apache Spark

- Sua arquitetura consiste em divisão para execução de uma tarefa entre Spark Executors (executam as tarefas) e Spark Driver (escalonam as tarefas para os executors). Conforme imagem:



- As tarefas do Spark podem ser executados com Yarn de dois modos: Modo cluster (Todas as tarefas são executadas no cluster, o Spark Driver é encapsulado no Application Master do Yarn), este é mais indicado para tarefas de processamento intenso (demoradas) e o Modo cliente (O Spark Driver é executado em um cliente, os Executors rodam no cluster), este é mais indicado para tarefas iterativas, mas as aplicações falham caso o cliente seja encerrado.

MongoDB

- O MongoDB é um SGBD NOSQL open-source e orientado a documentos. Ele possui diferenciais como:

  • Alto desempenho: documentos embutidos e índices atuando sobre eles;

  • Rica linguagem de consulta: permite operações CRUD, agregações de dados, busca por texto e consultas geoespaciais;

  • Alta disponibilidade: replica set;

  • Escalabilidade horizontal: sharding.

Aplicações Big Data

• Big data e Data mining: Mineração de dados (data mining) é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

• Big data na saúde: Atualmente, a pesquisa farmacêutica e os diversos dispositivos utilizados nos hospitais, tais como marcapassos e equipamentos sofisticados de diagnóstico da saúde, ampliam a quantidade de dados a um ritmo vertiginoso. Esses dados podem ser aproveitados para diagnosticar doenças e fazer prognósticos, melhorar o conhecimento sobre o estado de saúde de um dado paciente, prevenir epidemias e cuidar de modo mais apropriado das enfermidades, personalizando cada tratamento.

• Big data no setor financeiro: Complementando os dados provenientes dos modelos transacionais com dados sobre o comportamento dos clientes em canais múltiplos (correio eletrônico, pesquisas de qualidade de serviço, weblogs, call center, etc.), as instituições financeiras podem detectar mais facilmente novas oportunidades de negócios, compreender melhor as necessidades dos vários perfis de clientes, obtendo um leque de insights valiosos para abordá-los e expandir a sua clientela.

• Big data no marketing: Big data surge como uma ferramenta fundamental para ajudar os profissionais de marketing a lograr os objetivos de negócio. Entre as vantagens que oferece a análise de grandes volumes de dados estão: predizer comportamentos, segmentar campanhas e oferecer produtos e serviços personalizados.

Conclusão

Mais organizações estão armazenando, processando e extraindo valor de dados de todos os tipos e tamanhos. Os sistemas que oferecem suporte a grandes volumes de dados estruturados e não estruturados continuarão crescendo. Haverá uma demanda de mercado por plataformas que ajudem os administradores de dados a governar e proteger o Big Data e que permitam aos usuários analisar esses dados. Esses sistemas amadurecerão para operar de forma integrada com os padrões e sistemas de TI empresarial.

Ainda, outra assertiva sobre a necessidade, para engenheiros e cientistas da computação, de conhecer a área, ocorre pelo valor de mercado e utilização do Big Data. Isso, pois, grandes empresas como Google e Facebook estão investindo nesse tipo de conhecimento. Por fim, vemos que o desenvolvimento deste trabalho foi imprescindível para nossa formação acadêmica, pelos motivos citados acima.

Referências

Sites:

- Data Scientist vs Data Engineer, What’s the difference?
- https://cognitiveclass.ai/blog/data-scientist-vs-data-engineer/
- The Life of a Data Engineer
- http://www.mastersindatascience.org/careers/data-engineer/
- Dataquest
- https://www.dataquest.io/home
- DataCamp
- https://www.datacamp.com/courses/all
- Data Science Academy
- https://www.datascienceacademy.com.br/pages/todos-os-cursos-dsa

Referências

Livros:

•WHITE, Tom. Hadoop: The Definitive Guide. O'Reilly, 2015. 756 p.

• KARAU, Holden; KONWINSKI, Andy; WENDELL, Patrick; ZAHARIA, Matei. Learning Spark: lightning-fast data analysis. O'Reilly, 2015. 276 p.

•BENGFORT, Benjamin; KIM, Jenny. Analítica de Dados com Hadoop: uma introdução para cientistas de dados. Novatec, 2016. 352 p.

• HOWS, David; MEMBREY, Peter; PLUGGE, Eelco. Introdução ao Mongodb. Novatec, 2015. 168 p.