Universidade Federal do Rio Grande do Norte

DCA0133 -APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS - T01

Trabalho referente à Terceira Unidade
Questão 10 BIG DATA

Felipe Ferreira Barbosa - 20170008733
Vanessa Dantas de Souto Costa - 20170010188
Mariana Beatriz Fonseca Alvesz - 20160154335

June 15, 2018

1 Introdução

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.

Apenas para melhor contextualização, podemos afirmar que cerca de 90% de todos os dados gerados no planeta foram gerados nos últimos 2 anos. Aproximadamente 80% desses dados são não-estruturados, ou seja, estão em diferentes formatos e padrões, o que dificultam a análise (Textos em redes sociais, mensagens em sites, blogs, etc.). Atualmente, os modelos de análise de dados estruturados possuem limitações quando precisam tratar grandes volumes de dados (data warehousing), pois não foram projetados para lidar com dados não estruturados.

2 História

Embora o termo "big data" seja relativamente novo, o ato de coletar e armazenar grandes quantidades de informações para análises eventuais é muito antiga. O conceito ganhou força no começo dos anos 2000, quando o analista Doug Laney articulou a definição de big data em três Vs:

• Volume: Organizações coletam dados de fontes variadas, incluindo transações financeiras, redes sociais e informações de sensores ou dados transmitidos de máquina para máquina (na ordem de Zettabytes, 10^{21}).

• Velocidade: Os dados são transmitidos numa velocidade sem precedentes e devem ser tratados em tempo hábil. Etiquetas RFID, sensores e medições inteligentes estão impulsionando a necessidade de lidar com torrentes de dados praticamente em tempo real.

• 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).

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.

3 Desafios do Big Data

Trabalhar com Big Data fez com que surgissem diversos desafios a considermos. Entre eles, temos que 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.

Isso, porque podemos afirmar não é escalável (escalabilidade por definição é a capacidade de manipular uma porção crescente de trabalho de forma uniforme), bem como o consumo de energia consiste em um fator limitante, há limite físico de espaço para tantos discos rígidos.

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.

No entanto, existem ferramentas pensadas para trabalhar exclusivamente com Big Data, exemplos que iremos abordar:

- Apache Hadoop.
- Apache Spark
- MongoDB

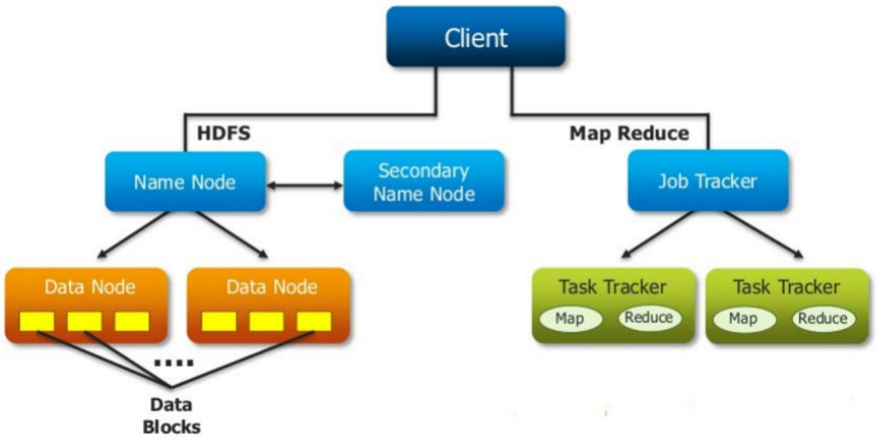
4 Apache Hadoop

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.

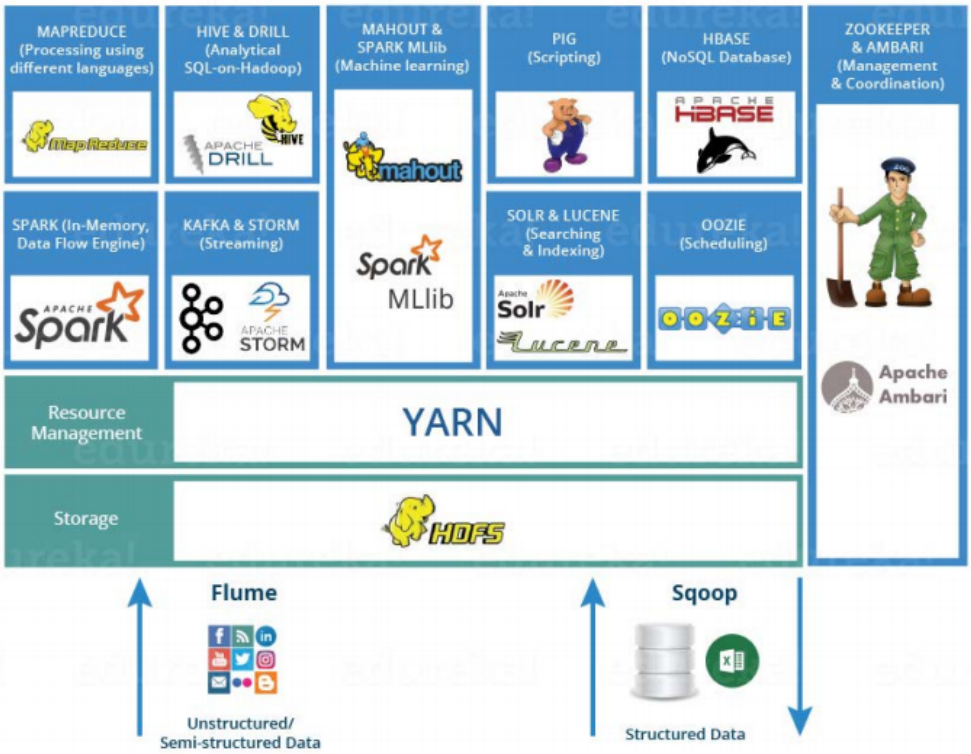
Algumas aplicações comuns do Hadoop: Processamento de texto em larga escala, aprendizado de máquina e mineração de dados, análise de dados em larga escala de redes sociais.

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.

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:



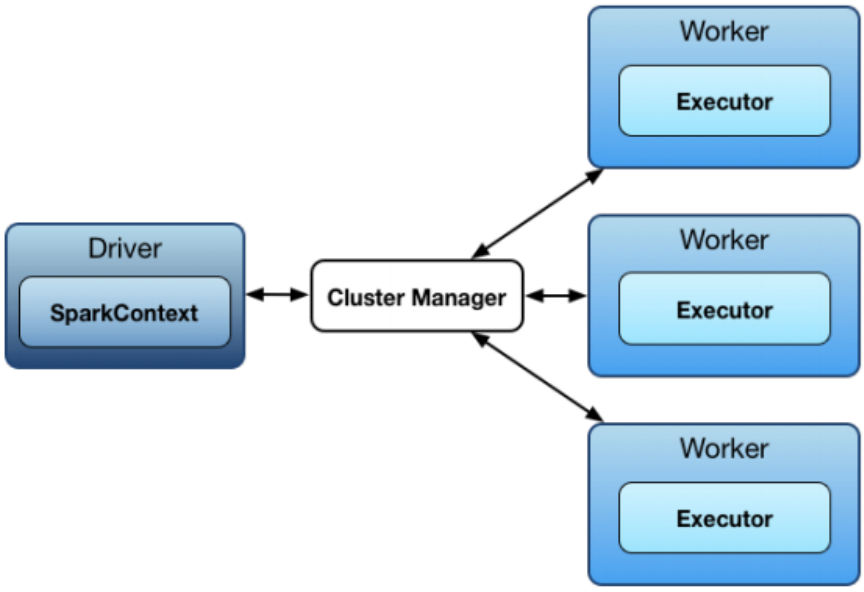
Nessa outra imagem, vemos o Ecossistema do Hadoop



5 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).

Sua arquitetura consiste em divisão para exucução 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.

6 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.

7 Aplicações do Big Data

Big data tem influenciado vários setores, transformando a sua realidade de modo significativo. A seguir, apresentam-se alguns casos de uso.

• 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. No campo da administração, a mineração de dados é o uso da tecnologia da informação para descobrir regras, identificar fatores e tendências-chave, descobrir padrões e relacionamentos ocultos em grandes bancos de dados para auxiliar a tomada de decisões sobre estratégia e vantagens competitivas.

• 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.

A enorme quantidade de dados gerados sobre os pacientes pode ser digitalizada e armazenada para análises e estudos compartilhados por organizações da saúde, hospitais e centros de pesquisa em nível mundial. Com o aproveitamento de recursos tecnológicos como big data, a pesquisa farmacêutica pode contar com conjuntos abrangentes de dados e melhorar a eficácia das drogas, ajudando a enfrentar enfermidades como diabetes e câncer.

• 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.

Os possíveis usos de big data na indústria financeira incluem análises com o objetivo de evitar a perda de clientes, mitigar riscos financeiros, detectar fraudes, obter maior rentabilidade, realizar campanhas mais efetivas de produtos e serviços e migrar para canais de marketing mais adequados para acessar uma dada clientela ou menos onerosos para a instituição.

• 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.

8 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.

Logo, a importância de entender o que é, como usar e quando utilizar Big Data se dá pelo imenso crescimento que esta abordagem tem experimentado nos últimos anos. Além disso, Big Data e Deep Learning têm uma "forte" ligação, pois este último é a base da análise preditiva ("predicitive analytics" or "predictive modelling").

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 impresas como a Google e Facebook estão investindo nesse tipo de conhecimento.

Por fim, vemos que o desenvolvimento desse trabalho foi imprescindível para nossa formação acadêmica, pelos motivos citados acima.

9 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

• Tendências e Aplicações Big Data:

• https://www.tableau.com/sites/default/files/whitepapers/849188_big_data_trends__slideshare_edits_pt-br.pdf?ref=lpsignin=c8e5308b0fc8ab7b1864fd9fdbdb3d5f

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.