Interações entre Big data e cloud computing

Gerenciamento de *Big Data*: coleta e processamento de dados

Gerenciamento dos dados e armazenamento

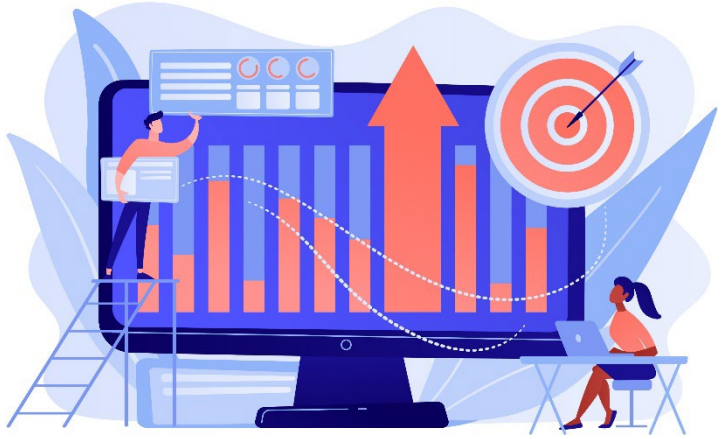
Bloco 1

Deivid Sardinha

Gerenciamento dos dados

- Lidar com dados é preciso ter cuidado em selecionar e entender o funcionamento do local de armazenamento. Os dados armazenados nos sistemas tradicionais passam por um rigor de tratamento até chegar à base de dados.

Figura 1 – Representação da análise de dados



Fonte: vectorjuice/Freepik.com.

Relatórios e *dashboard*

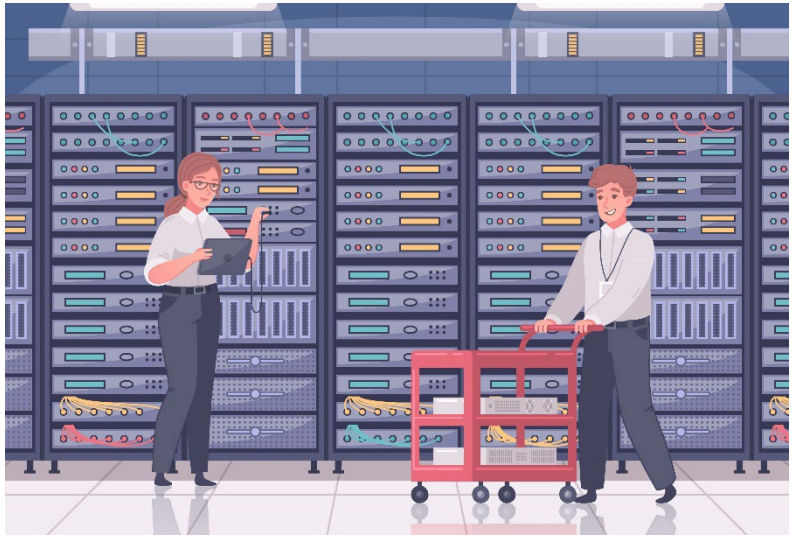
- Uma plataforma de *Big Data* pode armazenar todas as transações de um negócio para depois tentar obter valor por meio de processamento, embora, às vezes, essa estratégia não seja recomendada.

- É interessante deixar os dados adormecidos por um tempo no sistema *Big Data* e, quando descobrir um valor comprovado, confiável e sustentado, realizar a migração para o *warehouse*.

Cluster

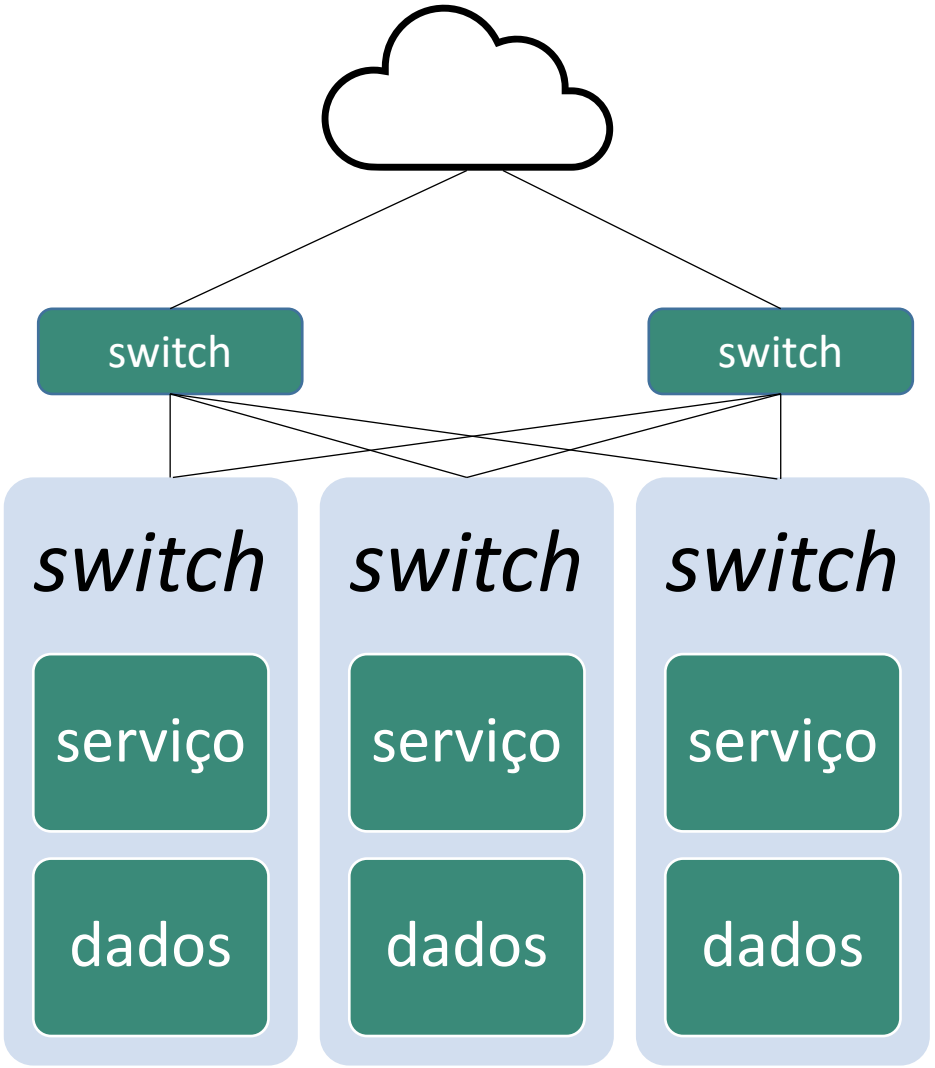
- *Clusters* de computadores são cada vez mais usados em *Big Data*, o que nos permite armazenar e processar por meio de diversos servidores.

Figura 2 – Representação de cluster de computadores



Fonte: macrovector/Freepik.com.

Armazenamento paralelo



Gerenciamento de *Big Data*: coleta e processamento de dados

Modelo de programação MapReduce e componentes do Hadoop

Bloco 2

Deivid Sardinha

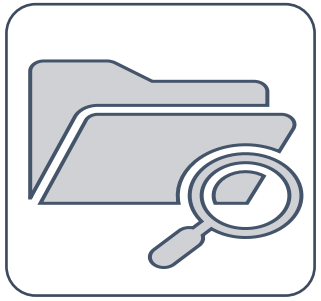
Hadoop

Figura 3 –Hadoop

Fonte: captura de tela de Hadoop.

Hadoop (MapReduce)

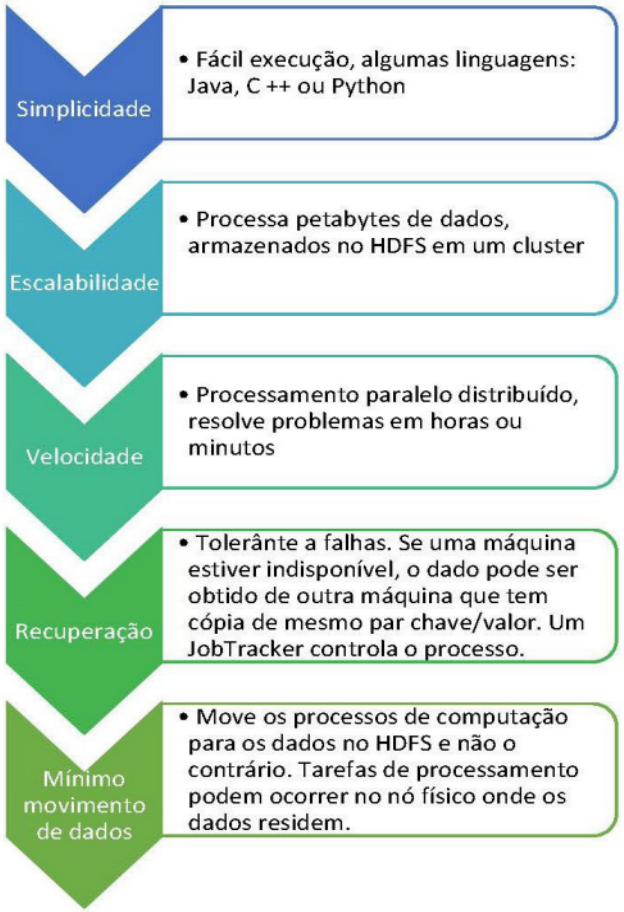
MapReduce



HDFS

MapReduce

Figura 4 – Caracterização do MapReduce



Fonte: adaptada de Hortoworks (2019).

Teoria em Prática

Bloco 3

Deivid Sardinha

Reflita sobre a seguinte situação

O Hadoop é importante atualmente, pois pode auxiliar as organizações nas suas decisões baseadas em dados em tempo real, pois possibilita execução de atividades com vários formatos de dados, sejam eles relacionamento e sentimento de mídia social, fluxo de cliques, streaming de vídeo e áudio e outros, podendo ser também um dado semi e não estruturado. Também impulsiona o futuro da ciência de dados, um campo interdisciplinar que combina aprendizado de máquina, estatística, análise avançada e programação. Permite transformar os dados do *data warehouse* local para um armazenamento distribuído baseado em Hadoop, consolidar os dados em toda a organização para aumentar a acessibilidade, diminuir os custos e acelerar as decisões com mais precisão (IBM, 2019). Faça uma reflexão sobre a importância da arquitetura Hadoop, se necessário, realize nova leitura do material de aula.

Componentes do Hadoop

- Hyve.
- Hadoop Stream.
- Avro.
- Flume.
- Lucene.
- ZooKeeper.

Dicas do(a) Professor(a)

Bloco 4

Deivid Sardinha

Leitura Fundamental

Prezado aluno, as indicações a seguir podem estar disponíveis em algum dos parceiros da nossa Biblioteca Virtual (faça o login através do seu AVA). Algumas indicações também podem estar disponíveis em sites acadêmicos como o Scielo, repositórios de instituições públicas, órgãos públicos, anais de eventos científicos ou periódicos científicos, acessíveis pela internet.

Isso não significa que o protagonismo da sua jornada de autodesenvolvimento deva mudar de foco. Reconhecemos que você é a autoridade máxima da sua própria vida e deve, portanto, assumir uma postura autônoma nos estudos e na construção da sua carreira profissional.

Por isso, te convidamos a explorar todas as possibilidades da nossa Biblioteca Virtual e além! Sucesso!

Indicação de leitura 1

A Apache Software Foundation desenvolve e mantém vários softwares de código aberto, e o Hadoop é um deles. Por meio da navegação pelo site, recomendo a leitura da seção *Overview* para explorar a configuração de um nó e de um *cluster*, conhecer comandos e sistema de arquivo. É sugerida especialmente a leitura da seção “HDFS” para entendimento do sistema de arquivo distribuído do Hadoop e a seção do Mapreduce, a estrutura de software tolerante à falha para desenvolvimento de aplicativo que trata grandes quantidades de dados em processamento paralelo em milhares de clusters de hardware.

Referência:

HADOOP. **Apache Hadoop 3.1.2**. [s.d.].

Indicação de leitura 2

Saber como funciona o sistema de armazenamento distribuído em massa do Hadoop é importante e desafiante. O artigo apresenta uma visão geral do HDFS, como ele trabalha para armazenar e recuperar os dados, além de explicar por que é tolerante a falhas etc. Investiga as interfaces com aplicativos e mostra os vários objetivos do HDFS. Descreve sua arquitetura e as relações entre os nós name node e data node, bem como a replicação e a integridade dos dados.

Referência:

HANSON, J. J. An introduction to the Hadoop Distributed File System. Explore HDFS framework and subsystems. **IBM**, 2011.

Dica do(a) Professor(a)

Com as ferramentas do Apache é possível iniciar rapidamente estudos com *Big Data*, seguindo tutoriais de casos reais.

Referências

PERRY, J. S. What is Big Data? More than volume, velocity and variety…. **IBM**, Endicott, 22 de maio de 2017. Disponível em: https://developer.ibm.com/blogs/what-is-big-data-more-than-volume-velocity-and-variety/. Acesso em: 3 jun. 2019.

PROVOST, F.; FAWCETT, T. Data Science for Business: What you need to know about Data Mining and Data-Analytic think. **PDF Drive**, [s.d.], 2013. Disponível em: https://www.pdfdrive.com/data-science-for-business-what-you-need-to-know-about-data-mining-anddata-analytic-thinking-d170193185.html. Acesso em: 30 ago. 2019.

Bons estudos!