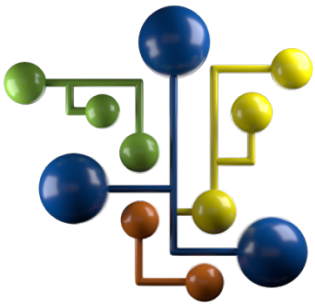


Data Science Academy

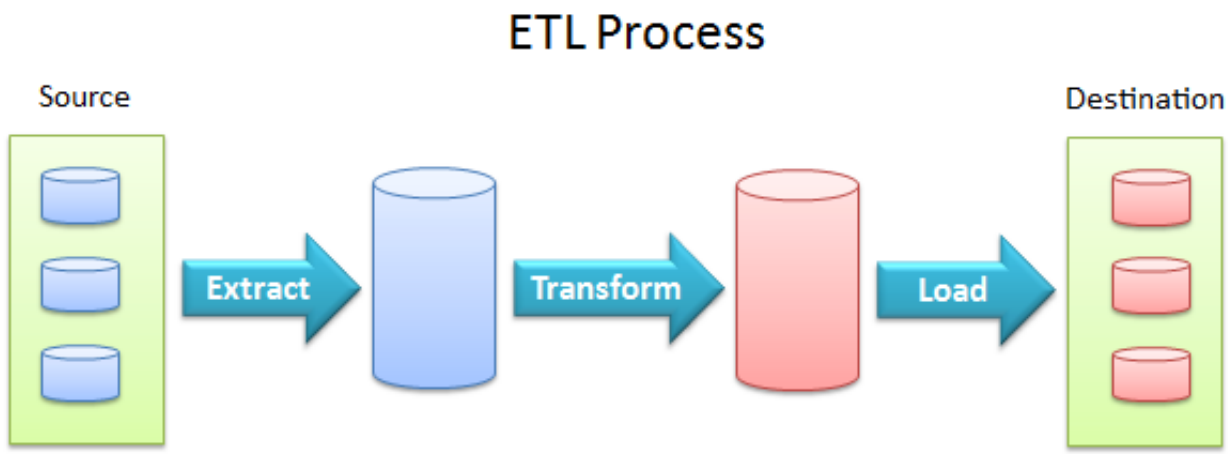
www.datascienceacademy.com.br

Microsoft Power BI Para Data Science, Versão 2.0

Apache Spark e Big Data

O Big Data está por todos os lugares. Dados são gerados por pessoas e máquinas de forma nunca antes vista pela humanidade. Cliques em web sites, pesquisas na web, cadastros, posts em redes sociais, mensagens de rede entre servidores, sensores ou até mesmo um simples pause que você dá em vídeo no Youtube. Tudo isso são dados gerados por bilhões de pessoas em todo mundo. E boa parte destes dados são gerados de forma não estruturada. Segundo estimativas, cerca de 80% dos dados gerados pela humanidade são não estruturados.

Para tratar esses dados não estruturados, utilizamos processos de ETL para coleta, transformação e carga de dados, para que possamos disponibilizar os dados para análise. Ou ainda utilizamos estruturas de Enterprise Data Hub para armazenar todo esse volume de dados, gerado em alta velocidade e com alta variedade, as propriedades do Big Data.



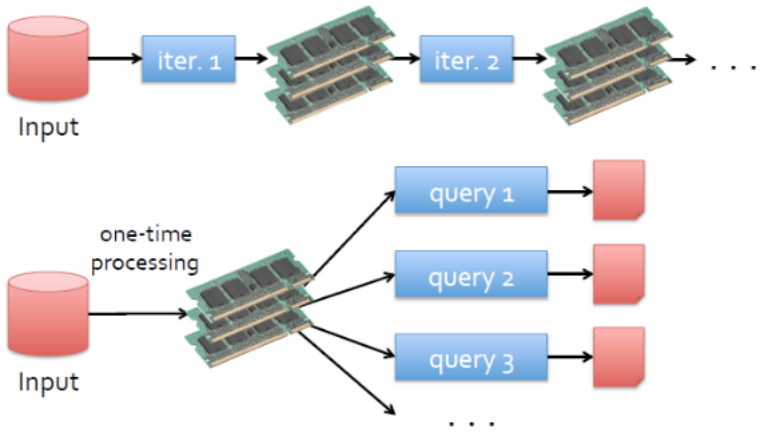
Mas aí nós temos um problema...

Como armazenar e processar todos esses dados, se o volume aumenta de forma exponencial? É praticamente impossível armazenar todos esses dados em apenas uma máquina, em apenas um servidor.

Por esse motivo, utilizamos cada vez mais clusters. Clusters são conjuntos de computadores (servidores) conectados, que executam como se fossem um único sistema. Cada computador no cluster é chamado node e cada node realiza a mesma tarefa, sendo controlado por software. Normalmente cada componente de um cluster é conectado através de redes locais (LAN's) e cada node executa sua própria instância de sistema operacional. Em nossa seção de links úteis você encontra o link para vídeo com o datacenter do Google, que mostra como essas a empresa armazena seus dados em grandes clusters de computadores. Vale a pena conferir.

Apache Spark é um sistema de análise de dados distribuído e altamente escalável que permite processamento em memória, e o desenvolvimento de aplicações em Java, Scala e Python, assim como linguagem R. É atualmente um dos principais projetos da Apache Foundation. O Spark estende as funcionalidades do MapReduce, suportando tarefas mais eficientes de computação como queries iterativas e processamento de streams de dados. Velocidade é importante quando processamos grandes conjuntos de dados e uma das principais características do Spark é exatamente sua velocidade, permitindo o processamento de dados em memória e ainda é bastante eficiente quando precisa processar dados em disco.

O Spark utiliza a memória distribuída através de diversos computadores, os nodes de um cluster. O preço da memória dos computadores vem caindo ano após ano e como o processamento em memória é muito mais rápido que o processamento em disco, o Spark é uma tecnologia realmente promissora. Enquanto o Hadoop armazena os resultados intermediários do processamento, em disco, o Spark armazena os resultados intermediários em memória. Esse é basicamente o grande diferencial do Spark.



Apache Spark é um framework open-source para processamento de Big Data construído para ser veloz, fácil de usar e para análises sofisticadas. Apache Spark é uma ferramenta de análise de Big Data, escalável e eficiente. O Spark é escrito na linguagem Scala e executa em uma máquina virtual Java. Atualmente, suporta como linguagens para o desenvolvimento de aplicativos, as linguagens: Scala, Java, Python e R.

A exemplo do Hadoop, o Spark pode ser integrado a diversas outras ferramentas, permitindo a criação de uma poderosa e gratuita solução para processamento de Big Data. Além da API do Spark, existem bibliotecas adicionais que fazem parte do seu ecossistema e fornecem capacidades adicionais para as áreas de análise de Big Data e aprendizado de máquina.

Para aprender mais sobre o Apache Spark, de forma 100% online e 100% em português, confira:

Big Data Real-Time Analytics Com Python e Spark
https://www.datascienceacademy.com.br/pages/curso-big-data-com-python-e-spark

Engenharia de Dados com Hadoop e Spark
https://www.datascienceacademy.com.br/pages/curso-engenharia-de-dados-com-hadoop

Machine Learning e IA em Ambientes Distribuídos
https://www.datascienceacademy.com.br/course?courseid=machine-learning-ia-ambientes-distribuidos