SISTEMA GERENCIADOR DE BANCO DE DADOS

AULA 6

Prof. Leonel da Rocha

CONVERSA INICIAL

Nesta etapa, veremos alguns assuntos extras relacionados a banco de dados. Por exemplo, os sistemas de *banco de dados distribuído* (BDD), que consiste em uma relação de nós ou servidores, em que cada um pode participar na execução de transações que acessam dados em um ou mais nós. Próximo assunto é sobre *data warehouse*, que é um repositório central de dados. Os dados surgem das transações, dos bancos de dados relacionais e de diversas outras fontes para o *data warehouse*. Veremos o que é um *data lake*, que é um lugar para armazenar dados estruturados e não estruturados, bem como um método para organizar grandes volumes de dados altamente diversificados de diversas fontes. Trataremos ainda sobre NoSQL, que é um padrão de armazenamento de dados alternativo ao SQL, oferecendo uma robustez e escalabilidade melhores. Para finalizar este estudo, estudaremos sobre ciência de dados, ou *data science*, que é um tema muito discutido entre profissionais e organizações que se concentram na coleta de dados e na elaboração de interpretações significativas para auxiliar no crescimento dos negócios.

TEMA 1 – BANCO DE DADOS DISTRIBUÍDOS

Um banco de dados distribuído (BDD) é composto por um conjunto de servidores que poderão ser acessados de forma ordenada e aleatória. No sistema de BDD, os dados poderão estar armazenados em servidores distintos, podendo ser acessados pelos sistemas que possuem permissão. Esses servidores de banco de dados são chamados de *nós* nessa tecnologia. Como acontece em outros sistemas conectados em uma rede de computadores, a comunicação dos BDD utiliza vários métodos para a sua interligação. Comparando o sistema centralizado com o distribuído, eles diferem pelo número de servidores que participam do processo. No *centralizado*, apenas um servidor de banco de dados e no *distribuído*, vários servidores poderão participar.

Os processadores que participam do sistema distribuído podem variar em tamanho e função, e o parque de *hardware* pode incluir servidores, microcomputadores, estações de trabalho e sistemas de uso genérico. Esses processadores são geralmente chamados de *nós*. O termo *nó* (lugar, posição) é utilizado para enfatizar a distribuição física desses sistemas.

Figura 1 – Sistema de Banco de Dados Distribuído

Crédito: Istel/Shutterstock.

O armazenamento distribuído de dados possui algumas características que trataremos agora. Uma relação r, como são conhecidas suas tabelas, possui diversos enfoques para o armazenamento em BDD:

- Replicação: os dados são replicados entre os servidores participantes do domínio;
- Fragmentação: os dados são particionados em tamanhos menores, podendo ser acessado em qualquer servidor do domínio;
- Replicação e fragmentação: nesse caso os dados são particionados em tamanhos menores e replicados entre os servidores do domínio.

A *replicação* de dados em um BDD, significa que suas relações podem ser armazenadas em vários servidores e seus dados replicados entre elas. Portanto a replicação em um BDD é uma funcionalidade primordial.

A *fragmentação*, por sua vez, é uma situação em que os dados não serão repetidos nas relações participantes da replicação e sim, serão divididos em volumes menores e distribuídos nas relações, permitindo por sua vez a reconstrução da relação original.

A fragmentação em um BDD pode ser realizada de duas maneiras:

- Fragmentação horizontal: os dados são divididos em partes menores sendo distribuídos em várias relações (linhas);
- Fragmentação vertical: divide a relação pelos seus atributos e os mantém em servidores diferentes (colunas).

Fragmentação e replicação podem ser aplicadas várias vezes em uma mesma relação, podendo um fragmento ser replicado e suas réplicas podem ser fragmentadas inúmeras vezes.

Um BDD mostra uma facilidade maior na leitura do que na atualização dos dados. Essa característica se dá pelo fato da maior dificuldade em garantir que todas as réplicas e fragmentos sejam atualizados após a realização de uma alteração. Todas as réplicas e fragmentos devem ser atualizados, sem exceção.

Comparativamente com um banco de dados centralizado, que apresenta como ponto chave do seu desempenho o acesso aos discos, os BDD apresentam outro problema característico da sua arquitetura, que além do acesso aos discos, que é a comunicação entre os servidores que formam os nós, e precisa ser tratado para que isso não se torne um impeditivo para a replicação dos dados e consequentemente para o seu não funcionamento.

Como já visto anteriormente, um BDD pode apresentar problemas de acesso aos discos como um agravante das falhas de comunicação. Outro fator que pode comprometer o seu funcionamento adequado é a perda de dados no momento da transmissão, ocasionando uma inconsistência na base de dados. A falta de comunicação e a perda de dados devem ser consideradas tanto no projeto de construção do BDD quanto no de recuperação. Um sistema de BDD, para ser robusto, deve detectar essas falhas e reconfigurar-se após sua recuperação.

Figura 2 – Sistema de banco de dados distribuídos, equipamentos participantes

Crédito: Net Vector/Shutterstock.

TEMA 2 – DATA WAREHOUSE

Um *data warehouse* é um repositório central de dados, que são gerados a partir das transações de diversas fontes. Profissionais da área de TI e os tomadores de decisão acessam esses repositórios por meio de ferramentas de *Business Intelligence*, que são sistemas especialistas que proporcionam um tratamento gerencial aos dados.

A análise de dados é uma ação importantíssima para que as empresas se mantenham competitivas. Para a realização da análise de dados, é possível contar com relatórios e *dashboards* para extrair *insights*, que auxiliam no monitoramento da *performance* dos negócios e também na tomada de decisões. Um *data warehouse* é fonte de consulta para essas ferramentas de análise, em que o armazenamento dos dados é realizado de forma eficiente, tendo como

característica a baixa carga de dados, sendo que estes são disponibilizados apenas para consulta, sem acesso para alterações.

Figura 3 – *Data warehouse*

Crédito: Nicescene/Shutterstock.

Um *data warehouse* é composto por dados que podem ser carregados de vários bancos de dados, que podem ser centralizados ou distribuídos. A grande característica de um *data warehouse* é que ele é um repositório de leitura de dados, não sendo realizada manutenção após a realização da carga de dados que geralmente é agendada em momentos de pouca utilização nas bases de dados de origem.

A utilização de um *data warehouse* apresenta os seguintes pontos:

- Permite a leitura de um grande conjunto de dados que contribuirão para a tomada de decisões;
- Os dados são carregados de várias bases de dados;
- Como tem uma grande massa de dados, permite analisar o histórico das operações;
- Permite ter dados com qualidade, consistência e precisão;
- Evita a leitura dos bancos de dados transacionais para análise.

Um *data warehouse* tem como principal objetivo a análise de dados em grandes volumes, possibilitando dessa maneira entender suas relações e tendências. Um banco de dados é utilizado para inserção e armazenamento de dados, e o *data warehouse*, por sua vez, é carregado com esses dados, que poderão ser lidos sem que alterações sejam feitas neles até a próxima carga, permitindo uma melhor *performance*.

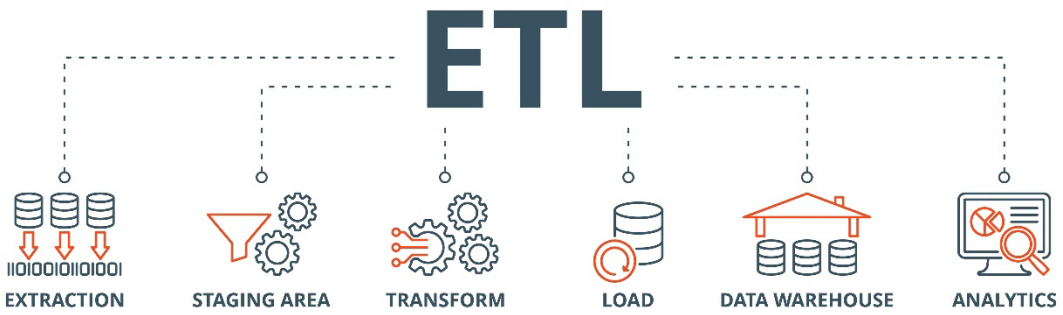
Em um *data warehouse,* os dados são armazenados e organizados em relações. O formato relacional é obrigatório para que se possa utilizar o SQL para as consultas, embora alguns aplicativos, como implementações de *Big Data*, pesquisas de texto e *machine learning,* conseguem acessar os dados mesmo que eles sejam *semiestruturados* ou completamente *não estruturados*.

O ETL é um tipo de integração de dados, que é realizado em três etapas: *extração*, *transformação* e *carregamento*. É utilizado para carregar dados de diversas fontes. Pode ser utilizado para construir um *data warehouse*. Nesse processo, os dados são extraídos de uma fonte, transformados para que sejam analisados e carregados em um *data warehouse*. Temos outro tipo de integração de dados, o ELT, que foi projetado para fazer a transformação dos dados depois do carregamento, melhorando dessa maneira a performance.

O processo de ETL é comumente utilizado para obter uma visão consolidada dos dados, possibilitando uma melhor tomada de decisões. Esse método que permite integrar dados de várias fontes compõe as ferramentas de integração de dados de uma organização:

- Permite acesso a dados históricos;
- Facilita a análise de dados e criação de relatórios;
- Melhora a produtividade analítica dos gestores;
- Suporta *streaming data*.

Figura 4 – ETL



Crédito: GaluhSekar/Shutterstock.

TEMA 3 – *DATA LAKE*

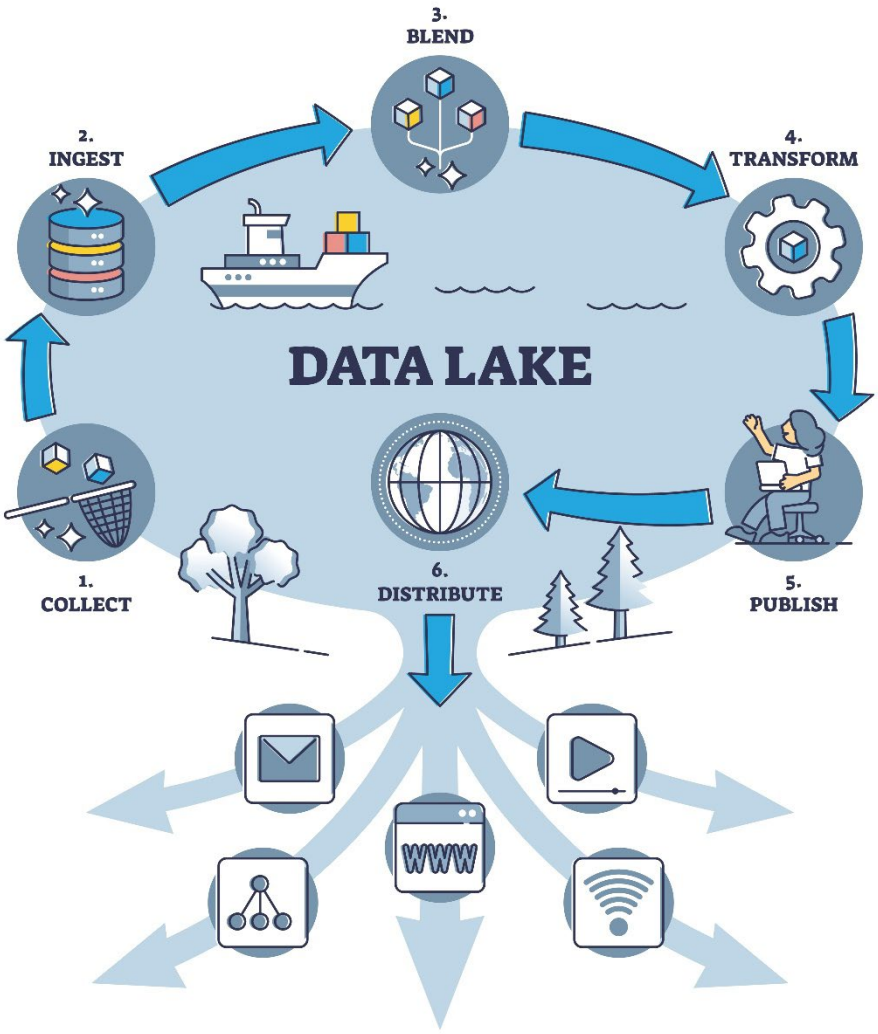
Um *data lake* é um conceito que descreve um local para o armazenamento de dados, sendo ainda um método para organizar grandes volumes de dados diversificados e de várias fontes. Esses locais de armazenamento estão cada vez sendo utilizados à medida que organizações exploram informações gerenciais em base de dados com um grande volume. Fazer uma concentração de dados ou a sua maioria em um único local torna essa exploração mais simples. O conceito *data lake* pode lidar com muitas estruturas de dados, não estruturados e multiestruturados, e pode auxiliar na extração de informações gerenciais para a tomada de decisão.

A diferença entre um *data lake* e um *data warehouse* é que no *data lake* os dados são armazenados mais rapidamente e preparados à medida que os usuários fazem o acesso. Em um *data warehouse*, os dados são preparados antes de serem armazenados.

Em um *data lake*, os dados são utilizados na sua forma original, principalmente em função do desempenho, mas também para análises mais avançadas que dependem de dados detalhados de forma original. Essa análise é baseada em qualquer tipo de mineração de dados, seja:

- Mineração de texto;
- Análise estatística;
- *Clusters*;
- Análise avançada de grafos.

Figura 5 – Data Lake



Crédito: VectorMine/Shutterstock.

Uma solução utilizando *data lakes* deve ter características para implementar as melhores maneiras de:

- Transformar diferentes tipos e formatos de dados;
- Garantir a segurança dos dados e que sejam acessados conforme necessidades;
- Utilizar a ciência de dados permitindo insights e tendências nos dados.

Usar um *data lake* para estender um *data warehouse* é comum no *marketing* multicanal, sendo uma abordagem integrada de uma marca para cada ponto de contato com o cliente nos canais disponíveis. A forma de pensar sobre

todo o conjunto de dados que compõe o *marketing* é que cada canal pode ter seu banco de dados e isso pode acontecer também com cada ponto de contato.

Dados também podem ser adquiridos, como dados demográficos e adicionais sobre clientes e futuros clientes, permitindo dessa maneira que o profissional de *marketing* enxergue de forma completa cada cliente, possibilitando a criação de campanhas de *marketing* personalizadas e direcionadas.

Esse universo de dados é bastante complexo e a todo tempo vem crescendo em volume e complexidade. O *data lake* é trazido muitas vezes para capturar dados que vêm de vários canais e pontos de contato. Aplicativos para *smartphone* podem transmitir dados, para as empresas, em tempo real conforme são utilizados. Isso permite que o departamento de *marketing* monitore de maneira granular o negócio e crie especialidades, incentivos, descontos e micro campanhas.

Os suprimentos digitais são igualmente diversificados e o *data lake* pode auxiliar no processo de junção dos dados, particularmente quando o *data lake* está no *Hadoop*. O *Hadoop* é em grande parte um sistema baseado em arquivos porque foi originalmente projetado para arquivos de *log* muito grandes e altamente numerosos provenientes de servidores *web*. Na cadeia de suprimentos, geralmente há uma grande quantidade de dados baseados em arquivo. Pense em dados baseados em arquivos e em documentos de sistemas EDI, XML e JSONs muito fortes na cadeia de suprimentos digital.

Informações de chão de fábrica, da expedição e do faturamento, que são altamente relevantes para a cadeia de suprimentos, podem ser reunidas por meio de um *data lake* para que se possa gerenciá-las de uma maneira baseada em arquivos.

A Internet das Coisas cria novas fontes de dados diariamente. Conforme essas fontes se diversificam, mais dados são criados. Existem mais sensores em máquinas e veículos, por exemplo, em caminhões de carga existe uma grande quantidade de sensores para que seja possível rastreá-los fisicamente, além da sua operação, verificando se está sendo feita com segurança e de forma econômica em relação ao consumo de combustível. Como podemos observar, existe uma grande quantidade de dados vindo desses dispositivos e sensores, e o *data lake* fornece um repositório para todos esses dados.

Existem outras abordagens que envolvem o departamento de tecnologia da informação fornecer um grande *data lake* multilocatário, podendo ser utilizado por diversos departamentos, unidades de negócios e sistemas diferentes. À medida que nos acostumamos com a utilização do *data lake*, descobrimos como otimizá-lo para diversos fins, operações e análises.

O *data lake* pode ser usado de várias formas além de muitas plataformas que podem estar sob ele. Aqui podemos citar o *Hadoop*, que é a plataforma mais comum, não sendo a única. A plataforma *Hadoop* é bastante interessante e já provou que tem escalabilidade linear. Possui custo baixo para a escalabilidade, se comparado com um banco de dados relacional. Mas o *Hadoop* não é apenas armazenamento barato. É também uma plataforma de processamento avançada. E para aqueles que fazem análises algorítmicas, o *Hadoop* pode ser muito útil.

Figura 6 – *Hadoop*

Crédito: Maslakhatul Khasanah/Shutterstock.

Um bom e velho sistema de gerenciamento de banco de dados relacional pode ser utilizado como uma plataforma para o *data lake*, pois algumas pessoas têm grandes quantidades de dados que desejam colocar no *data lake* estruturado e relacional. Portanto, se os dados forem relacionais, uma abordagem do tipo *Database Management System*, que é um sistema que auxilia no gerenciamento, segurança e manipulação das informações dentro do banco de dados, de forma centralizada e organizada, para o *data lake* faria todo sentido. Além disso, para os casos de uso em que se deseje fazer funcionalidade

relacional, como SQL ou junções de tabelas complexas, o RDBMS (sistema de gerenciamento de banco de dados relacional) é muito indicado para a implantação de um *data lake*.

TEMA 4 – NoSQL

NoSQL é um sistema gerenciador de bando de dados não estruturados, que se apresenta como alternativa de mercado aos sistemas gerenciadores de banco de dados estruturados. Esse termo foi utilizado pela primeira vez como o nome de um sistema de gerenciador de banco de dados não relacional de código aberto. O responsável por esse primeiro gerenciador, Carlo Strozzi, defende que o movimento NoSQL é totalmente diferente do modelo que implementa as tabelas relacionadas, defendendo que ele deveria ser chamado de uma maneira mais apropriada de *NoREL*, que significa *sem relacionamento*, que é uma funcionalidade dos sistemas gerenciadores de banco de dados relacionais.

Figura 7 – Banco de dados NoSQL

Crédito: Fatmawati Achmad Zaenuri/Shutterstock.

Com o avanço da tecnologia da informação, houve um incremento na produção e armazenamento de dados, tornando o tratamento deles mais difícil e encarecendo sua manutenção. O artigo BigTable: A Distributed Storage

System for Structured Data, publicada pelo Google em 2006, reviveu o conceito NoSQL. Em 2009, o termo *NoSQL* novamente aparece por meio de um funcionário do Rackspace, Eric Evans, quando Johan Oskarson da Last.fm tratou bancos de dados *open source* distribuídos em um evento.

O termo *NoSQL* foi utilizado para identificar os bancos de dados não relacionais, aproveitando a utilização nos nomes dos bancos de dados relacionais, como MySQL, MS SQL e PostgreSQL, que utilizam o codinome *SQL* em seus nomes comerciais. Como esses bancos não relacionais vão em outra direção dos relacionais, optou-se por utilizar esse termo, como que contrariando a utilização do termo SQL. Com essas características diferentes entre esses bancos de dados, o mercado adotou esse termo e passou a chamar os bancos de dados não relacionais de *NoSQL*.

Com o crescimento das mídias sociais e o aumento da geração de dados, o trabalho de armazenamento de dados para utilização em ferramentas analíticas esbarrou em questões de escalabilidade e custos de manutenção, e nessa situação os sistemas gerenciadores de banco de dados NoSQL tiveram um amplo caminho para o crescimento, pois os sistemas de banco de dados relacionais escalam, porém, quanto maior o tamanho, mais caro é essa escalabilidade. Já os sistemas de banco de dados não relacionais permitem escalar de uma forma mais barata e descomplicada, não exigindo equipamentos superdimensionados, além de um número menor de profissionais para a manutenção do sistema.

Os sistemas gerenciadores bancos de dados não relacionais estão se popularizando porque possuem funcionalidades que permitem trabalhar com dados semiestruturados ou não tratados, com origens de diversas fontes, tais como registros de *log*, *websites*, multimídia e mídias sociais.

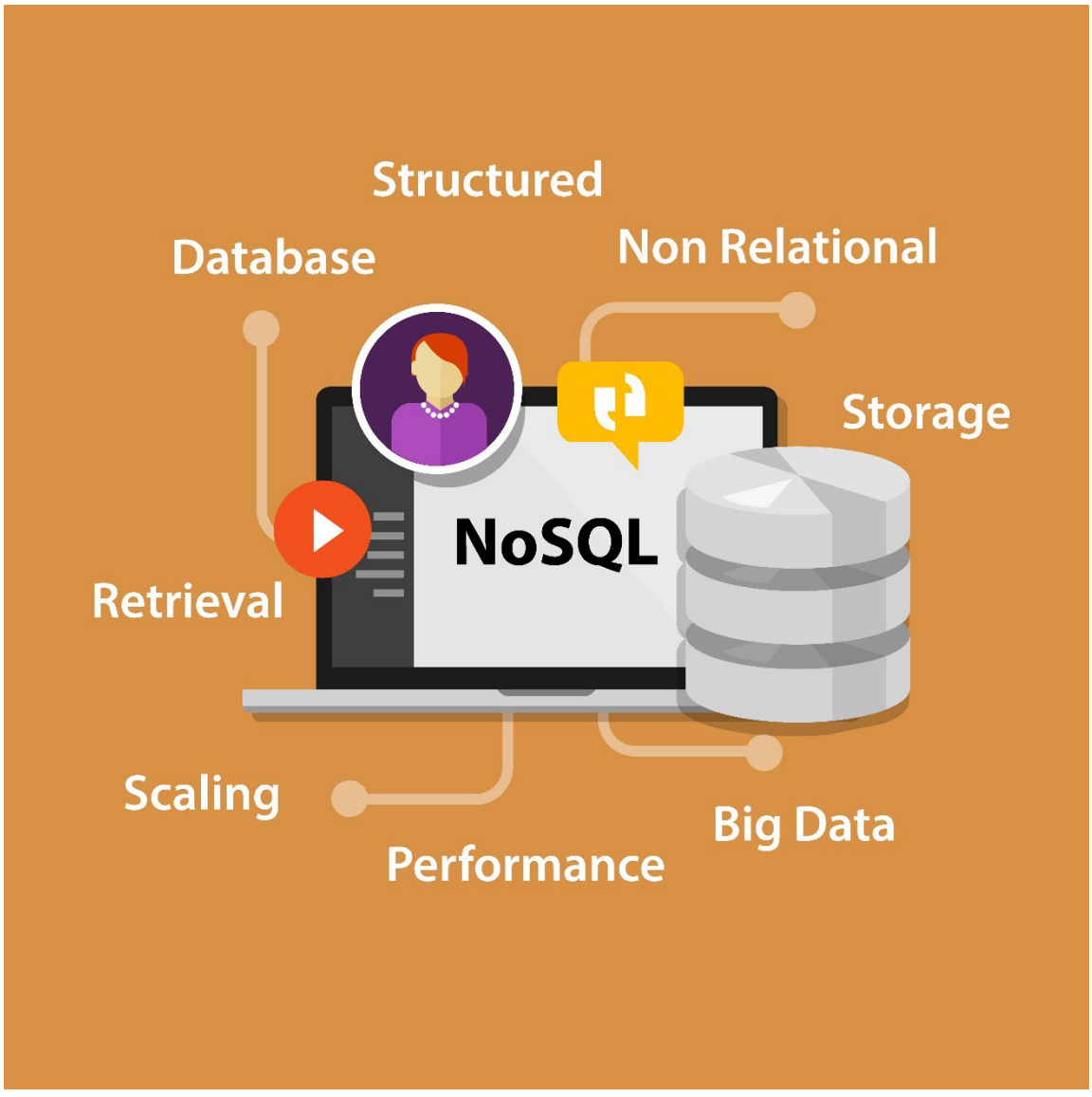
O uso do processamento paralelo dos dados, com o objetivo de atingir *performance* no processamento de um grande volume de dados, torna-se mais eficiente quando as tarefas são divididas em partes menores, podendo executá-las simultaneamente e distribuídas pelos processadores disponíveis. Para que isso aconteça, os sistemas precisam atingir um alto grau de maturidade no processamento paralelo.

A utilização de multiprocessadores, além de melhorar a *performance*, também é uma alternativa mais barata, permitindo escalar o sistema horizontalmente. Para isso é necessário somente adicionar hardware a estrutura

de armazenamento e gerenciamento de dados e ainda possibilita uma independência dos fornecedores de hardware.

A implantação dos sistemas gerenciadores de banco de dados não relacionais pode ser feita em escala global, sendo interessante, pois, para atender os usuários de forma eficiente, são utilizados vários *data centers*, podendo ser implantados em diversas localidades, trazendo questões de disponibilidade e *performance* para o desenvolvimento de sistemas. Essa distribuição na instalação, em conjunto com *hardwares* mais acessíveis, obrigando o sistema a ser robusto o suficiente para suportar falhas constantes, sejam de *hardware* ou de infraestrutura de comunicação.

Figura 8 – NoSQL



Crédito: Bakhtiar Zein/Shutterstock.

TEMA 5 – CIÊNCIA DE DADOS

Ciência de dados é a atividade que trata da coleta de dados e da criação de avaliações dos dados visando o crescimento dos negócios. Em qualquer setor comercial, as informações são um ativo primordial para traçar o rumo dos negócios, porém isso só terá valor se forem analisadas de forma eficiente e que de alguma maneira auxiliem a tomada de decisões.

Não é possível processar e analisar uma quantidade grande de dados não estruturados com as ferramentas normais. A ciência de dados, entretanto, permite processar e analisar volumes expressivos de dados, independentes da fonte, tais como transações financeiras e comerciais, *marketing* direto, sensores de vários elementos naturais, instrumentos de medição e arquivos de texto e multimídia.

A ciência de dados tem como objetivo principal o processamento e análise de dados para aprender sobre o assunto em questão. Ela oferece inúmeras possibilidades de aplicação em conjunto com inteligência artificial, aprendizado de máquina e processos analíticos.

A ciência de dados abrange um número grande de ferramentas, tendo com isso um funcionamento bastante variado. Seu núcleo é processar e analisar dados um grande volume de dados e extrair informações importantes para a tomada de decisão. A ciência de dados trata os dados desde a sua coleta, o seu armazenamento, além da preparação dos dados para que a análise seja feita de uma maneira eficiente. Essas atividades de coleta e preparação são uma competência dos administradores de banco de dados, e a análise e a divulgação dos resultados são de responsabilidade do cientista de dados. Ele deve responder a perguntas gerenciais sobre os dados que foram levantadas previamente na análise do ambiente que está sendo tratado.

A ciência de dados tem uma abrangência grande e cuja aplicação tem o apoio dos engenheiros de dados e pelos analistas de dados. São posições diferentes, porém complementares dentro do universo de dados e suas aplicações desenvolvidas para analisar e propor soluções baseadas em informações extraídas desses dados.

Figura 9 – Data Science

Crédito: Buffaloboy/Shutterstock.

O cientista de dados participa de forma intensa na direção que uma organização precisa tomar. Com base na análise de dados, ele tentará prever como os negócios poderão se comportar no futuro e quais as ações que devem ser tomadas ou não para a mudança de direção.

É uma atividade com posicionamento estratégico para os negócios e muito importante para a tomada de decisões, principalmente quando se lida com uma grande quantidade de dados. O cientista de dados precisa oferecer as respostas corretas, além de levantar questionamentos sobre os dados a serem analisados.

Baseado na análise dos dados, o cientista é capaz de oferecer respostas alinhadas com os questionamentos que foram levantados no planejamento estratégico da organização. Quando isso não ocorre, é preciso gerar novos dados e questões adequados para a tomada de decisões sobre o negócio.

Seguindo sobre os objetivos da ciência de dados, podemos citar um deles como o melhoramento de serviços e produtos, de modo que as organizações tenham uma vantagem competitiva real no mercado.

Podemos citar como itens que podem ser utilizados por organizações para a melhora de serviços na área meteorológica, que é uma contribuição da ciência de dados, o caso específico de previsão do tempo, em que os dados que são coletados de satélites, radares, navios e aeronaves, permitem construir modelos para prever o tempo e emitir alertas sobre possíveis calamidades naturais iminentes.

A análise de clientes com o objetivo de identificar possíveis cancelamentos de serviços é outra possibilidade de atuação da ciência de dados, pois permite analisar o relacionamento com o cliente e tomar as medidas

necessárias para que ele continue com a empresa. A ciência de dados permite ao *marketing* oferecer produtos e serviços sob medida para as preferências do consumidor, baseados nas operações efetuadas anteriormente, além de dados como idade, classe social, profissão e gênero.

Para operações da área da saúde é possível utilizar a ciência de dados para analisar dados nos exames clínicos, auxiliando os médicos a fazerem diagnósticos mais assertivos e precoces, permitindo dessa maneira que os pacientes sejam tratados de forma eficaz, minimizando os custos de tratamento.

Na área de logística e mobilidade, a ciência de dados é utilizada para análise do tráfego, condições climáticas e otimização de rotas, melhorando dessa forma as condições e o tempo das entregas, reduzindo os custos operacionais e aumentando a competitividade.

A ciência de dados utiliza várias ferramentas para analisar os dados. As principais que são utilizadas pelos profissionais especializados são as seguintes:

- R (linguagem de programação);
- Julia (linguagem de programação);
- Python (linguagem de programação);
- SQL (Padrão de banco de dados estruturado);
- MongoDB (Padrão de banco de dados não-estruturado).

Com o crescimento da ciência de dados, as aplicações utilizando essa tecnologia se multiplicaram. A seguir vamos ver alguns exemplos dessas aplicações.

Os principais mecanismos de busca na internet utilizam a ciência de dados em conjunto com o aprendizado de máquina para encontrar o objeto a busca mais preciso em frações de segundos. Essa rapidez nos motores de busca só é possível em virtude da utilização da ciência de dados.

O conteúdo de *marketing* digital na sua grande maioria é escolhido por algoritmos que utilizam ciência de dados. Dessa maneira, as empresas obtêm resultados muito satisfatórios e melhores do que o *marketing* convencional, porque os anúncios são criados conforme o histórico do usuário.

Fazendo uso de algoritmos de reconhecimento de imagens, várias aplicações, como *QR code*, permitem escanear uma imagem com um *smartphone* para poder utilizar o *Whatsapp Web*, por exemplo, ou até recursos automáticos de *tags* para marcar amigos em fotos postadas em redes sociais.

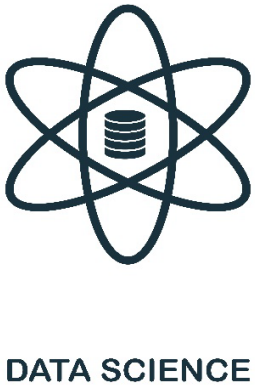
O cruzamento de dados da utilização das mídias sociais por um usuário e do seu histórico de navegação e de buscas permite aos algoritmos de avaliação mapear as preferências de cada usuário e adequar da melhor maneira possível as sugestões que serão apresentadas. É possível entender esse tipo de ocorrência quando são apresentadas a um usuário de mídia social novas sugestões de amizades, além de sugestões de novos filmes em *sites* especializados e pensando em *marketing*, sugestões de produtos para incentivar compras.

Muitas empresas de *e-commerce* de viagens utilizam a ciência de dados para melhorar os resultados em seus motores de busca, mostrando para o cliente os resultados de pesquisas sobre hotéis ou voos. Também são sugeridos serviços complementares, relacionados ao destino da viagem selecionada, como reserva de carros, pacotes de passeios, seguro de viagens e entretenimentos no destino.

Com o grande volume de dados gerado pelas lojas virtuais, os *sites* de busca de produtos utilizam a ciência de dados para mostrar os menores preços do produto que o cliente está procurando.

Na área financeira, é possível utilizar as ferramentas da ciência de dados para análise dos gastos de clientes em instituições financeiras, podendo dessa maneira traçar o perfil de consumo e também de endividamento desses clientes. Isso permite projetar as probabilidades de inadimplência ou não para possíveis concessões de crédito.

Figura 10 – Ciência de dados



Crédito: Anton Shaparenko/Shutterstock.

FINALIZANDO

Nesta etapa, trabalhamos os assuntos extras relacionados a banco de dados. Vimos os sistemas de banco de dados distribuído (BDD) que consiste em uma relação de nós ou servidores, em que cada um pode participar na execução de transações que acessam dados em um ou mais nós. Estudamos sobre *data warehouse* e vimos que ele é um repositório central de dados. Os dados surgem das transações, dos bancos de dados relacionais e de diversas outras fontes para o *data warehouse*. Outro assunto abordado foi sobre *data lake*, que é um lugar para armazenar dados estruturados e não estruturados, bem como um método para organizar grandes volumes de dados altamente diversificados de diversas fontes. Tratamos ainda sobre NoSQL, que é um padrão de armazenamento de dados alternativo ao SQL, oferecendo uma robustez e escalabilidade melhores. Finalizando este estudo de Sistema Gerenciador de Banco de Dados, estudamos sobre ciência de dados, ou *data science*, que é um tema muito discutido entre profissionais e organizações que se concentram na coleta de dados e na elaboração de interpretações significativas para auxiliar no crescimento dos negócios.

REFERÊNCIAS

LAUDON, K. C.; LAUDON, J. P. **Sistemas de informação gerenciais**: administrando a empresa digital. 5. ed. São Paulo: Prentice Hall, 2004.

LE COADIC, Y.-F. **A ciência da informação**. Brasília: Briquet de Lemos, 1996

OLIVEIRA, F. O. **Sistemas de Informação**: um enfoque gerencial inserido no contexto empresarial e tecnológico. 3. ed. São Paulo: Érica 2002

SETZER, V. W. Dado, informação, conhecimento e competência. **DataGramaZero – Revista de Ciência da Informação**, n. 0, 1999.