

Trabalho Banco de Dados

Vinicius Tardelli Leite RA132152

Índice

Data Warehouse…………………………………………………………….......2

Ferramentas do Data Warehouse………………………………………........3

Arquitetura do Data Warehouse………………………………………............4

## Ferramentas que Utilizam Data Warehouse…………………..………….….6

## Beneficios…………………………………………………………...…….……. 8

Data Mining……………………………………………………………………...9

Ferramentas/Ramificações do Data Mining………………………..………...10

Business Intelligence…………………………………………………………...12

Como Funciona………………………………………………………………….13

Técnicas de Business Intelligence…………………………………………….14

Crescimento do Business Intelligence………………………………………..15

Banco de Dados NoSQL……………………………………………………….17

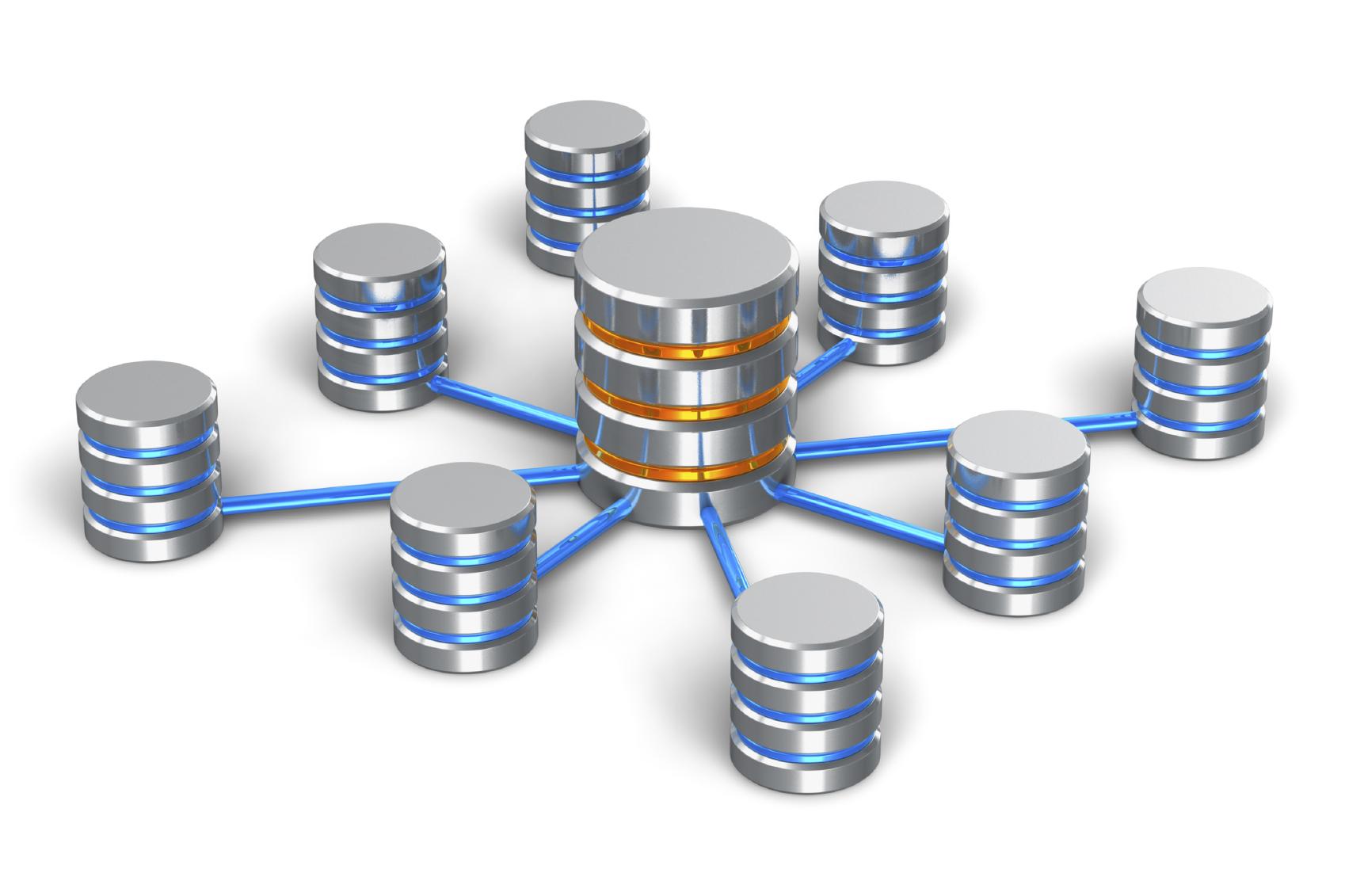
Características Bando de Dados NoSQL…………………………………….18

Tipos de Bancos NoSql…………………………………………...…………....19

#### Alguns exemplos de quando usar NoSQL…………………………………...22 Benefícios em se usar NoSql………………………………………………….23 SQL X NoSQL…………………………………………………………………...25 Dificuldades com NoSQL……………………………………………………….27 Big Data…………………………………………………………………………..28 OS Vs do Big Data……………………………………………………………..29 Importância do Big Data………………………………………………………..30 O Futuro do Big Data…………………………………………………………...31 Aplicações comerciais do Big Data…………………………………………..32

#### .

Data Warehouse



**O que é :** Data warehouse é um depósito de dados digitais que armazena informações detalhadas a uma empresa,criando e organizando relatórios através de históricos que são depois usados pela empresa para ajudar a tomar decisões importantes com base nos fatos apresentados ,ou seja , o *data warehouse* recolhe informações de uma empresa para que essa possa controlar melhor um determinado processo, disponibilizando uma maior flexibilidade nas pesquisas e nas informações que necessitam.

Além de manter um histórico de informações, o *Data Warehouse* cria padrões melhorando os dados analisados de todos os sistemas, corrigindo os erros e estruturando os dados sem afetar o sistema de operação, apresentando somente um modelo final e organizado para a análise.

*“Data Warehouse é um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões gerenciais”. Date(2004)*

Ferramentas do Data Warehouse

## **Data Mart**

U*m data mart é uma subdivisão ou subconjunto de um data warehouse. Os data marts são como pequenas fatias de data warehouse, que armazenam subconjuntos de dados,normalmente o data mart é direcionado para uma linha de negócios ou equipe, sendo que a sua informação costuma pertencer a um único departamento.*

## 

## **Data Mining**

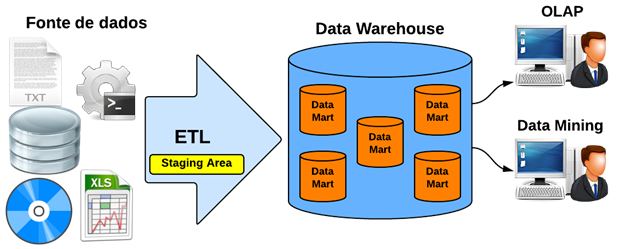
O Data Mining é uma ferramenta de filtragem a qual varre o Data Warehouse a fim de pesquisar tendências e padrões através de regras pré-definidas que dificilmente seriam encontrados em uma pesquisa comum. Vamos falar mais sobre essa ferramente mais a frente.

**Data Stage**

Composta por uma área de armazenagem e um conjunto de processos sua função é extrair os dados dos sistemas transacionais, proceder a limpeza, a transformação, combinação, de duplicação e preparação dos dados para o uso no Data Warehouse. Estes dados não são apresentados ao usuário final.

***Data Mart***

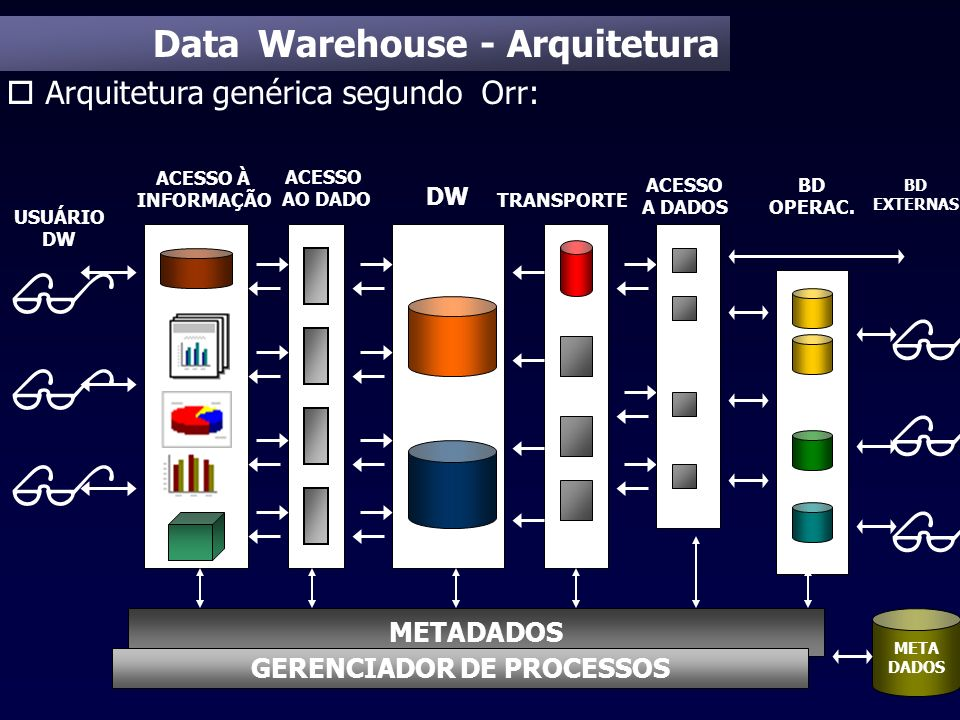
*S*ubconjunto lógico do Data Warehouse, geralmente divididos por departamento ou visões necessárias para os usuários.



**Arquitetura do Data Warehouse**

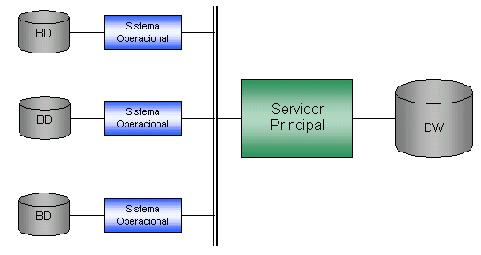
**Arquitetura Genérica**

A arquitetura genérica compreende a camada dos dados operacionais que serão acessados pela camada de acesso a dados. As camadas de gerenciamento de processos, transporte e Data Warehouse são responsáveis por distribuir os dados e estão no centro da arquitetura. A camada de acesso à informação possibilita a extração das informações do Data Warehouse utilizando um conjunto de ferramentas.



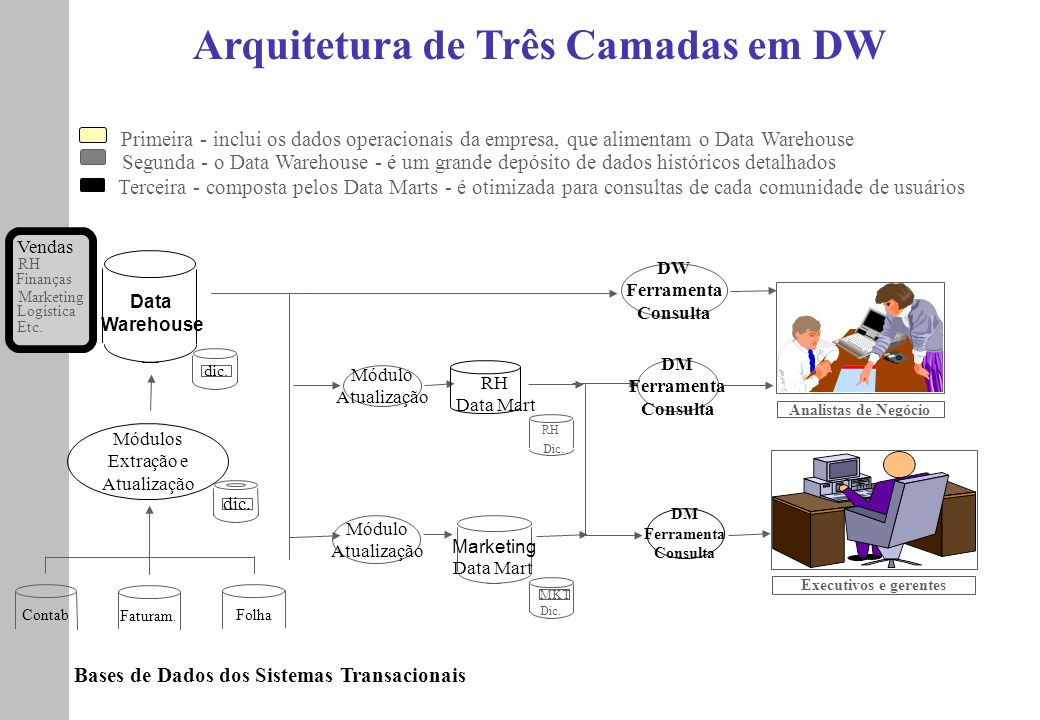
***Arquitetura de Duas Camadas***

A arquitetura de duas camadas utiliza um servidor, juntamente com aplicações front end, que são ferramentas que realizam operações sobre os dados consultados e os transformam em informações úteis para os usuários, os componentes back end são ferramentas responsáveis pela extração, limpeza e cargas dos dados, mais conhecidas como ETL também são utilizadas neste tipo de arquitetura.



**Arquitetura de Três Camadas**

A arquitetura de três camadas suporta vários usuários e serviços devido a sua flexibilidade, as informações ficam armazenadas em várias camadas. Na primeira camada estão as interfaces que trabalham com o usuário, onde geralmente são gráficas. Na segunda camada estão os servidores de banco de dados e aplicações e, por isso, têm a necessidade de ter um acesso eficiente e veloz aos dados compartilhados, e na última ficam armazenadas as fontes de dados. A arquitetura de três camadas é a mais utilizada pelos analistas.



## Ferramentas que Utilizam Data Warehouse

## OLAP On-Line Analytical Processing .

O OLAP e o Data Warehouse são destinados a trabalharem juntos, enquanto o data warehouse armazena as informações de forma eficiente, o OLAP deve recuperá-las com a mesma eficiência, porém com muita rapidez. As duas tecnologias se complementam, ao ponto de que um Data Warehouse para ser bem sucedido, já na sua concepção, deve levar em consideração o que se deseja apresentar na interface OLAP

O OLAP é uma interface com o usuário e não uma forma de armazenamento de dados, porém se utiliza do armazenamento para poder apresentar as informações.

Os métodos de armazenamento são:

**ROLAP (OLAP Relacional):**

Os dados são armazenados de forma relacional.

**MOLAP (OLAP Multidimensional):**

Os dados são armazenados de forma multidimensional.

**HOLAP (OLAP Híbrido):**

Uma combinação dos métodos ROLAP e MOLAP.

**DOLAP (OLAP Desktop):**

O conjunto de dados multidimensionais deve ser criado no servidor e transferido para o desktop. Permite portabilidade aos usuários OLAP que não possuem acesso direto ao servidor.

Os métodos mais comuns de armazenamento de dados utilizados pelos sistemas OLAP são ROLAP e MOLAP, a única diferença entre eles é a tecnologia de banco de dados. O ROLAP usa a tecnologia RDBMS (Relational DataBase Management System), na qual os dados são armazenados em uma série de tabelas e colunas. Enquanto o MOLAP usa a tecnologia MDDB (MultiDimensional Database), onde os dados são armazenados em arrays multidimensionais.

Os dois fornecem uma base sólida para análise e apresentam tanto vantagens quanto desvantagens. Para se escolher entre os dois métodos deve-se levar em consideração os requisitos e a abrangência do aplicativo a ser desenvolvido.

ROLAP é mais indicado para data warehouse pelo grande volume de dados, a necessidade de um maior número de funções e diversas regras de negócio a serem aplicadas.

MOLAP é mais indicado para DATA MARTS, onde os dados são mais específicos e o aplicativo será direcionado na análise com dimensionalidade limitada e pouco detalhamento das informações.

Para se fazer uma comparação básica entre os dois métodos, as regras mais importantes são desempenho da consulta e desempenho do carregamento.

**DESEMPENHO DE CONSULTA**

O MOLAP fornece uma resposta rápida para praticamente qualquer consulta, pois no modelo multidimensional são gerados previamente todas as combinações e resumos possíveis.

O ROLAP responde às consultas da mesma forma que os aplicativos RDBMSs, a velocidade da resposta depende da informação desejada, pois a maior parte do processamento é feito em tempo de execução tendo em vista que os dados pré-calculados e resumidos geralmente não atendem a todas as solicitações dos usuários.

**DESEMPENHO DE CARREGAMENTO**

O MOLAP necessita de um longo período para execução da carga de dados, raramente esta carga é diária devido ao grande volume de informações a serem atualizadas para possibilitar um retorno rápido às consultas da interface OLAP.

O ROLAP possibilita um carregamento mais rápido devido à estrutura de tabelas e colunas, menos complexa em comparação à estrutura de arrays utilizada pelo MOLAP. Outro fator importante na rapidez da carga é o número menor de informações pré-calculadas e resumidas.

Por fim, chegamos à conclusão de que não existe uma definição entre qual o melhor, ROLAP ou MOLAP, e que a tendência será a utilização do método HOLAP no qual é possível utilizar as vantagens dos dois modelos numa mesma arquitetura.

## 

## Beneficios

Um data warehouse permite pesquisas melhores, mais rápidas e mais fáceis de utilizar. Por ser focado somente em pesquisas e manter suas informações em uma estrutura separada do sistema de operação normal, um data warehouse pode estruturar seus dados de maneira diferente fazendo com que sejam mais rápidos de acessar. Por causa dessa dissociação com a fonte da informação é possível agregar dados de varias fontes, mesmo que não diretamente relacionadas, resolvendo conflitos entre os dados e permitindo uma visão mais ampla da empresa. Também é possível ter um histórico dos dados a serem analisados mesmo que o sistema que as coletou não guarde tais dados.

DATA MINING



**Oque é:** *Data Mining* consiste em um processo analítico projetado para explorar grandes quantidades de dados, na busca de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados. O processo consiste basicamente em 3 etapas: exploração, construção de modelo e validação.

*Data* Mining é uma argumentação ativa, isto é, em vez do usuário definir o problema, selecionar os dados e as ferramentas para analisar tais dados, as ferramentas do Data Mining pesquisam automaticamente os mesmos a procura de anomalias e possíveis relacionamentos, identificando assim problemas que não tinham sido identificados pelo usuário.

Por isso o Data Mining é um processo fundamental do Data Warehouse pois um está fortemente Vinculado com o outro.

O *Data Mining* descende fundamentalmente de 3 linhagens. A mais antiga delas é a estatística clássica. Sem a estatística não seria possível termos o Data Mining, visto que a mesma é a base da maioria das tecnologias a partir das quais o Data Mining é construído.

A segunda linhagem do Data Mining é a Inteligência Artificial, essa disciplina, que é construída a partir dos fundamentos da heurística, em oposto à estatística, tenta imitar a maneira como o homem pensa na resolução dos problemas estatísticos

E a terceira e última linhagem do Data Mining é a chamada *machine learning*, que pode ser melhor descrita como o casamento entre a estatística e a Inteligência Artificial. Enquanto a Inteligência Artificial não se transformava em sucesso comercial, suas técnicas foram sendo largamente cooptadas pela *machine learning*, que foi capaz de se valer das sempre crescentes taxas de preço/performance oferecidas pelos computadores nos anos 80 e 90, conseguindo mais e mais aplicações devido às suas combinações entre heurística e análise estatística. *Machine learning* é uma disciplina científica que se preocupa com o design e desenvolvimento de algoritmos que permitem que os computadores aprendam com base em dados, como a partir de dados do sensor ou bancos de dados. Um dos principais focos da Machine Learnig é automatizar o aprendizado para reconhecer padrões complexos e tomar decisões inteligentes baseadas em dados

Ferramentas/Ramificações do Data Mining

As ferramentas de *Data Mining*, baseadas em algoritmos que formam a construção de blocos de inteligência artificial, redes neurais, regras de indução, e lógica de predicados, facilitam e auxiliam o trabalho dos analistas de negócio das empresas, ajudando as mesmas a conseguirem serem mais competitivas.

**Redes neurais**

São sistemas computacionais baseados numa aproximação à computação baseada em ligações. Nós simples são interligados para formar uma rede de nós - daí o termo "rede neural". A inspiração original para esta técnica advém do exame das estruturas do cérebro, em particular do exame de neurônios. Exemplos de ferramentas: *SPSS Neural Connection, IBM Neural Network Utility, NeuralWare NeuralWork Predict.*

**Indução de regras**

A Indução de Regras, ou *Rule Induction*, refere-se à detecção de tendências dentro de grupos de dados, ou de “regras” sobre o dado. As regras são, então, apresentadas aos usuários como uma lista “não encomendada”. Exemplos de ferramentas: IDIS da *Information Discovey* e *Knowledge Seeker da Angoss Software*.

**Árvores de decisão**

Baseiam-se numa análise que trabalha testando automaticamente todos os valores do dado para identificar aqueles que são fortemente associados com os itens de saída selecionados para exame. Os valores que são encontrados com forte associação são os prognósticos chaves ou fatores explicativos, usualmente chamados de regras sobre o dado. Exemplos de ferramentas: *Alice d’Isoft*, *Business Objects BusinessMiner*, *DataMind.*

***Analise de séries temporais:***

A estatística é a mais antiga tecnologia em Data Mining, e é parte da fundação básica de todas as outras tecnologias. Ela incorpora um envolvimento muito forte do usuário, exigindo engenheiros experientes, para construir modelos que descrevem o comportamento do dado através dos métodos clássicos de matemática. Interpretar os resultados dos modelos requer “expertise”. O uso de técnicas de estatística também requer um trabalho muito forte de máquinas/engenheiros. A análise de séries temporais é um exemplo disso, apesar de freqüentemente ser confundida como um gênero mais simples de Data Mining chamado “forecasting” . Exemplos de ferramentas: S+, SAS, SPSS.

***Visualização:***

Mapeia o dado sendo minerado de acordo com dimensões especificadas. Nenhuma análise é executada pelo programa de Data Mining além de manipulação estatística básica. O usuário, então, interpreta o dado enquanto olha para o monitor. O analista pode pesquisar a ferramenta depois para obter diferentes visões ou outras dimensões. Exemplos de ferramentas: IBM Parallel Visual Explorer,SAS System, Advenced Visual Systems (AVS) Express Visualization Edition.

*Business Intelligence*

**

**Oque é:** O termo Business Intelligence , inteligência de negócios, refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios. É o conjunto de teorias, metodologias, processos, estruturas e tecnologias que transformam uma grande quantidade de dados brutos em informação útil para tomadas de decisões estratégicas.

Inicialmente, as primeiras fontes de informação são coletadas dentro das próprias empresas. Posteriormente, as segundas fontes de informações incluem as necessidades do consumidor, o processo de decisão do cliente, as pressões competitivas, as condições industriais relevantes, os aspectos econômicos e tecnológicos e as tendências culturais. Cada sistema de BI determina uma meta específica, tendo por base o objetivo organizacional e a visão da empresa, existindo em ambos objetivos, sejam eles de longo ou curto prazo, as disputas nos negócios, a coleta de informação.

Como Funciona

Tudo começa com a coleção de dados, Data Warehousing, a integração de dados de uma ou mais fontes e assim, cria um repositório central de dados, um Data Warehouse os armazéns de dados. Com essa imensidão de dados, Data Mining, aplica-se a mineração desses dados, o processo de explorar grandes quantidades de dados à procura de padrões consistentes para detectar relacionamentos e novos subconjuntos de dados a serem mapeados e extrair-se informações previlegiadas. Análises de minerações geram relatórios detalhados para fortalecer o esclarecimento do cenário.

## 

## 

## A quem o Business Intelligence se destina?

Business Intelligence se destina às empresas que, inseridas em um mundo competitivo, sabem que são necessários investimentos para que uma empresa possa crescer de forma consciente e estruturada.

Podendo ser estas de vários ramos diferentes , como por exemplo : Comércio , indústrias , laboratórios , logística , Prestação de Serviços , Produção de alimentos, construtoras, importadoras,corretoras,bancos...etc

Investir em Business intelligence significa investir em um sistema de organização de dados e relatórios gerenciais dinâmicos que possibilitarão rápida visualização do processo evolutivo da empresa.

Técnicas de Business Intelligence

Reengenharia de processos de negócio

Trata-se de uma estratégia de gestão de negócios para a análise e desenho dos fluxos de trabalho e dos processos de negócios visando a reestruturação organizacional, com foco no design de baixo para cima de processos de negócios dentro de uma organização

Análise de clientes

Uma versão específica do Business Intelligence em que informações sobre o comportamento de clientes são analisadas. Isso envolve a análise do comportamento passado do cliente para tomar decisões de negócios importantes e prever padrões de comportamento futuros.

Benchmarking

A busca das melhores práticas com o propósito de maximizar o desempenho. Onde uma empresa examina como realiza uma função específica a fim de melhorar como realizar a mesma ou uma função semelhante, ou seja, um processo de comparação do desempenho entre dois ou mais sistemas.

CRM

Um sistema baseado em tecnologia para o gerenciamento de relacionamentos com clientes potenciais e atuais. O objetivo da tecnologia de CRM é dar suporte a interações mais aprofundadas e rentáveis com os clientes.

## 

## Relatórios Analíticos

Ajudam a segmentar e identificar as informações e traduzir em uma avaliação precisa, colaborando com indicadores relevantes e relatórios que refletem realmente cada atividade.

Crescimento do Business Intelligence

Com ações mais eficientes e resultados mais significativos da gestão de dados para tomadas de decisões, empresas passaram a investir em meios digitais e ações de Business Intelligence para alcançar suas metas. Promover a compreensão do processo de captura e utilização de informações externas e internas à organização, para o desenvolvimento e monitoramento de estratégias coerentes ao momento competitivo. Com o crescimento exponencial do uso das redes sociais por grandes corporações nas suas estratégias de negócios, e Business Intelligence também precisou se reinventar. Várias empresas estão desenvolvendo software para ter à disposição o seu histórico de interações e relacionamento na Internet. Não somente as grandes empresas, mas também as corporações de pequeno, médio porte necessitam de Business Intelligence para auxiliá-las nas mais diferentes situações para a tomada de decisão, otimizar o trabalho da organização, reduzir custos, eliminar a duplicação de tarefas, permitir previsões de crescimento da empresa como um todo e contribuir para a elaboração de estratégias. Na grande maioria das vezes, para estas empresas enquadradas como pequenas, não precisam, necessariamente, de sistemas de alto investimento, com milhares de relatórios, gráficos para identificar que existe um problema específico numa linha de produto que não está sendo comercializado conforme a sua potencialidade. Basta o feeling da área comercial em se pensar numa outra estratégia de vendas para o mesmo.

Benefícios

No Brasil, soluções de Business Intelligence estão em instituições financeiras, empresas de telecomunicações, seguradoras e em toda instituição que perceba a tendência da economia globalizada, em que a informação precisa chegar de forma rápida, precisa e abundante. O principal benefício do BI para a empresa é a sua capacidade de fornecer informações precisas quando necessárias, incluindo uma visão em tempo real do desempenho corporativo geral e de suas partes individuais. A partir dos resultados de uma pesquisa entre 510 empresas, que os benefícios do BI são a economia de tempo, versão única da verdade, melhores estratégias e planos, melhores decisões táticas, processos mais eficientes e a economia de custos.

Ferramentas/Empresas Brasileiras Consagradas em consultoria para Business Intelligence

* Lucída
* +++Pasquali Solution Tech
* Gigatron
* B2T
* Qliktech

Banco de Dados NoSQL



**Oque é:**O termo noSQL é bastante abrangente, pois envolve diversas ferramentas, tecnologias, estruturas de dados e arquiteturas. Esta nova *buzzword* representa muito mais um movimento, ou uma nova escola de pensamento, do que alguma tecnologia em particular.

O NoSQL surgiu da necessidade de uma performance superior e de uma alta escalabilidade, os atuais bancos de dados relacionais são muito restritos a isso, sendo necessária a distribuição vertical de servidores, ou seja, quanto mais dados, mais memória e mais disco um servidor precisa.

O NoSQL tem uma grande facilidade na distribuição horizontal, ou seja, mais dados, mais servidores, não necessariamente de alta performance.

As tecnologias noSQL não têm como objetivo substituir os bancos de dados relacionais, mas apenas propor algumas soluções que em determinados cenários são mais adequadas. Desta forma, é possível trabalhar com tecnologias noSQL e banco de dados relacionais dentro de uma mesma aplicação.

Características Bando de Dados NoSQL

Modelo de Dados

Bancos de dados não relacionais (NoSQL) normalmente não aplicam um schema. Uma “chave hash” é normalmente usada para recuperar valores, conjuntos de colunas ou documentos semiestruturados JSON, XML ou outros que contenham atributos de itens relacionados.

Propriedades

Bancos de dados NoSQL normalmente trocam algumas propriedades ACID de sistemas de gerenciamento de bancos de dados relacionais (RDBMS) por um modelo de dados mais flexível que escala horizontalmente. Essas características fazem dos bancos de dados NoSQL uma excelente opção em situações em que os RDBMS deparam com desafios de arquitetura e precisam solucionar uma combinação de gargalos de desempenho, escalabilidade, complexidade operacional e custos crescentes de administração e suporte.

Desempenho

Desempenho geralmente é uma função do tamanho do cluster do hardware subjacente, da latência de rede e da aplicação que faz a chamada.

Escala

Projetado para aumentar a escala "horizontalmente" usando clusters distribuídos de hardware de baixo custo para aumentar a transferência sem aumentar a latência.

APIs

APIs baseadas em objetos permitem que desenvolvedores de aplicações armazenem e restaurem facilmente estruturas de dados na memória. “Chaves hash” permitem que as aplicações procurem pares de chave-valor, conjuntos de colunas ou documentos semiestruturados contendo objetos de atributos de aplicação serializados.

Tipos de Bancos NoSql

Chave Valor

A ideia principal aqui é usar uma tabela hash na qual há uma chave única e um indicador de um dado ou de um item em particular.

O modelo chave-valor é o mais simples e fácil de implementar. Mas ele é ineficiente quando você somente está interessado em consultar ou em atualizar parte de um valor, entre outras desvantagens,esses tipos de bancos de dados são o que tem a maior escalabilidade e os e os que suportam maior carga de dados.

Exemplos :

* Berkeley
* DB
* Tokyo
* Cabinet
* Project
* Voldermort
* MemcacheDB
* SimpleBD

Grafos

Diferentemente de outros tipos de bancos de dados NoSQL, esse está diretamente relacionado a um modelo de dados estabelecido, o modelo de grafos. A ideia desse modelo é representar os dados e/ou o esquema dos dados como grafos dirigidos, ou como estruturas que generalizem a noção de grafos .

O modelo de grafos é mais interessante que outros quando informações sobre a inter-conectividade ou a topologia dos dados são mais importantes, ou tão importante quantos os dados propriamente ditos .

O modelo orientado a grafos possui três componentes básicos: os nós (são os vértices do grafo), os relacionamentos (são as arestas) e as propriedades (ou atributos) dos nós e relacionamentos.

Neste caso, o banco de dados pode ser visto como um multigrafo rotulado e direcionado, onde cada par de nós pode ser conectado por mais de uma aresta.

Exemplos:

* Neo4j
* InfoGrid
* Sones
* HyperGraphDB

Documentos

Este modelo permite armazenar qualquer documento, sem ter a necessidade de definir previamente sua estrutura. O documento e composto por uma inúmeros campos, com tipos de dados diversos, inclusive um campo pode conter um outro, possui uma estrutura semelhante a um arquivo XML.

Exemplos:

* AmazonSimpleDb
* ApacheCouchdb
* MongoDb
* Riak

Colunas

Esse modelo se tornou popular através do paper BigTable do Google, com o objetivo de montar um sistema de armazenamento de dados distribuído, projetado para ter um alto grau de escalabilidade e de volume de dados é composto por três componentes

Keyspace tem como função agrupar um conjunto de Famílias de Colunas. Semelhante a um banco de dados relacional.

Família de Colunas organizas as colunas. faz o uso de uma chave única, que traz flexibilidade ao modelo sem poluir as linhas com colunas nulas. Semelhante a uma tabela no modelo relacional.

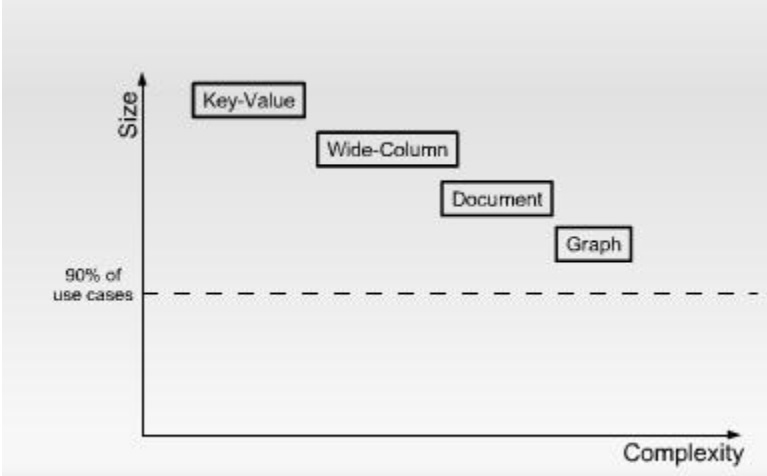
Coluna, que é uma tupla composta por nome,timestamp e valor, onde os dados são realmente armazenados.

Duas características adicionais e importantes deste modelo são a forma de consulta, que pode ser executada apenas através da chave das linhas de uma família de colunas, e a necessidade de definir previamente o conjunto de colunas que podem ser armazenadas em cada família.

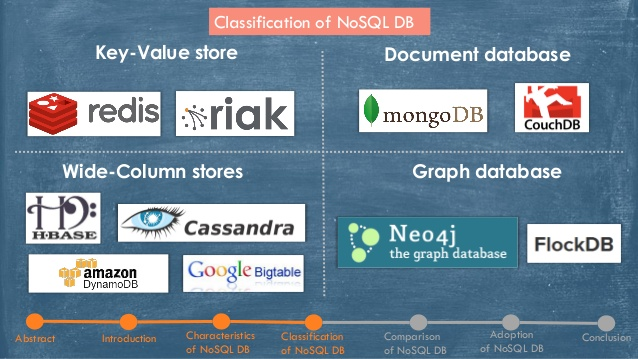
Exemplos:

* · Cassandra
* · GoogleBigTable
* · HBase
* · Hyperbase

Na imagem abaixo, podemos ver um gráfico demonstrando a diferença entre o tamanho da base de dados e a complexidade dos seus dados. Assim, podemos perceber que os bancos do tipo chave-valor conseguem suportam maior quantiadade dados, sendo que seus dados são mais simples, enquanto que os bancos do tipo grafo suportam menos, porem seus dados são mais complexos.



E na Imagem abaixo Temos a Logo de Alguns dos Banco de Dados NoSql apresentados anteriormente



#### Alguns exemplos de quando usar NoSQL

Logging/Arquivamento.

Ferramentas de Log-mining vêm a calhar porque elas conseguem acessar logs através de servidores, relacioná-los e analisá-los.

Insight em computação social.

Muitas empresas hoje fornecem a seus usuários a habilidade de atuar na computação social através de fóruns de mensagens blogs etc.

Integração de feed de dados externos.

Muitas empresas precisam integrar os dados oriundos de seus parceiros. Mesmo se as duas partes conduzirem várias discussões e negociações, as empresas têm pouco controle sobre o formato dos dados que chegam a elas.

Além disso, existem muitas situações em que esses formatos mudam muito frequentemente baseado na mudança das necessidades de negócios dos parceiros.

Sistemas de processamento de pedidos Front-end.

Hoje, o volume de pedidos, aplicações e serviços flutuando através de diferentes canais para comerciantes, banqueiros e seguradoras, fornecedores de entretenimento, logística etc é enorme. Esses pedidos precisam ser capturados sem nenhum interrupção sempre que um usuário final faz uma transação de qualquer parte do mundo. Depois disso, um sistema de reconciliação tipicamente os atualiza parasistemas de back-end, ao mesmo tempo que atualiza o usuário final com o seu status.

Serviço de gerenciamento de conteúdo empresarial.

O gerenciamento de conteúdo agora é utilizado através de diferentes grupos funcionais da empresa, como RH ou vendas. O desafio é congregar grupos diferentes utilizando estruturas diferentes de metadados em um serviço comum de gerenciamento de conteúdo.

Estatísticas/análises em tempo real.

Às vezes, é necessário usar o banco de dados como uma maneira de rastrear métricas de performance em tempo real para websites (visualizações de páginas, visitas únicas etc). Ferramentas como o Google Analytics são ótimas, mas não são em tempo real , às vezes, é util construir um sistema secundário que fornece estatísticas em tempo real. Outras alterativas, como [monitoramento de tráfego](http://portal.monitis.com/index.php/products/web-traffic-monitor) 24/7, são uma boa opção também.

Benefícios em se usar NoSql

## FLEXIBILIDADE

Estruturas de dados intuitivas e flexíveis são funcionalidades que mais atraem desenvolvedores que trabalham em times de desenvolvimento ágil, pois elimina a complexidade de mudanças dos bancos de dados gerando um bom suporte para adaptações rápidas. A maioria dos bancos de dados NoSQL tem essas qualidades.

## ESCALABILIDADE

A maioria dos bancos de dados NoSQL são construídos para escalar horizontalmente, distribuindo os dados por clusters melhor que os SGBDs relacionais, que sofrem muito em performance quando executa consultas com “joins” em ambientes clusterizados. Como o NoSQL evita “joins” naturalmente, a performance das consultas permanece alta.

A escalabilidade NoSQL aplica-se tanto ao crescimento dos dados quanto ao número de usuários agindo simultaneamente sobre os dados.

## DISPONIBILIDADE

A indisponibilidade de um banco de dados pode causar sérios prejuízos, incluindo perda de clientes. A maioria dos bancos de dados NoSQL oferecem eficientes arquiteturas de replicação de dados que proporciona aos NoSQLs maior disponibilidade. Ou seja, se um ou mais servidores (ou nós) cai um outro nó do cluster está apto a continuar o trabalho automaticamente sem perda de dados.

## RAÍZES OPEN SOURCE

Muitos bancos de dados NoSQL tem raízes na comunidade open source. Nota-se que as companhias que oferecem versões comerciais de bancos NoSQL com uma forte estrutura de suporte e serviços, estão ao mesmo tempo participando direta ou indiretamente de comunidades de bancos de dados NoSQL open source. Exemplos são o Cassandra da Apache, CouchDB da IBM, MongoDB da MongoDB open source, entre outros.

## 

## 

## BAIXO CUSTO OPERACIONAL

Devido ao peso do open source no NoSQL, o custo para iniciar a utilização desses bancos de dados também costuma muito baixo ou zero. É comum que a transição relacional para NoSQL diminua muito os custos enquanto obtém um desempenho melhor ou igual ao anterior.

Grandes bancos de dados relacionais requerem computadores ou mainframes caros. Com o NoSQL, esse custo também diminui, pois este foi desenvolvido para trabalhar em ambientes distribuídos.

## FUNCIONALIDADES ESPECIAIS

Também com o peso do open source sobre o NoSQL, muitos distribuidores de bancos de dados NoSQL oferecem algumas funcionalidades especiais para incentivar e atrair mais usuários. Índices específicos, capacidade de consulta de dados geoespaciais, replicação automática de dados, funcionalidades para sincronização, APIs RESTful são exemplos de funcionalidades especiais oferecidas pelos diferentes distribuidores NoSQL.

## 

## 

## SQL X NoSQL

Além de lidar com volumes extremamente grandes de dados dos mais variados tipos, soluções com o Big Data também precisam trabalhar com distribuição de processamento e *elasticidade*, isto é, suportar aplicações com volumes de dados que crescem substancialmente em pouco tempo.

O problema é que os bancos de dados "tradicionais", especialmente aqueles que exploram o modelo relacional, como o MySQL, o PostgreSQL e o Oracle, não se mostram adequados a estes requisitos, já que são menos flexíveis.

Isso acontece porque bancos de dados relacionais normalmente se baseiam em quatro propriedades que tornam a sua adoção segura e eficiente, razão pela qual soluções do tipo são tão populares: Atomicidade, Consistência, Isolamento e Durabilidade. Esta combinação é conhecida como *ACID*, sigla para o uso destes termos em inglês: *Atomicity*, *Consistency*, *Isolation* e *Durability*. Vejamos uma breve descrição de cada uma:

* Atomicidade: toda transação deve ser atômica, isto é, só pode ser considerada efetivada se executada completamente;
* Consistência: todas as regras aplicadas ao banco de dados devem ser seguidas;
* Isolamento: nenhuma transação pode interferir em outra que esteja em andamento ao mesmo tempo;
* Durabilidade: uma vez que a transação esteja concluída, os dados consequentes não podem ser perdidos.

O problema é que este conjunto de propriedades é por demais restritivo para uma solução de Big Data. A elasticidade, por exemplo, pode ser inviabilizada pela atomicidade e pela consistência. É neste ponto que entra em cena o conceito de **NoSQL**, denominação que muitos atribuem à expressão em inglês "*Not only SQL*", que em tradução livre significa "Não apenas SQL" (*SQL* -*Structured Query Language* - é, em poucas palavras, uma linguagem própria para se trabalhar com bancos de dados relacionais).

O NoSQL faz referência às soluções de bancos de dados que possibilitam armazenamento de diversas formas, não se limitando ao modelo relacional tradicional. Bancos do tipo são mais flexíveis, sendo inclusive compatíveis com um grupo de premissas que "compete" com as propriedades ACID: a *BASE* (*Basically Available*, *Soft state*, *Eventually consistency* - Basicamente disponível, Estado Leve, Eventualmente consistente).

Não é que bancos de dados relacionais tenham ficado ultrapassados - eles são e continuarão por muito tempo sendo úteis a uma série de aplicações. O que acontece é que, geralmente, quanto maior um banco de dados se torna, mais custoso e trabalhoso ele fica: é preciso otimizar, acrescentar novos servidores, empregar mais especialistas em sua manutenção, enfim.

Via de regra, escalar (torná-lo maior) um bancos de dados NoSQL é mais fácil e menos custoso. Isso é possível porque, além de contar com propriedades mais flexíveis, bancos do tipo já são otimizados para trabalhar com processamento paralelo, distribuição global (vários data centers), aumento imediato de sua capacidade e outros.

Além disso, há mais de uma categoria de banco de dados NoSQL, fazendo com que soluções do tipo possam atender à grande variedade de dados que existe, tanto estrurados, quanto não estruturados: bancos de dados orientados a documentos, bancos de dados chave/valor, bancos de dados de grafos, enfim.

Maiores Utilizadores dos Bancos de Dados NoSQL

* Twitter
* Facebook
* Digg
* Amazon
* LinkedIN
* Google
* Yahoo
* The New York Times

## 

## 

## Dificuldades com NoSQL

Há alguns problemas que ainda hoje existem nessas bases de dados. O primeiro destes problemas é o como não há um padrão (e também não faz sentido existir um) a migração de um banco de dados para outro dentro de uma mesma categoria pode se mostrar uma tarefa bem trabalhosa.

Outro problema é a ausência de ferramentas de alto nível como as que encontramos no modelo relacional.

A curva de aprendizado também pode ser um problema. Como cada banco de dados apresenta a sua própria linguagem de consulta e manipulação, o desenvolvedor precisará adicionar cada vez mais e mais linguagens no seu cinto de utilidades.

Big Data



**O que é:** Oriundo do termo de tecnologia da informação, o conceito de *Big Data* é focado no gigantesco armazenamento de dados, com enorme velocidade. O Big Data é baseado no conceito de 5*V:* Valor, Veracidade, Variedade, Volume e Velocidade. O Big Data é tido por muitos, como a solução de eventuais situações problemáticas da economia. E devido ao modelo economico adotado pela globalização, o termo *Just in time,* totalmente dependente da necessidade de uma expansão virtual, se tornou a palavra de ordem das negociações e forçou a ampliação de estrutura para armazenamento de dados.

Neste quesito, de acordo com a *IBM,* até o ano de 2008 já tinham sido produzidos mais de 2,5 quintilhões de bytes, sendo que aproximadamente 90% das informações armazenadas em centrais, foram produzidas nos últimos dois anos, fato este devido a inserção das empresas no meio online (Internet), além da difusão de dispositivos móveis por exemplo.A diferença entre um Big Data e o chamado Armazém de Dados *Data Warehouse,* é basicamente que o segundo se baseia em um conjunto de dados, com variação de tempo e a opção de auxiliar nas decisões de negócios; enquanto o *Big Data* se baseia em grande volume de dados integrados, porém com mais velocidade. Outra diferenciação é o software, que faz uma varredura de todos os dados fornecidos por uma empresa e os separa para que sejam utilizados em outra ocasião, tornando-o assim mais fácil de ser utilizado.

**OS Vs do Big Data**

**Volume**

Organizações coletam dados de uma grande variedade de fontes, incluindo transações comerciais, redes sociais e informações de sensores ou dados transmitidos de máquina a máquina. No passado, armazenar tamanha quantidade de informações teria sido um problema – mas novas tecnologias (como o Hadoop) têm aliviado a carga.

**Velocidade**

Os dados fluem em uma velocidade sem precedentes e devem ser tratados em tempo hábil. Tags de RFID, sensores, celulares e contadores inteligentes estão impulsionado a necessidade de lidar com imensas quantidades de dados em tempo real, ou quase real.

**Variedade**

Os dados são gerados em todos os tipos de formatos de dados estruturados, dados numéricos em bancos de dados tradicionais, até documentos de texto não estruturados, e-mail, vídeo, áudio, dados de cotações da bolsa e transações financeiras.

**Veracidade**

Também pode ser considerado, pois não adianta muita coisa lidar com a combinação "volume + velocidade + variedade" se houver dados não confiáveis. É necessário que haja processos que garantam o máximo possível a consistência dos dados. Voltando ao exemplo da operadora de cartão de crédito, imagine o problema que a empresa teria se o seu sistema bloqueasse uma transação genuína por analisar dados não condizentes com a realidade.

Importância do Big Data

*Informação é poder*, logo, se uma empresa souber como utilizar os dados que tem em mãos, poderá entender como melhorar um produto, como criar uma estratégia de marketing mais eficiente, como cortar gastos, como produzir mais em menos tempo, como evitar o desperdício de recursos, como superar um concorrente, como disponibilizar serviços para a um cliente especial de maneira satisfatória e assim por diante. Estamos falando de fatores que podem inclusive ser decisivos para o futuro de uma companhia. Mas, Big Data é um nome relativamente recente (ou, ao menos, começou a aparecer na mídia recentemente).

Há tempos que departamentos de TI contemplam aplicações de *Data Mining*, *Business Intelligence* e *CRM* , por exemplo, para tratar justamente de análise de dados, tomadas de decisões e outros aspectos relacionados ao negócio.

A proposta de uma solução de Big Data é a de oferecer uma abordagem ampla no tratamento do aspecto cada vez mais "caótico" dos dados para tornar as referidas aplicações e todas as outras mais eficientes e precisas. Para tanto, o conceito considera não somente grandes quantidades de dados, a velocidade de análise e a disponibilização destes, como também a relação com e entre os volumes.



O Futuro do Big Data

Lidamos com dados desde os primórdios da humanidade. Acontece que, nos tempos atuais, os avanços computacionais nos permitem guardar, organizar e analisar dados muito mais facilmente e com frequência muito maior.

Este panorama está longe de deixar de ser crescente. Basta imaginar, por exemplo, que vários dispositivos em nossas casas : geladeiras, TVs, lavadoras de roupa, cafeteiras, entre outros; Deverão estar conectados à internet em um futuro não muito distante. Esta previsão está dentro do que se conhece como *Internet das Coisas*.

Se olharmos para o que temos agora, já veremos uma grande mudança em relação às décadas anteriores: tomando como base apenas a internet, pense na quantidade de dados que são gerados diariamente somente nas redes sociais; repare na imensa quantidade de sites na Web; perceba que você é capaz de fazer compras on-line por meio até do seu celular, quando o máximo de informatização que as lojas tinham em um passado não muito distante eram sistemas isolados para gerenciar os seus estabelecimentos físicos.

As tecnologias atuais nos permitiram , e permitem , aumentar exponencialmente a quantidade de informações no mundo e, agora, empresas, governos e outras instituições precisam saber lidar com esta "explosão" de dados. O Big Data se propõe a ajudar nesta tarefa, uma vez que as ferramentas computacionais usadas até então para gestão de dados, por si só, já não podem fazê-lo satisfatoriamente.

A quantidade de dados gerada e armazenada diariamente chegou a tal ponto que, hoje, uma estrutura centralizada de processamento de dados já não faz mais sentido para a maioria das grandes entidades. O Google, por exemplo, possui vários *data centers* para dar conta de suas operações, mas trata todos de maneira integrada. Este "particionamento estrutural", é bom destacar, não é uma barreira para o Big Data , em tempos de [computação nas nuvens](http://www.infowester.com/cloudcomputing.php), nada mas trivial.

## 

## Aplicações comerciais do Big Data

lteryx

A empresa oferece um serviço self-service baseado em cloud que permite a analistas de negócio misturar dados, analisá-los e compartilhá-los para que decisões sejam tomadas. A Alteryx tem suas raízes em uma firma chamada SRC, formada em 1997, e afirma ter mais de 1 mil clientes ao redor do mundo

BigPanda

oferecendo um plataforma algorítmica que pode ajudar companhias a manterem serviços de tecnologia rodando perfeitamente.

Cogito

A ferramenta Cogito Dialog, que roda tanto em ambiente instalado quanto em nuvem, usa uma tecnologia analítica “comportamental” – que inclui análise de voz – para ajudar a melhorar o suporte telefônico em comunicações.

DataVisor

A promessa da solução é prover análise de big data para poder prever ações em sites de recomendação (como TripAdvisor e Yelp) e aplicativos móveis. A empresa afirma que seu motor analítico trabalha em conjunto com a plataforma Spark e é capaz de gerenciar “bilhões de usuários e trilhões de contas”.

Datos IO

Enquanto muitas startups focam sua abordagem em acesso veloz ou melhorar a visibilidade em grandes volumes de dados, a Datos IO se concentra em dar a certeza de que você não perderá a informação armazenada em meio a uma montanha de informações.

Fuzzy Logix

A solução foi desenhada para ajudar organizações a processarem informações sem processos separados de extração de dados em bancos, servidores analíticos intermediários ou ferramentas adicionais de storage. Fundada em 2007 por executivos e investidores oriundos do Bank of America, a companhia clama ter grandes nomes em sua lista de clientes.

napLogic

A empresa introduziu sua plataforma de integração corporativa como serviço em 2013, permitindo que empresas conectem melhor suas aplicações e dados através de nuvens híbridas. Há alguns meses, passou a oferecer suporte de sua solução para Spark, Cassandra e Microsoft Cortana Analytics.