PUC Minas Virtual

Arquiteturas e Serviços de Data Lakes e Data Warehousing

Apresentação

Apresentação do Professor

Ricardo Brito Alves

Formação Acadêmica:

- Graduado em Ciência da Computação pela Pontifícia Universidade Católica de Minas Gerais, em 1994.
- MBA em Gestão Estratégica de Projetos pela Una, em 2009.
- Mestrado em Engenharia Elétrica pela Pontifícia Universidade Católica de Minas Gerais, em 2018.
- Doutorando em Ciência da Computação pela Universidade Federal de Minas Gerais, em 2024.

Experiência Profissional:

- Desde 2002, atua com projetos de BI, Data mining.
- Atua há mais de 7 anos na área de Inteligência Artificial.

PUC Minas
Virtual

Unidade 1 – Dados e Estratégia

Data Driven

Data Driven

A gestão Data Driven é bem diferente dos modelos tradicionais, nos quais a tomada de decisão, geralmente, se baseava na intuição do dono ou nos “palpites” de especialistas, sem que houvesse dados reais para embasar essas atitudes.

Como Adotar o Data Driven?

O principal objetivo do Data Driven ***é entregar respostas mais precisas e confiáveis por meio de dados.***

A ideia é ***reduzir os "achismos" e entendam o valor dos dados.***

- **Transforme a sua cultura.**
- **Use boas soluções.**
- **Aprenda a entender os dados.**
- **Use os indicadores de performance.**

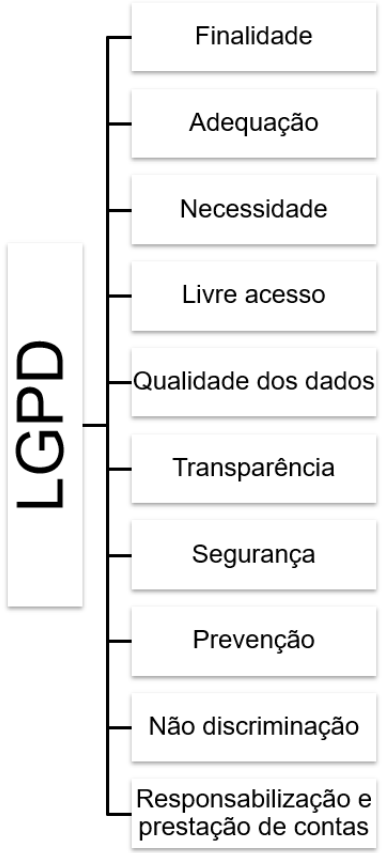
Data Driven

O Data Driven surgiu como extensão da Ciência de Dados utilizada para transformar dados estruturados e não estruturados em conhecimento.

Atualmente, no ambiente corporativo, isso é feito por meio de ferramentas como Big Data, Inteligência Artificial e Machine Learning, para obter insights a partir da coleta, cruzamento e interpretação de dados.

LGPD

LGPD



Lei Geral de Proteção de Dados – Lei 13.709/18

- Dispõe sobre o tratamento de dados pessoais – Nome, gênero, preferências de lazer, GPS, IP, etc.
- Conformidade / Compliance em agosto/2021.
- Regras para coletar, armazenar, tratar e compartilhar dados pessoais.
- Estende-se aos subcontratantes - fornecedores e parceiros.
- Multas podem chegar em 2% do faturamento / R$ 50 milhões (por infração).

Com a LGPD vou ter que apagar toda a minha base de contatos?

Caso a base de contatos seja formada pela base de clientes da empresa, não será necessário apagar. Contudo, o tratamento dos dados deve ser para a finalidade específica que justifique o seu uso de acordo com a base legal da respectiva finalidade.

No entanto, se a base de dados foi adquirida ou repassada por terceiros, provavelmente terá que apagá-la, caso os titulares dos dados não saibam que você trata estes dados pessoais. Uma forma de remediar esta questão seria entrar em contato e informá-los de modo a pegar o seu consentimento.

Como será comprovado que a empresa garante a proteção dos dados?

A comprovação se dará por meio dos documentos que devem compor o Relatório de Impacto à Proteção de Dados Pessoais, tais como: mapeamento do ciclo de vida do dado pessoal, mapeamento de risco, identificação dos agentes de tratamento em cada etapa ou processo de tratamento de dados, criação de políticas, códigos de conduta e comprovantes de treinamento, entre outros. Ademais, existem certificações mundiais sobre segurança da informação, tais como a ISO 27001 e 27002.

Qual o valor da multa caso minha empresa não tenha se adequado à LGPD ?

Em âmbito administrativo, as multas aplicadas poderão ser arbitradas pelos órgãos fiscalizadores, tais como PROCONs, Ministério Público, entre outros. Para cada multa será levado em consideração o grau de comprometimento da empresa com a segurança da informação e a proteção de dados pessoais, mediante a comprovação documental, a informação sobre o incidente aos titulares dos dados; o porte da empresa e seu faturamento.

LGPD

Toda minha empresa terá que fazer treinamento para a segurança dos dados pessoais?

O artigo 50 da LGPD determina que as empresas promovam ações educativas e de treinamento sobre o tratamento dos dados de acordo com a LGPD. Desta forma, é recomendável que a empresa faça treinamentos com profissionais especializados para conscientização sobre a lei bem como faça treinamentos específicos para os setores mais envolvidos no tratamento de dados pessoais de acordo com o modelo de negócio de cada empresa.

E o que são os Dados?

O Que é Dado?

- Os **dados** são elementos que constituem a matéria-prima da informação.
- Podemos defini-los como o **conhecimento bruto**, que ainda não foi devidamente tratado de forma a prover insights para uma organização.
- **Dados, de forma isolada, não conseguem transmitir uma mensagem clara.**

O Que é Dado?

- Exemplos: Manga, Roupa, Sapato.
- Definimos dado como uma sequência de símbolos quantificados ou quantificáveis. Portanto, um texto é um dado. De fato, as letras são símbolos quantificados, já que o alfabeto por si só constitui uma base numérica.

O Que é Dado?

- Também são dados **as imagens, sons e animação**, pois todos podem ser quantificados a ponto de alguém que entra em contato com eles ter eventualmente dificuldade de distinguir a sua reprodução, a partir da representação quantificada, com o original.
- É muito i**mportante notar-se que qualquer texto constitui um dado ou uma sequência de dados, mesmo que ele seja ininteligível** para o leitor.

Dados Quantitativos

- **Dados quantitativos ou numéricos** que podem ser:
  - **Discretos**
    - número de ovos postos pela tartaruga marinha.
    - número de ataques de asma no ano passado.
  - **Contínuos**
    - Volume, área, peso, massa.
    - Velocidade do carro.

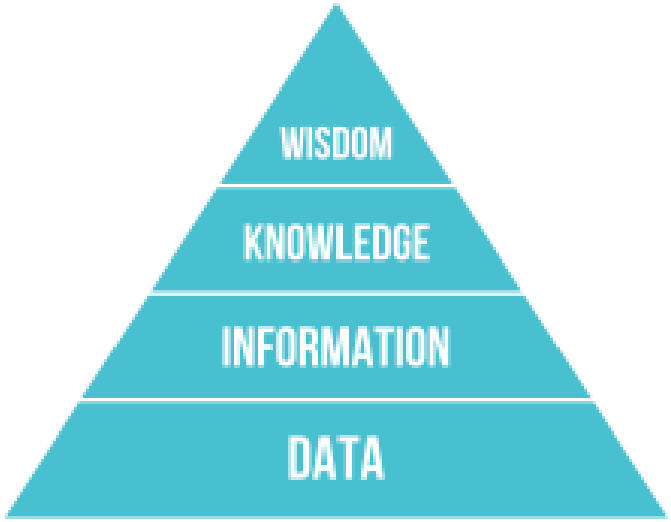
Dados Qualitativos

- **Dados qualitativos ou categóricos** podem ser:
  - **Nominais**
    - Cor dos olhos: castanho, verde, azul.
    - Fumante / Não Fumante.
    - Doente / Sadio.
  - **Ordinais**
    - Grau de periculosidade: baixa, média, alta.
    - Escolaridade: 1º grau, 2º grau, 3º grau.
    - Estágio do tratamento: inicial, intermediário, final.

Pirâmide DIKW

DIKW

- A Hierarquia ou pirâmide DIKW é o modelo utilizado para discussão de dados, informações, conhecimento, sabedoria e suas inter-relações.



- Fonte: Wikimedia Commons

DIKW

- A Hierarquia ou pirâmide **DIKW** é o modelo utilizado para discussão de **dados, informações, conhecimento, sabedoria** e suas **inter-relações**.
- Em **1955**, o economista e educador anglo-americano Kenneth Boulding apresentou uma variação na hierarquia que consiste em "sinais, mensagens, informações e conhecimento". No entanto, o primeiro autor a distinguir entre dados, informação e conhecimento **e também empregar o termo "gestão do conhecimento"** pode ter sido o educador americano Nicholas L. Henry, em um artigo de jornal de **1974**.

DIKW - Dados

- No contexto do DIKW, os dados são concebidos como fato, símbolos ou sinais, representando estímulos ou sinais, **que são "inúteis.**
- **Dados como fato:** fatos ou observações discretas e objetivas, que são desorganizadas e não processadas.
- **Dados como sinal:** estímulos sensoriais, que percebemos por meio de nossos “sentidos" ou "leituras de sinais", incluindo "leituras sensoriais e/ou sensoriais de luz, som, cheiro, sabor e toque”.
- **Dados como símbolos:** "símbolos“ ou "conjuntos de sinais que representam estímulos ou percepções empíricas" de "uma propriedade de um objeto, um evento ou de seu ambiente".

DIKW - Informação

- A informação está contida nas descrições e é diferenciada dos dados por ser "útil".
- **Estrutural x Funcional**
  - São "dados organizados ou estruturados, que foram processados de tal forma que a informação agora tem relevância para um propósito ou contexto específico e, portanto, é significativo , valioso, útil e relevante.
- **Simbólico x Subjetivo**
  - A informação pode ser concebida nos modelos DIKW como:
    - Universal, existindo como símbolos e signos;
    - Subjetivo, o significado ao qual os símbolos atribuem;
    - Ambos.

DIKW - Conhecimento

- **Conhecimento como Processado**
  - É a síntese de múltiplas fontes de informação ao longo do tempo“ "organização e processado para transmitir compreensão, experiência, aprendizado acumulando uma mistura de informações contextuais, valores, experiência e regras.
  - Responde a perguntas “como”.
  - Caracterizado pela crença justificável do indivíduo de que é verdade - contextuais, valores, experiência e regras.

DIKW - Sabedoria

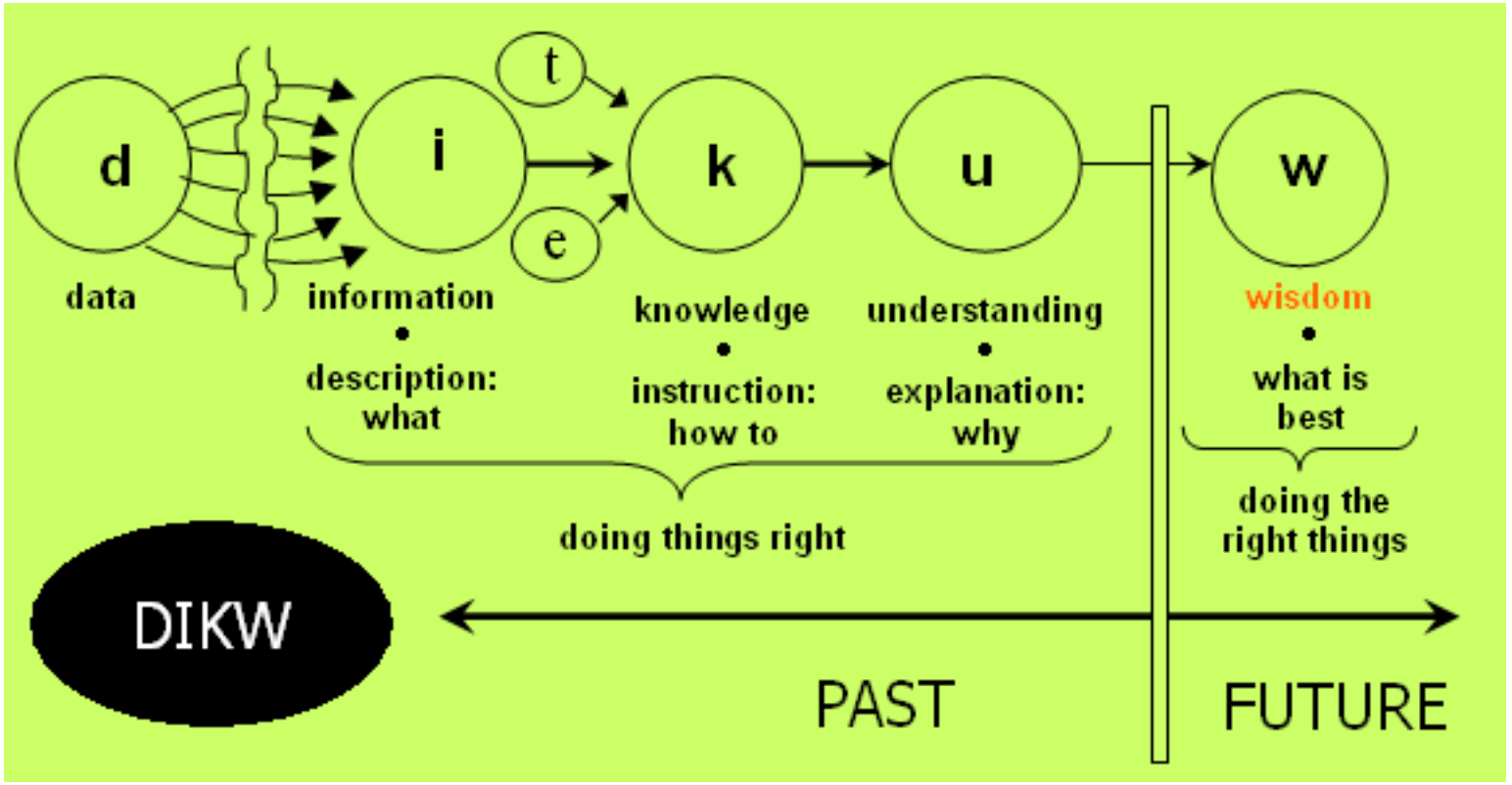
- Simplesmente é o **"conhecimento integrado—informação super útil"**.
- A sabedoria como "saber as coisas certas a fazer" e "a **capacidade de fazer julgamentos e decisões sensatas aparentemente sem pensar**".

Pirâmide DIKW - Fluxo

DIKW - Sabedoria

- A Pirâmide DIKW representa as relações entre dados, informações, conhecimento e sabedoria. Cada bloco de construção é um passo em direção a um nível superior - primeiro vêm os dados, depois as informações, depois o conhecimento e, finalmente, a sabedoria.
- Cada etapa responde a diferentes perguntas sobre os dados iniciais e agrega valor a eles.
- Quanto mais enriquecemos nossos dados com significado e contexto, mais conhecimento e percepções obtemos deles para que possamos tomar decisões melhores, informadas e baseadas em dados.

DIKW – Diagrama de Fluxo, GI e GC



- Fonte: Creative Commons Attribution 4.0 International

Gestão da Informação - GI

O gerenciamento de informações combina processos de negócios, procedimentos e tecnologia para organizar, proteger e acessar os dados de uma organização independentemente do formato, incluindo dados digitais, documentos em papel e arquivos de áudio e vídeo.

- A informação consiste em dados que um usuário inseriu em um computador, processou e colocou em contexto.
- As empresas devem gerenciar essas informações com ferramentas como Microsoft SharePoint ou Alfresco; um sistema de gerenciamento de documentos, como Microsoft OneDrive ou Google Drive.

Gestão do Conhecimento - GC

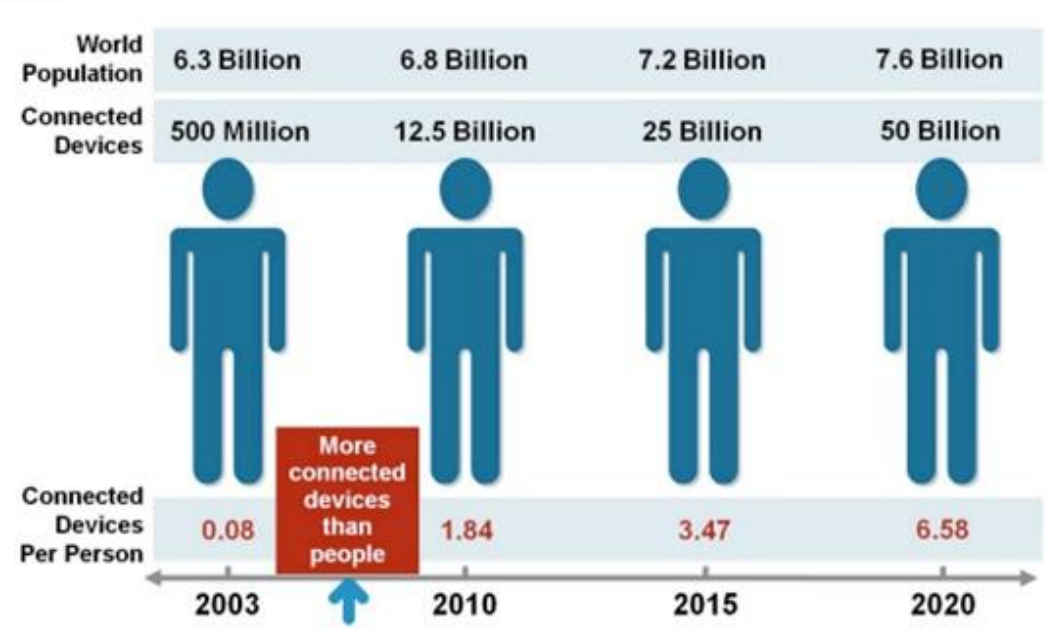
A gestão do conhecimento usa processos e ferramentas para transmitir sabedoria e compreensão de diferentes assuntos.

- Quando a informação é colocada no contexto de ser usada para maior compreensão de um assunto, ela se torna conhecimento.
- Esse conhecimento ajuda os funcionários a realizarem seus trabalhos, muitas vezes tornando-os mais eficientes. Também pode beneficiar os clientes de uma organização.
- **A gestão do conhecimento envolve a coleta, organização e compartilhamento do conhecimento.**
- Esse conhecimento pode estar na forma de documentos, vídeos e outros recursos destinados a ensinar as pessoas sobre um assunto específico.

Big Data

Big Data

- Qual é o tamanho da web?

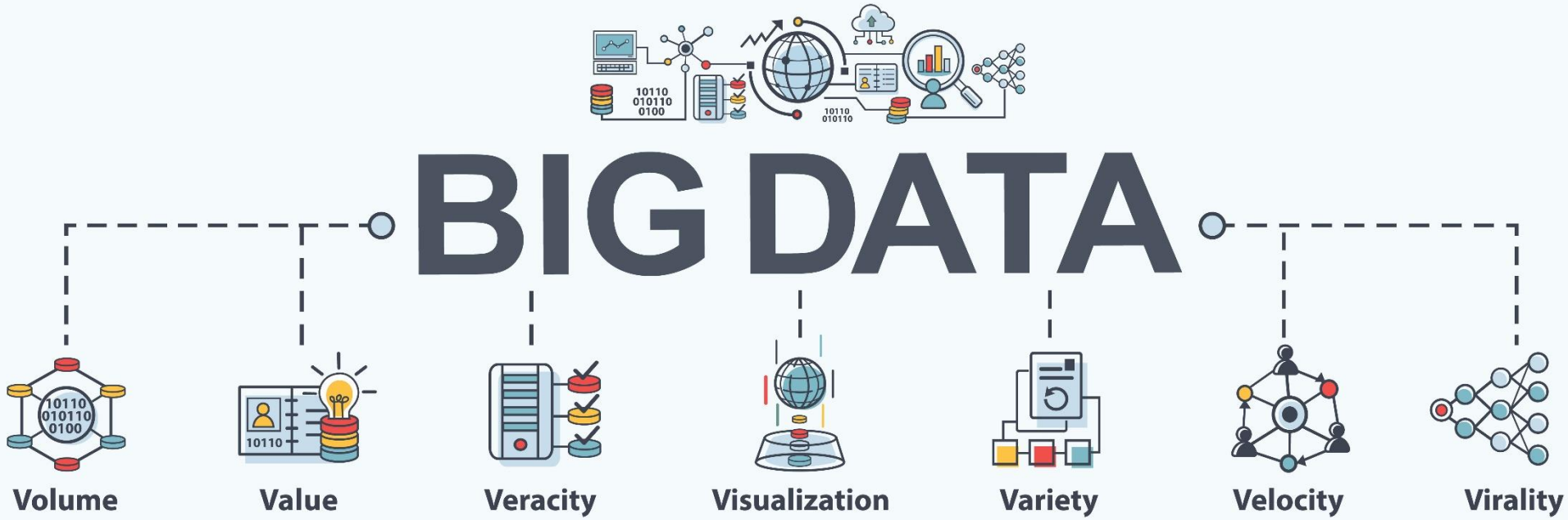


- Razões para usar o Big Data:
  - Entender padrões;
  - Prever situações;
  - Criar fronteiras;
  - Informar coleções de dados;
  - Estimar parâmetros escondidos;
  - Calibrar.

Fonte: https://www.researchgate.net/publication/316994150_THE_INTERNET_OF_THINGS_EVOLUTION

Big Data

- Os 3Vs
- Os 5Vs
- Os 7Vs



Os Vs do Big Data

- **Volume**: são os dados gerados a cada segundo. O volume define a quantidade de dados que temos – o que costumávamos medir em Gigabytes agora é medido em Zettabytes (ZB) ou mesmo Yottabytes (YB). A Internet das Coisas (IoT) cria um crescimento exponencial de dados. As projeções mostram o volume de dados mudando significativamente nos próximos anos.

Os Vs do Big Data

- **Velocidade**: um dos grandes desafios do Big Data. A velocidade representa a velocidade com que os dados são processados e se tornam acessíveis. Hoje, se a entrega não for em tempo real, geralmente não é rápida o suficiente.

Os Vs do Big Data

- **Variedade**: quanto mais dados e fontes, maior a complexidade, mas também é maior a chance de gerar informações úteis. A variedade descreve um dos maiores desafios do big data. Os insights podem vir sem estrutura. O ativo total pode incluir muitos tipos de dados, de XML a vídeo e SMS. Organizar os dados de maneira significativa não é tarefa simples quando os próprios dados mudam rapidamente.

Os Vs do Big Data

- **Variabilidade**: a variabilidade é diferente de variedade. Uma cafeteria pode oferecer seis misturas diferentes de café, mas se você obtém a mesma mistura todos os dias e tem um sabor diferente a cada dia, isso é variabilidade. O mesmo acontece com os dados. Se o significado mudar constantemente, isso pode afetar significativamente a homogeneização de seus dados.

Os Vs do Big Data

- **Veracidade**: o quanto a informação é verdadeira. A veracidade garante que os dados sejam precisos, o que requer processos para evitar que dados insuficientes se acumulem em seus sistemas. O exemplo mais simples é quando os contatos entram em seu sistema de automação de marketing com nomes falsos e informações de contato imprecisas. Quantas vezes você viu o Saci Pererê ou Mickey Mouse em seu banco de dados? É o clássico desafio de “entrar lixo, sair lixo”.

Os Vs do Big Data

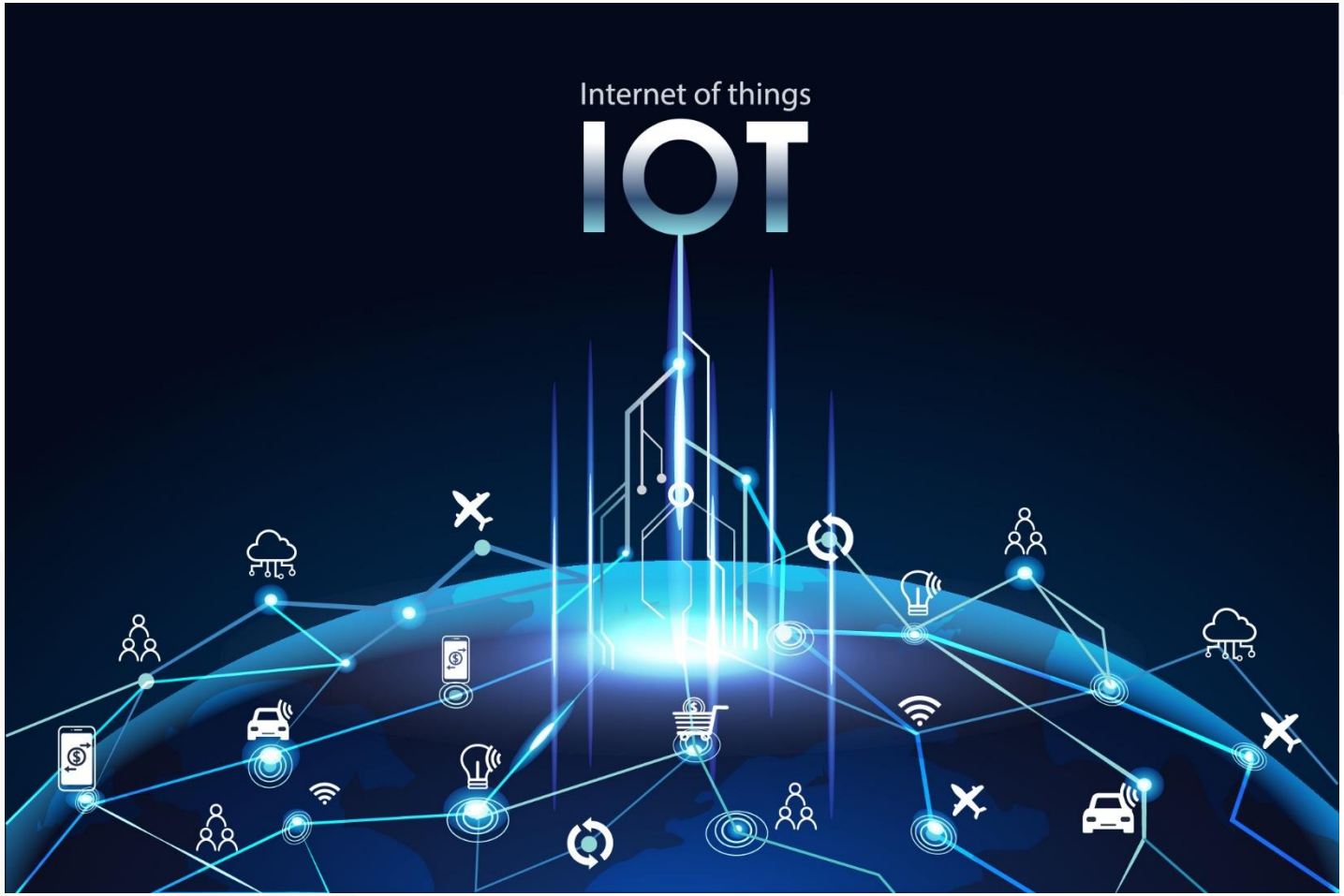
- **Visualização**: a visualização é fundamental no mundo de hoje. Usar tabelas e gráficos para visualizar grandes quantidades de dados complexos é muito mais eficaz em transmitir significado do que planilhas e relatórios repletos de números e fórmulas.

Os Vs do Big Data

- **Valor**: deve-se gerar informações com valor. Depois de abordar volume, velocidade, variedade, variabilidade, veracidade e visualização – o que leva muito tempo, esforço e recursos –, você quer ter certeza de que sua organização está obtendo valor dos dados.

IoT

IoT - Internet das Coisas



IoT – Internet of Things

Internet das Coisas ou, pela sigla **“IoT” (Internet of Things)**, traz um conceito atual e transformador sobre a conexão entre objetos físicos utilizando sensores, chips e softwares.

Quando falamos de “coisas” em IoT, estamos *nos referindo a qualquer objeto que teve a implementação de sensores e outros sistemas digitais para funcionar de forma mais inteligente por meio da troca de informações com pessoas e outros objetos*, não necessariamente utilizando conexões de internet.

Essas trocas de informações *podem ocorrer via radiofrequência (RFID), WiFi, Ethernet, Bluetooth*, entre outras formas de conexão existentes atualmente, e os sistemas de redes de comunicação existem em diferentes proporções, podendo ser conectados à web mundial ou apenas à rede de casa ou ao carro do usuário, por exemplo.

IoT - Internet das Coisas

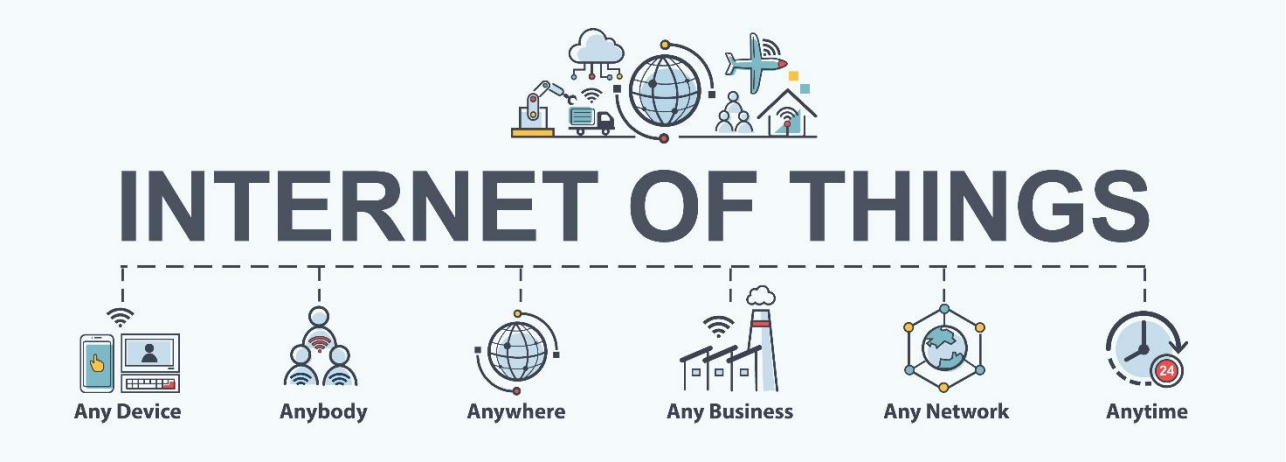
O termo foi utilizado pela primeira vez em 1999 para descrever um sistema onde os objetos poderiam ser conectados à internet.

Por causa do avanço do IoT, houve a necessidade de muito mais endereços de IP(IPv4 para IPv6 – 128 bits).

O fato de ter tantos dispositivos conectados à rede literalmente tem revolucionado o modo como vivemos.

IoT - Internet das Coisas

- 32 bilhões de coisas vão estar conectadas a internet.
- 10% de todos os dados serão gerados por sistemas embarcados (versus 2% atualmente).
- 21% dos mais valiosos dados serão gerados por sistemas embarcados (versus 8% atualmente).
- Dados de telemetria - semi-estruturados e contínuos - representam um desafio para bancos de dados relacionais, que exigem um esquema fixo e dados estruturados.



Cidades Inteligentes

- Poluição Sonora.
- Otimizar coleta de lixo.
- Controle de tráfego.
- Controle de distribuição de energia elétrica.
- Segurança pública.

Casas Inteligentes

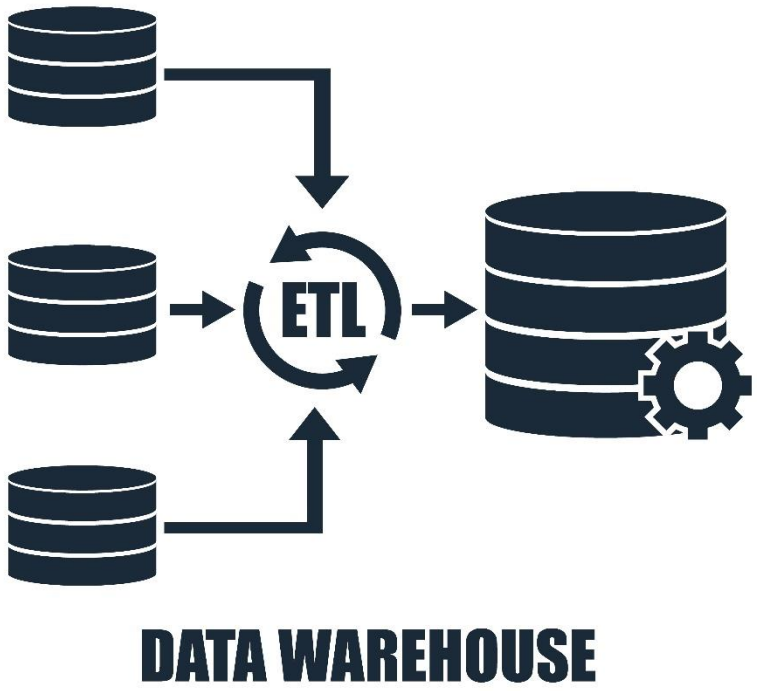
Inteligência Empresarial

Inteligência Empresarial

A gestão pode ser mais eficiente com o uso das métricas corretas.

Inteligência Empresarial

Processo de *coleta*, *organização*, *análise*, *compartilhamento* e *monitoramento* de informações para suporte a gestão de negócios.



Inteligência Empresarial

Data Driven

O Data Driven se baseia no uso de ferramentas tecnológicas capazes de coletar e analisar dados diferentes da sua empresa.

Esses dados, por sua vez, podem ser compilados por meio de BI ou Inteligência Artificial e ajudam o gestor a ter uma ideia mais precisa do seu negócio, facilitando a tomada de decisão estratégica.

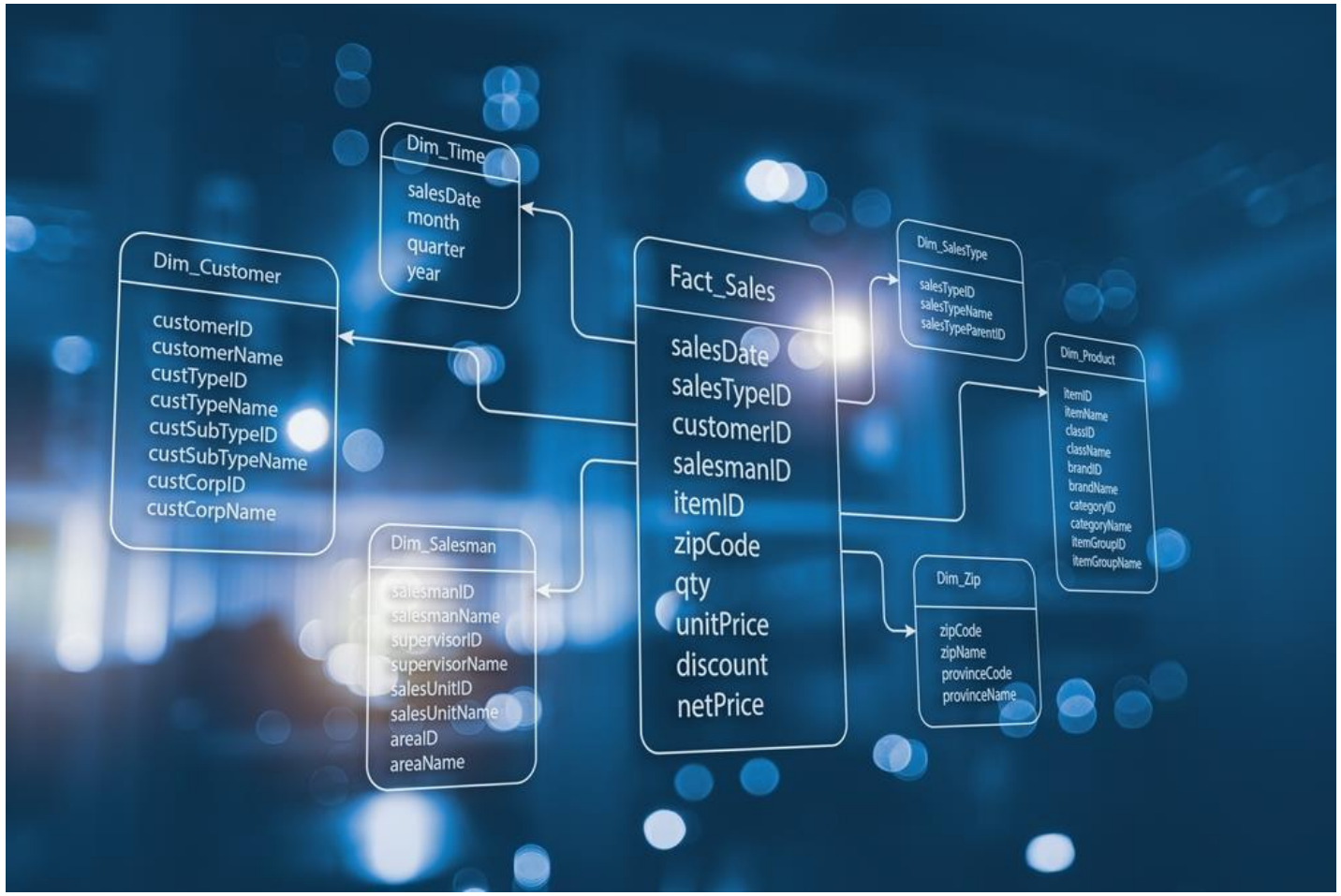
Dados Estruturados x Dados Não Estruturados

Dados Estruturados

Os dados estruturados, em essência, seguem uma estrutura rígida no qual foram armazenados.

Dados estruturados podem ser vistos como registros (ou transações) em um ambiente de banco de dados; por exemplo, linhas em uma tabela de um banco de dados SQL.

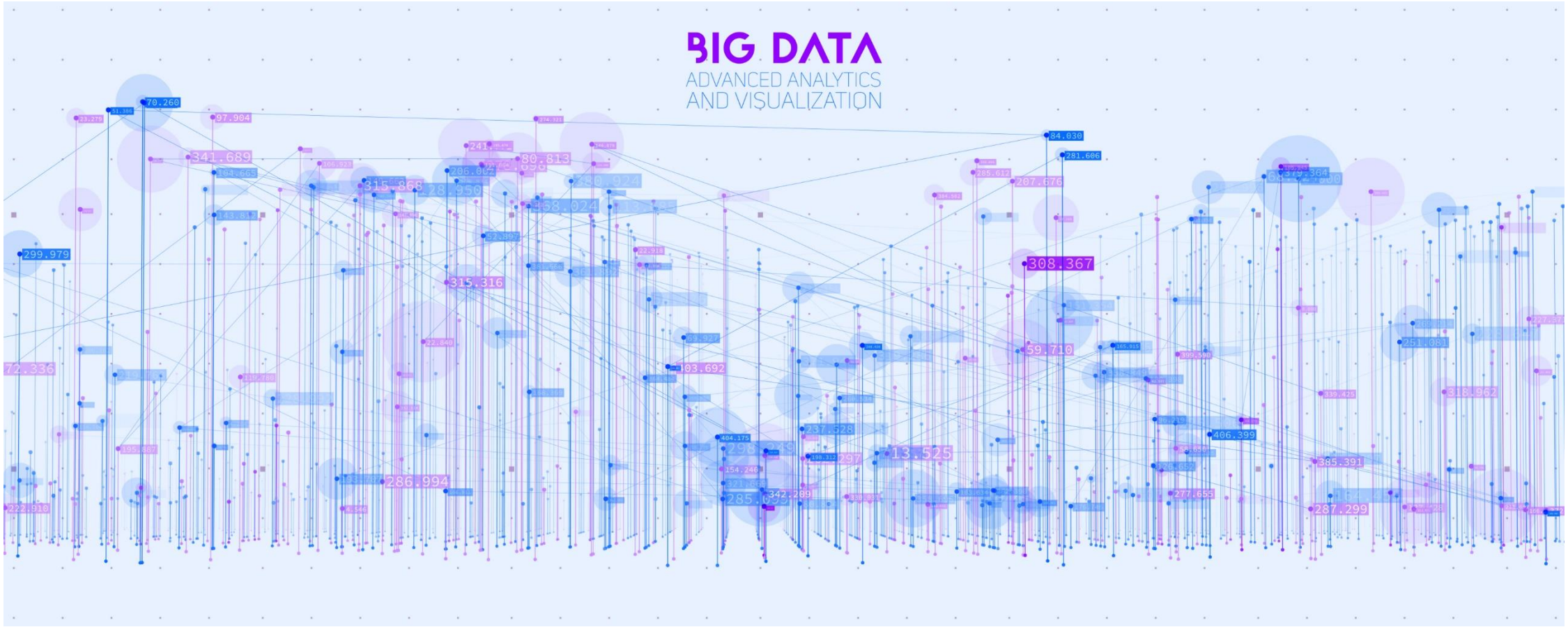
Dados Estruturados



Dados Não Estruturados

Os dados não estruturados vêm de fontes diversas, tais como internet e redes sociais. Esses dados são coletados nos mais variados formatos – texto, imagem, som, vídeos – e precisam passar por um processamento antes de serem cruzados com outros dados e analisados.

Dados Não Estruturados



Dados Estruturados

Os dados não estruturados são mais abundantes do que os dados estruturados.

Exemplos de dados não estruturados são:

Dados de mídia e entretenimento, dados de vigilância, dados geoespaciais, áudio, dados meteorológicos, e-mails, Internet das Coisas (IoT).

#2 Data Analytics

2 Data Analytics

Analytics

Big Data e o Data Analytics

- O **Big Data** traz novos desafios na gestão de dados, *como a manutenção de uma linearidade dos dados, sua integridade e qualidade,* a fim de que eles possam ser transformados em **informação útil.**
- Soluções de **Inteligência Operacional** podem *correlacionar e analisar dados de fontes variadas em várias latências (desde o batch, até o tempo real), para revelar informações importantes.*
- O **Data Analytics** *pode ser usado em vários segmentos de mercado*. Os bancos usam essa estratégia para evitar possíveis fraudes. Na educação, você pode medir o progresso dos alunos e avaliar a eficácia do sistema. No varejo, o principal uso é rastrear as características sociais e comportamentais dos clientes, de modo a prever tendências e hábitos.

Data Analytics

Geralmente, o processo de análise do fluxo de dados segue as seguintes etapas:

- ✓ Coleta.
- ✓ Ingestão e transformação.
- ✓ Armazenamento.
- ✓ Análise.
- ✓ Desenvolvimento de algoritmos.
- ✓ Visualização.

É importante avaliar *quais são os principais insights desejados na etapa de visualização, ou até mesmo o levantamento de quais problemas de negócio você gostaria de resolver.*

Data Driven x Analytics Driven

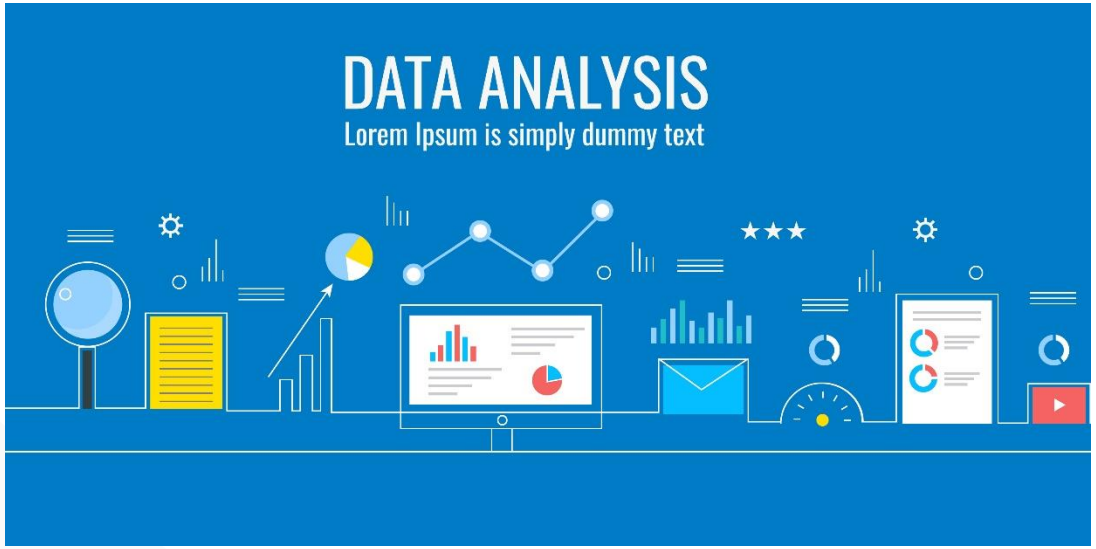
Embora ambos os conceitos derivem da Ciência de Dados, existem algumas diferenças fundamentais.

- O processo **Data Driven** possui uma abordagem mais quantitativa, uma vez ***que se baseia em números e modelos preditivos.***
- O **Analytics Driven** também considera o aspecto qualitativo, estabelecendo padrões e correlações entre os dados. Podemos dizer que ***o Analytics Driven vai além da análise de dados, interpretando, também, o contexto e outras variáveis ligadas a essas informações que podem impactar nos resultados.***

Papel do Cientista de Dados

Analista de Dados

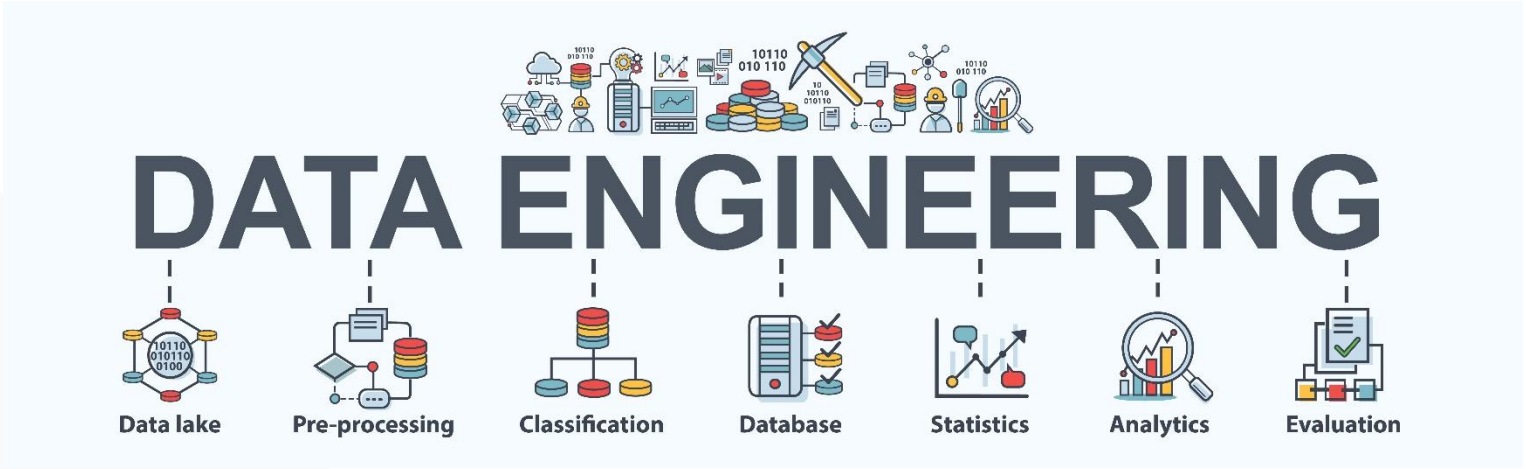
Um Analista de Dados é o profissional dedicado a ***garimpar, catalogar, analisar e interpretar o enorme volume de informações e dados que são obtidos por meios digitais ou analógicos.***



Engenheiro de Dados

Um Engenheiro de Dados é o profissional dedicado ao ***desenvolvimento, construção, teste e manutenção de arquiteturas, como um sistema de processamento em grande escala.***

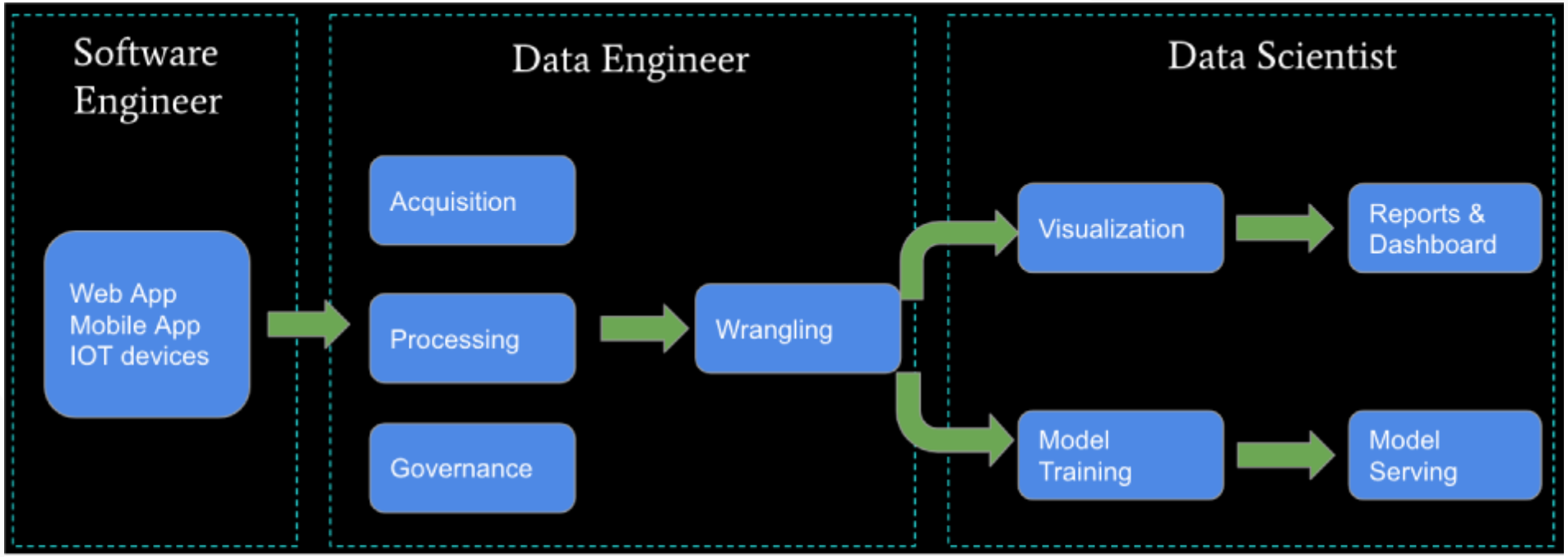
O Engenheiro de Dados é responsável por criar o pipeline dos dados, ***desde a coleta até a entrega*** para análise ou para alimentar um produto ou serviço baseado em análise preditiva já em produção.



Cientista de Dados

Cientistas de Dados ***recebem uma enorme massa de dados (estruturados e não estruturados)*** e usam suas habilidades em Matemática, Estatística, Ciência da Computação e Programação para ***limpar, tratar e organizar os dados.***

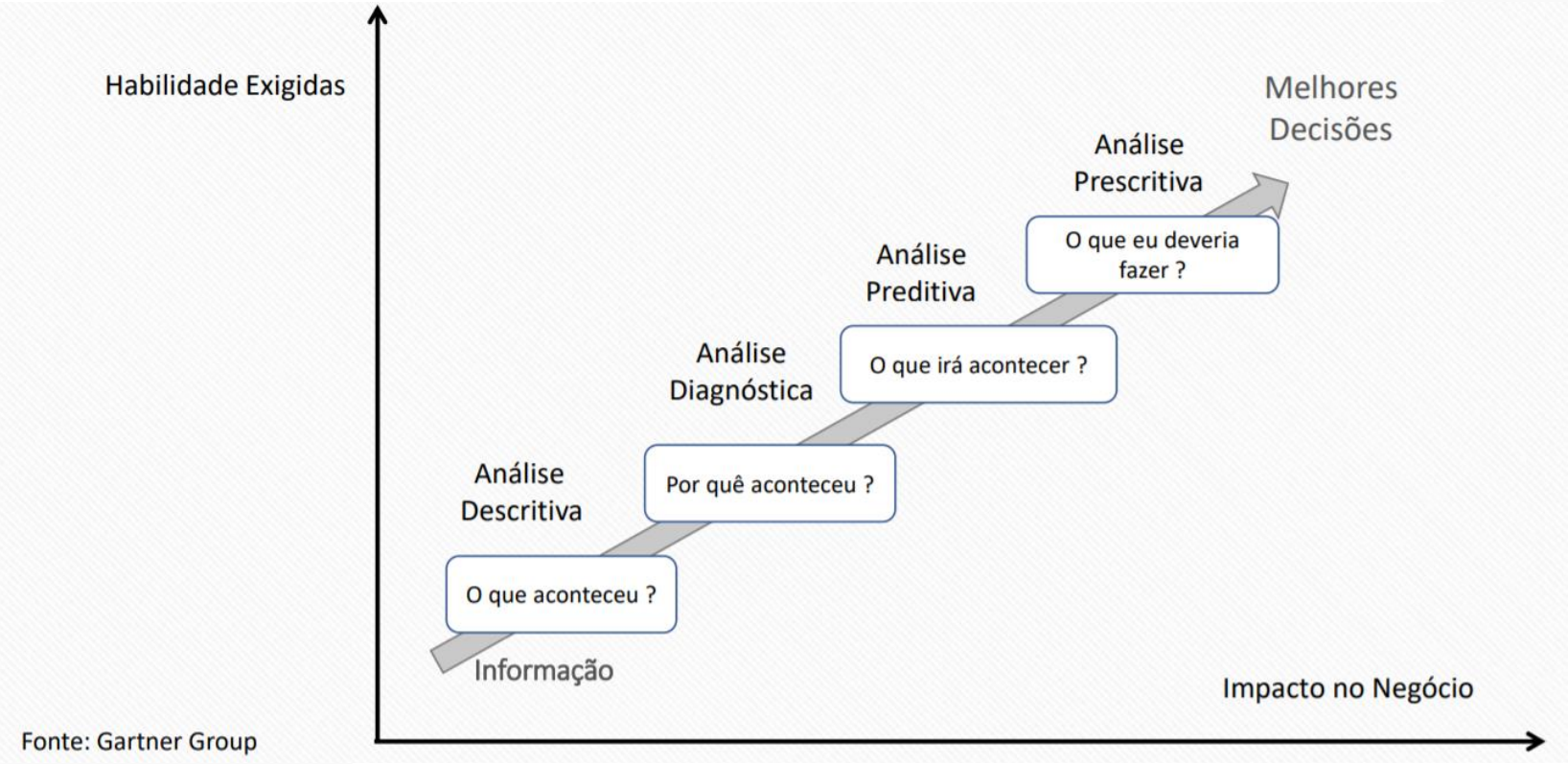
Cientista de Dados



Source: https://towardsdatascience.com/voicing-for-data-engineering-the-unsung-hero-b91b6ef39dcd

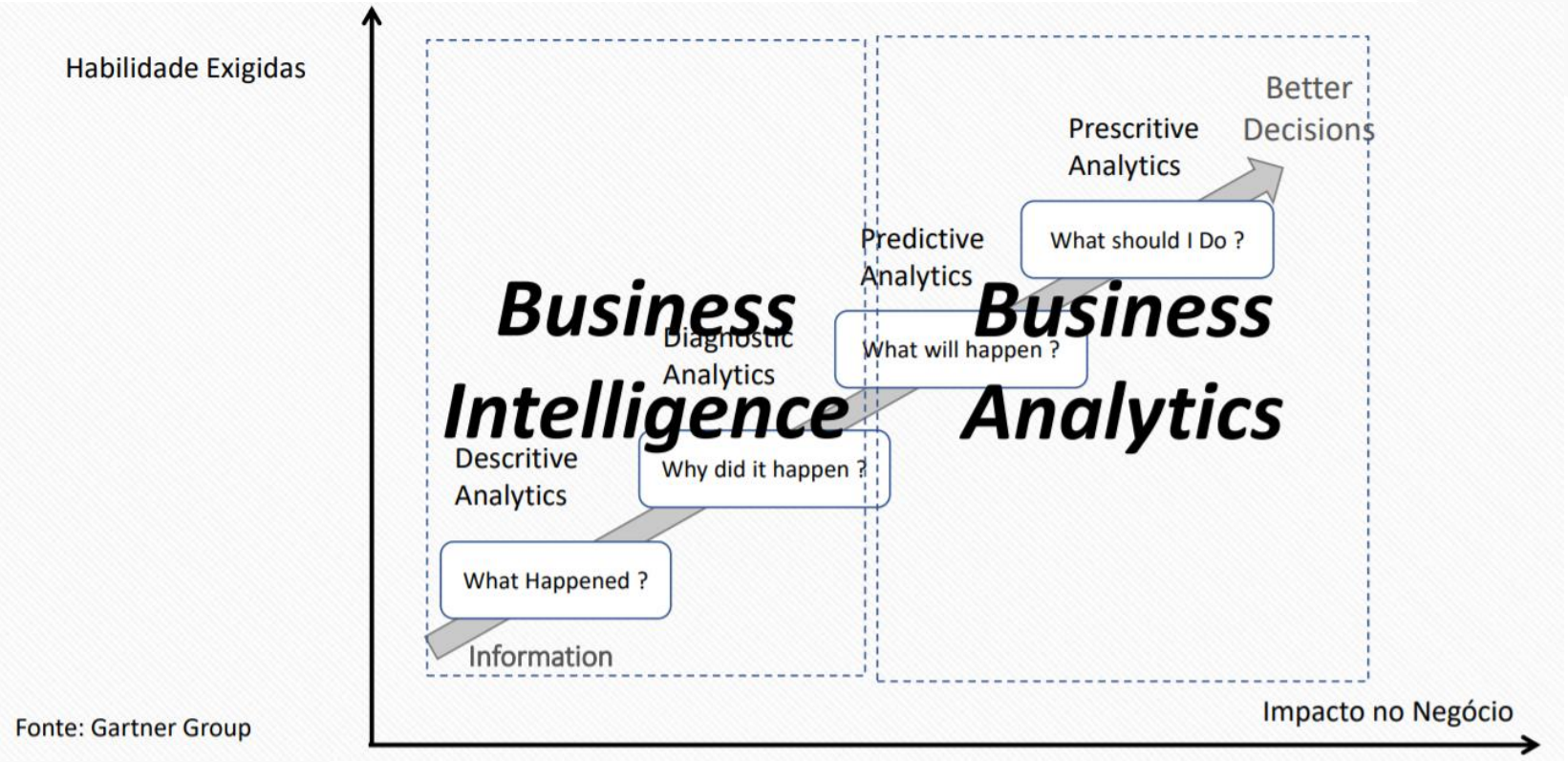
BI x BA

Business Intelligence x Business Analytics



Fonte: Gartner Group

Business Intelligence x Business Analytics



Fonte: Gartner Group

Informação Estratégica

Inteligência Empresarial

Poluição Sonora.
Otimizar a coleta de lixo.
Controle de tráfego.
Controle de distribuição de energia elétrica.
Segurança pública.

Inteligência Empresarial

Análise de Rotas
Geomarketing
Análise de Crimes
Análise de Doenças
Análise de Desastres

Inteligência de Dados

Inteligência de dados é o processo de organizar informações, sejam estruturadas ou não, e integrá-las a sistemas de softwares para que sejam melhor analisadas e interpretadas e possam contribuir para tomadas de decisões mais assertivas.

Dados Ajudando as Empresas

Competitividade: é possível converter os dados da organização em rentabilidade. O cruzamento das informações podem gerar relatórios importantes para as decisões estratégicas, o que aumenta a competitiva da empresa no mercado.

Inovação: a inteligência dos dados permite que as corporações se tornem mais sensitivas em relação às mudanças no comportamento dos clientes no mercado. Por meio das informações certas, a empresa consegue investir no fortalecimento de seu relacionamento com os clientes, explorar novas demandas e lançar tendências.

Dados Ajudando as Empresas

Desempenho: outra forma que a inteligência de dados auxilia as empresas é na utilização de informações importantes para avaliar os seus KPIs (Key Performance Indicators), os quais avaliam o desempenho dos processos executados para que possam ser otimizados e alinhados aos objetivos corporativos.

Estratégia: uma análise efetiva de dados serve justamente para realizar o alinhamento estratégico da organização, focando-o nas tendências e realidade do mercado.

OKR

OKR - Objectives and Key-Results ou Objetivos e Resultados-Chave: metodologia que estabelece uma direção clara de onde o negócio ou produto quer chegar com base na definição de objetivos (metas e intenções) e resultados (indicadores de progresso).

KPI x OKR

- Os **KPIs** demonstram um comportamento esperado do seu produto ou negócio. Você observa uma fotografia dos KPIs para saber a posição do seu produto ou negócio e identificar se ele está dentro de um padrão esperado.
- Já os **OKRs** são planejados para elevar o patamar do seu produto ou negócio.

Informação Estratégica

Copa do Mundo 2014: Alemanha campeã.

A utilização de BI foi um dos fatores responsáveis pela vitória da seleção alemã na Copa do Mundo de 2014. Ele serviu como ferramenta para impulsionar a produtividade do time.

Com o BI, foi possível analisar dados relevantes para o desempenho dos jogadores, tais como:

- Número de passes;
- Velocidade em campo;
- Finalizações;
- Defesas;
- Penalidades.

Com base nas informações obtidas, foi possível identificar quais atletas tinham melhor rendimento e, assim, escalar o melhor time titular.


PUC Minas
Virtual