

Análise de Dados com Base em Processamento Massivo em Paralelo

Aula 1: Introdução

Cristina Dutra de Aguiar Ciferri
ICMC/USP
cdac@icmc.usp.br

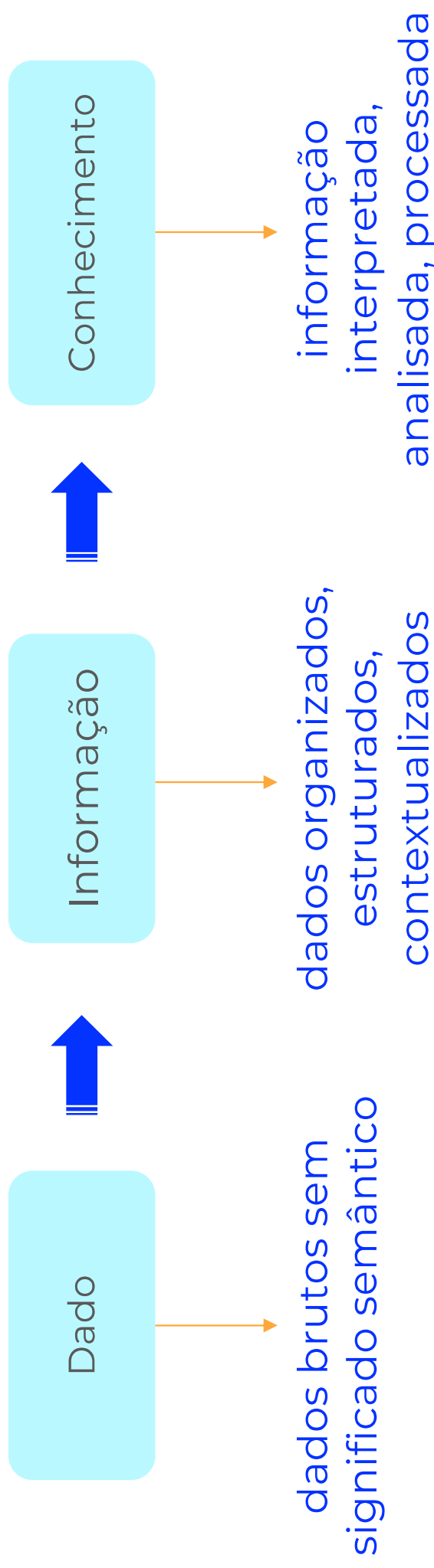


Agenda

- Business Intelligence
- Data Warehousing
- Diferenças entre os Ambientes Operacional e Informacional

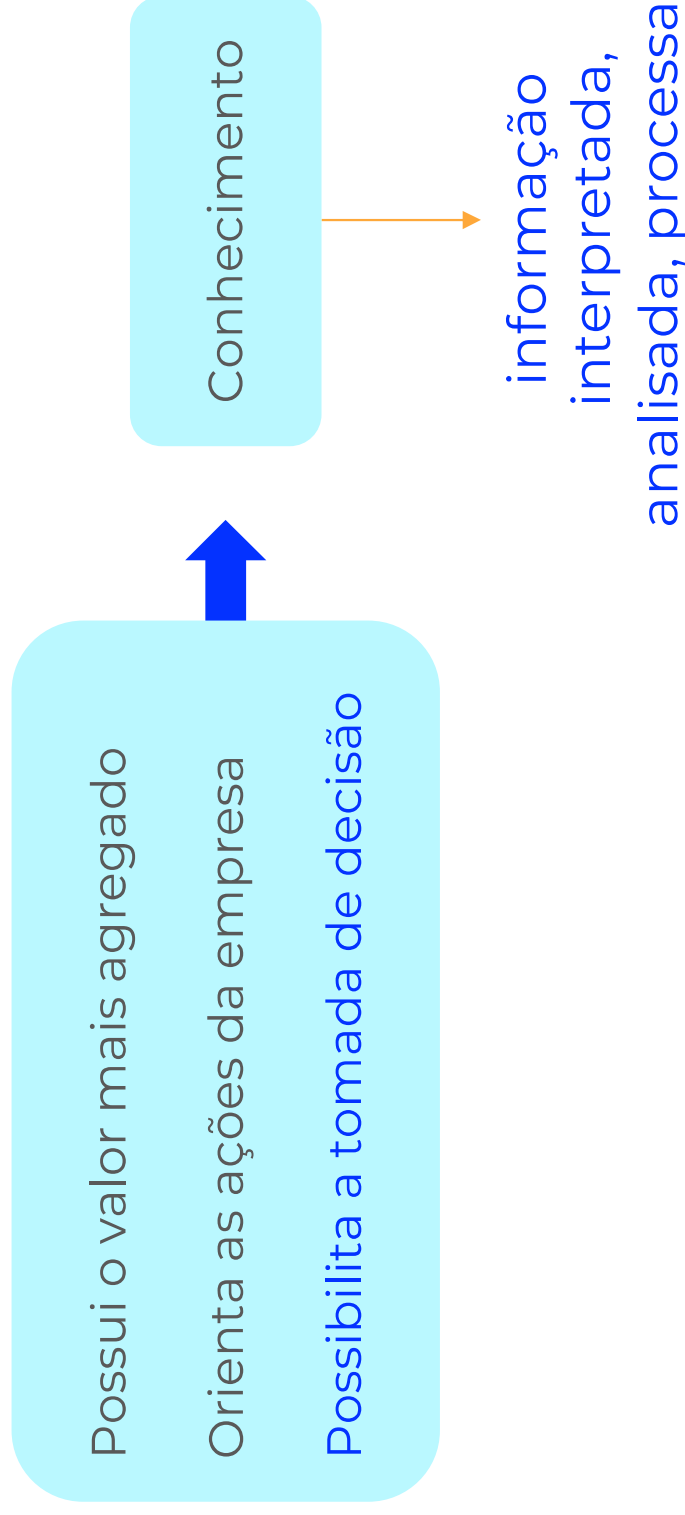
Business Intelligence (BI)

- Processo de transformação dos dados em **informação** e depois em **conhecimento**



Business Intelligence (BI)

- Processo de transformação dos dados em **informação** e depois em **conhecimento**



Objetivos

- Satisfazer às necessidades dos usuários de sistemas de suporte à decisão
 - Analisar de forma eficiente e eficaz os dados corporativos
 - Compreender melhor a situação do negócio
 - Melhorar o processo de tomada de decisão estratégica
- Fornecer um conjunto de processos para
 - Produzir a [informação certa](#), para a [pessoa certa](#), na [hora certa](#)

Pensamento Motivacional

A obtenção de informações estratégicas, relativas ao contexto de tomada de decisão, é de suma importância para o sucesso de uma empresa.

Tais informações permitem à empresa um **planejamento rápido** frente às mudanças nas condições do negócio, essencial na atual conjuntura de um mercado globalizado.

Tarefas

- Criação de **medidas** (métricas) que indiquem o progresso da empresa com relação às suas metas
- Geração de **relatórios** que possibilitem análises complexas e que possuam visualização apropriada
- Uso **exploratório** das informações com possibilidade de identificar tendências e realizar previsões
- Uso de ferramentas que possibilitem o **trabalho colaborativo** e que ofereçam suporte desde a obtenção dos dados até a geração do conhecimento
- **Gerenciamento do conhecimento** para realizar a tomada de decisão estratégica bem fundamentada, resultando em ações bem sucedidas que garantam um maior retorno sobre o investimento

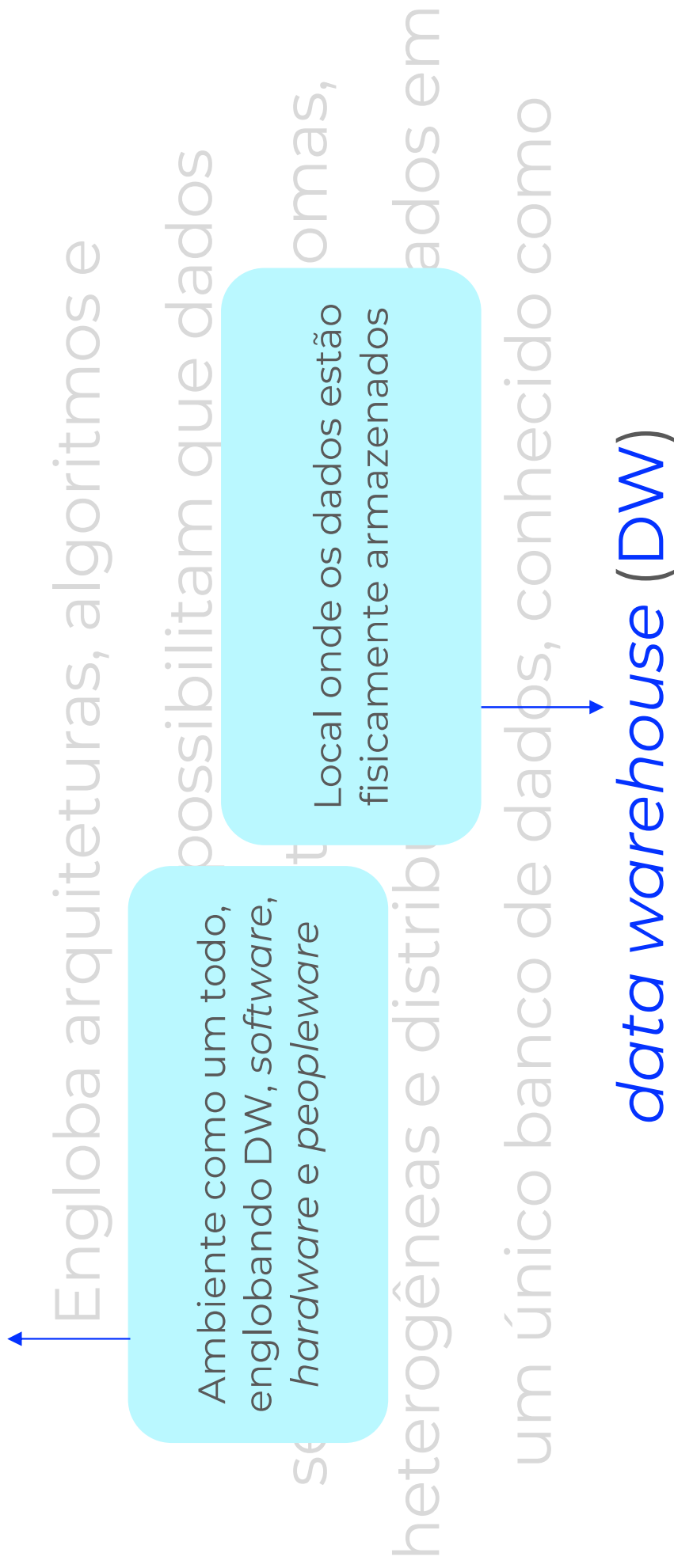
Agenda

- Business Intelligence
- Data Warehousing
- Diferenças entre os Ambientes Operacional e Informacional

Data Warehousing

Engloba arquiteturas, algoritmos e ferramentas que possibilitam que dados selecionados de fontes de dados autônomas, heterogêneas e distribuídas sejam integrados em um único banco de dados, conhecido como *data warehouse (DW)*

Data Warehousing



Acesso às Informações

- **Etapa ETL** (extração, transformação e carga)
 - Dados de interesse de cada fonte de dados são extraídos previamente, devendo ser traduzidos, filtrados, integrados aos dados relevantes de outras fontes e finalmente armazenados no **DW**
- **Etapa de análise e consulta**
 - As consultas, quando realizadas, são executadas diretamente no **DW**, sem acessar as fontes de dados originais

Aplicação 1: Área Médica

- Foco em **número** de pacientes
 - Dados integrados de **pacientes**, tipos de **exame**, **hospitais** nos quais os exames foram feitos e **datas** de coleta dos exames
- Exemplos de **análise**
 - Qual o número de pacientes que testaram positivo para a COVID-19 por mês?
 - Qual o número de pacientes de cada faixa etária que tiveram complicações devido à COVID-19, considerando cada um dos estados do Brasil?
 - Qual a porcentagem de pacientes que vieram a falecer devido às complicações causadas pela COVID-19 de janeiro a agosto de 2020?

Aplicação 1: Área Médica

- Foco em **número** de pacientes
 - Dados integrados de **pacientes**, tipos de **exame**, **hospitais** nos quais os exames foram feitos e **datas** de coleta dos exames
- Exemplos de **conhecimento**
 - Curva de evolução de uma determinada doença ao longo dos meses
 - Características dos pacientes (por exemplo, tipo sanguíneo, faixa etária, faixa salarial) mais suscetíveis a uma determinada doença
 - Localidades geográficas que podem ser consideradas como epicentros

Aplicação 2: Cadeia de Supermercados

- Foco em **unidades** vendidas de produtos e seus **lucros**
 - Dados integrados de **produtos** vendidos, **promoções** realizadas, **filiais** nas quais os produtos foram vendidos e **datas** das vendas
- Exemplos de **análise**
 - Quais as vendas mensais dos produtos de uma determinada marca nos últimos três anos?
 - Quais as vendas diárias dos produtos nas diferentes filiais, de acordo com as promoções realizadas no período do dia dos namorados e do dia das mães?
 - Quais os lucros obtidos nas vendas de produtos para tratamento estético?

Aplicação 2: Cadeia de Supermercados

- Foco em **unidades** vendidas de produtos e seus **lucros**
 - Dados integrados de **produtos** vendidos, **promoções** realizadas, **filiais** nas quais os produtos foram vendidos e **datas** das vendas
- Exemplos de **conhecimento**
 - Produtos mais vendidos e menos vendidos e os lucros ou prejuízos associados
 - Impacto das promoções realizadas na venda dos produtos e nos lucros obtidos
 - Filiais deficitárias que precisam ser fechadas ou remodeladas

Aplicação 3: BI Solutions

- Empresa exemplo que será usada ao longo da disciplina



BI Solutions

Razão social:

BI Solutions

Slogan:

Desenvolvimento de soluções inteligentes para o seu negócio

Sobre a empresa:

A BI Solutions é uma empresa de desenvolvimento de *software* totalmente brasileira e com alcance internacional, que implementa soluções inteligentes para atender os clientes dos mais diversos setores de negócio.

Aplicação 3: Folha de Pagamento da BI Solutions

- Foco nos **salários** dos funcionários e na **quantidade** de lançamentos
 - Dados integrados de **funcionários**, **cargos** ocupados por estes, **filiais** nas quais os funcionários trabalham e **datas** de pagamento
- Exemplos de **análise**
 - Quais os gastos mensais em salários dos funcionários?
 - Quais as filiais que possuem o maior gasto anual em salários de funcionários?
 - Qual a média salarial dos funcionários ocupantes de cargos de nível superior em uma determinada filial no primeiro trimestre de 2019?

Aplicação 3: Folha de Pagamento da BI Solutions

- Foco nos **salários** dos funcionários e na **quantidade** de lançamentos
 - Dados integrados de **funcionários**, **cargos** ocupados por estes, **filiais** nas quais os funcionários trabalham e **datas** de pagamento
- Exemplos de **conhecimento**
 - Cargos que receberam a maior soma de salários e filiais relacionadas
 - Graus de escolaridade dos funcionários e seus impactos nas médias salariais dos mesmos, bem como nos cargos ocupados
 - Curvas de gastos em salários dos funcionários por mês nos últimos anos

Questionamento

- Esses tipos de análise são possíveis de serem realizados usando os sistemas existentes?
 - Aplicações de banco de dados *stand-alone*
 - Aplicações desenvolvidas de forma centralizada
 - Sistemas legados
 - Uso de planilhas
- Limitação
 - Análises muito custosas com tempos de respostas proibitivos para a produção da [informação certa](#), na [hora certa](#), para a [pessoa certa](#)

Análise usando Sistemas Existentes

- Exemplos de desafios
 - Dados de interesse de análise encontram-se espalhados nos diferentes sistemas, assumem diferentes formatos e requerem processos de limpeza acurados
 - Aplicações encontram-se projetadas com foco em normalização, visando diminuir ou até mesmo eliminar a redundância
 - O foco em normalização impacta a complexidade de se especificar consultas analíticas
 - A complexidade das consultas impacta no desempenho das mesmas
 - O tratamento de dados temporais usualmente é incipiente

Vantagens do Data Warehousing

- Análises podem ser realizadas eficientemente
 - DW contém dados integrados
 - DW é projetado com foco em assuntos de interesse
 - DW modela explicitamente o aspecto temporal
- Maior disponibilidade dos dados
 - Consultas são executadas diretamente no DW sem acessar as fontes originais
- Autonomia das fontes de dados originais
 - Processamento local nas fontes de dados originais não é afetado por causa da participação destes no ambiente de data warehousing

Vantagens do Data Warehousing

- Análises podem ser realizadas eficientemente
 - DW contém dados integrados, cuja heterogeneidade já foi eliminada
 - DW é projetado com foco em assuntos de interesse
 - DW modela explicitamente as fontes de dados
- Maior disponibilidade
 - Consultas são executadas diretamente no DW, sem a necessidade de acessar as fontes originais
- Autonomia das fontes de dados originais
 - Processamento local nas fontes de dados originais não é afetado por causa da participação destes no ambiente de *data warehousing*

Agenda

- Inteligência do Negócio
- Data Warehousing
- Diferenças entre os Ambientes Operacional e Informacional

Separação entre os Ambientes

- Ambientes fundamentalmente diferentes
 - Dados
 - Tecnologias
 - Usuários
 - Necessidades de processamento
 - Necessidades de segurança
 - Requisitos de desempenho das aplicações

Ambientes Operacional e Informacional

- Ambiente Operacional
 - Constituído por aplicações que oferecem **suporte ao dia a dia** do negócio
 - Sistemas existentes
- Ambiente Informacional
 - Constituído por aplicações que analisam o negócio
 - *Data warehousing*

Ambientes Operacional e Informacional

- Ambiente Operacional
 - Constituído por aplicações que oferecem suporte ao dia a dia do negócio
 - Sistemas existentes
- Ambiente Informacional
 - Constituído por aplicações que **analisam** o negócio
 - *Data warehousing*

Ambientes Operacional e Informacional

- Ambiente Operacional
 - Constituído por aplicações que oferecem suporte ao dia a dia do negócio
 - Sistemas existentes
- Ambiente Informacional
 - Constituído por aplicações que analisam o negócio
 - *Data warehousing*

DW é mantido
separadamente dos
bancos de dados
operacionais

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Principal Característica	voltado ao processamento de transações (OLTP)	voltado ao processamento de consultas (OLAP)
Tipos de Operação mais Frequentes	inserção remoção atualização	leitura (consulta)
Foco do Desempenho	produtividade das transações	produtividade das consultas

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Principal Característica	voltado ao processamento de transações (OLTP)	voltado ao processamento de consultas (OLAP)
Tipos de Operação mais Frequentes	inserção remoção atualização	leitura (consulta)
Foco do Desempenho	produtividade das transações	produtividade das consultas

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Tipos de Usuários	administradores do sistema, projetistas, usuários finais	usuários de SSD (ex.: executivos, analistas, gerentes)
Número de Usuários Concorrentes	grande	relativamente pequeno
Interações com os Usuários	estáticas, predefinidas	dinâmicas, exploratórias

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Tipos de Usuários	administradores do sistema, projetistas, usuários finais	usuários de SSD (ex.: executivos, analistas, gerentes)
Número de Usuários Concorrentes	grande	relativamente pequeno
Interações com os Usuários	estáticas, predefinidas	dinâmicas, exploratórias

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Volume das Operações	relativamente alto	relativamente baixo
Características das Operações	mais simples, acessando menos registros por vez	mais complexas, acessando muitos registros por vez

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Volume das Operações	relativamente alto	relativamente baixo
Características das Operações	mais simples, acessando menos registros por vez	mais complexas, acessando muitos registros por vez

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Projeto do Banco de Dados	normalizado	multidimensional
Granularidade dos Dados	nível de detalhe específico	diferentes níveis de detalhe
Volume de Dados	megabytes a gigabytes	<i>gigabytes a terabytes a petabytes</i>

Diferenças entre os Ambientes

	Ambiente Operacional	Ambiente Informacional
Projeto do Banco de Dados	normalizado	multidimensional
Granularidade dos Dados	nível de detalhe específico	diferentes níveis de detalhe
Volume de Dados	<i>megabytes a gigabytes</i>	<i>gigabytes a terabytes a petabytes</i>

Diferenças entre os Ambientes

Exemplos de Aplicação	Ambiente Operacional	Ambiente Informacional
	transações bancárias	planejamento de <i>marketing</i>
	empréstimos de livros	análise financeira
	contas a pagar	tomada de decisão
	matrículas em cursos	planejamento estratégico