



LABORATÓRIO DE BANCO DE DADOS

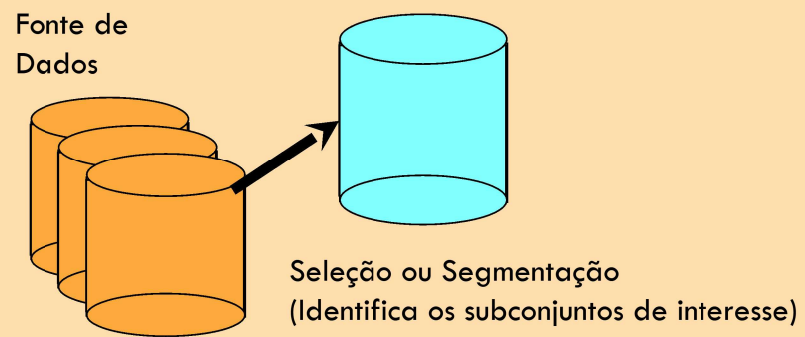
IBD100

- Data Mining
 - Cenário
 - Exemplo
 - Fases de Execução
 - Vantagens e Desvantagens
- Data Warehouse – Depósito de Dados
 - Aproximação...
 - Software ou Conceito?
 - Método ou Software?
 - Os programas que podem ser DWH
- Big Data
 - Conceituação
 - Os 5 'Vs'
 - Exemplos
 - Big Data x DW x DM?
-

- Cenário inicial
- Década de 1970-1990
- Aumenta a utilização de Sistemas de informação como ferramenta de apoio à decisão, dando ênfase aos programas do sistemas, pois é onde estavam implementadas 'as regras do negócio'.
- O volume de dados cresce muito rápido ao mesmo tempo que os custos de armazenamento diminuem: consequência: aumenta o volume de dados armazenados e aumenta a necessidade de analisar grandes volumes de dados.
- Inicia-se o estudo sobre as técnicas de *agrupamento* e *tratamento* de dados para *identificação* de *comportamentos* do ambiente que gerou os dados.
 - Qual é o comportamento do mercado ao consumo de certo produto e como este mercado vai reagir ao lançamento de outro produto similar?
 - Quais são as queixas latentes do consumidores? (Aqueles que podem rejeitar um produto ou uma empresa)
 - Como customizar as campanhas de marketing e propagandas de produtos 'fidelizando' meus clientes?
- Estes são exemplos de perguntas que impulsionam um novo ramo de estudo sobre bancos de dados: O Data Mining (ou mineração de dados).

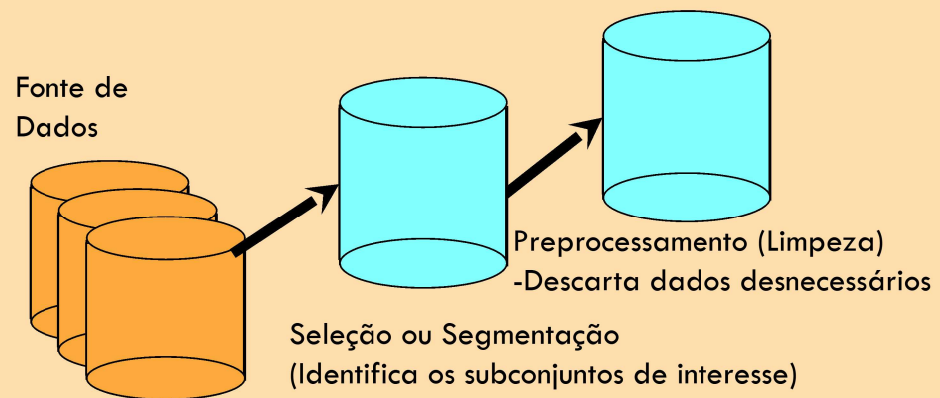
- Exemplo
 - Análise de Vestibular:
Perfil: Universidade Particular de uma cidade rica.
 - Regra percebida: 65% dos candidatos do sexo feminino que trabalha e teve aprovação com boas notas no vestibular **NÃO** efetivava a matrícula.
 - Mulher em idade de vestibular, se trabalha → Alto Empenho. Inscrição em escola paga deve ser a 2ª opção (1ª opção → Escola Pública – mesmo distante). Notas boas no vestibular → deve ter passado também na 1ª opção, sendo assim escolhe viajar a não fazer escola paga.
 - Exceções: pessoas mais velhas, de alto poder aquisitivo e que voltaram a estudar por outras razões que ter uma profissão, etc..
 - Diante do cenário: Como fazer a candidata ficar na cidade e PAGAR a escola? DECISÃO: Oferecer um desconto alto, por prazo determinado para mulheres com altas notas no vestibular.
- Data Mining – Técnica de processamento de dados até o ponto onde é possível se aplicar algum algoritmo que estabelece as 'regras de *correlação*', identificando o comportamento da realidade geradora dos dados com uma margem de precisão pré-estabelecida.
 - O processamento se dá em etapas... Processamentos de dados que melhoram os dados para o processamento dos algoritmos.

- Fases do Data Mining



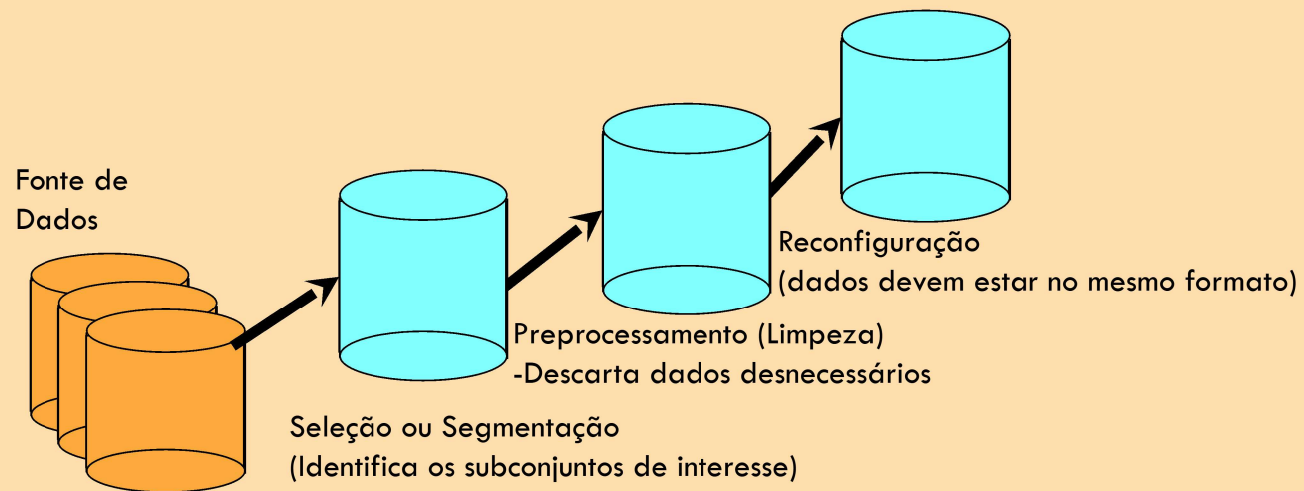
- Fases do Data Mining.
- Seleção ou segmentação:
 - Dados são filtrados de acordo com critérios definidos
 - Determinam-se subconjuntos que sejam de interesse na análise:
 - Exemplos:
 - * (Empresa de Seguros): Todas as pessoas que são proprietárias de carros é um subconjunto de dados determinado.
 - * (Rede de Vídeo-Locadora): Todos os clientes que movimentaram filmes (quais filmes e quais tipos de filmes), nos últimos 5 anos.
 - O analista escreve SQL conhecendo *profundamente* o modelo de dados e direcionando a segmentação para os passos seguintes do DM.

- Fases do Data Mining.



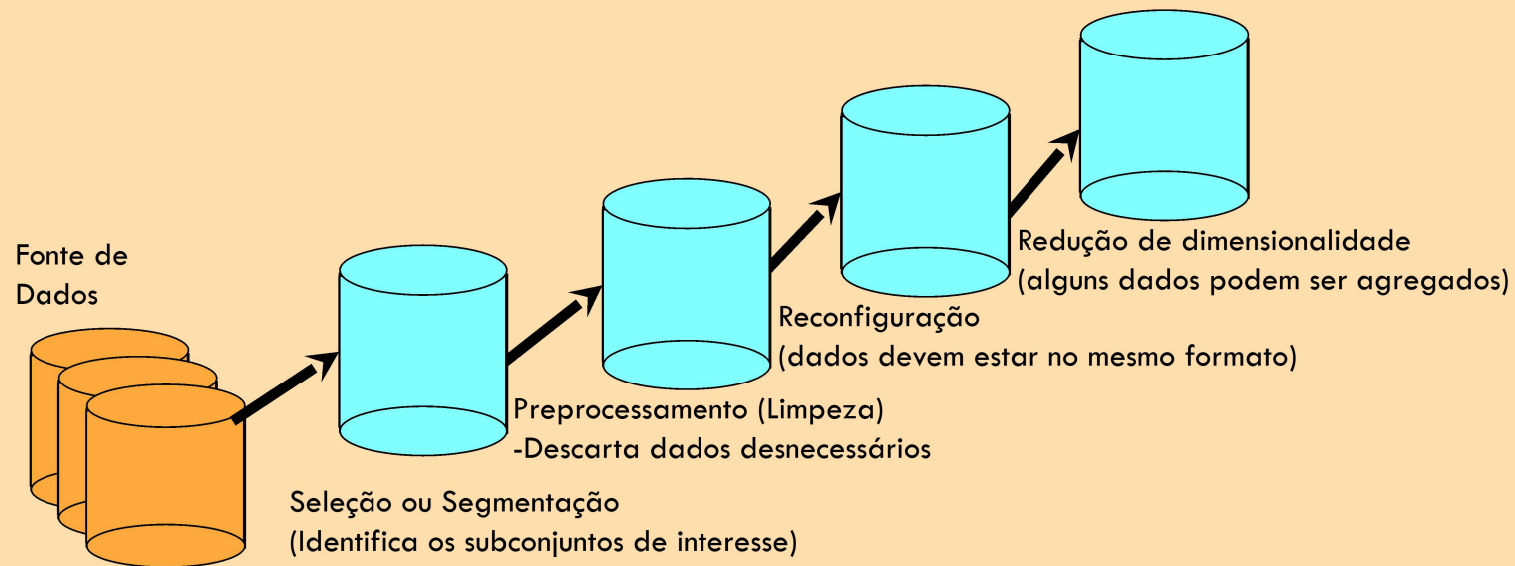
- Fases do Data Mining.
- Pré-Processamento – Limpeza dos dados.
- Dados desnecessários são descartados.
 - Exemplo:
 - * Entre os correntistas de um banco, na pesquisa sobre possíveis investidores, desconsiderar público com menos de 20 anos de trabalho em análise de planos de previdência.
 - * Desconsiderar pessoas com título de Mestre ou Doutor na pesquisa sobre público alvo para divulgação de cursos de extensão universitária.
- Usam-se junções abertas para indicar dados a serem excluídos do DM

- Fases do Data Mining.



- Fases do Data Mining.
- Reconfiguração de dados
- Dados que podem ser provenientes de fontes diferentes, mas devem ter a mesma forma de apresentação e processamento
 - Exemplo:
 - * datas no formato "DD-MM-AAAA" ou "MM-DD-AAAA" deve ser convertido para "AAAA-MM-DD"
 - * Ponto decimal de números sendo o sinal ',' em parte do banco e em outra o sinal '.'
 - * os tipos de dados char e varchar devem ser os mesmos para campos de texto de uma parte e outra do banco de dados.

- Fases do Data Mining.

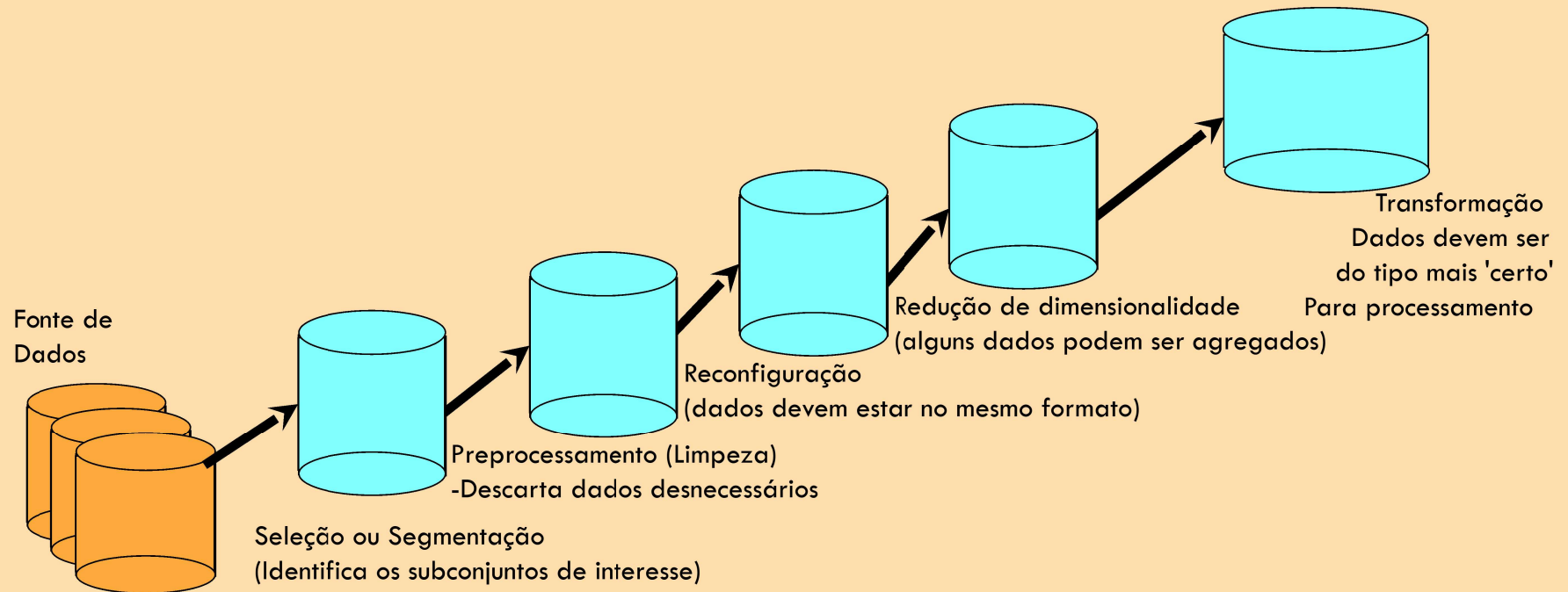


- Redução de dimensionalidade
- Nesta etapa executa-se a combinação de valores.
- Dados que sempre irão estar relacionados nos resultados das análises podem ser combinados para poupar tempo de processamento e melhorar a interpretação dos dados.
 - Exemplo:

Concatenar valores de data de retirada de DVDs e gênero de filme (se a análise for condensar os dados sempre por estes dois valores).

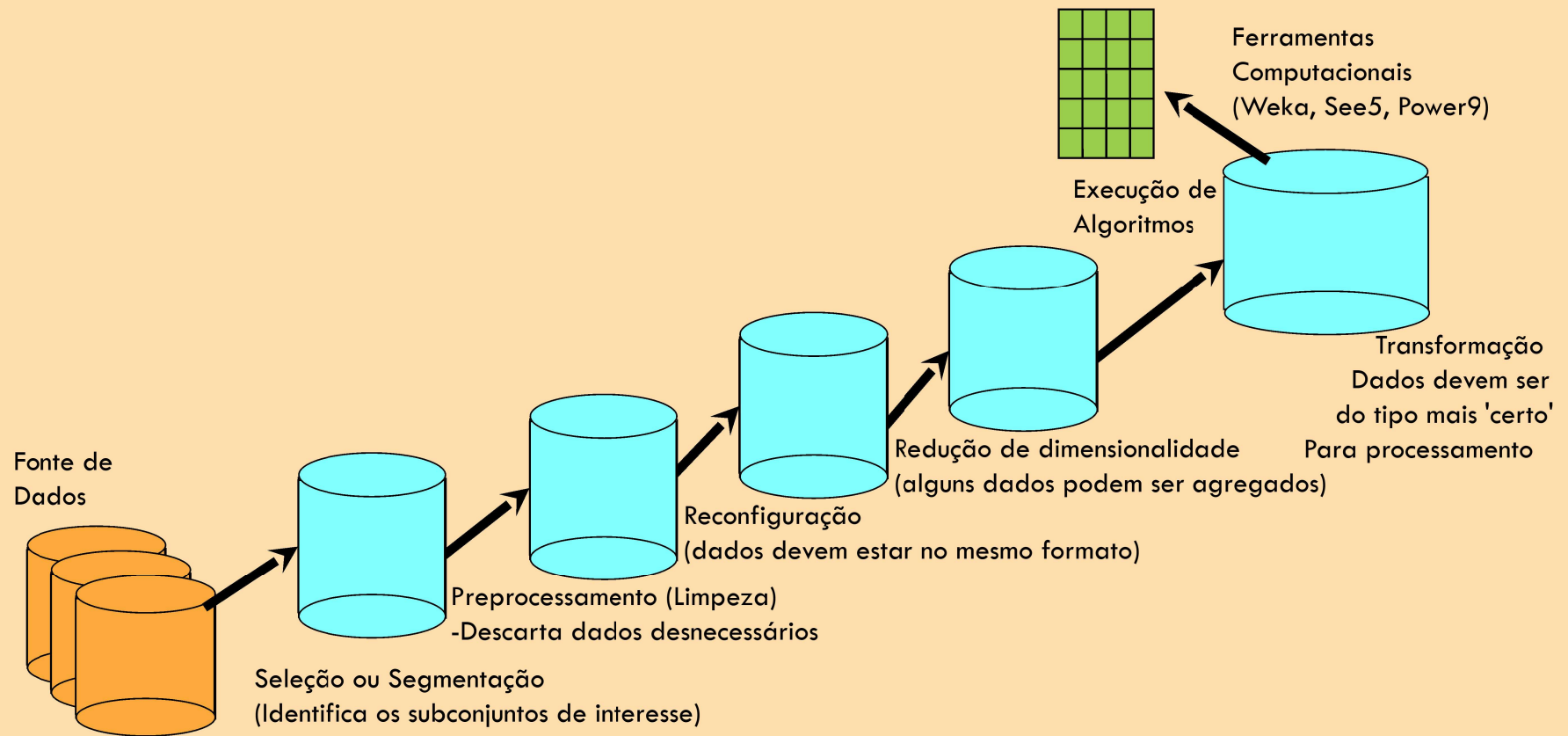
Concatenar Nome do Cidade com Logradouro de moradia, pois formam um só valor de dado que simplesmente podem surgir relacionados à uma distribuição geográfica dos dados.

- Fases do Data Mining.



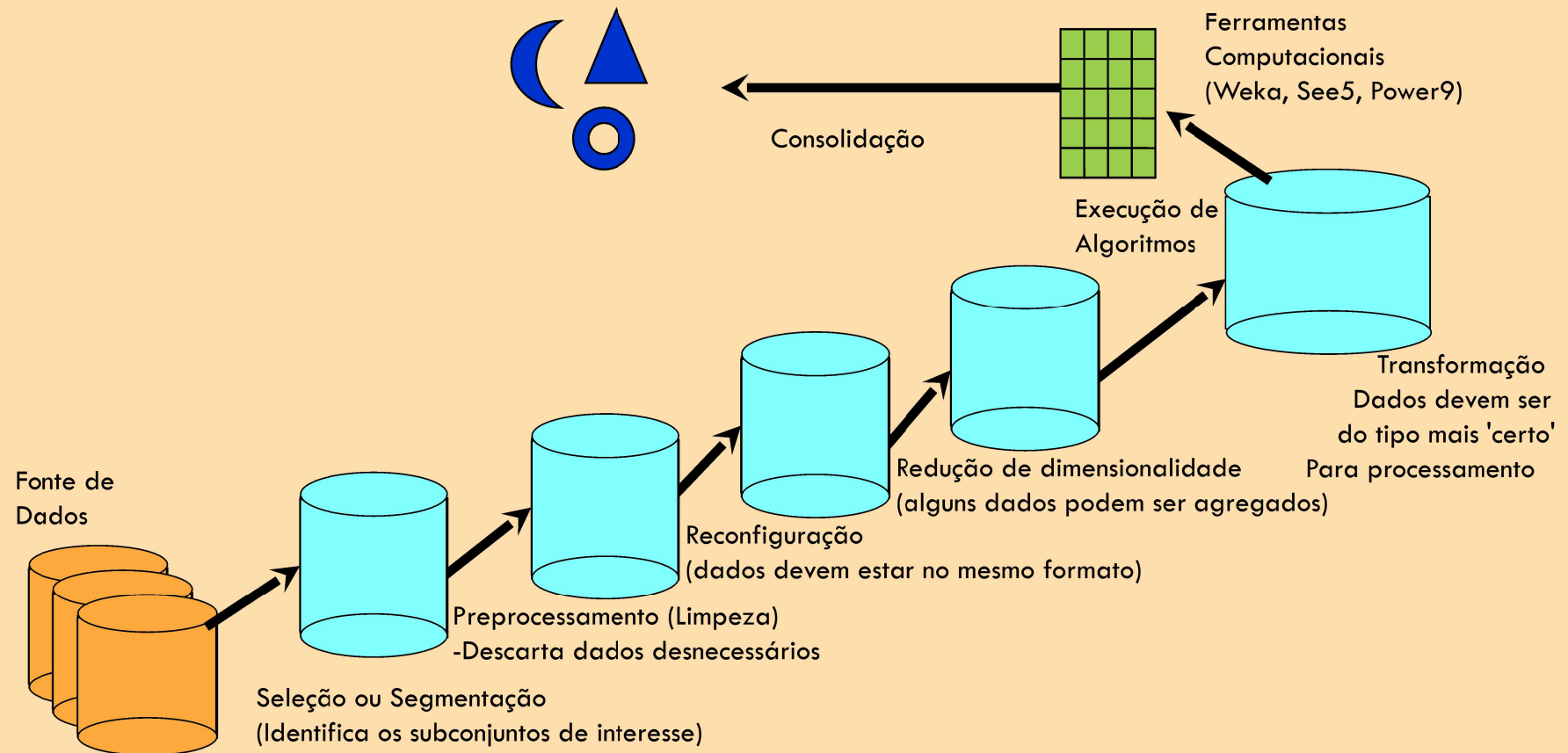
- Fases do Data Mining.
- Transformação
- Mudança dos tipos de dados para formatos mais adequados para o processamento dos algoritmos de correlação
- Algoritmos de correlação trabalham em geral sobre dados de um tipo (ou dois), em função da linguagem de programação que implementa o algoritmo.
 - Exemplo:
 - * Alterar datas e números para formato texto
 - * converter todos os tipos inteiros em float.
 - Note:
 - O tipo de dado destino depende do algoritmo que vai se aplicar na análise de correlação.
 - A escolha para o destino da transformação vai depender de uma análise prévia com base em alguns processamentos (em SQL) de comandos que resumem os dados e permitem indicar a escolha do algoritmo mais adequado ao processamento.

- Fases do Data Mining.



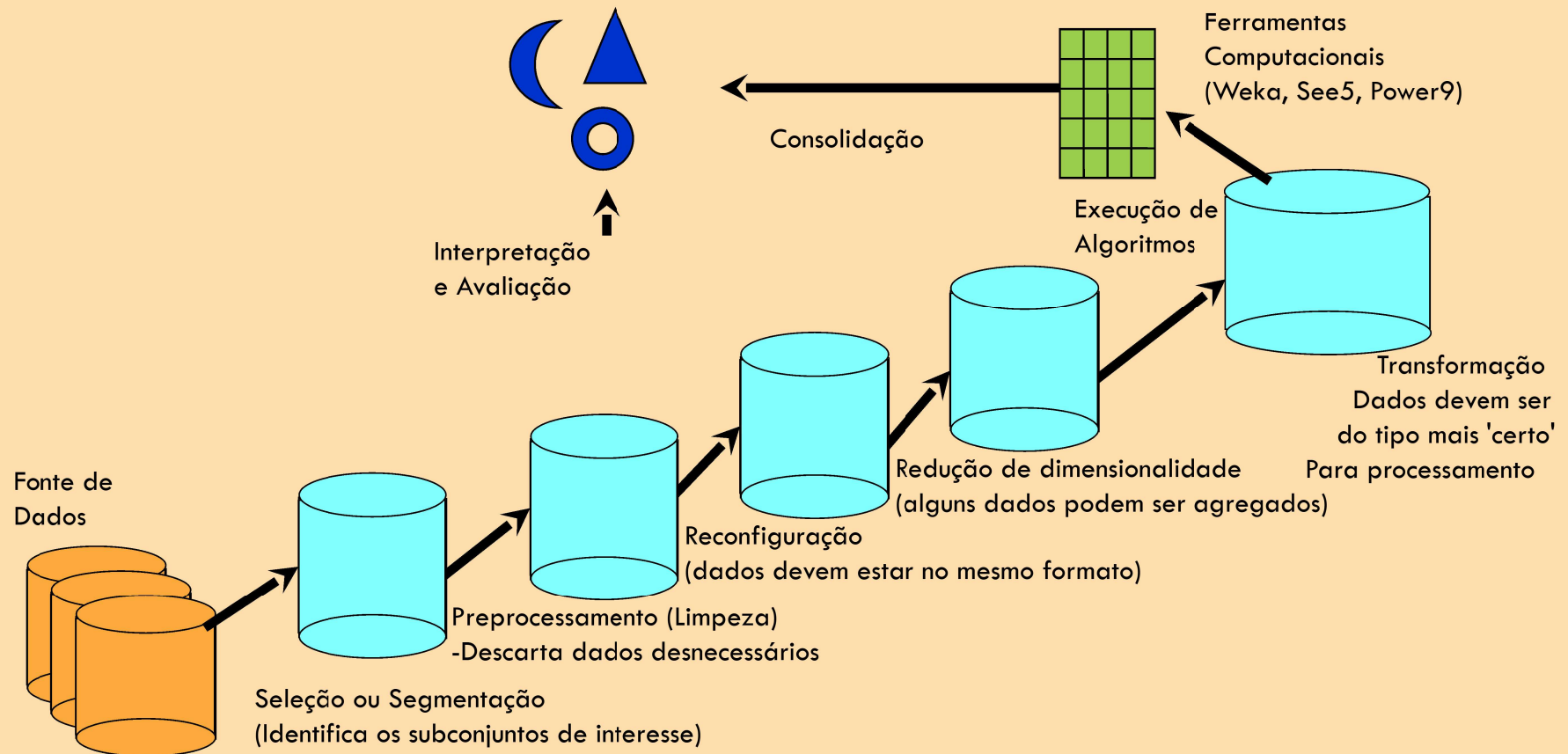
- Fases do Data Mining.
- Execução dos algoritmos
- Os dados são movidos (copiados) para possibilitar o processamento (o 'destino' pode ser em formato de arquivos textos ou planilhas eletrônicas ou sobre um determinado formato de armazenamento de um SGBD).
- Segue-se: Processamento dos dados para percepção de regras de comportamento.
- São executados algoritmos de redes neurais, adequação à modelos estatísticos, até o momento que se escrevem as regras de correlação com diferentes *significâncias*.
- Exemplos:
 - See5 (Excell, texto)
 - Weka (PostgreSQL, texto)
 - IBM Power9 (DB2)

- Fases do Data Mining.



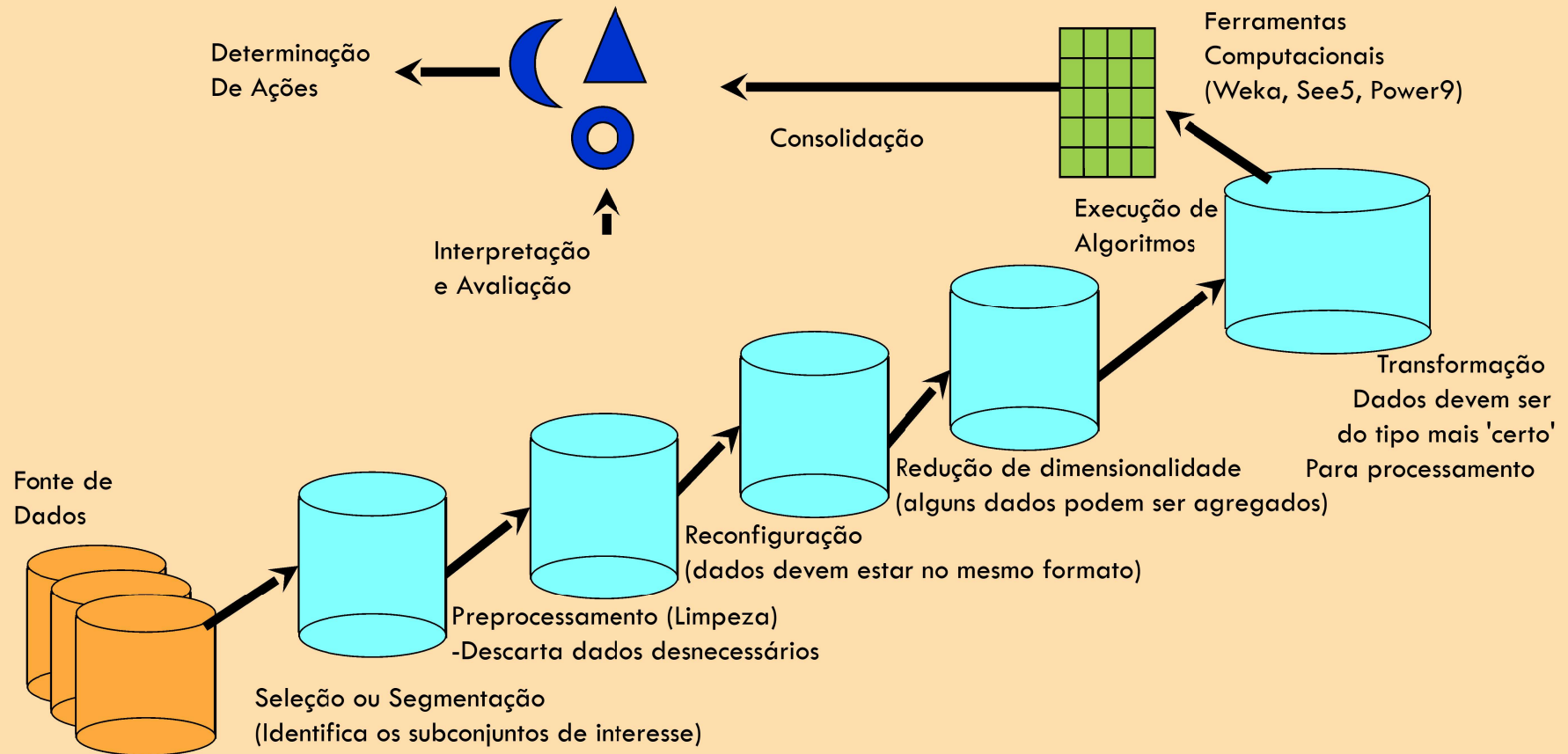
- s • Fases do Data Mining.
- Consolidação
- Seleção de regras mais relevantes (com significâncias altas → Quanto? Bom senso ajuda a escolher o nível de significância), armazenamento e disponibilização destas regras.
 - Ex.: Considerar regras aplicadas a mais de 80% dos dados e que envolvam 4 atributos na correlação.
- As regras de correlação são escolhidas por critérios subjetivos (humanos) ou podem ser pré-estabelecidos nas ferramentas de processamento (em geral com intervalos maiores de separação das regras geradas em relação as escolhidas).

- s • Fases do Data Mining.



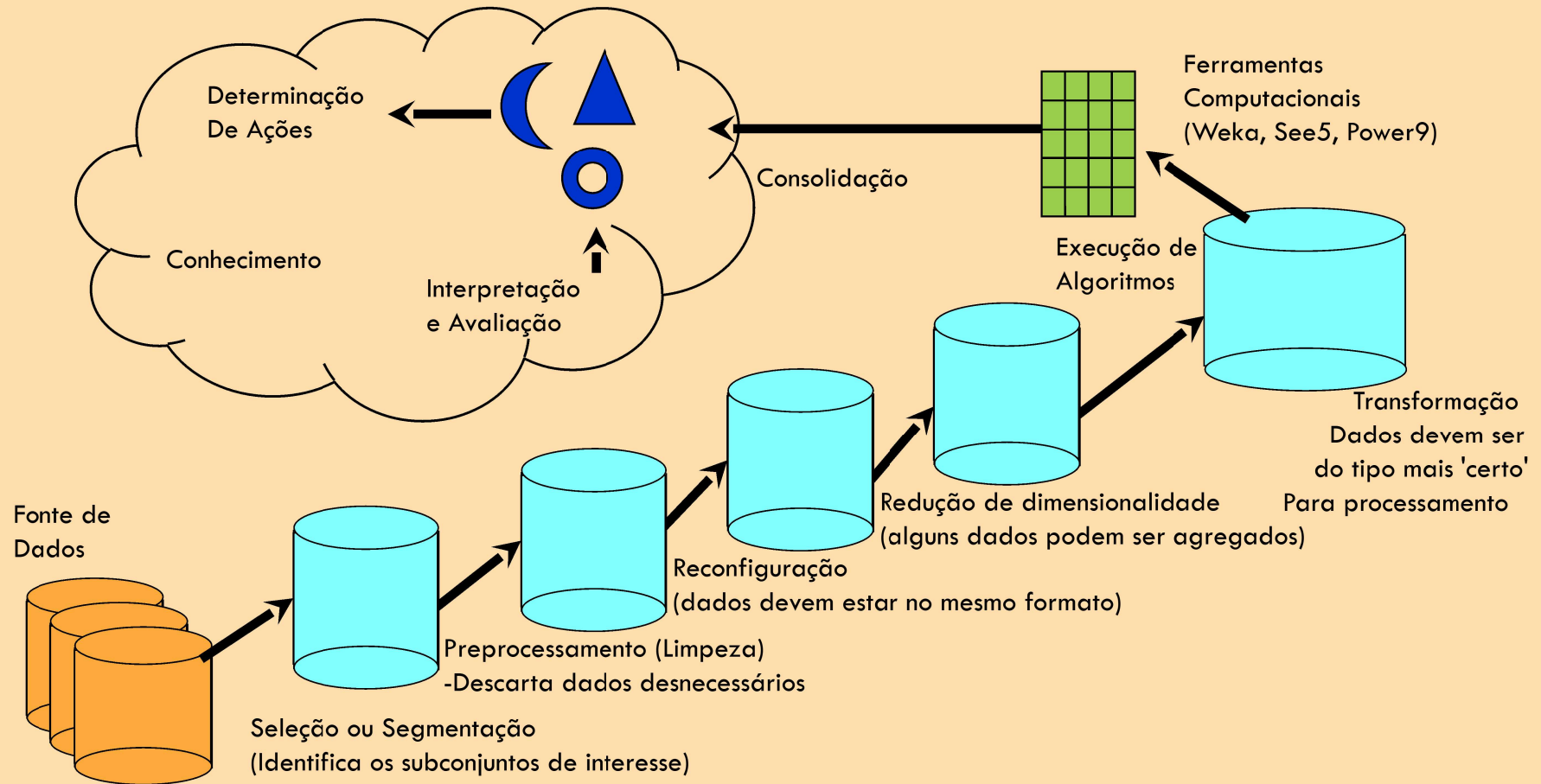
- s • Fases do Data Mining.
- Interpretação e avaliação.
 - Ação humana e nível gerencial
 - Leitura das regras de correlação desenvolvidas
 - Classificação das regras produzidas (ordem de percentagem de incidência dos dados ou direcionamento estratégico com alguma diretriz da empresa – por exemplo: melhorar a imagem ecológica da empresa, isso vai direcionar a ordem das regras de correlação independente da percentagem de incidência das regras de correlação).
 - Para as Regras de correlação (já ordenadas ou agrupadas), se estabelece o Plano de Ações mínimas para cada regra – o que fazer para cada Regra percebida – estima-se: custo, prazo, métrica de acompanhamento e forma de avaliação se execução foi bem executada.

• s • Fases do Data Mining.



- Fases do Data Mining
- Determinação de ações decorrentes
 - Cada regra indicada tem suas ações detalhadas
 - Cada ação é executada e seus resultados devem ser monitorados, coletados e aprendidos.
 - As ações administrativas devem ser executadas com o comprometimento de todos os envolvidos no processo de DM.
- A aplicação efetiva das ações das regras e seu acompanhamento geram o conhecimento mais aprimorado sobre o contexto onde a organização executora do Data Mining atua.

- Fases do Data Mining



- Como todo o Processo executado no contexto empresarial, o Data Mining apresenta suas vantagens e desvantagens.

VANTAGENS

- Tomada de decisão
- Descoberta de aspectos inesperados sobre os dados da organização
- Análise de grandes bases de dados
- Pode indicar tendências de mercado.

& DESVANTAGENS

- Alto custo de implantação;
- Resultado a longo prazo;
- Necessidade de profissionais especializados.
- Alta Sazonalidade (intervalo de tempo até ações demonstrarem resultados). Depende do contexto: 6 meses ou mais.

- Tendências...
 - Aumento do Volume de dados nas empresas
 - Aumenta a necessidade de uso de Data Mining.
Ok! Vamos fazer um 'monte' de DM!!! Mas...
 - Existem Sistemas com modelos inadequados para o DM
 - Implementações inadequadas (realidade modelada com falhas- conceituais).
 - Sistemas com falhas na Documentação (tanto na especificação quanto na versão 'construída')
 - Em muitos sistemas ainda existem Modelos Relacionais não normalizados
 - Desafio: Usar DM em pequenas e médias empresas (As empresas duram em média 6~7 anos).
 - Para viabilizar DM em P&M devemos produzir bons sistemas → É uma questão de postura profissional.
- Mas e se o contexto for um tanto diferente?

- Situação:
 - Uma multinacional do setor de laticínios monta uma subsidiária em um país e resolve "aglomerar" vários laticínios menores por um tempo 'determinado'.
 - Esta expansão ocorre com a empresa montando um “casamento” por tempo determinado. Durante este tempo a empresa mandatária acessa os dados da subsidiária.
 - Quando o contrato se encerra as empresas ficam autônomas e independentes.
- Então (perguntas não tão simples...)
 - Como fazer a integração do modo menos “traumático” para as empresas 'casadas'?
 - Como proteger e garantir o legado dos dados feitos por ambas as empresas?
 - Como fazer a separação das empresas e garantir que cada uma mantenha seus dados?
 - Como assegurar que as empresas não acessem dados de outras empresas da holding?
- Parece que é preciso 'integrar' mantendo 'compartimentalizado' (separado)
 - 'caixas' de onde se EXTRAEM os dados necessários para uma 'visão global'.

- Considerações iniciais...

- Do ponto de vista operacional podemos dizer que um SGBD oferece um acesso de modo controlado aos dados para os usuários através de programas aplicativos.
- Na década de 70/80 as empresas passaram por um processo de "crescimento por aquisição" → Holding.
- As empresas necessitaram de mecanismos que combinassem os Sistemas de Informação adquiridos. Os Bancos de Dados tiveram que ser também combinados (e em alguns casos mantidos os 'originais').
- Surge o conceito e a implementação do Data WAREHOUSE.
- DWH → Ferramenta que permite combinar e visualizar dados de Sistemas de Informação distintos permitindo processamento e aquisição de informação dos dados integrados.
- Pré-Requisito:
 - Padronização das especificações de tabelas (esquemas de tabelas com estrutura semelhantes).
 - Instalação e configuração de ferramentas de INTERFACE que permitam "migrar" os dados dos SIs para uma 'esfera' ou 'instância' mais alta (resumindo os dados – ou não!!! Se for possível).

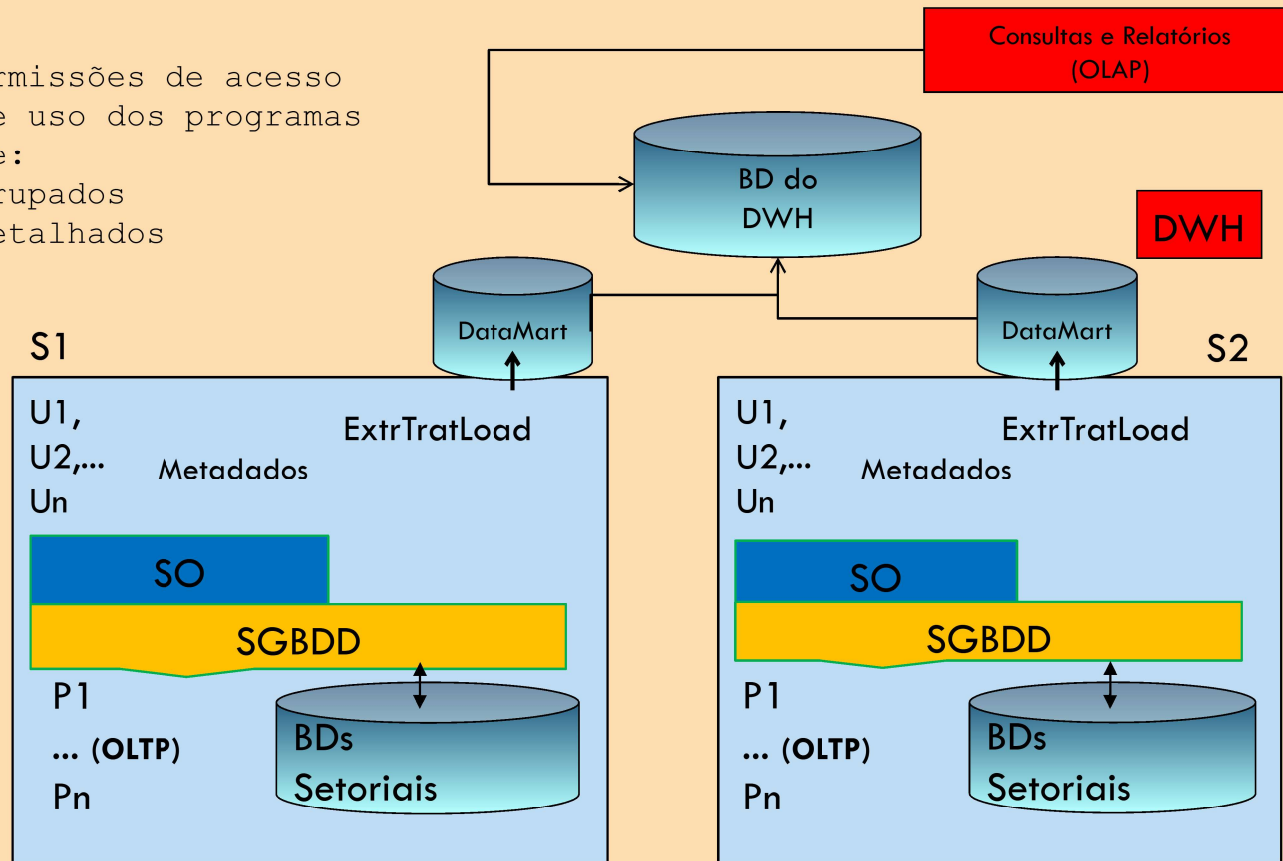
• Esquemáticamente...

Defini-se:

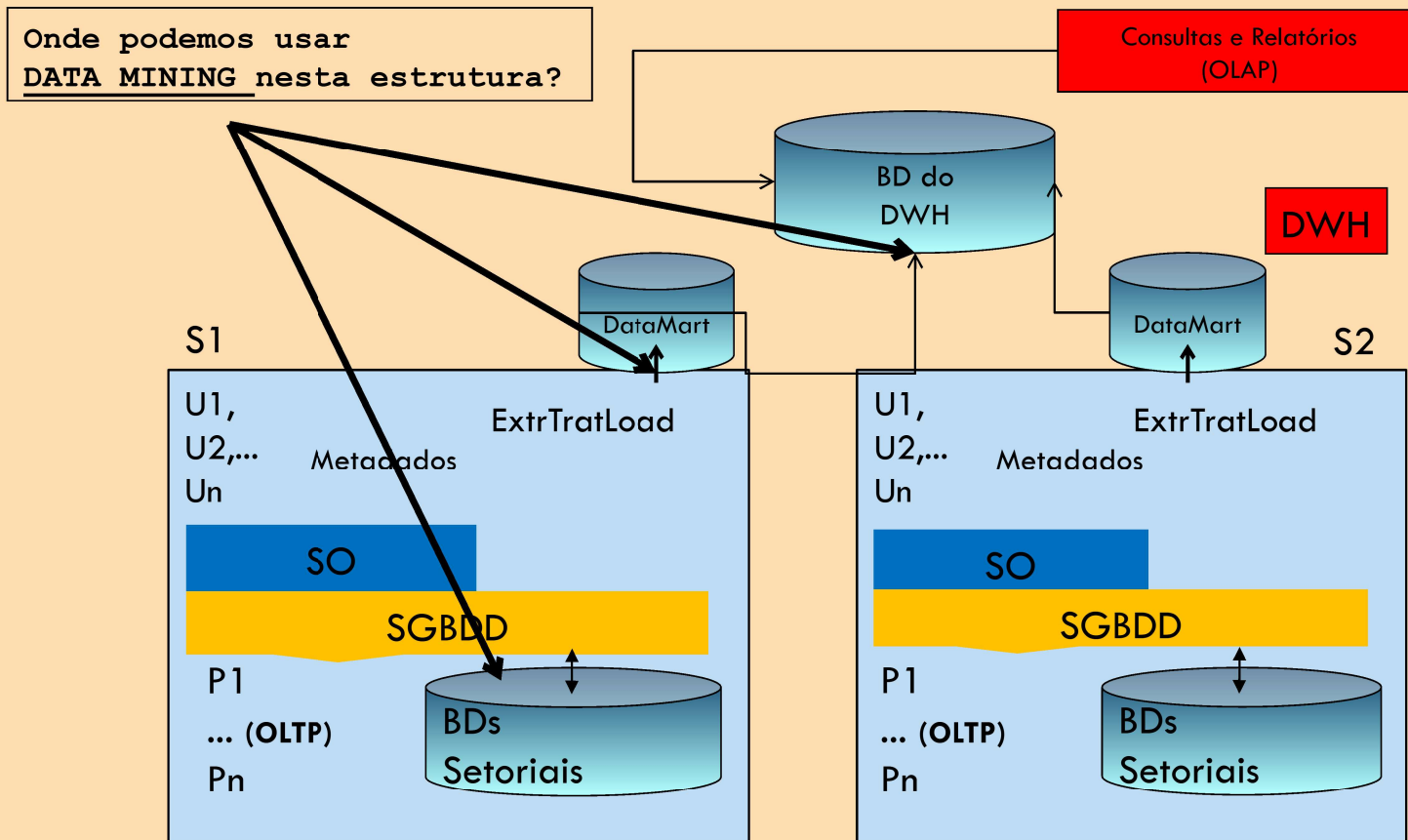
- Usuário e Permissões de acesso
- Frequência de uso dos programas
- Granularidade:

ALTA – Dados Agrupados

BAIXA – Dados Detalhados



- Esquematicamente...



- Para a integração acontecer devemos:
 - Montar um Dicionário de Tradução dos componentes da Base de Dados "hospedeira"
 - Neste DD devemos ter a descrição dos campos de todas as tabelas do Sistema de informação.
 - Instalar os programas de interface (Extractor, Transacter e Loader – ETL).
 - Em geral, é executado no Servidor Setorial (origem dos dados)
 - Os ETLs geram os níveis superiores de dados (Data Mart), que podem ser estratificados em várias camadas.
 - O último nível de tratamento é o BD do Data Warehouse (nível mais alto na compartimentalização – Caixa com os dados próprios para rodar um PA-OLAP (On-Line Analytical Processing)).

- Ferramentas existentes (pesquise: Data Warehouse- software)
 - A primeira foi a "Forest & Trees" (Lá pelos ~1990)
 - IBM- *InfoSphere Information Server*
 - Oracle- *Oracle Exadata*
 - Idera- *WherScape RED*
 - ... E mais outras...
- Tendências de desenvolvimento
 - O conceito DWH pode ser aplicado com qualquer ferramenta de TI (Plataformas ≠ e SGBD ≠).
 - Os Bancos de Dados dos Sistemas continuam sendo Relacionais... (embora comecem a existir Sis com os 'NoSQL')
 - Os grandes fabricantes de SGBD vão desenvolver e oferecer as "*suas*" soluções já configuradas para se "comunicar" com *seus* produtos.
 - Isso é uma tendência forte para outros tipos de ferramentas como ferramentas CASE (usando ou não abordagem Orientada a Objeto ou SGBD NoSQL).

- Conceitos iniciais...
- O SGBD é um programa que tem dois objetivos
 - Criar *Transpor as definições do Modelo Físico para o computador*
e
Administrar *Controlar acessos (privativos e concorrentes)*
Bancos de Dados
 - Disponibilizar um *Ter ou entender uma Linguagem de Programação*
ambiente de desenvolvimento
de
Sistemas de Aplicativos

De modo Geral:
Dar suporte ao desenvolvimento e uso de
Sistemas de Informação!!

- Existem situações onde um grande volume de dados precisa ser processado (mas talvez não armazenado), seguindo uma análise que deve ser confiável, tratando de dados de origens e tipos diferentes e agregando algum valor a uma tomada de decisão... Nesta situação estamos diante de uma Aplicação BIG DATA.
- Talvez o termo tenha se originado pela necessidade de processamento (mesmo que somente leitura) de grandes volumes de dados. Mas o que está em foco são os 5 Vs:
 - Velocidade (de processamento dos dados em tempo adequado para a análise e decisão);
 - Variedade (de tipos de dados – texto, imagens, sons, etc);
 - Volume (de dados – que não são gerados no sistema);
 - Veracidade (confiabilidade, significância dos dados) e
 - Valor (o que é importante na análise).
- Agora... como se implementa? Existem ferramentas que rodam em servidores, estes servidores devem ter duas fortes características: CONECTIVIDADE e PODER COMPUTACIONAL. (note que *armazenamento* não é pré requisito).
- Finalmente, definindo...

- Big DATA é um sistema de acesso a diversos tipos de dados podendo ou não executar uma *Coleta, Armazenamento, Processamento, Disponibilização* de dados diversos interpretados em um *tempo de resposta* satisfatório agregando ou mantendo algum *valor*...
- Exemplos de aplicação Big DATA...
- Um supermercado envia uma propaganda recomendando produtos para bebês para uma adolescente de 17 anos. O pai vai na sede do supermercado de onde saiu a correspondência e briga com o gerente: "Vocês estão sugerindo que minha filha tenha uma criança?". O Gerente pediu desculpas e alegou que o sistema havia montado a propaganda sem a supervisão de uma pessoa.
- Passadas três semanas o Gerente recebe uma ligação do pai (antes indignado). O pai se desculpou pela forma como tratou o Gerente da loja. Disse ainda que ao conversar com a filha e com mais calma acabou sabendo da gravidez de 3 meses.

- Mas... Como o sistema BD percebeu isso?
- O Sistema de BD NÃO Percebeu NADA disso!!!
- Os dados foram **analizados** por um ALGORITMO que:
 1. Monitorou as compras do cartão de crédito dos clientes da rede de supermercado. Isso é feito 'consultando' a base de dados de cartões de crédito.
 2. Detectou que uma compra foi feita em farmácias e 'consultou' a base de dados da rede de farmácias e coletou (registrou) o item comprado.
 3. Quando a ocorrência se repetiu (3 vezes) disparou a 'consulta' no Cartório de Registro Civil e detectou que o dono do cartão era casado, a idade da esposa beirava 65 anos e que tinham uma filha. O *algoritmo* determinou que a chance da moça estar grávida era maior que do chance da mãe dela estar grávida.
 4. O *Algoritmo* consultou os itens do supermercado relacionados à crianças recém nascidas e montou o folder de propaganda, enviando-o logo em seguida.

- O Sistema de BD NÃO Percebeu isso!!!
- Veja os 5 Vs nesta situação.
 - **Volume**: Dados de diversas origens que foram coletados, processados e armazenados em computadores do supermercado (os resumos).
 - **Variedade**: Diversas formas de armazenamento diferentes (documentos que mantinha a integridade da família – pai+mãe+filha, consumo com o cartão do pai associado a produtos para mulher)
 - **Veracidade**: Conformidade dos dados à realidade estudada
 - **Velocidade**: o tempo de resposta foi muito bom (a última compra de teste de gravidez aconteceu 3 horas antes do envio da propaganda)
 - **Valor**: para o supermercado ter este perfil detalhado do consumidor ajuda a adotar estratégias de venda mais eficientes.

- Outro exemplo:
- Série de Fatos (alinhados no tempo):
 - Moça passa por consulta médica (de manhã) e sai com receita para medicamento para sua rinite alérgica.
 - No final da tarde, no mercado, compra ingredientes para um jantar (cogumelos, peixes, etc.).
 - A seguir, passa na farmácia e compra o medicamento (no servidor está rodando o Big Data).
 - A moça fecha a compra e vai para casa preparar o jantar.
- Durante o preparo do jantar a moça recebe uma ligação no celular.
 - Seu médico!! Ele recebeu uma mensagem (SMS) no telefone **ALERTANDO** que a moça fez a compra do remédio e tinha comprado os ingredientes que podem causar um surto alérgico se misturados ao remédio.
 - O médico, então, recomenda o início do tratamento para o dia seguinte.
- Onde e como atuou o algoritmo?
- Percebe os 5'V's?

- Onde e como atuou o algoritmo?
 - O Algoritmo está ativo no servidor da Drogaria (caso verídico).
 - Ao passar o Cartão no Caixa da Drogaria o Algoritmo consultou a base do laboratório do produto vendido e recebe a indicação que existem compostos que podem causar surtos alérgicos.
 - O Algoritmo 'consulta' a base do cartão e vê que a moça fez compras na parte da tarde (mercado pago com cartão). Indicando que foi num supermercado.
 - O algoritmo consulta a base do supermercado e determina os itens comprados e compara com a base de compostos recuperados da consulta no laboratório.
 - ➔ Existe risco de surto alérgico!!
 - Mas a moça já havia fechado a compra!!
 - Com a receita escaneada o Algoritmo **identifica** (OCR) o CRM do médico, 'consulta' a base de médicos no sistema de saúde e obtém o número do celular do médico. **Monta** a mensagem para o médico e a envia.
- Percebe onde estão os 5 Vs? Qual V pode ser melhorado?
- Existe ainda o exemplo da LAPD... Polêmico, invasivo, errado??

- Comparando as Estruturas
- O Conceito Big Data precisa de MUITO *poder de processamento*.
- Os algoritmos de pesquisa e processamento devem ser muito 'inteligentes'. (*Parece que no BIG DATA roda um 'Mini' Data Mining*).
- DM gera regras que podem ser inesperadas (Já o BigD é todo previsível, todas as decisões são previstas).
- DWH processa os dados sempre em fluxo unidirecional com relatórios pré-definidos. Sempre é dos BD Setoriais para o BD do Warehouse.
- BigD tem todos os 'caminhos' previstos, portanto precisa que o profissional conheça muito do contexto a ser analisado.

- Comparando as estruturas...
- Data Mining, Data Warehouse e Big Data são conceitos e implementações que tem diferenças e semelhanças.
- Semelhanças:
 - Podem 'viver' em um mesmo ambiente
 - Vão exigir muito processamento.
 - São ferramentas interessantes para auxiliar as empresas a uma nova postura perante mercados que podem apresentar comportamentos diversos.
- Diferenças
 - Propósito para se iniciarem:
Data Mining: Determinar as regras de correlação para indicar novas posturas da empresa (ofertar novos serviços)
Data Warehouse: aplicar um fluxo unidirecional dos dados para execução de programas OLAP (relatórios gerenciais).

Big Data: a rota de decisão é bem determinada e o algoritmo consulta muitas bases para tomada de decisão.



LABORATÓRIO DE BANCO DE DADOS

IBD100