SUPORTE À DECISÃO

José Machado

Evolução dos Sistemas de Apoio à decisão - Problemas

- Custos
- Pouca flexibilidade
- Limitações de recursos
- Partilha da informação
- Desempenho dos relatórios
  - Resumo da actividade operacional
  - Falta duma visão da evolução temporal
  - Dúvidas sobre quém é o dono da informação
  - Segurança dos dados
- Necessidade da construção de um repositório central

Apoio à decisão

- Sistemas de Transações vs Sistemas Analíticos
- Arquitectura
- Conceitos

Ciclo de Vida de um SAD

- Fontes dos dados
- Interface humana
- Análise e modelação
- Dimensões
- Desempenho
  - Agregações
  - Indexação

Extracção de Dados

- Fontes heterogéneas
- Processamento de dados
- Selecção de dados
- Transformação de dados (conversão, mapeamento, desduplicação)
- Tratamento de valores nulos
- Normalização (distribuições e unidades de grandeza)
- Regras de integridade
- Regras de negócio
- Registos em dimensões (mudanças)
- Agregações (cálculo)
- Tratamento de erros
- Integração de dados

Data Mining

- Definição
- Categorias algorítmicas (classificação, estimação, previsão, descoberta de afinidades, agrupamentos homogéneos ou clustering, descrição, alternativas)
- Perfis dos utilizadores

OLAP

- OLAP e Data Mining
- Navegação em OLAP
- Estruturas de dados
- OLAP nas organizações (e.g., finanças, comercial, marketing, recursos humanos)
- Integração em ERP

Previsão em Séries Temporais

- Base temporal
- Caracterização dos dados das séries
- Modelos de previsão (média móvel, alisamento exponencial simples e linear, sazonalidades)
- Medição do erro de previsão

Indução de Árvores de Decisão

- Introdução
- CART
- CHAID
- ID3
- C4.5
- Treino, teste e previsão
- Erro, overfitting e poda de árvores
- Dedução de regras

Descoberta de afinidades em transacções

- Cesto de compras do hipermercado
- Negação e dissociação de regras
- Taxionomias na generalização de items
- Artigos virtuais
- Temporalidade

Clustering
(partição padronizada de dados)

- Análise vectorial de dados
- Algorítmo de partição
- Partição probabilística (fuzzy)

Redes Neuronais

- Redes neuronais artificiais
- Topologias
- Combinação, transferência e activação
- Entrada e interpretação de resultados

Algorítmos Genéticos

- Codificação
- Função de avaliação
- Selecção para a evolução das espécies
- Renovação das populações

Desempenho

- Desempenho dos modelos
- Selecção de ferramentas
- Medição do desempenho

Resultados

- Sistemas de informação
- Sistemas de informação executivos
- Servidores de business intelligence
- Aplicações OLAP
- Disponibilização e integração de plataformas

Qualidade dos dados

- Impacto
- Data cleaning
- Registos duplicados
- Campos de texto livre
- Qualidade da análise
- Ambiguidade, nebulosidade, sensibilidade e informação incompleta

Problemas

- Custo elevado das soluções
- Pouca flexibilidade (relatórios)
- Limitação de recursos
- Os relatórios são informação resumida da actividade operacional
- Há falta de visão da evolução temporal
- Há dúvidas sobre quem é dono da informação

Sistemas operacionais vs sistemas analíticos

- Incapacidade de integração
- Desempenho do sistema operacional encontra-se optimizado para dar resposta a conjuntos de interrogações bem definidos
- Incapacidade de lidar com informação provenientes de outras fontes de dados

Sistemas operacionais vs sistemas analíticos

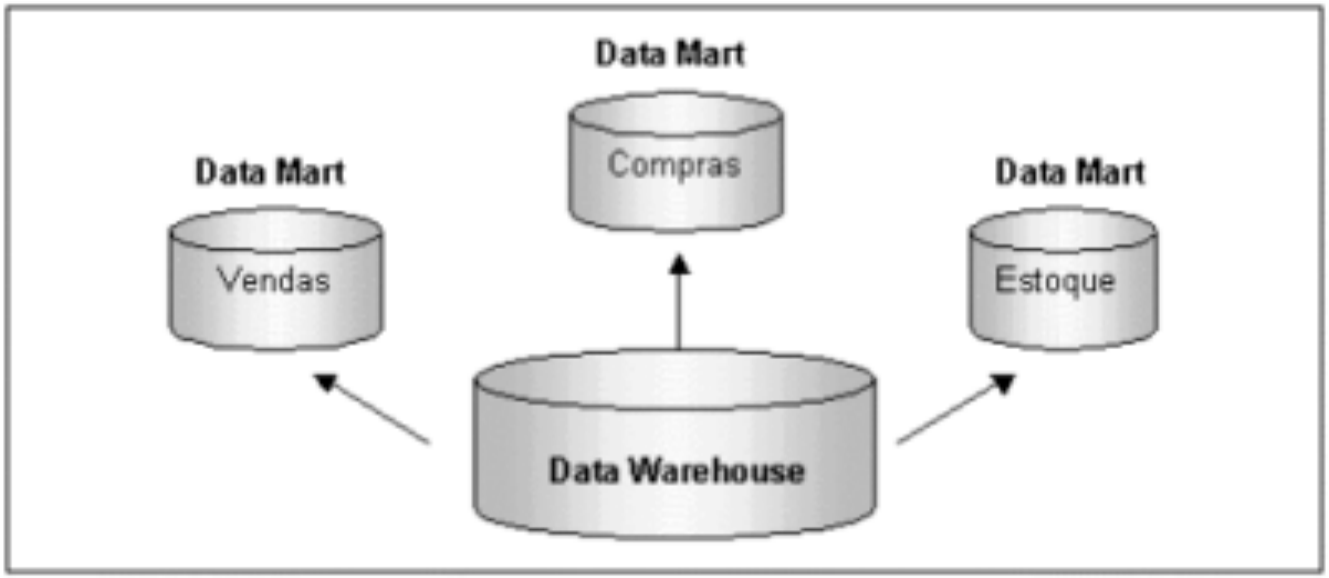
- Diferentes níveis de detalhe
- Actualização de dados
- Diferentes comunidades de utilizadores
- Instruções SQL típicas
- Necessidades de recursos
- Formas de acesso a dados e normalização relacional
- Disponibilidade dos sistemas
- Desempenho e recursos

Datawarehouse

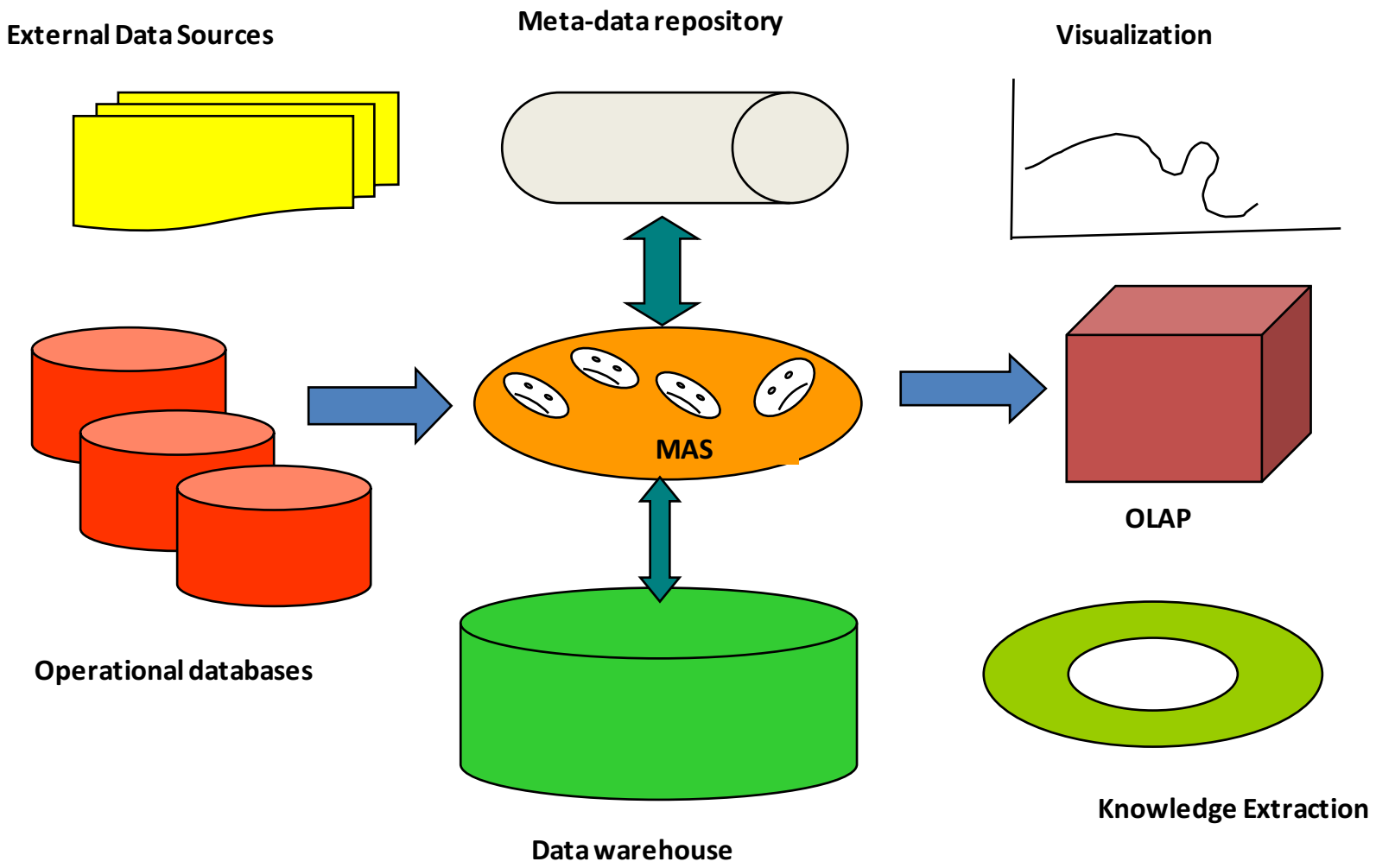
- Uma forma de acesso centralizado mais facilitada e de forma legível a toda a informação
- Consistência em toda a informação que circula
- Capacidade de lidar com requisitos flexíveis e com a própria dinâmica da organização
- Segurança no acesso e na circulação da informação (monitorização)

Data Marts

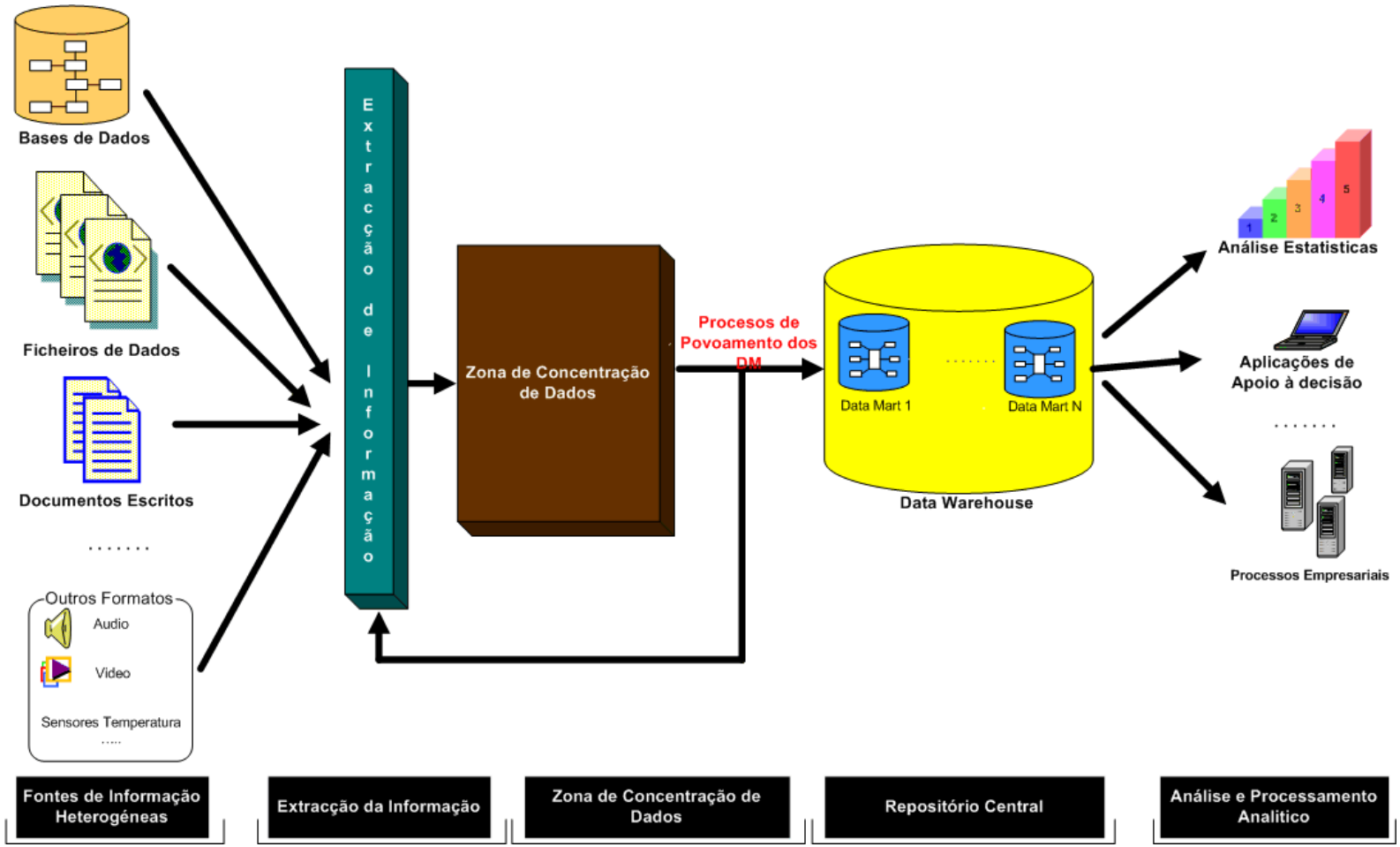
- Um DW é uma colecção de Data Marts
- A divisão entre DM e DW não é uma divisão tecnológica mas funcional
- Interligação de Data Marts



Datawarehousing



Data Warehouse



Datawarehousing

- As organizações analisam data correntes e históricos para identificar padrões e suportar estratégias de negócio.

- A análise é complexa e interactiva e considera volumes elevados de dados integrados a partir de fontes variadas.

- A diferença entre OLAP (On-Line Analytic Processing) e OLTP (On-line Transaction Processing) é que a actualização da informação é substituída por interrogações frequentemente longas.

3 Áreas Complementares

- Data Warehousing: consolida dados a partir de muitas fontes num repositório mais vasto, considera o carregamento de dados, a sincronização periódica entre réplicas e a integração semântica.

- OLAP: Interrogações complexas em SQL e “views”; interrogações baseadas em operações do tipo folha de cálculo e visualização multidimensional dos dados; Interrogações interractivas e em tempo real.

- Data Mining: procura exploratória de tendências interessantes e de anomalias.

Datawarehouse

- Dados integrados que medem períodos de tempo longos;
- Adição frequente de informação sumária;
- Volume medido em gigabytes ou até terabytes;
- Melhoria dos tempos de resposta interactiva para perguntas complexas;
- Não é habitual fazer “updates” ad hoc.

Datawarehouse

- Integração semântica: quando são processados dados de fontes múltiplas devem-se eliminar erros, e.g., moedas diferentes, esquemas, ….
- Fontes heterogéneas: os dados devem ser acedidos a partir de formatos de dados e repositórios variados (as potencialidades da replicação podem ser exploradas aqui).
- Integração, Actualização e Eliminação: os dados integrados devem ser periodicamente actualizados e os dados obsoletos devem ser eliminados.
- Gestão de meta-dados: devem ser guardados os parâmetros sobre a fonte da informação e os tempos de carregamento, assim como outros parâmetros relevantes sobre os dados.

O que é um Data Warehouse?

-Estrutura de armazenamento em que a informação :

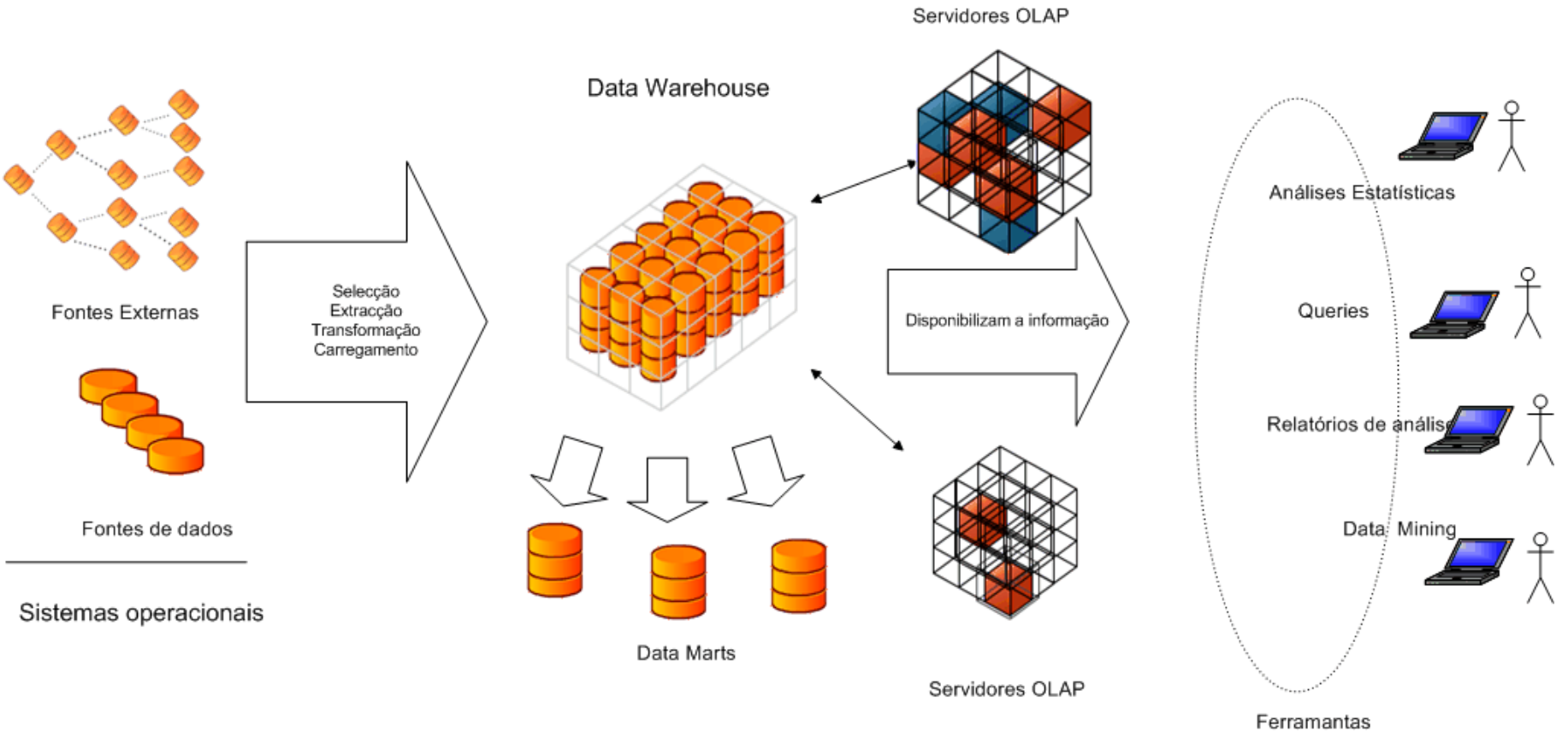
- É orientada à área de negócio/interesse;
- Está integrada (centralizada);
- A informação é não volátil;
- A informação está etiquetada temporalmente;
- Todos os dados pretendidos já estão calculados;

-A informação está estruturada segundo o modelo dimensional (tabela de factos, dimensões e medidas):

- Modelo designado de:
  - Estrutura em estrela (gastámos mais espaço, mas temos mais eficiência);
  - Estrutura em floco de neve (gasta menos espaço, mas é menos eficiente na pesquisa de informação);

-A dimensão mais importante de um DW é a dimensão Tempo;

Data Warehouse



Data Warehouse

Problema:

- Os sistemas de informação das organizações têm a sua génese nas necessidades de suporte operacional (OLTP) e armazenamento dos dados;

Revelam fraquezas perante a necessidade da análise expedita da informação dos dados segundo várias vertentes e perspectivas de negócio das organizações;

Data Warehouse

Solução:
- As **ferramentas OLAP** disponibilizam de um modo rápido e flexível mecanismos de anáise da informação conjugando várias variáveis de negócio;

As ferramentas OLAP assentam sobre estruturas de dados DW;
As estruturas multidimensionais (**Cubos**) computam a partir desses DWs todos os possíveis agrupamentos, gerando sub-cuboides, interligando-os entre si através das várias dimensões;

O objectivo dos servidores OLAP é computar todos os agrupamentos e disponibilizar respostas rápidas sobre todas as possíveis queries agregadas sobre os cuboides com diferentes atributos de agrupamento;

Necessidade de um Data Warehouse:

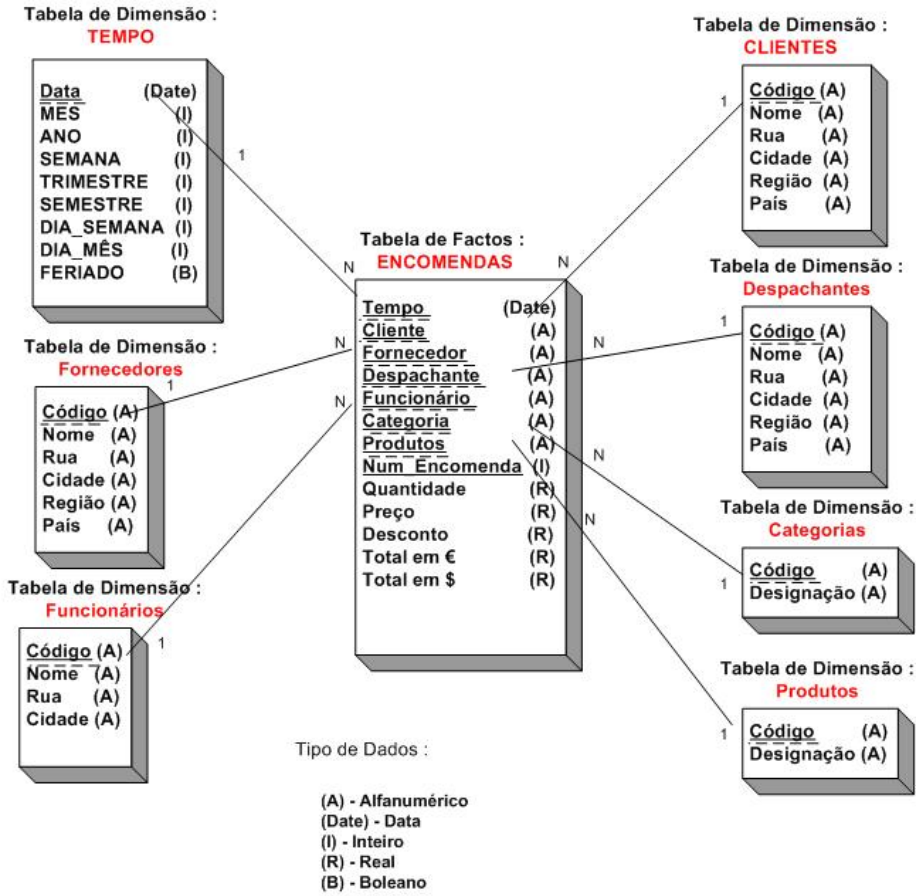
- Qualquer organização em que o problema é o desempenho na análise dos sistemas de suporte à decisão, uma vez que potencialmente e de acordo com os dados existentes, qualquer query de análise será materializada no menor tempo possível sejam quais forem os dados;

- A **necessidade de descoberta de tendências de negócio**, ou padrões de interligação entre entidades da organização;

- As fontes de dados heterogéneas (potencialmente e se for possível) serão materializadas numa estrutura de dados que a organização pretende, uma vez que está orientada por assuntos;

- O desempenho dos sistemas de DW são autónomos (caso MOLAP e HOLAP) não afectando o desempenho dos sistemas operacionais.

Data Mart (estrutura em estrela) das encomendas:



Passos para o desenho dimensional de um DW:

1) Analise de requisitos da actividade da organização e da necessidade dos agentes de decisão, de modo a determinar a informação mais importante a ser disponibilizada de imediato aos agentes de decisão (Construção dos Data Marts) ;

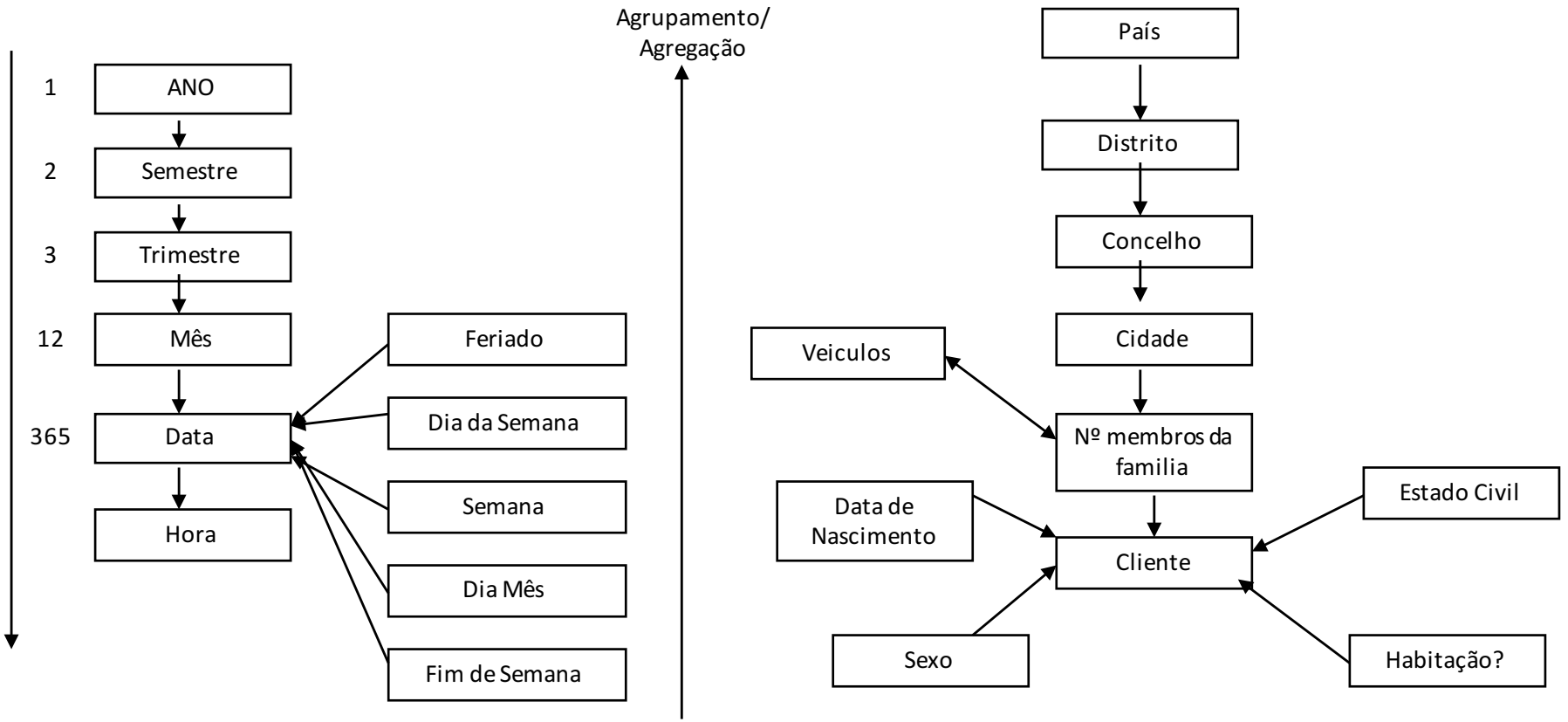
2) Especificar a granularidade da informação (na dimensão tempo e nas outras dimensões) que os agentes de decisão pretendem. É através deste refinamento que se irão efectuar os refinamentos e agregações nos critérios de selecção de dados;

3) Especificar as dimensões envolvidas. A dimensão mais importante é a dimensão **Tempo**, de modo a analisar temporalmente que informação se possuí;

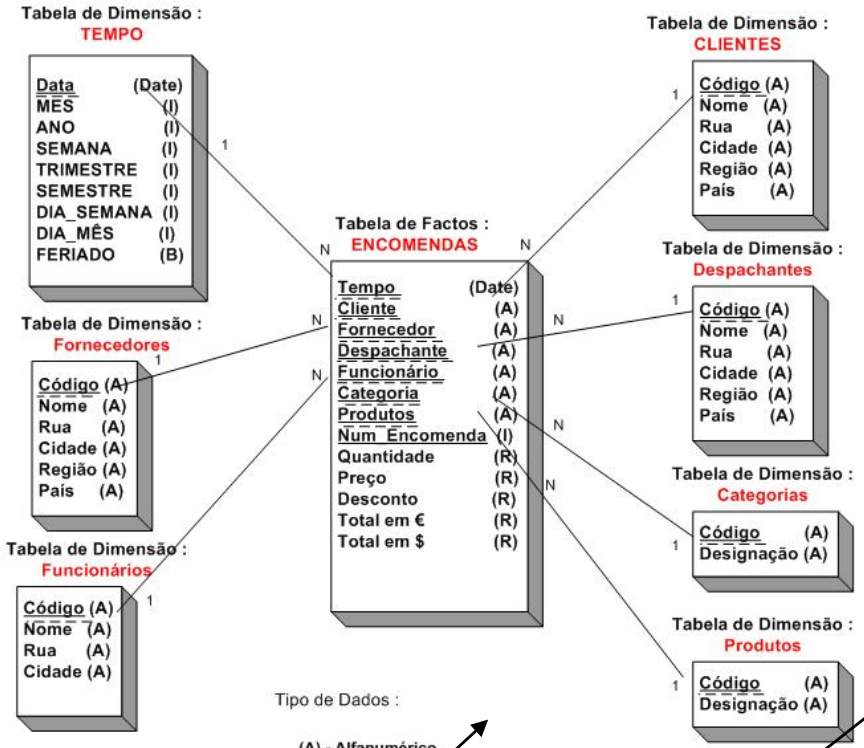
4) Especificar as medidas (Atributos derivados). Caso não houvessem apenas se poderíam efectuar contagens (teríamos uma tabela de factos sem factos "*factless*")

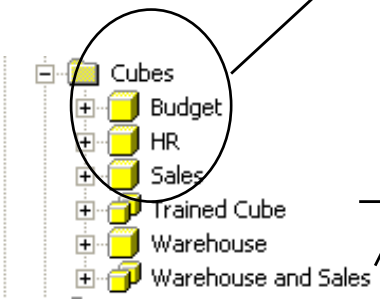
NOTA: Um DW é constituído por 1 ou mais Data Marts

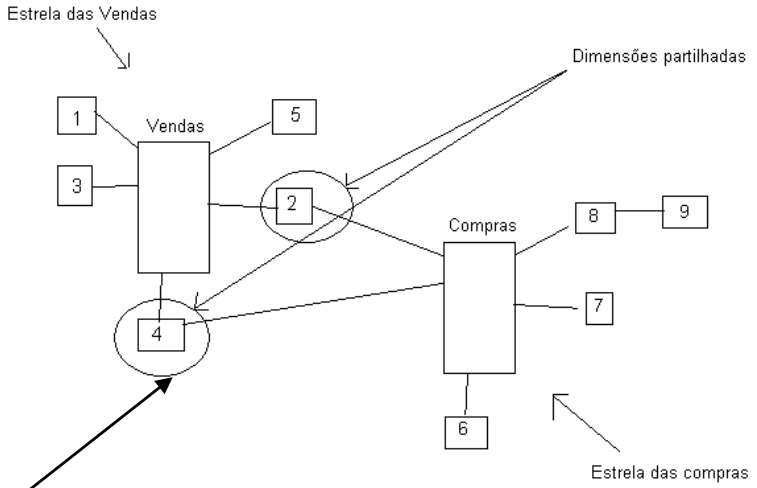
Granularidade da informação:



Cubos OLAP







NOTA:

Se apartir da dimensão Tempo, colocásse-mos a dimensão a DATA_DE_DESPACHO das encomendas, poderíamos colocar mais uma medida TEMPO_DE_SATISFAÇÂO da Encomenda

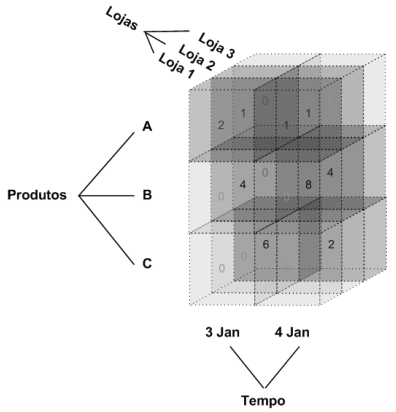
Terímos uma Dimensão (DATA_DE_DESPACHO) **Degenerada**

Cubos OLAP

	Produto	Loja	Tempo	qtd
1	A	Loja 1	3 Jan	1.0
2	A	Loja 1	3 Jan	1.0
3	B	Loja 2	3 Jan	4.0
4	C	Loja 3	3 Jan	2.0
5	C	Loja 3	3 Jan	2.0
6	C	Loja 3	3 Jan	2.0
7	A	Loja 1	4 Jan	1.0
8	A	Loja 2	4 Jan	1.0
9	B	Loja 2	4 Jan	4.0
10	B	Loja 2	4 Jan	4.0
11	B	Loja 3	4 Jan	4.0
12	C	Loja 3	4 Jan	2.0
13	A	Loja 2	3 Jan	1.0

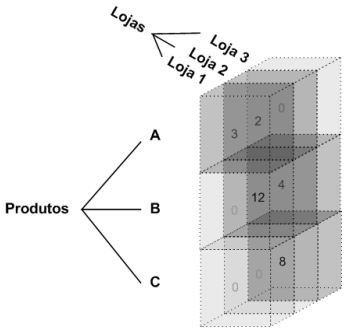
Agrupamento - PLT

	Produto	Loja	Tempo	qtd
1	A	Loja 1	3 Jan	2.0
2	A	Loja 2	3 Jan	1.0
3	B	Loja 2	3 Jan	4.0
4	C	Loja 3	3 Jan	6.0
5	A	Loja 1	4 Jan	1.0
6	A	Loja 2	4 Jan	1.0
7	B	Loja 2	4 Jan	8.0
8	B	Loja 3	4 Jan	4.0
9	C	Loja 3	4 Jan	2.0



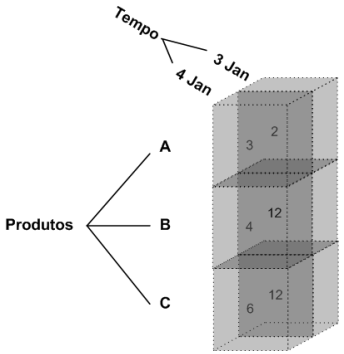
Agrupamento - PT

	Produto	Loja	qtd
1	A	Loja 1	3.0
2	A	Loja 2	2.0
3	B	Loja 2	12.0
4	B	Loja 3	4.0
5	C	Loja 3	8.0



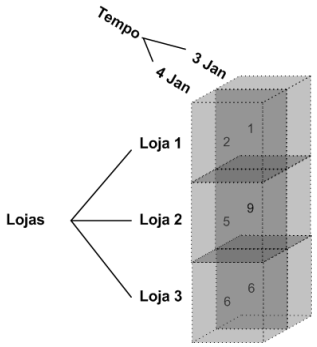
Agrupamento - LT

	Produto	Tempo	qtd
1	A	3 Jan	3.0
2	B	3 Jan	4.0
3	C	3 Jan	6.0
4	A	4 Jan	2.0
5	B	4 Jan	12.0
6	C	4 Jan	2.0

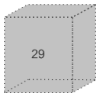


Agrupamento - LT

	Loja	Tempo	qtd
1	Loja 1	3 Jan	2.0
2	Loja 2	3 Jan	5.0
3	Loja 3	3 Jan	6.0
4	Loja 1	4 Jan	1.0
5	Loja 2	4 Jan	9.0
6	Loja 3	4 Jan	6.0

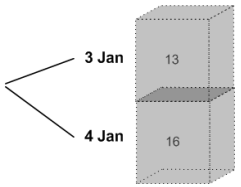


Agrupamento - ALL



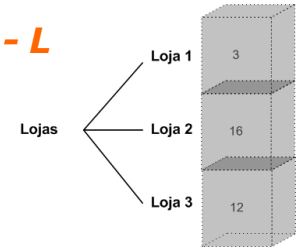
Agrupamento - T

	Tempo	qtd
1	3 Jan	13.0
2	4 Jan	16.0



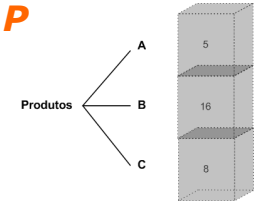
Agrupamento - L

	Loja	qtd
1	Loja 1	3.0
2	Loja 2	14.0
3	Loja 3	12.0



Agrupamento - P

	Produto	qtd
1	A	5.0
2	B	16.0
3	C	8.0



Metadados

Segundo Inmon os metadados englobam o DW e mantém as informações sobre o que está e onde. Define ainda quais que informações os metadados mantém:

- A estrutura dos dados segundo a visão do programador;
- A estrutura dos dados segundo a visão dos analista de SAD;
- A fonte de dados que alimenta o DW;
- A transformação sofrida pelos dados no momento de sua migração para o DW;
- O modelo de dados;
- O relacionamento entre o modelo de dados e o DW;
- O histórico das extrações de dados;
- Acrescentamos ainda os dados referentes aos relatórios que são gerados pelas ferramentas OLAP assim como os que são gerados nas camadas semânticas.

OLAP

- **Consultas ad-hoc**
  consultas com acesso casual único e tratamento dos dados segundo parâmetros nunca antes utilizados, geralmente executado de forma iterativa e heurística

- **Slice-and-Dice**
  analisar informações de diferentes prismas limitados somente pela imaginação. Utilizando esta tecnologia conseguimos ver a informação sobre ângulos que anteriormente inexistiam sem um DW e sem OLAP.

- **Drill Down/Up**
  exploração em diferentes níveis de detalhe das informações, "subir ou descer" dentro do detalhe dos dados, como por exemplo analisar uma informação diariamente ou anualmente, partindo da mesma origem de dados.

- **Geração de Queries**

  A geração de queries no OLAP é simples, amigável e transparente para o utilizador final, que precisa de ter um conhecimento mínimo de informática para obter as informações que deseja.

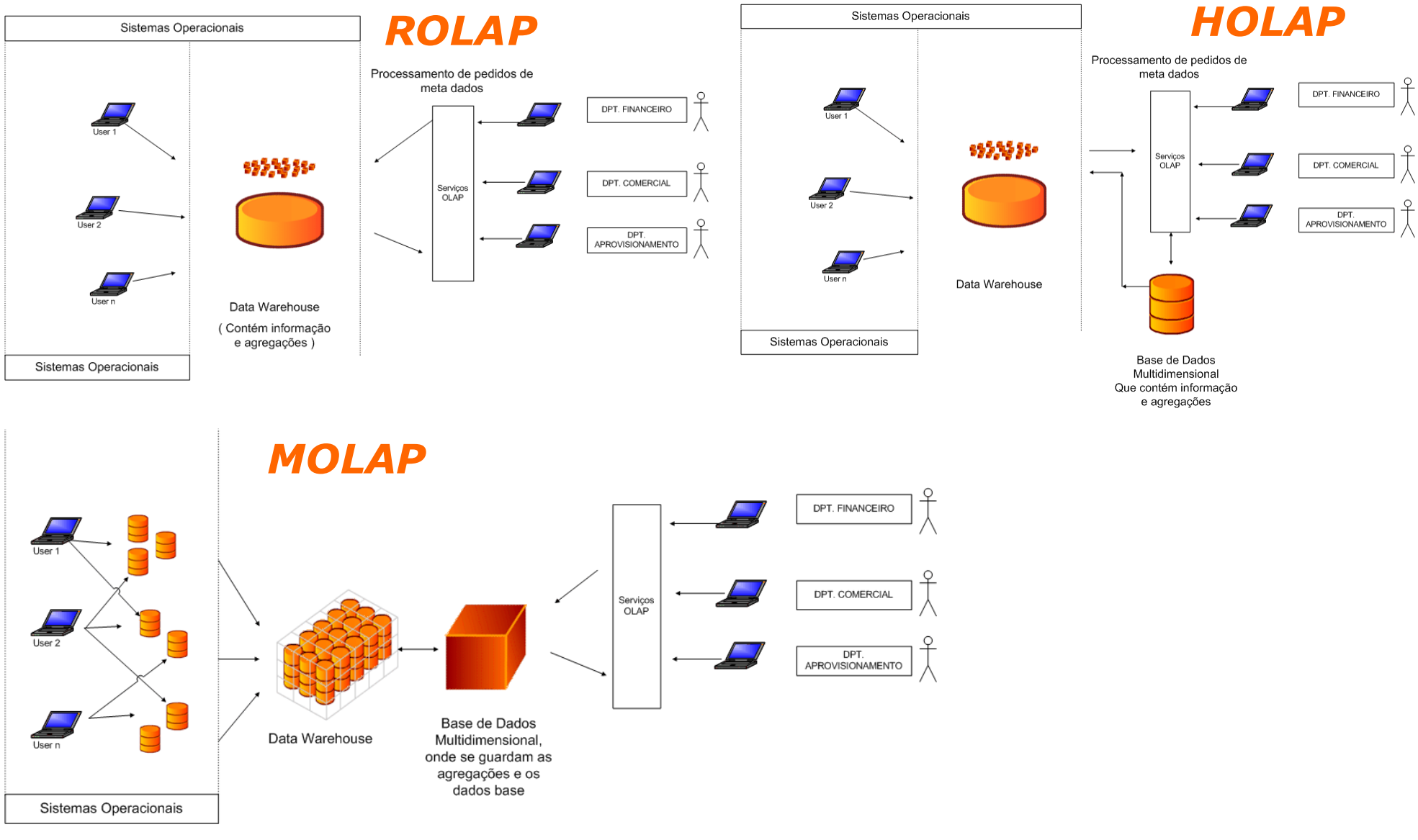
Ferramentas

Para a visualização dos dados, existe uma classe específica de ferramentas, conhecidas como ferramentas OLAP.

Há várias sub-classes como

- ROLAP (Relational OLAP),
- DOLAP (Desktop OLAP),
- MOLAP (Multidimensional OLAP) e
- HOLAP (hybrid OLAP) = DOLAP OU ROLAP + MOLAP.

Estruturas Dimensionais



Ferramentas

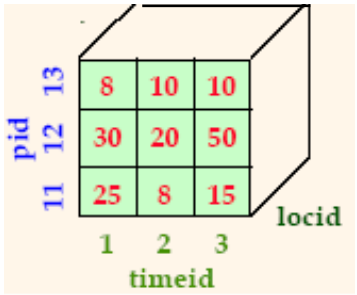
- ROLAP – OLAP fica num servidor dedicado, que armazena os vários “cubos” de informação. O utilizador acede aos vários cubos, e analisa as informações com o processamento OLAP sendo realizado no servidor. Pode trazer problemas de escalabilidade (número de utilizadores) e de tráfego de rede. Por outro lado, permite a análise de grandes volumes de dados.

- DOLAP – O processamento OLAP é feito na máquina cliente, sem tráfego de rede nem problemas de escalabilidade. Contudo, pode trazer problemas em alguns relatórios, quando o volume de dados fica muito grande, apesar das boas ferramentas tratarem os dados de maneira compactada.

- MOLAP – no SGBD ficam os dados num formato simples, e no Servidor MOLAP, ficam os dados consolidados. O utilizador visualiza directamente o Servidor MOLAP, usando os módulos de consulta desta ferramenta.

- HOLAP – junta-se uma ferramenta ROLAP ao MOLAP. É um sistema extremamente completo, contudo é o mais caros de todos, sendo que muitas vezes a análise custo/benefício inviabiliza esta opção.

Vista tri-dimensional



Locid = 1

pid	timeid	locid	vendas
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

Comparação com SQL

```
SELECT SUM(S.sales)
FROM  Sales S, Times T, Locations L
WHERE  S.timeid=T.timeid AND S.timeid=L.timeid GROUP BY T.year, L.state
```

```
SELECT SUM(S.sales)
FROM  Sales S, Times T
WHERE  S.timeid=T.timeid
GROUP BY T.year
```

```
SELECT SUM(S.sales)
FROM  Sales S, Location L
WHERE  S.timeid=L.timeid
GROUP BY L.state
```

O operador CUBE

Generalizando a partir do exemplo, sendo k o número de dimensões, existem 2^k interrogações do tipo GROUP BY que podem ser geradas "girando" em cada sub-conjunto de dados associados às dimensões:

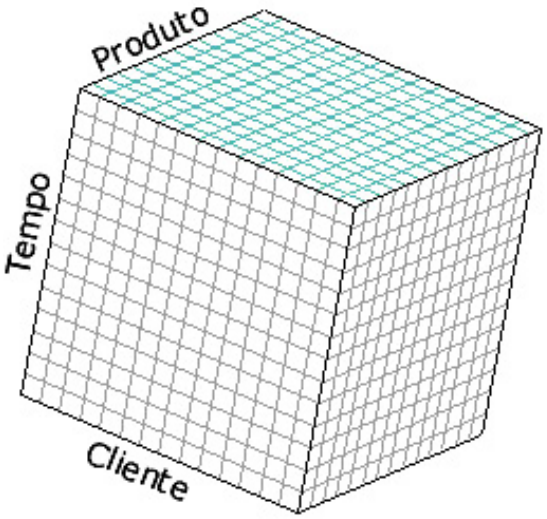
CUBE pid, locid, timeid BY SUM sales

é equivalente a correr Sales em cada um dos 8 subconjuntos de {pid, locid, timeid}. Cada parte corresponde a uma interrogação SQL do tipo GROUP BY.

Tabelas Dimensionais

Chave	Produto	Tipo de Venda
01	Laranja	Caixa
02	Laranja	Saco
03	Maçã	Unidade
04	Maçã	Caixa
05	Maçã	Saco
06	Morango	Caixa
...	...	...

Chave	Produto
01	Laranja
02	Laranja
03	Maçã
04	Maçã
05	Maçã
06	Morango
...	...



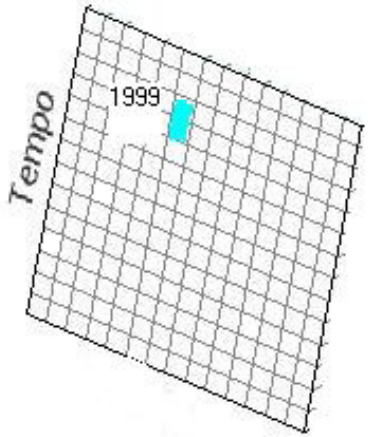
Dimensão

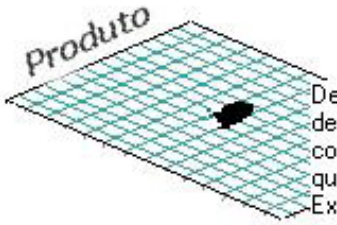
A técnica de se criar bases de dados simples e compreensíveis foi aperfeiçoada para o surgimento da modelação dimensional.

A capacidade de poder observar uma base de dados no formato de um cubo, contendo duas, três, quatro ou até mais dimensões, permite fatiá-lo em qualquer uma de suas dimensões.

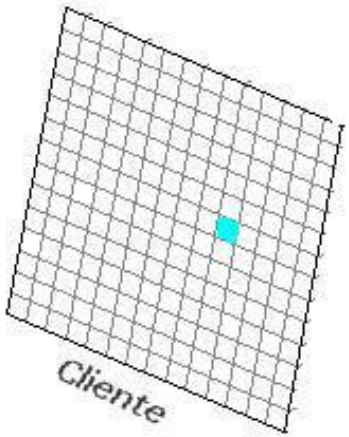
A Modelação dimensional irá permitir analisar um negócio em qualquer uma das suas dimensões (visões de negócio).

Dimensões

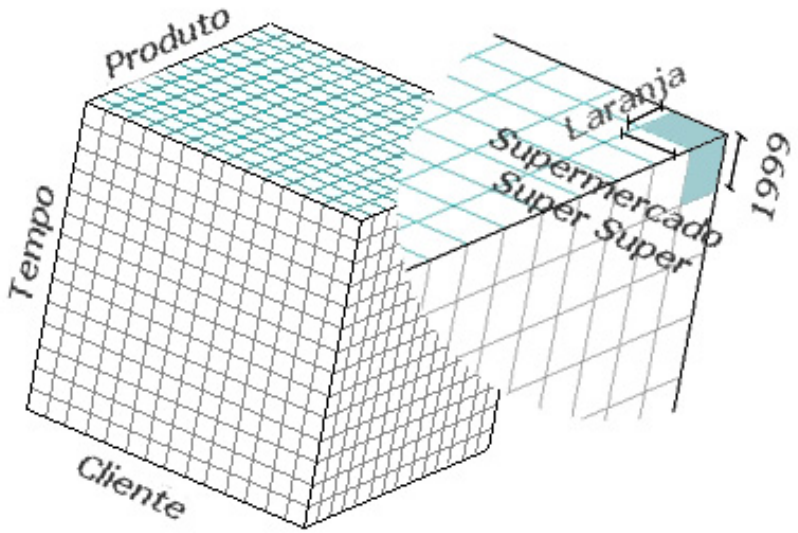




Dentro de uma variedade enorme de produtos só será capturado como informação o ou os produtos que o usuário escolheu.
Ex. Laranja



Em muitas ocorrencias, somente é capturada a informação desejada pelo cliente.



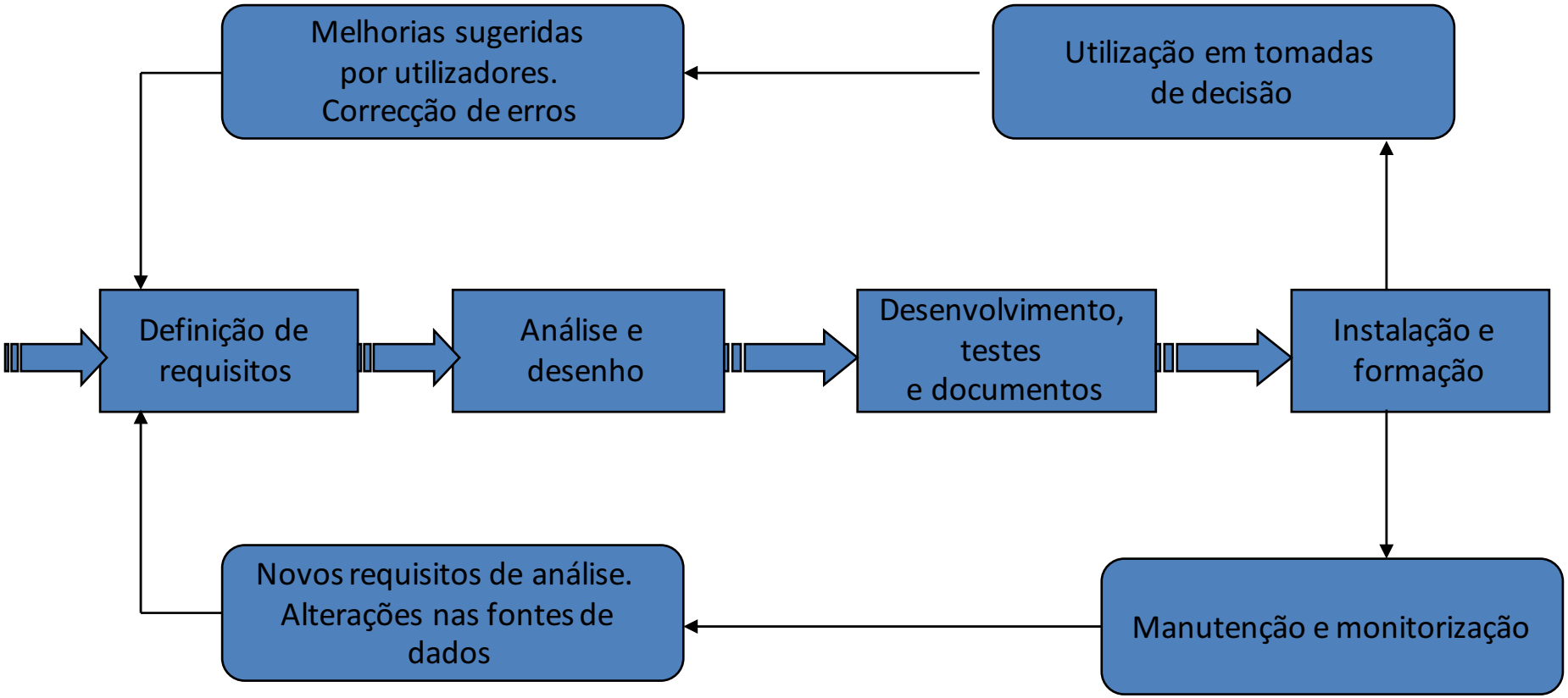
Extracção de Dados

- Extracção primária;
- Identificação dos registos modificados;
- Generalização de chaves para dimensões em modificações;
- Transformação em imagens de registo de carga;
- Migração do sistema legado para o sistema DDW;
- Classificação e construção de agregados;
- Generalização de chaves para agregados;
- Carregamento;
- Processamento de excepções;
- Garantia de qualidade e,
- Publicação.

5 operações principais de back-end

1. Extracção dos dados de fontes internas e externas;
2. Limpeza dos dados extraídos;
3. Transformação;
4. Carga no DW e,
5. Actualizações (refresh).

Ciclo de vida de um SAD



Pessoas a envolver

- Administração (adjudicação e financiamento);
- Gestor de projecto de suporte à decisão;
- Gestores de áreas de negócio;
- Utilizadores finais;
- Administradores de bases de dados

Linhas a explorar

- Quem são e o que fazem as pessoas;
- Missão e objectivos dos grupos ou departamentos;
- Directivas principais;
- Medidas de lucro e custos;
- Áreas sensíveis da organização;
- Opiniões sobre os clientes;
- Informação a disponibilizar no repositório analítico;
- Avaliação do nível de detalhe da informação;
- Relevância da informação passada, tempo de ponderação e período de análise;
- Janela temporal para as actualizações (frequência da integração operacional).

Questões

- Fontes de dados?
- Interacções inter-sistemas?
- Manutenções e configurações à medida?
- Documentação interna?

Abordagem à modelação dimensional

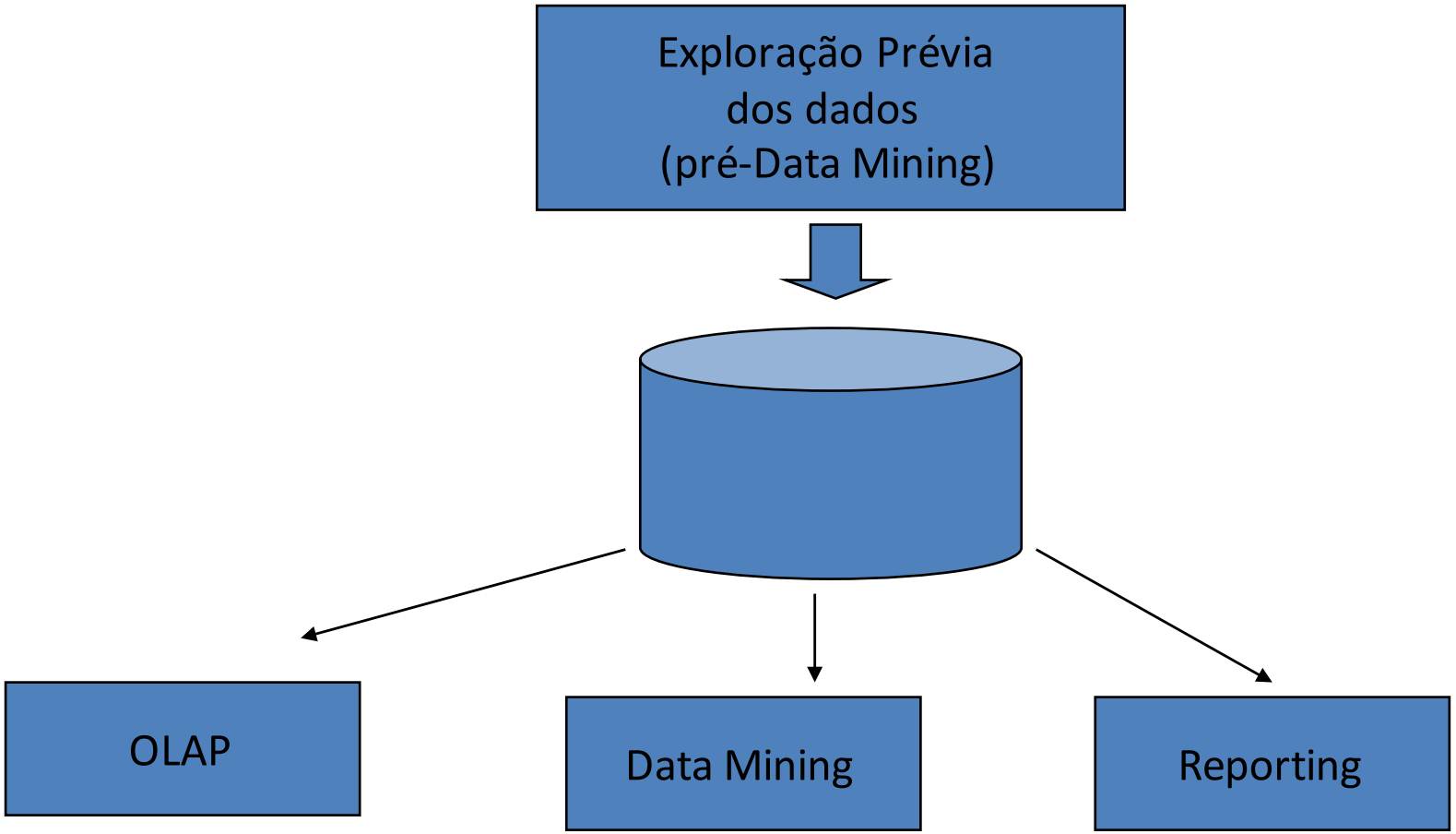
- Processos funcionais e de negócio
- Granularidade
- Dimensões necessárias
- Medidas e valores pré-calculados
- Atributos das dimensões
- Detecção e tratamento de dimensões variáveis no tempo
- Níveis de agregação, dimensões heterogéneas, hierarquias e outras definições de agregação
- Tempo de vida dos dados e históricos
- Disponibilidade dos dados.

Selecção, extracção, transformação e integração dos dados

- Fontes heterogéneas
- Área de processamento de dados
- Processo de selecção de dados
- Processo de extracção de dados (ferramentas, etiquetagem)
- Transformação de dados (conversões, mapeamento, desduplicação de registos, valores nulos, normalização de distribuições e unidades de grandeza, regras de integridade, regras de negócio, mudanças em registo de dimensões, agregações, erros)
- Integração de dados (validação e desempenho, monitorização e reposição de cópias de segurança).

Data Mining

DM = Extracção de Conhecimento + Teste de Hipóteses



Problemas

- Classificação de entidades
- Estimação de propriedades
- Previsão temporal
- Descoberta de afinidades
- Clustering (agrupamentos homogéneos)
- Descrição de comportamentos organizacionais.