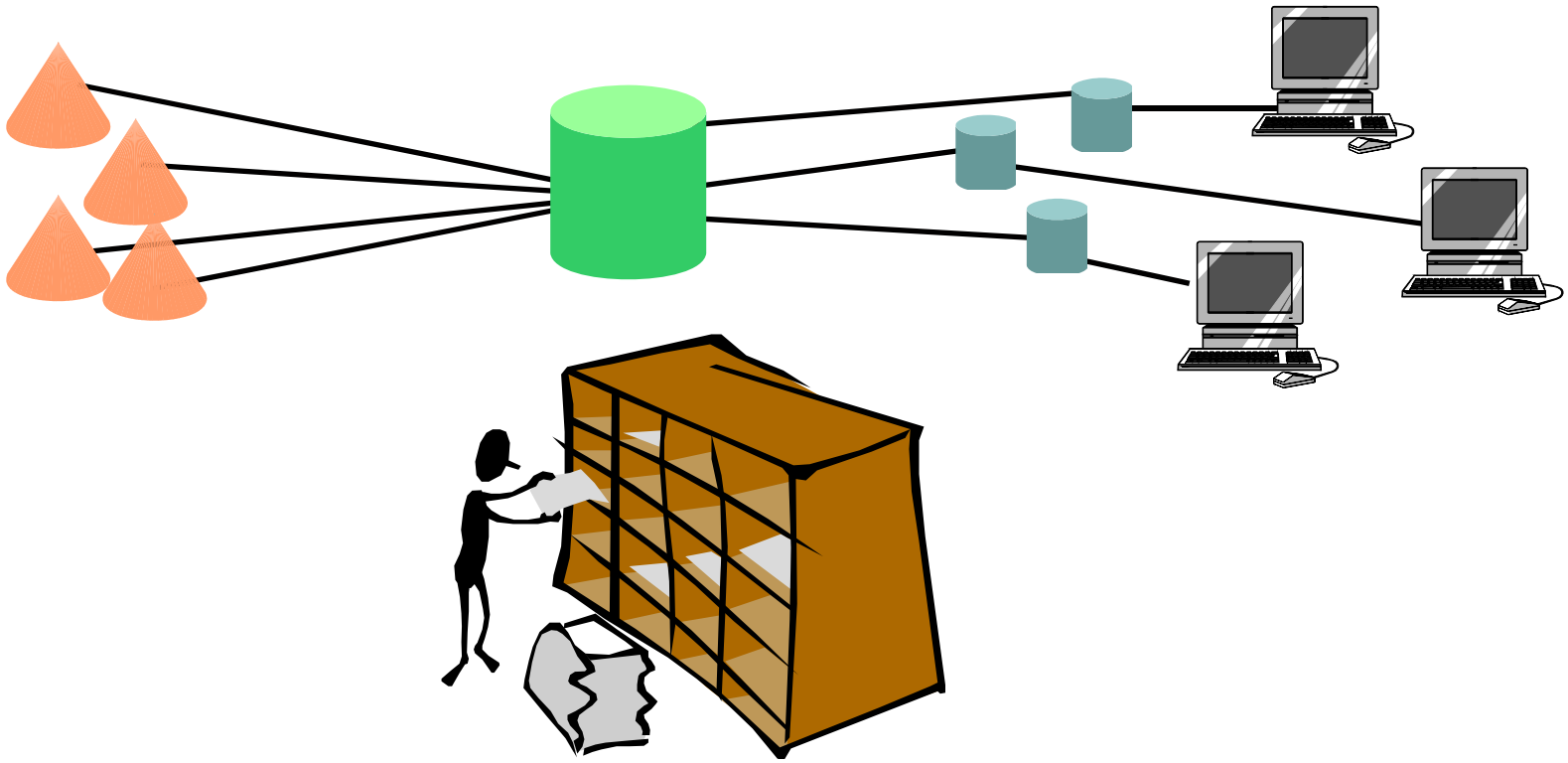
MBA em Gestão da Informação e *Business Intelligence*

Arquitetura de *Data Warehouse*

Aula 03 - Ferramentas de ETLM

Extração, Transformação, Carga e Metadados: ETLM

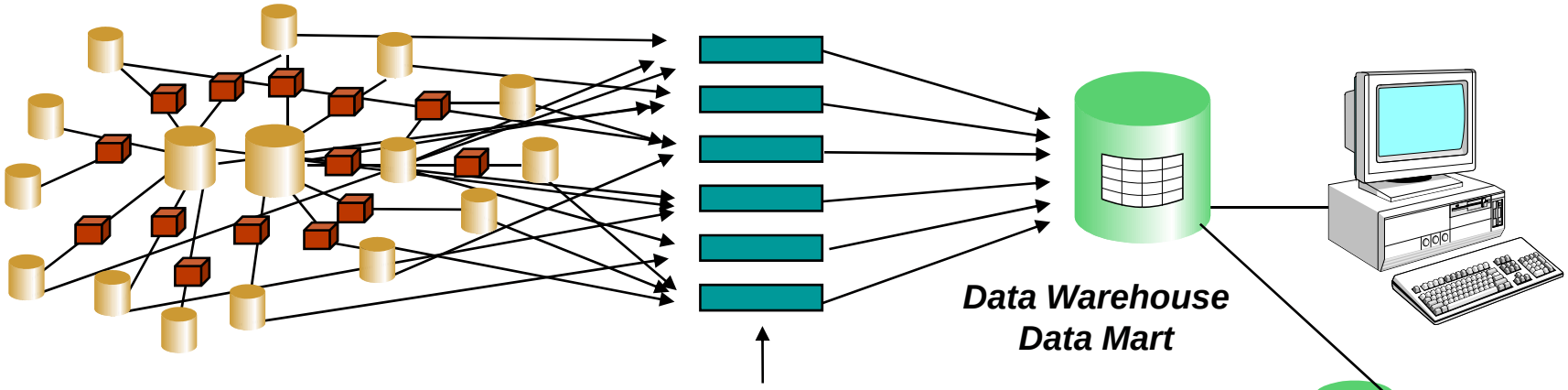


70% do esforço em *data warehousing* é gasto com a definição de fontes, mapeamentos, regras, *scheduling*, e manutenção dos processos de ETLM

Transformando Dados em Informação

Ambiente *Legacy*:
Ilhas de Teias de Aranha de Dados

Processamento Informacional



Data Warehouse
Data Mart

Transformando Dados em Informação:

- Extrair
- Integrar
- Sumarizar
- Filtrar
- Converter
- Valores default
- Reestruturar
- Reformatar
- Criar consistencia
- Regras de Negócio

Metadados

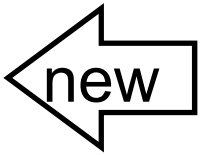
4 Tipos de atividades

- Monitoração: dos dados que vem das fontes
- Integração: Limpeza de dados, Carga, ...
- Gerência: Metadado, Projeto, ...
- Processamento*: Query processing, indexing*, ...

Monitoração

- Tipos de Fontes: tabelas, arquivos, IMS, VSAM, IDMS, WWW, IBGE, REUTERS, …
- Incremental vs. *Refresh* (tudo)

customer	id	name	address	city
	53	joe	10 main	sfo
	81	fred	12 main	sfo
	111	**sally**	**80 willow**	**la**



Técnicas de Monitoração

- *Snapshots* periódicos
- *Database triggers*
- *Log shipping* (envio de *log*)
- *Data shipping (replication service)*
- *Transaction shipping*
- *Polling* (*queries* nas fontes)
- Recortes de telas
- ....

Questões na Monitoração

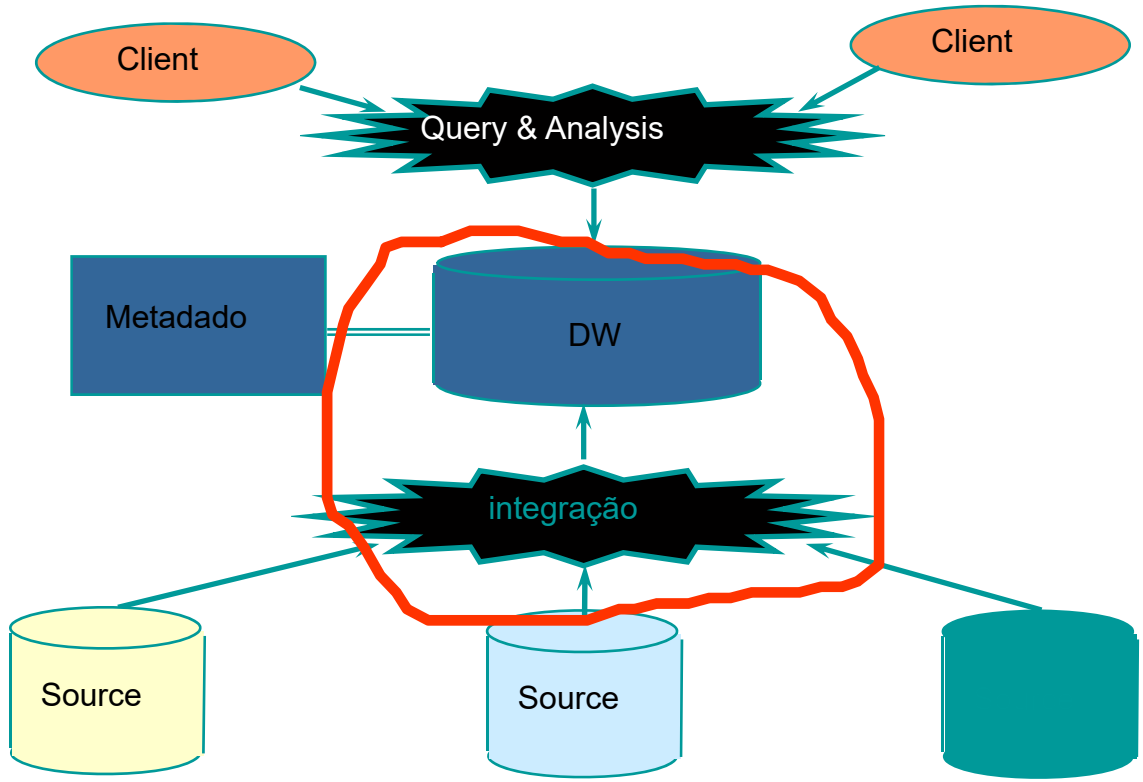
- Frequência
  - periódica: diária, semanal, …
  - *triggered*: quando ocorre uma **grande** mudança, muitas mudanças, ...

- Transformação de Dados
  - converte dados (formato uniforme)
  - remove & add campos(ex., *add date* => *history*)

- Uso de Padrões (ex., ODBC)
- *Gateways*

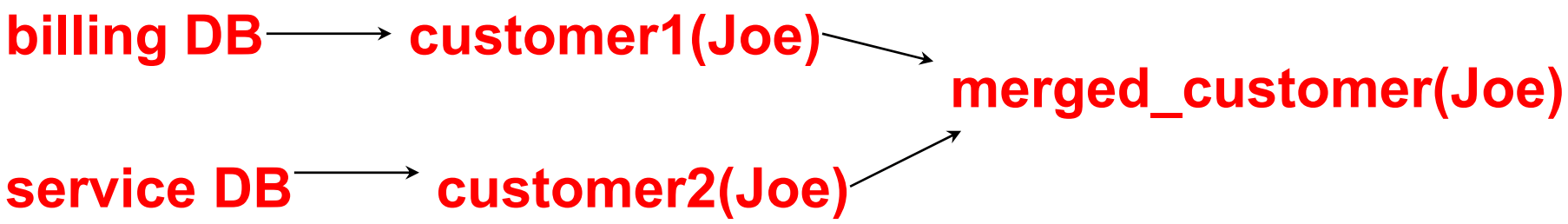
Questões na Integração

- Limpeza de Dados
- Carga de Dados
- Dados Derivados



Limpeza de Dados

- Migração (ex., yen → dollar)
- Scrubbing: uso de conhecimento em domínio específico (ex., números de CPFs)
- Fusão (ex., lista de correio, casar dados clientes)



- *Auditing*: descobrir regras & relacionamentos (ex. *data mining*)

Carga de Dados

- Incremental vs. *Refresh*
- *Off-line* vs. *on-line*
- Frequência de carga
  - A noite, 1x p/sem/mês, continuamente

- Carga Paralela/Particionada

Dados Derivados

- Dados Derivados no DW
  - Índices
  - Agregados
  - *Views* materializadas

- Quando atualizar dados derivados?
- Incremental vs. *Refresh*...

Carga de DW

- Para aumentar a performance, DWs frequentemente armazenam resumos calculados e visões predefinidas

- Informação adicional de fontes externas também podem ser incluídas no DW

Views Materializadas

- Define nova tabela no DW usando SQL

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

product	id	name	price
	p1	bolt	10
	p2	nut	5

joinTb	prodId	name	price	storeId	date	amt
	p1	bolt	10	c1	1	12
	p2	nut	5	c1	1	11
	p1	bolt	10	c3	1	50
	p2	nut	5	c2	1	8
	p1	bolt	10	c1	2	44
	p1	bolt	10	c2	2	4

Nova tabela
Com base em fontes
diferentes

Carga de DW

- Leitura de dados de fontes variadas
- Qualidade de dados é crítica
- Precisamos de cargas eficientes, flexíveis
- Cargas noturnas (limites)

Realidades sobre Qualidade de Dados

- DW vem de múltiplas fontes “sujas”
  - Legacy systems não documentados
  - Sistemas de produção sem verificações de integridade
  - Fontes externas com procedimentos de qualidade questionáveis

- Decisões e recomendações com segurança precisam de dados com qualidade

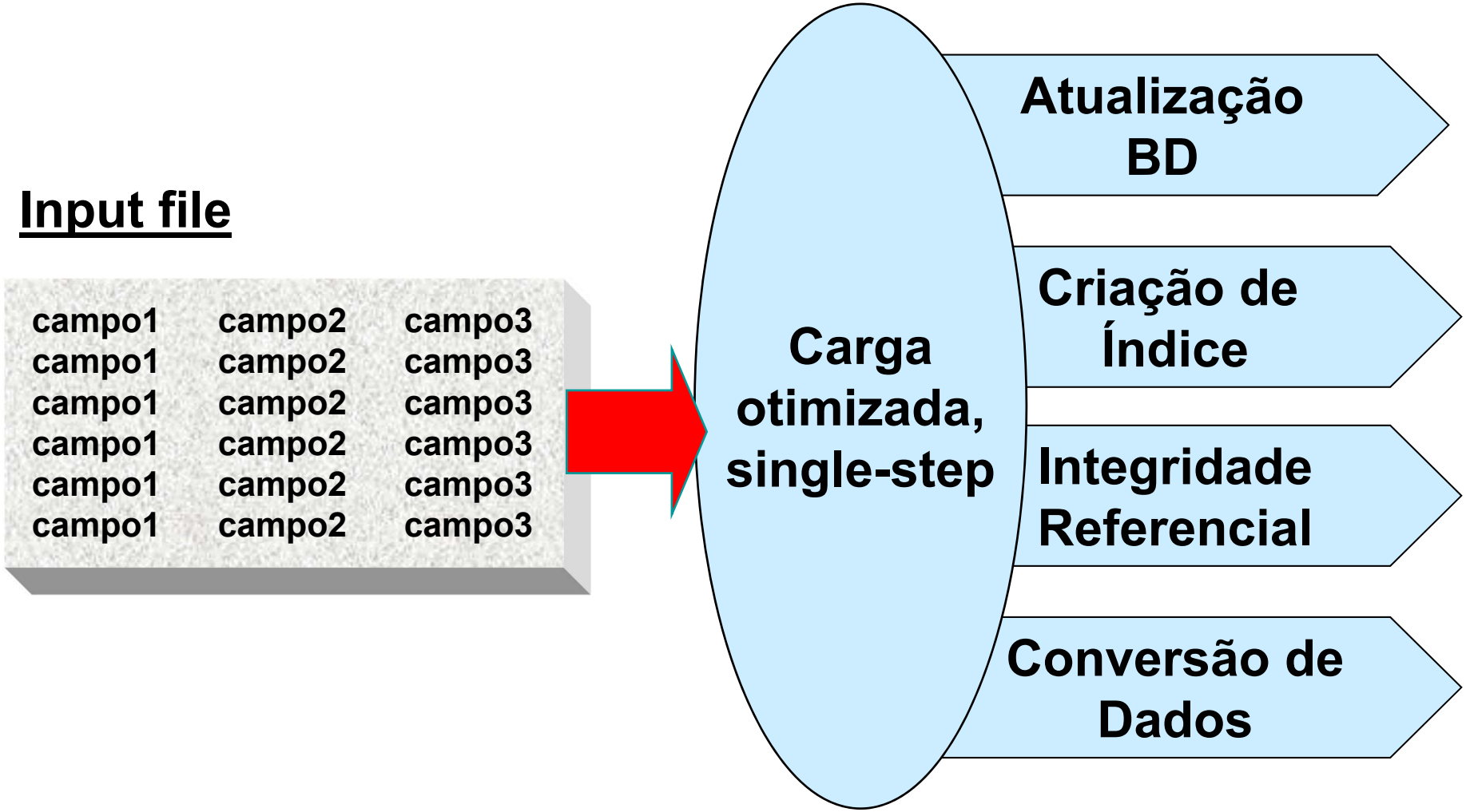
Cargas Eficientes, Flexíveis

- Processamento de único passo
  - Ler, ajustar e reformatar a entrada
  - Detectar dados sujos, incluindo violações de integridade referencial
  - Agregar, armazenar e indexar dados
  - Nesse caso sempre uma carga total

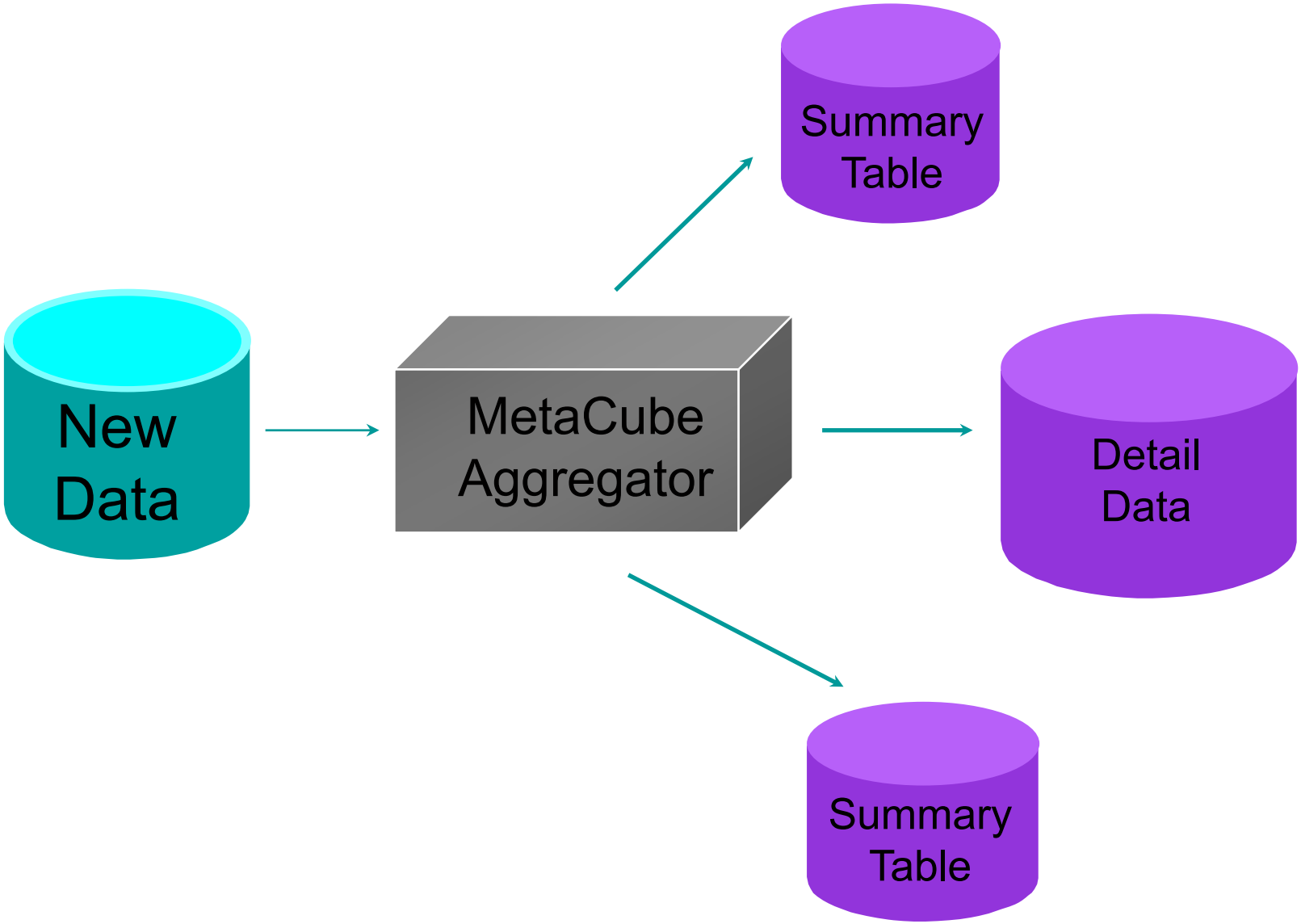
- Cargas multi-função
  - Insert, append, update, modify, replace
  - Carga total e incremental

- Cargas on-line e off-line

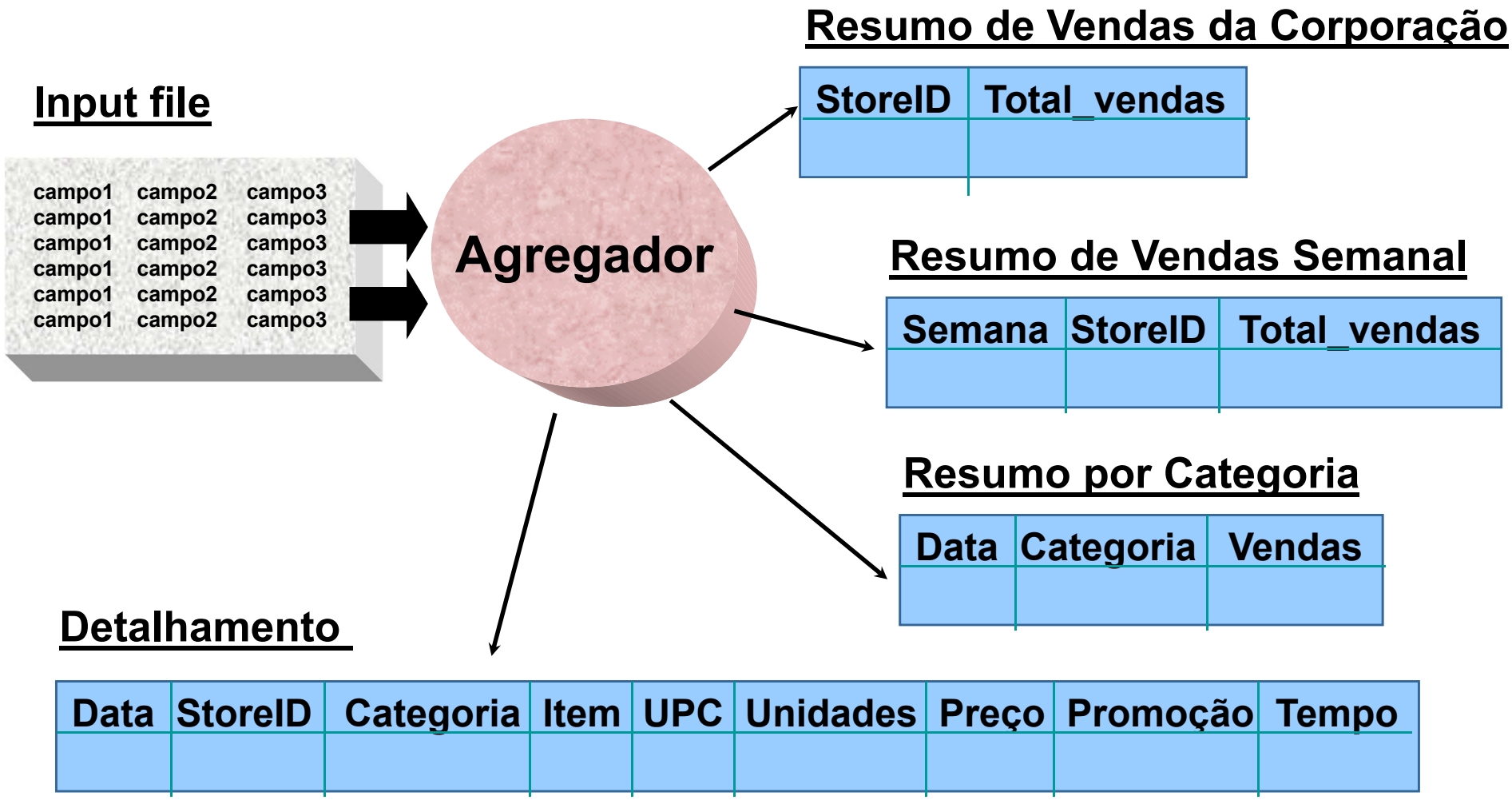
Processamento de Carga



Existem ferramentas para agregar dados



Processamento de Carga



Armazenamento de Dados no DW

- DW > 100 GB estão tornando-se comuns
- “52% dos DWs passarão de 20GB para terabytes nos próximos anos” (META Group 97)
- Nível de detalhe requerido pelo negócio determina volumes de dados armazenados

Cargas Noturnas

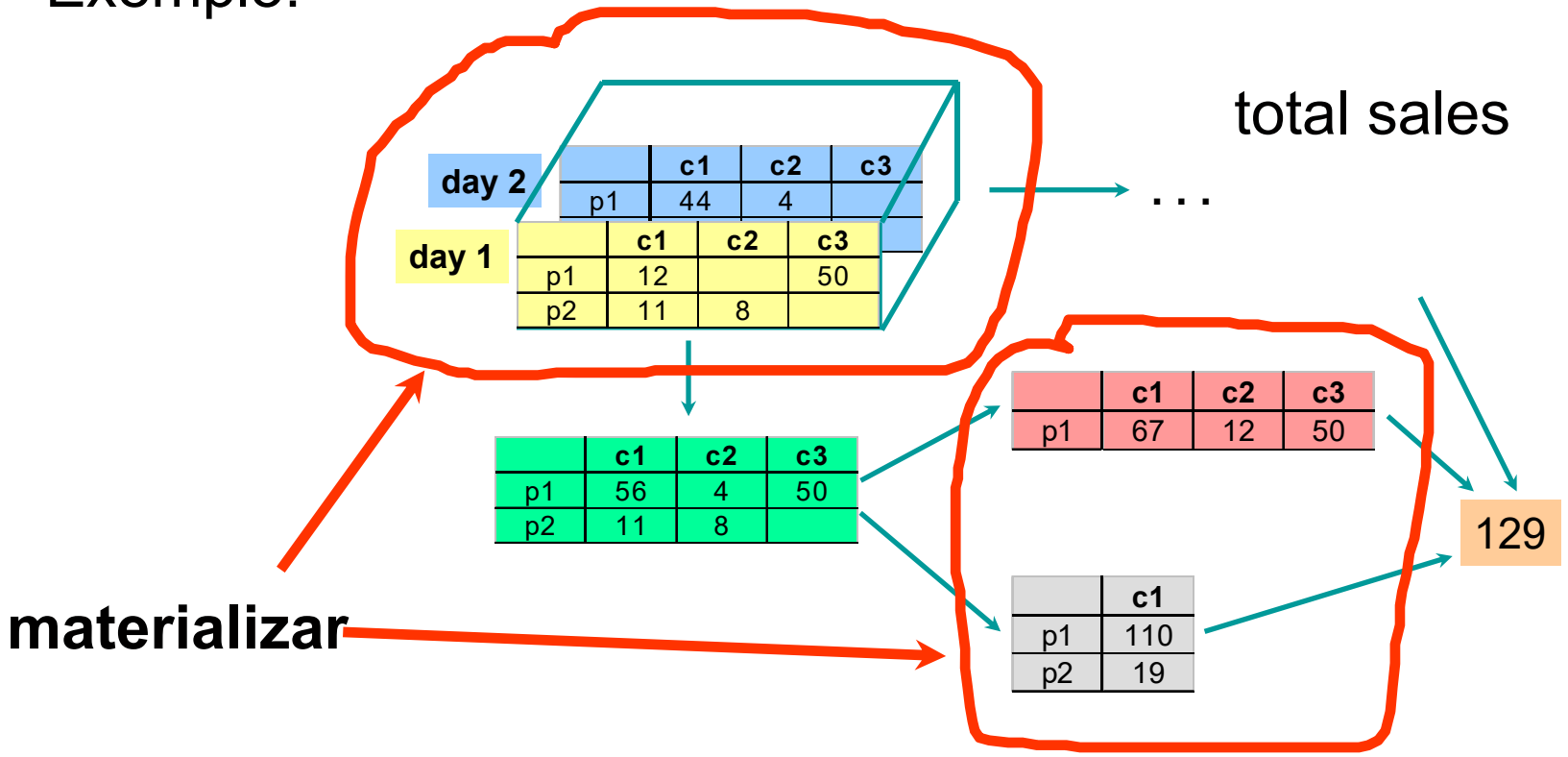
- As “janelas” da operação já não dão para carregar os dados de DW
- Volumes de dados operacionais crescem
- Medidas em gigabytes por hora...
- Limite ? 70-100 Gb / noite

Administração

- São necessárias novas formas de gerência de BD para os grandes volumes de dados dos DWs de hoje
- “Resiliência” de BD é chave para gerência
- Grande número de usuários => precisa de baixo custo de administração
- Grandes BDs tendem a ter mais falhas de hardware... (Segmentar? Por mês, Por Produto,...)

O que Materializar?

- Gravar no DW resumos e agregados úteis p/ queries mais comuns
- Exemplo:



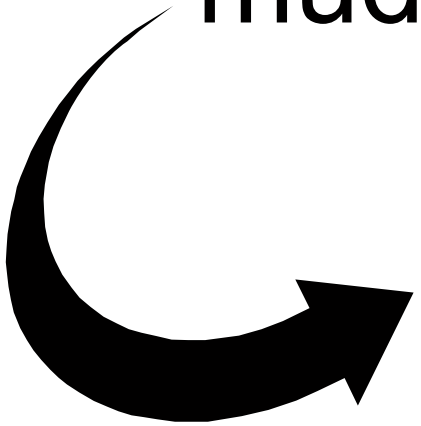
Fatores p/ Materialização

- Tipo/frequência de queries
- Tempo de resposta de Queries
- Custo de armazenamento
- Custo de atualização

Qual é a **funcionalidade** necessária para uma ferramenta de Extração e Transformação de dados operacionais para carga de DW?

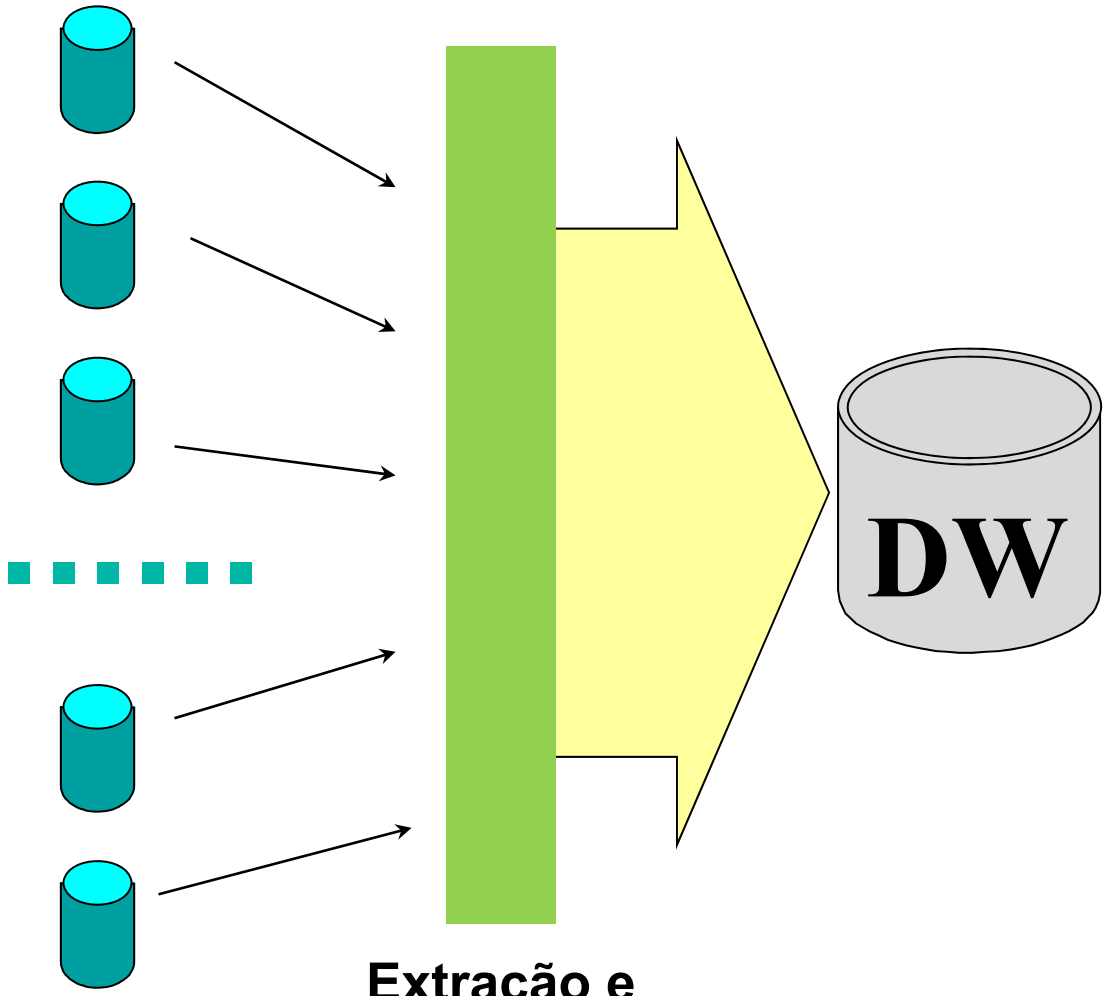
Extração e Transformação

A extração de dados do ambiente operacional para o ambiente DW requer uma mudança de **tecnologia**!



- Leitura com um SGBD operacional
- Gravação por meio de um SGBD de DW (com SQL estendida p/ DW)

Extração e Transformação



À medida que os dados vão sendo extraídos e transformados, vão sendo carregados no DW (e gerando **metadados**)

Exemplo de Passos de ETLM

- Extração primária (leitura dos arquivos operacionais)
- Identificação dos registros alterados
- Generalização de chaves das dimensões
- Transformação em registros para carga
- Migração dos dados do ambiente operacional para o ambiente de DW

Exemplo de Passos de ETLM

- Construção dos agregados
- Generalização de chaves para os agregados (Tabelas resumos etc.)
- Carga
- Processamento de exceções
- Garantia de qualidade
- Documentação e publicação

ETLM: Desenvolvimento Manual

Características

- Codificação Manual
- Performance Depende da Linguagem Usada e do Ambiente
- Linguagens 3GL / 4GL (Cobol, C, Natural, VB, Easytrieve, PL/SQL,
- Transact/SQL, Shell Scripts)

Vantagens

- Pouco Investimento Inicial
- Aproveitamento de Equipes Treinadas e Metodologias Consolidadas (se existentes), bem como de Recursos de Mercado
- Menor Dependência de Fornecedores

Desvantagens

- Qualidade Depende dos Programadores (Difícil Padronização)
- Difícil Manutenção/Entendimento
- Não Integração a Execução / Transporte / Scheduling
- Inexistência de Templates ERP / CRM
- Não Captura de Metadados

ETLM: Ferramentas de 1a. / 2a. Geração

Características

- Geradores de Código ou Frameworks de Código (ETI Extract, Oracle Warehouse Builder, CA/Platinum Decision Base, Natquery)
- Desempenho Depende da Linguagem Gerada e do Ambiente
- Principais Linguagens Geradas (Cobol, C, Natural, PL/SQL, Extensões de SQL)

Vantagens

- Aproveitamento de equipes existentes e recursos de mercado relativos às linguagens
- Dependência de Fornecedores é Atenuada pelo Código Fonte Gerado
- Maior Facilidade de Desenvolvimento e Manutenção
- Captura de Metadados

Desvantagens

- Investimento Inicial
- Menor produtividade que 3a. Geração
- Não Integração a Compilação / Transporte / Scheduling
- Necessidade de Código Manual Adicional
- Inexistência de Templates ERP / CRM

ETLM: Ferramentas de 3a. Geração

Características

- Tem como Base um “Engine” que gera apenas Código Interno (também chamado de “codeless”)
- Escalabilidade e Performance
- Dependem da Tecnologia do Engine e do Ambiente
- Principais Produtos no Mercado (Acta - ActaWorks, Ascential - DataStage, Cognos - DecisionStream, DataJunction - Integration Studio, IBM - Warehouse Manager, Informatica - PowerMart/PowerCenter, Microsoft - DTS, Sagent - Solution Data Load Server)

Vantagens

- Integração a Pré-compilação / Transporte / Scheduling
- Recursos Avançados (Debuggger,
- Scheduling, Metadados)
- Maior Inteligência / Extensibilidade
- Maior Produtividade
- Templates ERP / CRM Disponíveis
- Captura de Metadados
- Otimização do Desenvolvimento e
- Manutenção

Desvantagens

- Investimento Inicial
- Maior Dependência de Fornecedores

Requisitos Desejáveis em ETL para Ferramentas de 3a. Geração

- Interface Gráfica de Fácil Uso
- Engine Escalável e com Boa Performance
- Biblioteca de Funções (Quantidade e Funcionalidade)
- Suporte a Joins Heterogeneos
- Tabelas de Lookup em Memória
- Geradores de Números Seqüenciais
- Chamada e Inclusão de Stored procedures e Código Externo Especial nas Bibliotecas de Funções
- Suporte a Agregação Incremental

Requisitos Desejáveis em ETL para Ferramentas de 3a. Geração

- Criação e Schedulagem de Sessões de ETL
- Batches para Seqüencias / Dependências de Carga
- Monitoração de Performance em Tempo Real
- Recuperação de Erros
- Metricas de Performance de Carga e para Refinamento
- Suporte a Processamentos Pré e Pós Sessão
- Notificação Automática de Resultados via e-mail
- Disponibilidade de Plataformas
- Opções na Linha de Produtos com Escalabilidade do Investimento

Escolha de Ferramentas de 3a. Geração

Avaliar

- Volumes de Dados
- Periodicidade dos Processos
- Complexidade das Transformações
- Estratégia de Atualização
- Variedade de Fontes e Alvos
- Ambiente de H/W, S/W, Rede
- Necessidade de Integração a Pacotes

✓ Desde “Custo Zero” até Centenas de Milhares de US$

✓ Desde um Revólver 22 até um Lança Mísseis

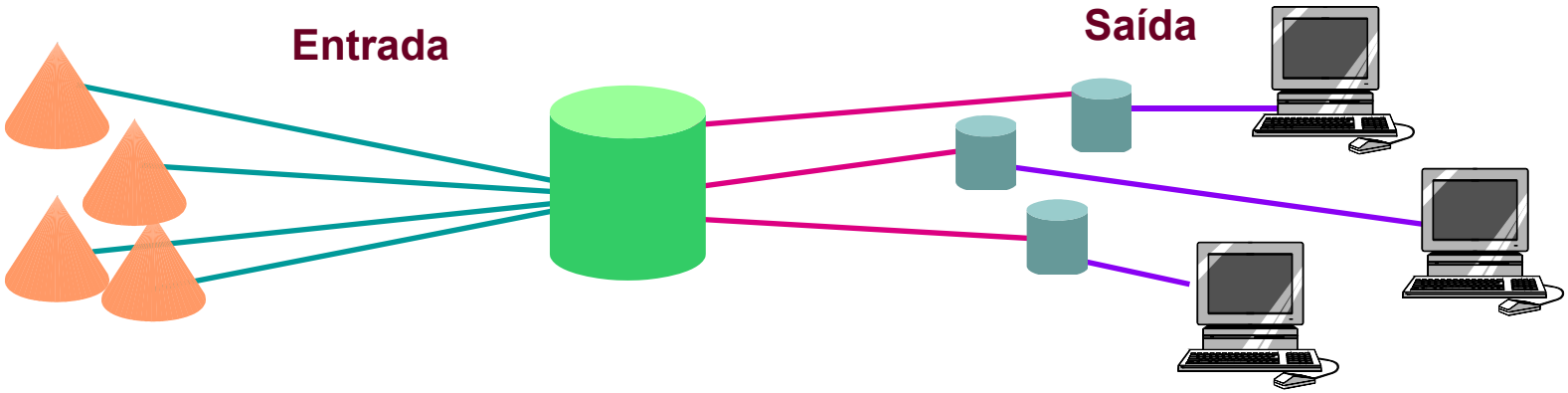
Extratores de Dados

Os fornecedores que oferecem “DW solutions”, em geral, também oferecem ou tem parcerias para uso de produtos como:

- ETI: Unix - gera C, Cobol, etc. - extrai de DB2, Oracle, IMS, Cobol etc.
- Prism: Gera Cobol - para os sistemas comuns (Oracle, Sybase,DB2 etc)
- Passport e outros mais.

Desempenho em ETLM

- O desempenho da saida é muito mais crítico que o desempenho da entrada no DW!



- Desempenho é um conceito relativo que deve ser analisado à luz de arquitetura, modelagem, volumes, recursos de hardware, software e rede, etc
- Codificação Manual/Geração de Código Não Significam Maior Desempenho que Ferramentas ETL com Engine
- Monitoração e refinamento constante são necessários para refletir as mudanças do ambiente do DW (fontes, regras de negócio e alvo)

Sobre o Tamanho dos DWs

- Os DWs estão crescendo demais
  - Terabytes! VLDB! Big Data!

- “O meu DW é maior do que o seu”

- Se é de graça, os usuários querem todas as informações
  - 2 anos? 5 anos?
  - Diária? Mensal?

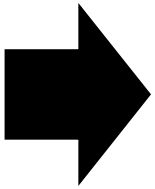
Sobre o Tamanho dos DWs

- Falta de metodologia para extração de dados → VLDW
- Exemplo
  - Código “M” “Masculino”
  - Código 0315 “Vacinação”

- Na extração aumenta o DW

Sobre o Tamanho dos DWs

- Replicação em DMs
- Precisamos de pesquisas para abordar esse problema porque

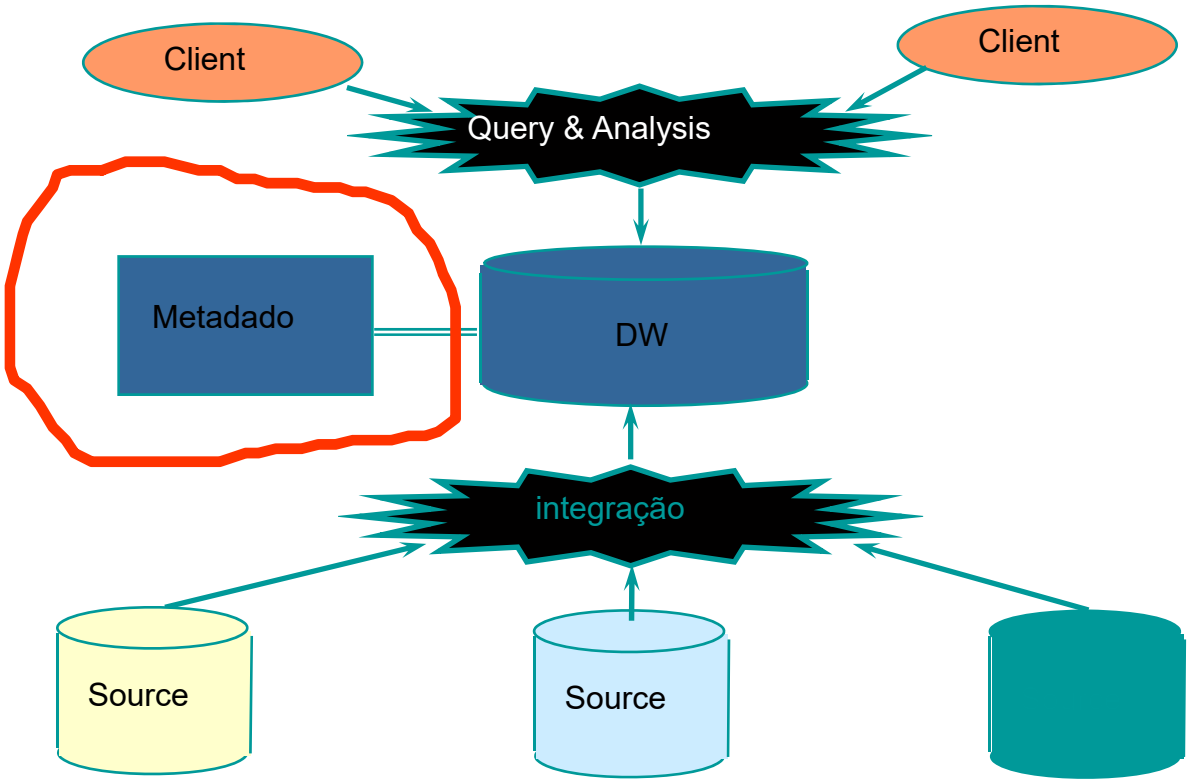
VLDW 

baixo desempenho
alto custo
maior risco de não disponibilidade
usuários menos felizes

Necessidade de ADM de DW!

Gerência de DW

- Gerência dos Metadados
- Gerência do Projeto de DW
- Gerência das Ferramentas



Resumo de Questões do Projeto

- Que dados são necessários?
- De onde vêm (orígem, fontes)?
- Como “limpá-los”, sincronizá-los?
- Como representá-los em DW (schema)?
- O Que sumarizar?
- O Que materializar?
- O Que indexar?

Resumo: *Data Warehousing*

Data Warehousing não é apenas desenvolver um super BD disponibilizado para Análise de Negócios. É uma estratégia que inclue uma arquitetura, uma metodologia de desenvolvimento, um conjunto de ferramentas, um modelo de dados, um BD, um “padrinho” de negócios e um ciclo de vida.

Os 7 Pecados Capitais em *Data Warehousing*

1) Falta de planejamento
2) Descaso com a Arquitetura
3) Pouca importância à documentação
4) Descaso com metodologia e ferramentas
5) Desrespeito ao ciclo de vida do DW
6) Descaso com a resolução de conflitos
7) Falta de aprendizado com erros passados

Administração do DW
Administrando o Crescimento

- Duas das principais causas de crescimento são:
  - Novos dados históricos adicionados de forma composta
  - Adição de dados sumarizados
- Assim, o seguinte paradoxo ocorre:
  - O custo do data warehouse CRESCE!
  - O desempenho do data warehouse DIMINUI!
- Para controlar custos e melhorar performance, o Administrador do Data Warehouse necessita:
  - Otimizar investimentos em hardware (principalmente em discos, memória e processadores)
  - Otimizar investimentos em software
  - Melhorar a performance das queries para atender às necessidades de produtividade dos usuários finais

Metadados

- São os dados que definem os dados
- Metadados: técnicos e semânticos
- Usuários podem examinar o repositório de metadados para
  - a seleção de subconjuntos apropriados de dados, em suas consultas; ou,
  - validações do significado de dados em seus relatórios

Metadados

- De Negócio
  - termos & definições do negócio
  - posse do dado, cobranças etc.

- Operacional
  - origem do dado (fonte)
  - status do dado (ex., ativo, arquivado, "*purged*")
  - uso de estatísticas, relatórios de erro, *audit trails* etc.

Metadados

- Administrativo
  - definição de fontes, *tools*, ...
  - *schemas*, Hierarquias de Dimensão, …
  - regras para extração, limpeza, …
  - políticas de refresh, exclusão (*purging*)
  - perfis de usuários, *access control*, ...

Metadados - Exemplo de usos

- Uso por ferramenta de consulta que automaticamente lê o catálogo de um BD (metadados), acessa os dados desejados e apresenta aos usuários informação sobre negócios

- Quando o usuário faz "*drill down*" em resumos de dados em um BD (usa metadados) para detalhar dados em uma certa analise

- As ferramentas (ETL) de extração / transformação automaticamente usa os metadados na tarefa de mapeamento dos dados "*legacy*" para a carga de DW

- Etc.

Metadados (resumo)

Contém (pelo menos):

- A estrutura do dados
- Os algoritmos usados para os resumos e derivação de dados
- O mapeamento do ambiente operacional para o DW

Metadados (resumo)

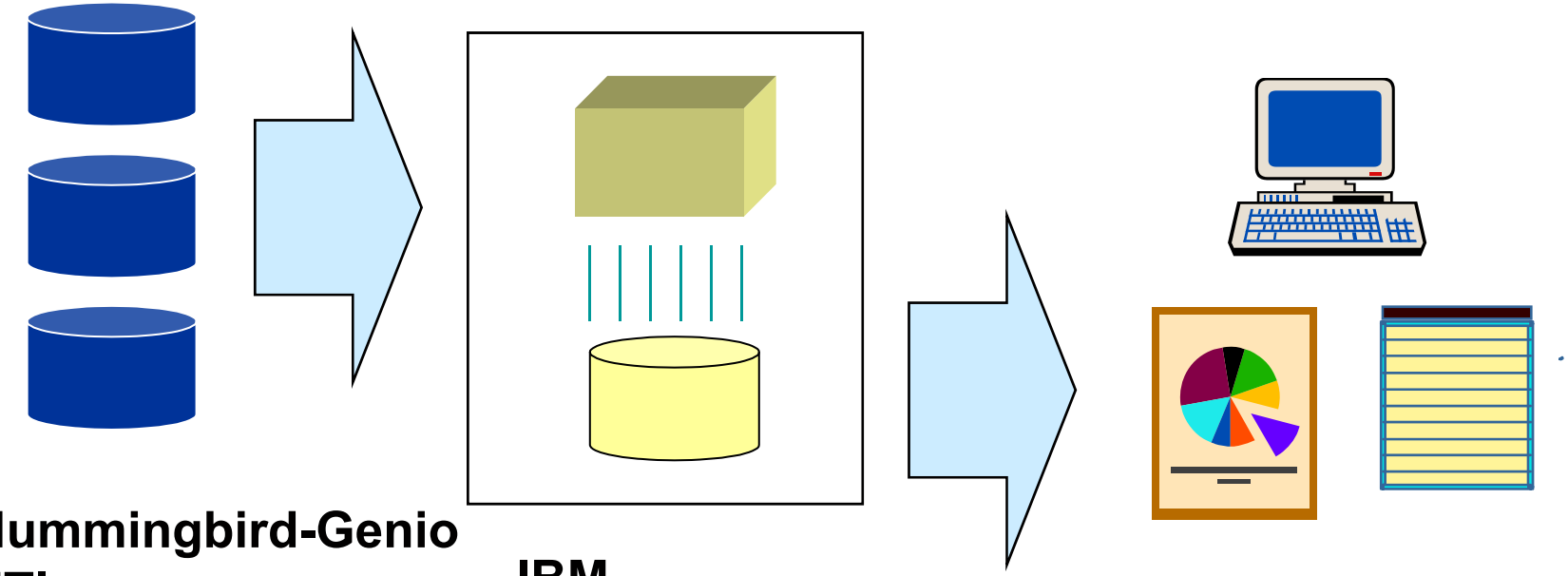
É usado como:

- Um diretório para ajudar o analista de OLAP a localizar o conteúdo do DW
- Um guia para o mapeamento de dados, do ambiente operacional para o ambiente warehouse
- Um guia para os algoritmos usados no processo de agregação e resumos de dados

Requisitos de Metadados para Ferramentas de 3a. Geração

- Geração e Atualização Automática de Metadados
- Visualização de Metadados via Web
- Metadados Técnicos, Operacionais e de Negócio
- Análise de Dependências
- Armazenamento dos Metadados em Repositório Contido em SGBDR Aberto
- Integração de Metadados Técnicos a Metadados Operacionais e a Metadados de Negócio

Ferramentas



Hummingbird-Genio
ETI
Sagent
Informatica
....

IBM
(Informix)
(Red Brick)
Microsoft
Oracle
Sybase
(Tandem)
Teradata
.....

Brio
Business Object
Cognos
MicroStrategy
INF Advantage
.....

Ferramentas de DW

- de Desenvolvimento
  - o design & edit: schemas, views, scripts, rules, queries, reports
- de Planejamento & Análise
  - o Cenários what-if (mudança de schema, períodos de refresh), capacity plan etc.
- de Gerência de DW
  - o monitoração de performance, padrões de uso, relatórios de exceção etc
- de Gerência de Sistema & Network
  - o mede tráfego (fontes => DW => clientes)
- de Gerência de Workflow
  - o Scripts para “limpar” & analisar dados, executar tarefas etc.

Situação do Mercado

- Extração e integração feitas *off-line*
  - em grandes e lentos processos em *batch*

- Tudo vai para o DW
  - Não é seletivo sobre o que deve ir ao DW
  - Benefício de *Query* vs custo de *storage & update*

- Query optimization (dbms) ainda de OLTP
  - => alto *throughput* em vez de rapidez
  - pois processa toda a *query* antes de mostrar alguma coisa...

Check-list de Arquitetura para o DW

- **Arquitetura Informacional "Multi-camada"**
  - Informação consistente para a corporação (DW), para cada departamento e para os usuários/unidades (DMs)
  - Informação necessária, formato e nível de detalhe adequado para os diversos tipos de usuários
  - Estrutura de dados adequada para cada tipo de usuário
  - Performance de acesso otimizada para cada tipo de usuário
- **Arquitetura de ETL (ETLM) em Camadas**
  - Minimizando o impacto nos sistemas "*legacy*" - performance otimizada
  - Assegurando qualidade dos dados dentro do DW
  - Coordenando a captura de metadados
  - Minimizando o esforço de desenvolvimento
  - Baixo impacto, manutenção simplificada - fácil adaptação a mudanças

O Balanço Adequado dos Ingredientes

Ferramentas de Software

Extração/Transformação/Carga
Qualidade/Limpeza de Dados
Gerenciamento de Metadados
Scheduling e Transporte
Acesso OLAP / Data Mining
Monitoração e Adm.
. . .

Consultoria e Serviços

Especialistas - Negócio
Especialistas - Ferram.
Especialistas – Plataform.
Arquitetos/Modeladores
Gerentes de Projeto
Adm de dados/metadados
. . .

Infra-estrutura de Hardware e Rede
Metodologia
Best-practices
Arquitetura
Modelos Genéricos
...

Obrigado!

...e agora suas perguntas?

ricardo.avila@outlook.com.br

@theavila