

INE 5643

Data Warehouse

Aula 2 - Evolução dos Sistemas de Apoio a Decisão

Prof. Mateus Grellert

Prof. Renato Fileto

Créditos: Prof. Tite Todesco (slides originais, adaptados por Mateus e Fileto)

Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)

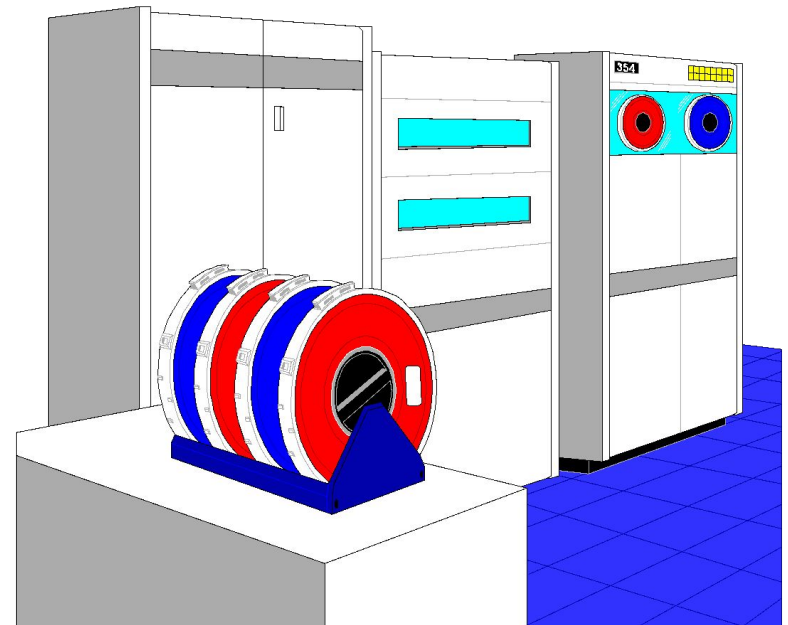
TÓPICOS

1. Evolução e Histórico

2. Conceitos Básicos

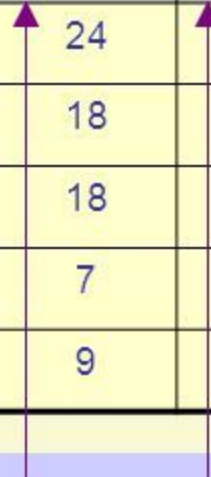
EVOLUÇÃO E HISTÓRICO

1960 - uso de arquivos mestres e relatórios. A computação consistia na criação de **aplicações individuais** que eram executadas sobre **arquivos mestres**. As aplicações eram caracterizadas por relatórios e programas, geralmente em COBOL.



Master File Example

- A shop selling clothes may have a master file like this...(only 5 records are displayed)



Code	Description	Price	Supplier Code	Reorder Level	Stock Level	Sold this Year
T101	T-Shirt(Red)	£4.99	S156	20	24	57
T102	T-Shirt(Blue)	£4.99	S156	20	18	64
T256	Shorts	£6.99	S315	10	18	9
T259	Skirt (Black)	£12.99	S156	5	7	19
T262	Skirt (Blue)	£12.99	S156	5	9	8

Which fields may need updating regularly?

The shop needs to order some more of which item?

EVOLUÇÃO E HISTÓRICO

1965 - O crescimento dos arquivos mestres e das fitas magnéticas explodiu, havia arquivos mestres por toda parte, surgindo assim enormes quantidades de dados redundantes. Alguns problemas surgiram:



- A necessidade de **sincronizar** dados a serem atualizados.
- A complexidade de **manutenção** de programas.
- A **complexidade** de desenvolvimento de novos programas.
- A quantidade de **hardware** necessária para manter todos os arquivos mestres.

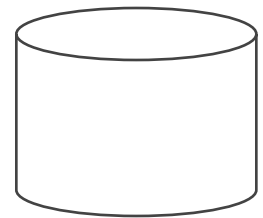
EVOLUÇÃO E HISTÓRICO

1970 - A década de 70 presenciou o advento do armazenamento em disco, ou **DASD** (direct access storage device). O acesso aos dados se tornou muito mais rápido, não sendo mais o seqüencial.



Começou a usar-se o termo **Banco de Dados**

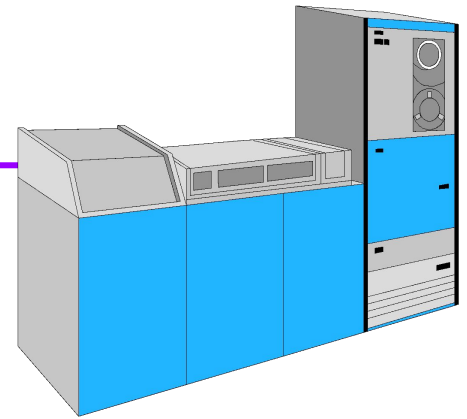
A partir do DASD surgiu um novo tipo de software conhecido como **SGBD**, que tinha como objetivo tornar o armazenamento e o acesso a dados no DASD mais fáceis para o programador.



DASD
SGBD

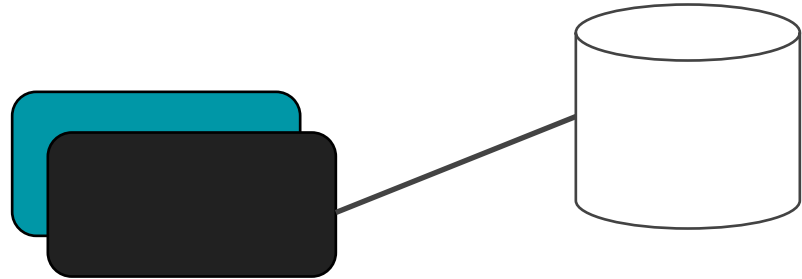
EVOLUÇÃO E HISTÓRICO

Banco de Dados



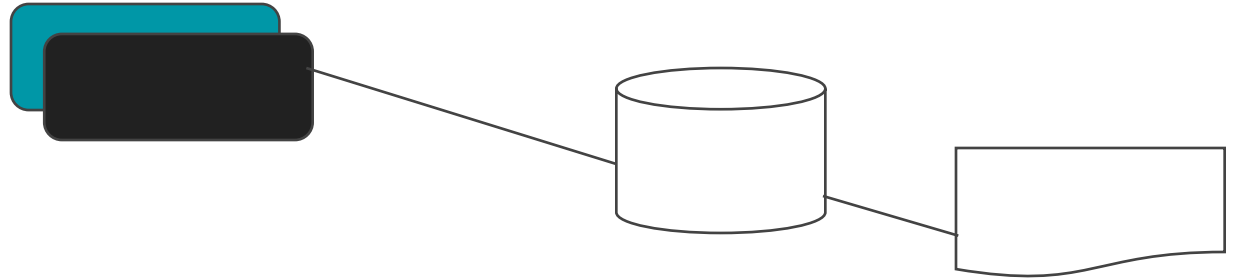
- Na década de 70, os bancos de dados eram fundamentalmente para o processamento **operacional** - geralmente transacional.
- Nos últimos anos, além do processamento operacional também vem sendo usado para atender necessidades informacionais ou **analíticas**.

EVOLUÇÃO E HISTÓRICO



1975 - Em meados de 70, o processamento de transações online começou a ser feito sobre banco de dados. Com um terminal e o software apropriado, os técnicos descobriram que um acesso mais rápido aos dados era possível. Sistemas de reservas online, sistemas de caixas bancários, sistemas de controle de produção e outros similares puderam ser construídos e usados.

EVOLUÇÃO E HISTÓRICO

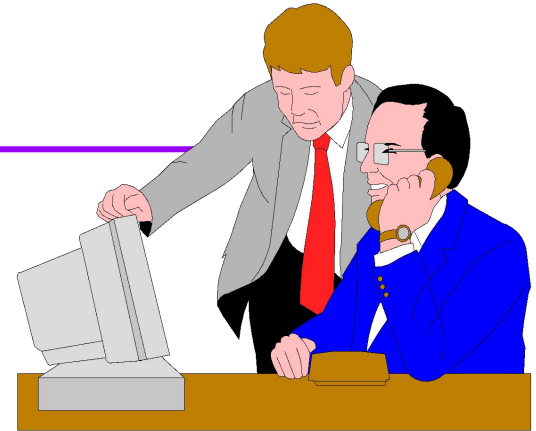


1980 - No início da década de 80, novas tecnologias, como os PCs e as linguagens de 4ª geração começaram a aparecer.

O usuário final passou a assumir um papel que anteriormente não era possível, controlando diretamente os sistemas e os dados, fora do domínio do processamento de dados.

Assim, surgiram os Sistemas de Informação Gerencial (SIG ou Management Information Systems - MIS) e, **mais tarde os Sistemas de Apoio à Decisão, SADs.**

EVOLUÇÃO E HISTÓRICO



O processamento informacional ou analítico é aquele que atende às necessidades dos gerentes durante o processo de tomada de decisões, conhecido como **SAD (Sistema de Apoio à Decisão)**.

EVOLUÇÃO E HISTÓRICO

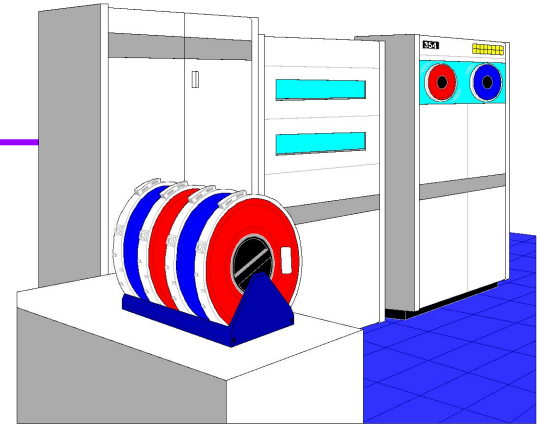


O processamento **analítico** examina amplos conjuntos de dados para detectar **tendências**, em vez de considerar um ou dois registros de dados (como ocorre no processamento operacional).

EVOLUÇÃO E HISTÓRICO

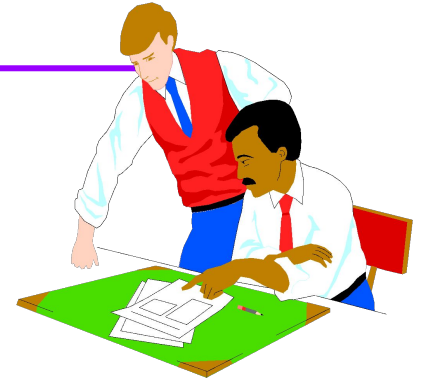
A divisão de banco de dados operacionais e informacionais ocorre por várias razões:

1. Os dados que atendem a necessidades operacionais são **fisicamente diferentes** dos dados que atendem a necessidades informacionais ou analíticas;



EVOLUÇÃO E HISTÓRICO

2. A **tecnologia** de suporte ao processamento operacional é fundamentalmente **diferente** da tecnologia utilizada para prestar suporte a necessidades informacionais ou analíticas.



3. A **comunidade de usuários** dos dados operacionais é **diferente** da que é atendida pelos dados informacionais ou analíticos. As características de processamento do ambiente operacional e do ambiente informacional são, também diferentes.

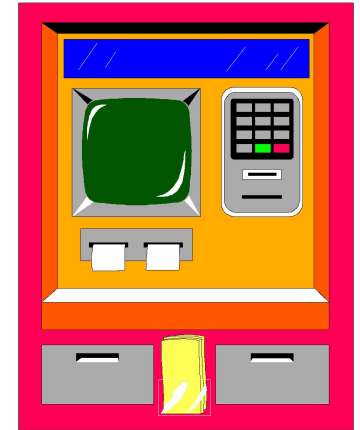


EVOLUÇÃO E HISTÓRICO

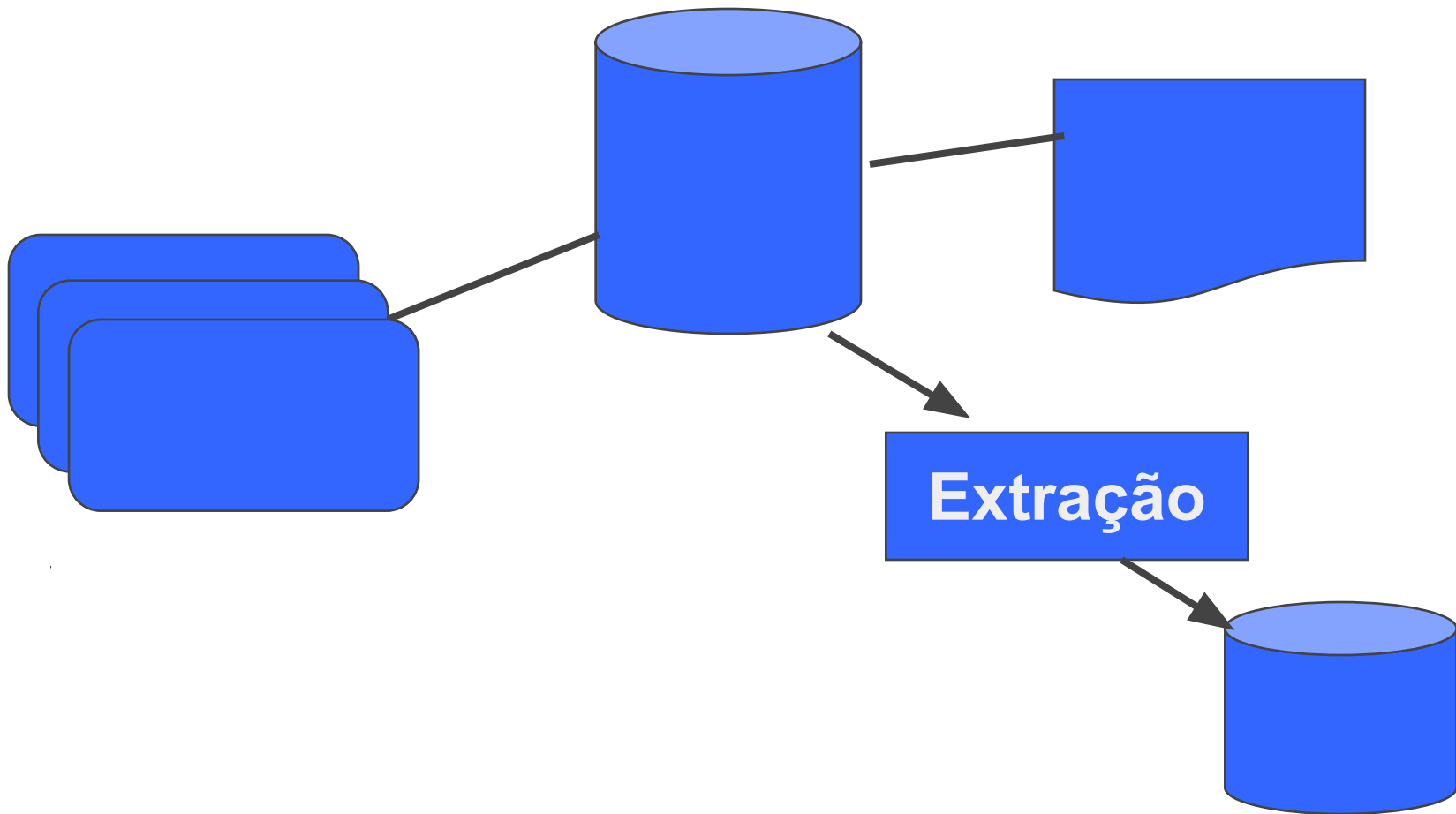
Após o advento das transações online de alta performance em massa, surge um programa chamado de processamento de **extração**.

Este programa varre um arquivo ou banco de dados, usa alguns critérios de seleção, e, ao encontrar dados que atendam aos critérios, transporta os dados para outro arquivo ou banco de dados.

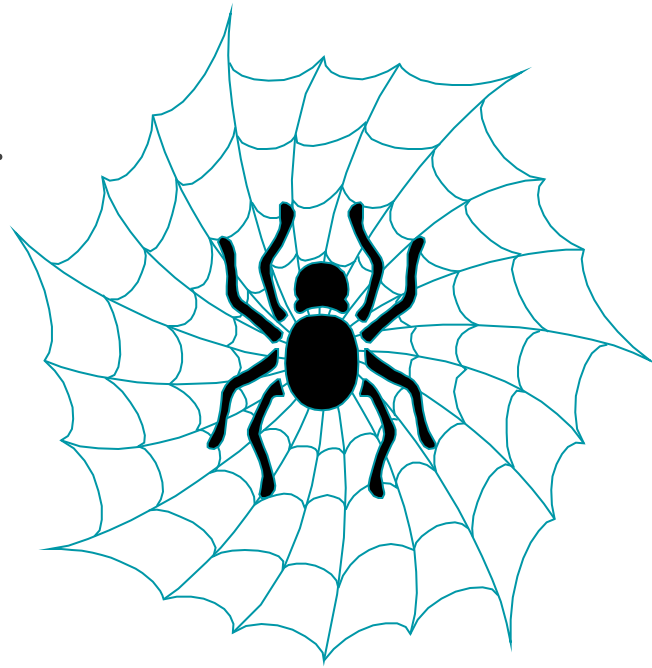
Estes programas foram muito utilizados, sendo que até a década de 90 havia muitos programas.



PROGRAMA DE EXTRAÇÃO



Uma “**teia de aranha**” começou a se formar. Primeiro, havia extrações, depois extrações das extrações, e assim por diante (formando uma arquitetura de desenvolvimento espontâneo ou sistemas herdados).



Alguns problemas ocorreram como:

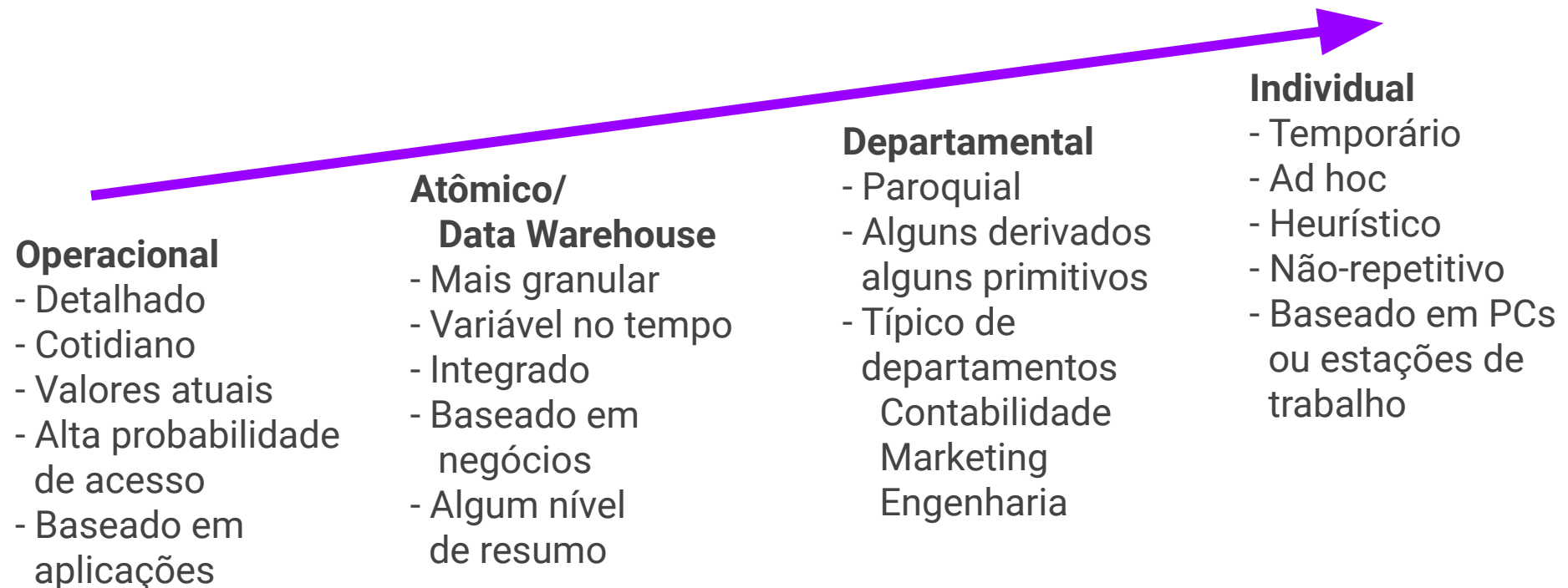
- **credibilidade** dos dados,
- **produtividade** e
- Impossibilidade de **transformar** dados em informações.

Uma Mudança de Enfoque

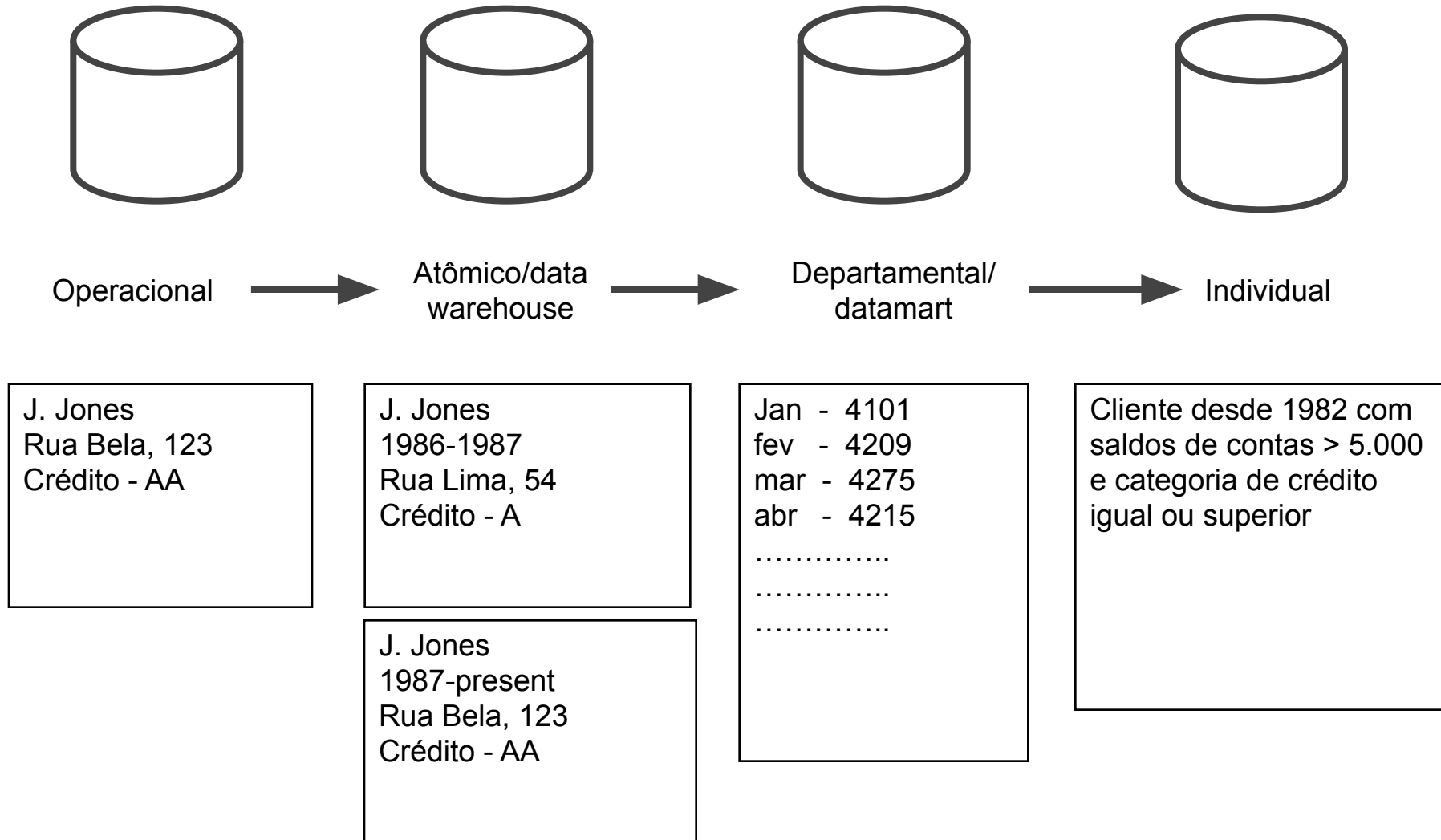
- O *status quo* da arquitetura de desenvolvimento espontâneo, no qual se encontra, atualmente, a maioria das organizações, simplesmente não bastava para atender às necessidades do futuro.
- O que se faz necessário é uma mudança de arquitetura, que faça surgir um ambiente projetado de data warehouse.
- No cerne desse ambiente "projetado" está a percepção de que há fundamentalmente duas espécies de dados:
 - dados primitivos e
 - dados derivados.

O Ambiente Projetado

As extensões naturalmente resultantes da separação dos dados, causada pela diferença entre dados primitivos e dados derivados, são apresentados abaixo segundo os níveis de arquitetura:



Um exemplo elementar - um cliente



TÓPICOS

1. Evolução e Histórico

2. Conceitos Básicos

Taxonomia de Anthony

Categorias de atividades gerenciais (1965)

- Tipos diferentes de planejamento e controle
- Apoio de SI com características distintas



Características da Informação

Estratégico

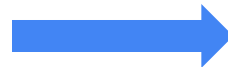
- quais produtos, mercados ?
- faturamento, crescimento

Tático

- vigiar vendas
- custo de estoque
- promoções

Operacional

- registra venda
- atualiza estoque
- contata fornecedor



- Menos precisa, agregada
- qualitativa
- fontes internas, externas, abrangência geral e indefinida
- uso infrequente, ad hoc
- dados históricos, futuro

- Precisa, detalhada
- quantitativa
- fontes internas, abrangências específicas
- uso freqüente, pré-definido
- dados instantâneos, presente

Sistemas de Informação

- **OLTP**: automatizar os processos, melhorar o desempenho e confiabilidade
- **SAD**: sistemas que ajudam decisores a tomar decisões em situações onde o julgamento humano é uma contribuição importante ao processo de resolução, mas existe uma limitação humana para processar informações

**Sistemas de Apoio a
Decisão (SAD)**

**Sistemas
Transacionais
(OLTP)**

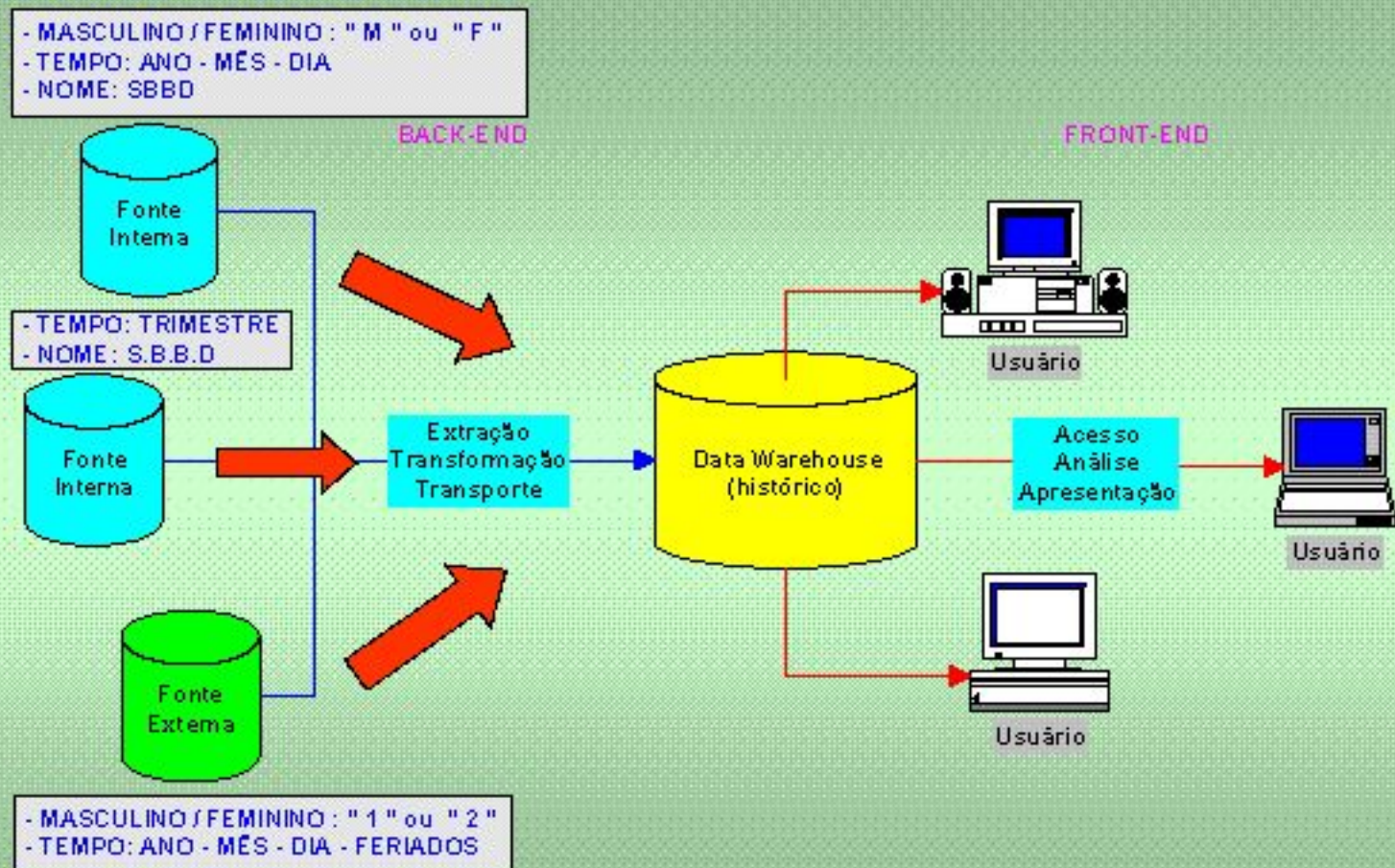


Sistema de Apoio à Decisão - SAD

- Um assistente para quem o decisor delega atividades envolvendo recuperação, computação e divulgação de informações (Keen, 1981)
 - recuperação ad hoc (filtros, agregações, resumos, etc);
 - apresentação de informação (relatórios, mapas, gráficos, animações, visualização, etc);
 - manipulação de modelos (estatísticos, matemáticos, de simulação, econométricos, IA, etc);
 - outros tipos de apoio (escolha, estruturação do processo, comunicação, negociação, etc).
- **Data Warehouse é um tipo de SAD**

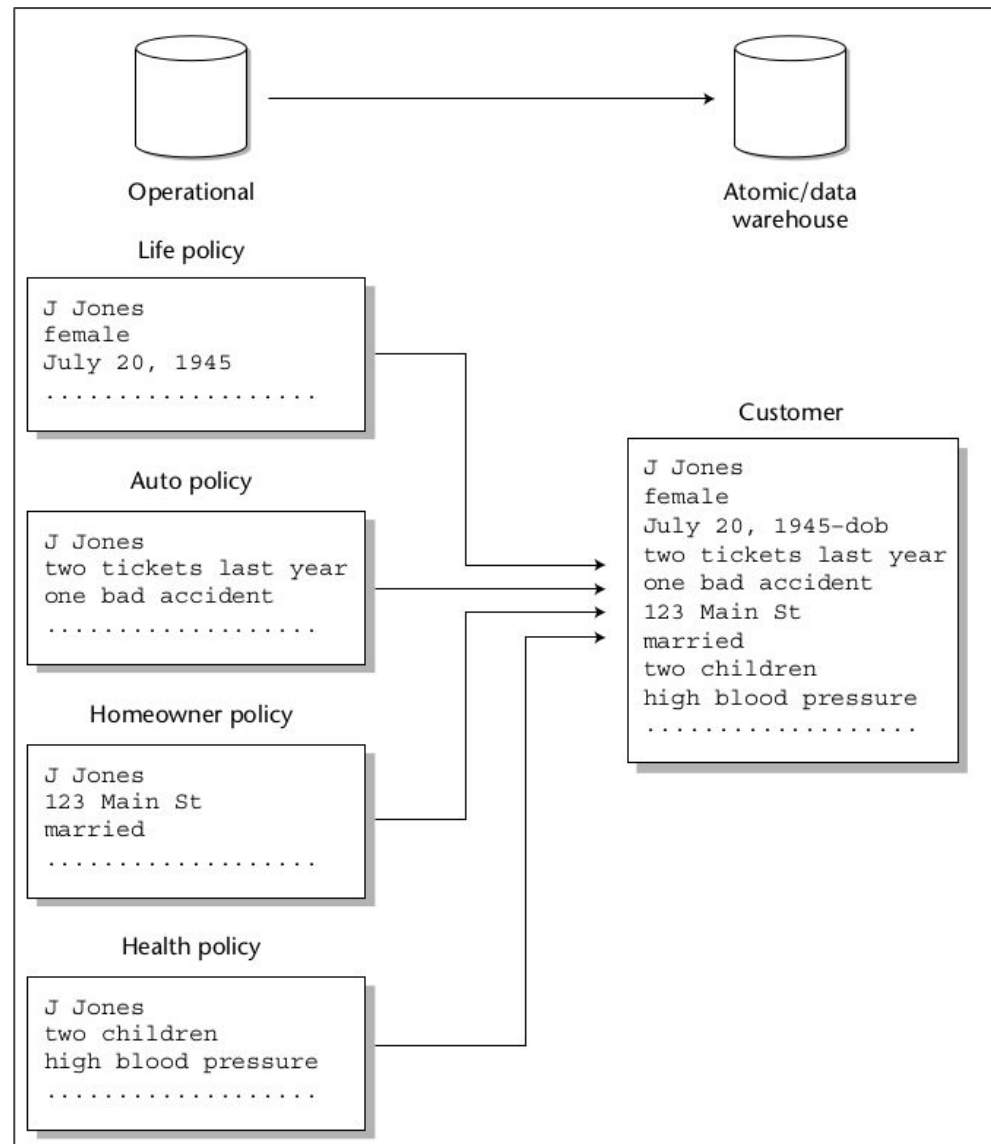
Integrado

Os dados fonte de sistemas OLTP são modificados e convertidos para um **estado uniforme** de modo a permitir a carga no DW.



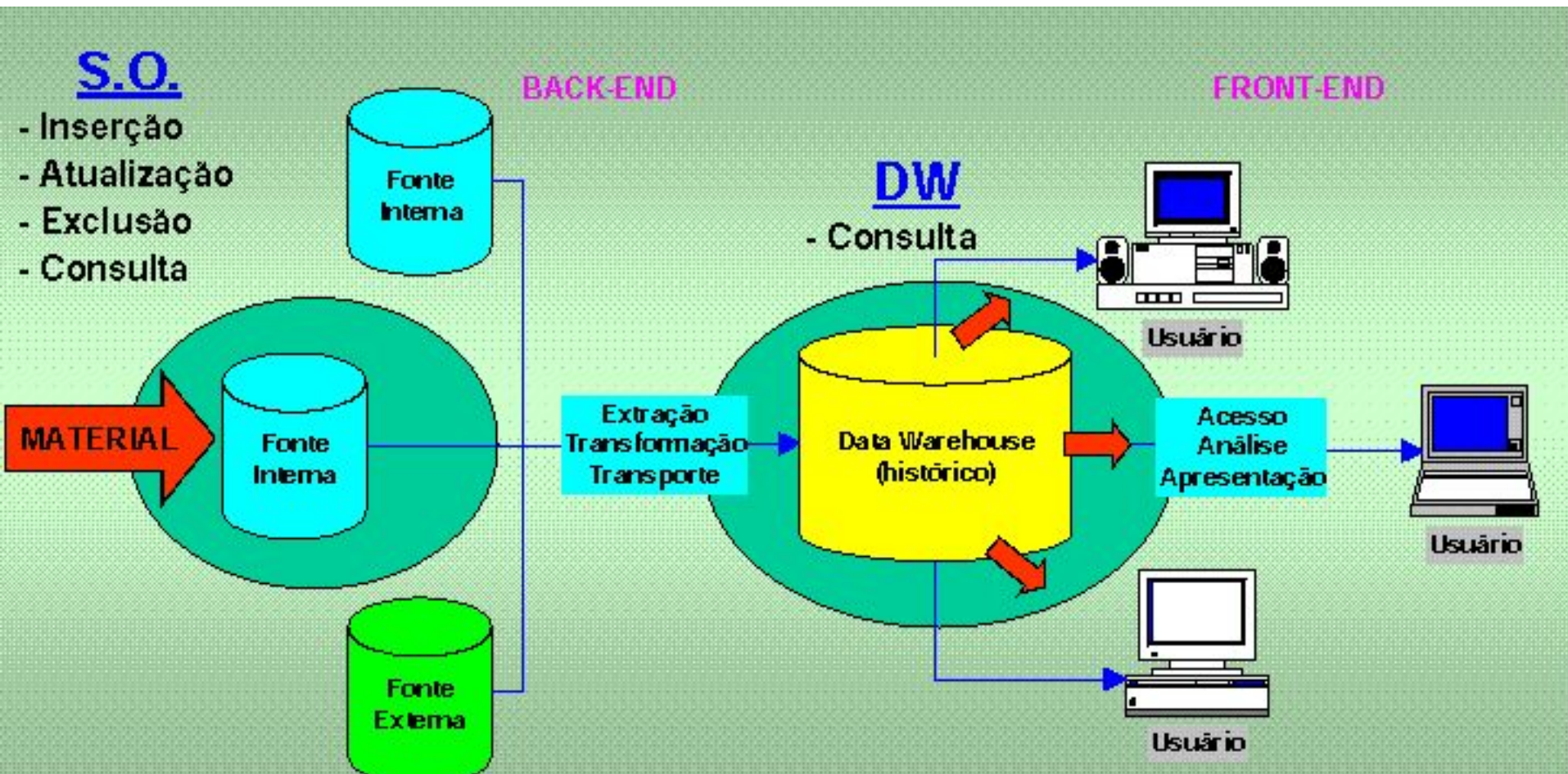
Integrado

Exemplo simples de integração



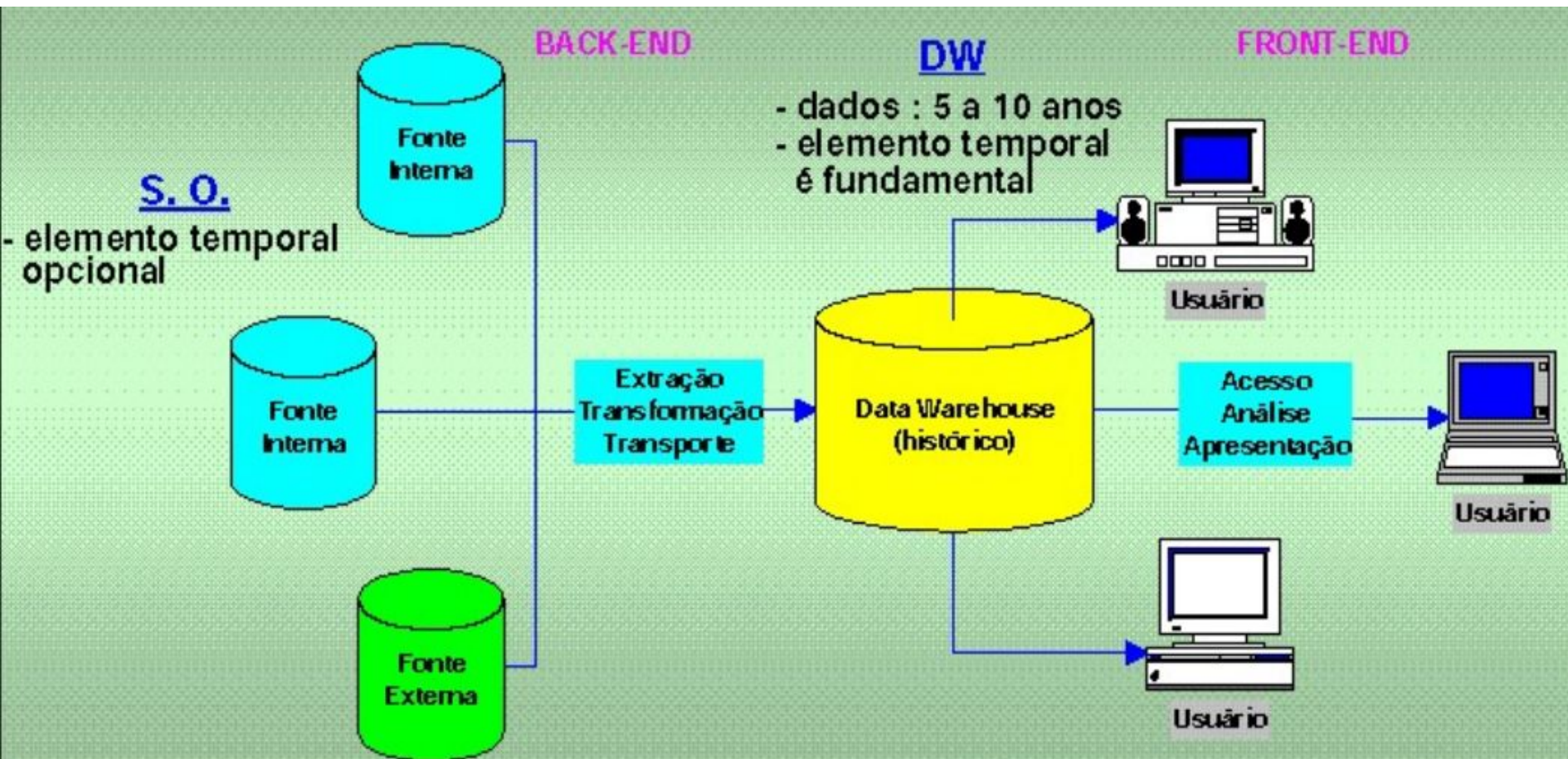
Não Volátil

Os dados após serem extraídos, transformados e transportados para o DW estão disponíveis aos usuários **somente para consulta.**



Variável no Tempo

Os DW armazenam dados por um período de tempo de 5 a 10 anos. O elemento **tempo** é fundamental.



Definição

da·ta ware·house:

(subs. masc.)

um repositório de dados integrado, não volátil, variável em relação ao tempo usado como apoio a tomada de decisão.

E nos dias de hoje?

- A criação de DWs democratizou o acesso a dados importantes dos processos de negócio nas empresas e permitiu a criação de ferramentas poderosas de **BI** como Tableau, Talend e Pentaho
- Mas **Cientistas de Dados** normalmente querem derivar conclusões com o auxílio de modelagem preditiva utilizando estatística avançada e Machine Learning

E nos dias de hoje?

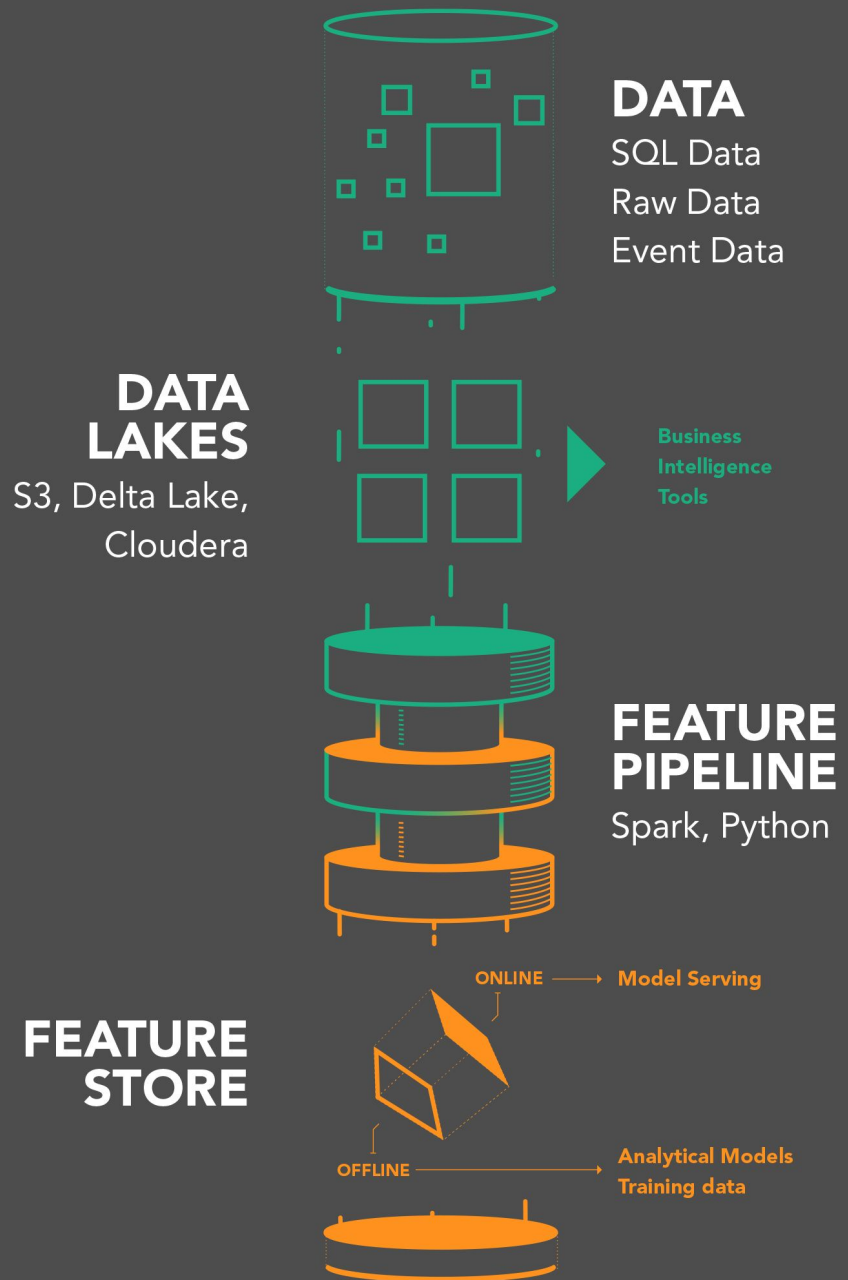
- Com a explosão de aplicações baseadas em Data Mining e Machine Learning, os valores **derivados e agregados** do DW passam a ser menos interessantes para as empresas
- No entanto, ainda deseja-se manter algumas das demais características, como integração, uso de metadados, etc
- Além disso, o acesso cada vez maior a dados heterogêneos e em grande quantidade pode trazer gargalos no processo de ETL
- Com isso surgem conceitos como **Data Lake, ELT over ETL e Feature Store**

Feature Stores

Repositório central para armazenar dados (features) tratados, documentados e de acesso controlado, os quais podem ser utilizados por diversos modelos de ML.

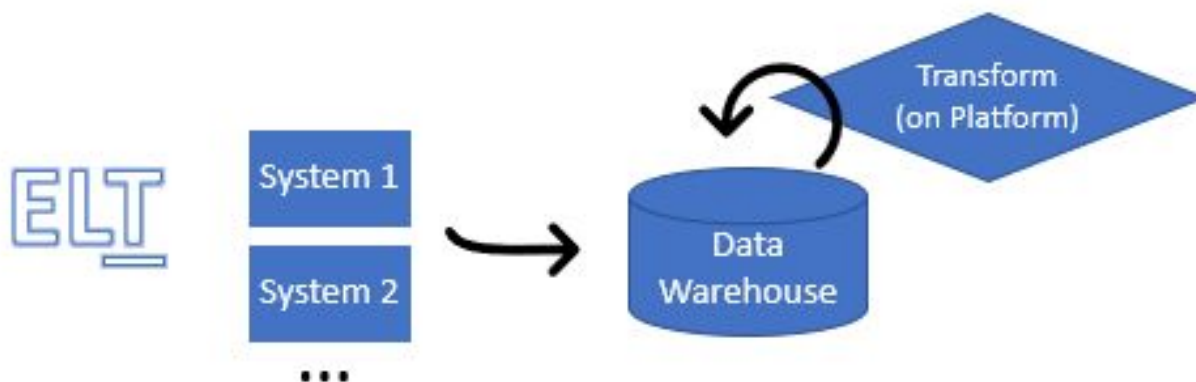
A Feature Store recebe dados de diversas fontes, aplicando operações de transformação, validação e até mesmo agregação quando necessário.

Fonte: <https://www.kdnuggets.com/2020/12/feature-store-vs-data-warehouse.html>



ETL vs. ELT

- Nem sempre temos tempo/conhecimento necessário para transformar os dados, mas também não queremos que eles se percam.
- As principais vantagens de arquiteturas baseadas em ELT são a **carga rápida** e a **flexibilidade**. A desvantagem é o **tempo extra** nas consultas das aplicações BI.



Bibliografia

- [BON98] Bontempo, Charles & Zagelow, George. The IBM - Data Warehouse Architecture. Communications of the ACM, 41 (9): 38-48. Sept. 1998.
- [DEV97] Devlin, Barry. Data warehouse: from architecture to implementation. Addison Wesley Longman, 1997.
- [FAY96] Fayyad, Usama; Piatetsky-Shapiro, Gregory & Pandhraic, Smyth. >From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge and Data Mining. Califórnia, AAAI Press, 1996.
- [GAR98] Gardner, Stephen R. Building the Data Warehouse. Communications of the ACM, 41 (9): 52-60. Sept. 1998.
- [GRA98] Gray, Paul & Watson, Hugh J. Decision Support in the Data Warehouse. New Jersey, Prentice Hall PTR, 1998.
- [INM97] Inmon, William H. Como construir o data warehouse. Rio de Janeiro, Editora Campus, 1997.
- [KEE78] Keen, Peter G. W. & Morton, Michael S. Scott. Decision Support Systems: on organizational perspective. Addison-Wesley Publishing Company, 1978.
- [KIM98] Kimball, Ralph; Reeves, Laura; Ross, Margy & Thornthwaite, Warren. The Data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. New York, John Wiley & Sons, 1998.
- [MIC98a] Microsoft Corporation. Microsoft SQL Server 7.0 OLAP Services. 1998.
<http://www.microsoft.com/sql/70/gen/whatsnew.htm>.
- [MIC98b] Microsoft Corporation. Microsoft SQL Server 7.0 Data Warehousing Framework. 1998.
<http://www.microsoft.com/sql/70/gen/whatsnew.htm>.
- [POE98] Poe, Vidette; Klauer, Patricia & Brobst, Stephen. Building a data warehouse for decision support. New Jersey, Prentice Hall PTR, 1998.
- [SEN98] Sen, Aru & Jacob, Varghese S. Industrial - Strenght Data Warehousing. Communications of the ACM, 41 (9): 29-31. Sept. 1998.