

Procedência dos Dados

- Conjunto de metadados para identificar
 - As fontes
 - Os processos de transformação

Desde a criação até o estado atual dos dados

Desafios de um Modelo de Procedência



Quais dados armazenar?



Como coletar?



Como armazenar?



Como consultar?

Quatro aspectos

Quais dados de procedência armazenar?

- Definição dos dados de procedência que são necessários para uma determinada aplicação





Quais dados de procedência armazenar?

- Os dados de procedência recebem diferentes classificações na literatura
 - *Source e Transformation Provenance*
 - *Why e Where Provenance*
 - *Provenance e Process Meta-Information*
 - *Perspective e Retrospective Provenance*



Quais dados de procedência armazenar?

- Procedência “BD Integrado”
 - Fonte e transformação

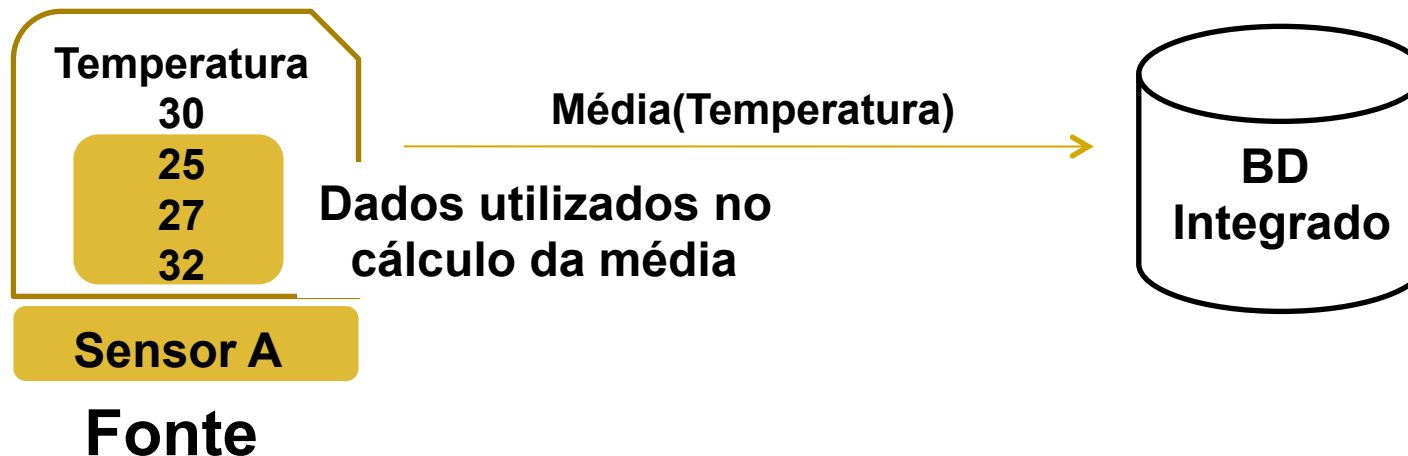


Source e transformation provenance



Quais dados de procedência armazenar?

- ▶ Procedência “BD Integrado”
 - Detalhamento da fonte

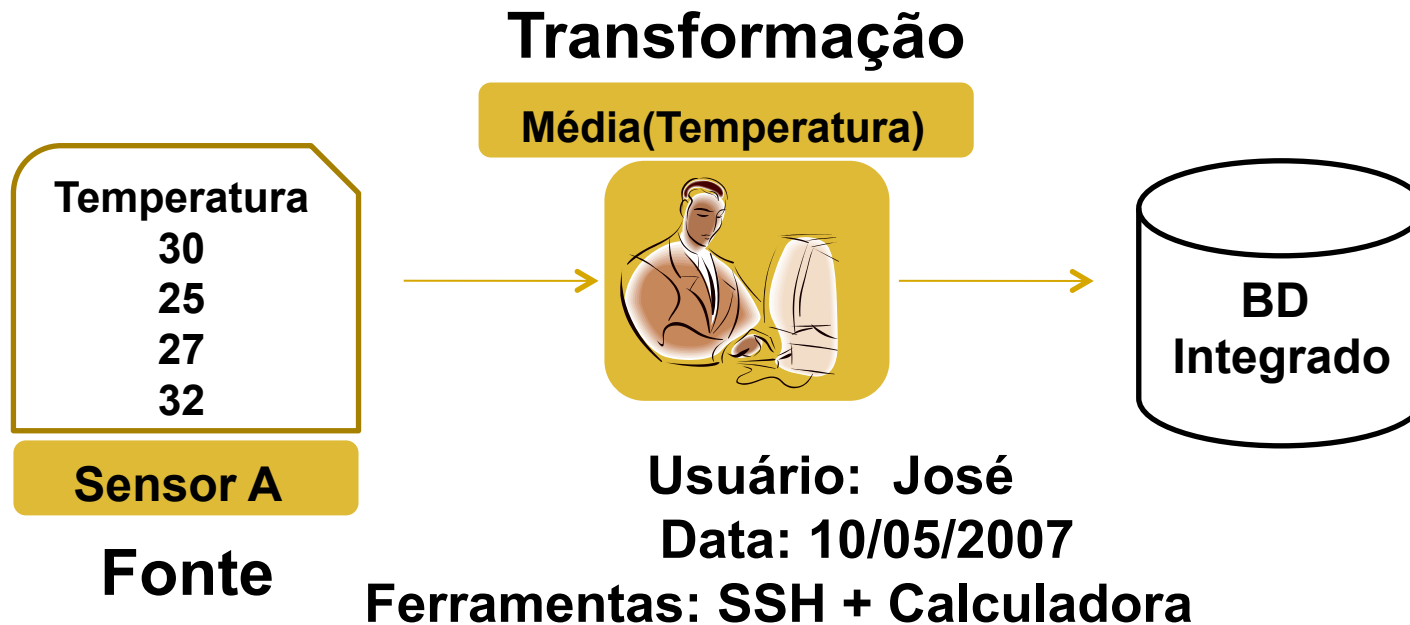


Why e Where Provenance



Quais dados de procedência armazenar?

- Procedência “BD Integrado”
 - Fonte + Transformação + Informações do ambiente



*Process e Provenance Meta-Information
Prospective e Retrospective Provenance*



Granularidade dos dados

- Identifica o nível de detalhe dos dados

- Quanto menor a granularidade
 - Maior o custo de coleta e armazenamento
 - Grande variedade de consultas podem ser respondidas

- Quanto maior a granularidade
 - Menor o custo de coleta e armazenamento
 - Pouca variedade de consultas podem ser respondidas



Granularidade dos dados

- Ponderar o custo-benefício de armazenar um dado
 - Identificar os dados que devem ter a procedência armazenada
 - Identificar os dados de procedência que devem ser armazenados



Granularidade dos dados

- Ponderar o custo-benefício de armazenar um dado
 - ❑ Identificar os dados que devem ter a procedência armazenada
 - ❑ Identificar os dados de procedência que devem ser armazenados

A granularidade depende dos objetivos para o qual a procedência está sendo armazenada

Como coletar os dados de procedência?

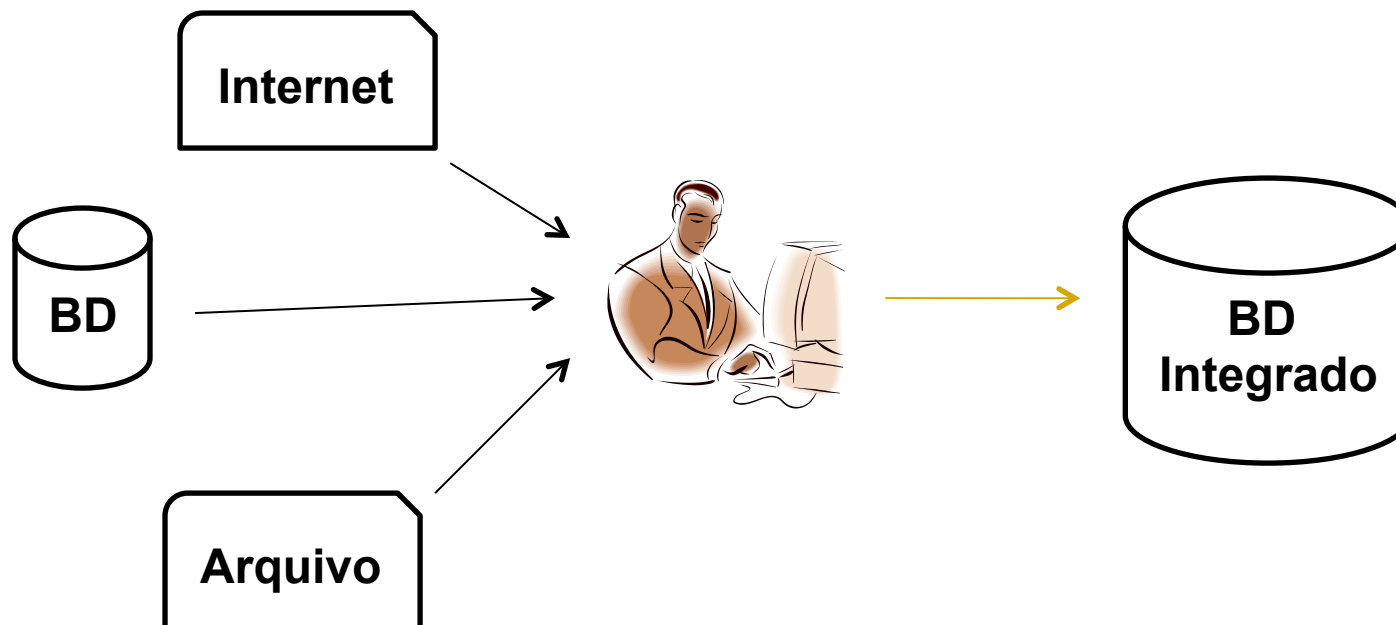
- Como
 - Manual x Automática
- Quando
 - *Lazy* x *Eager*





Manual

- Requer mais de tempo do usuário
 - ❑ Ferramentas sem suporte à procedência
 - ❑ Bancos de dados acurados manualmente

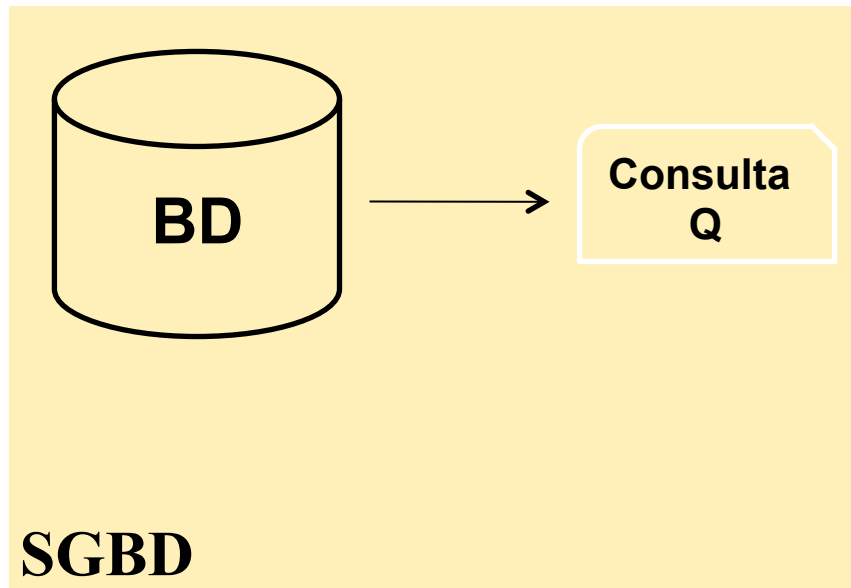




Automática

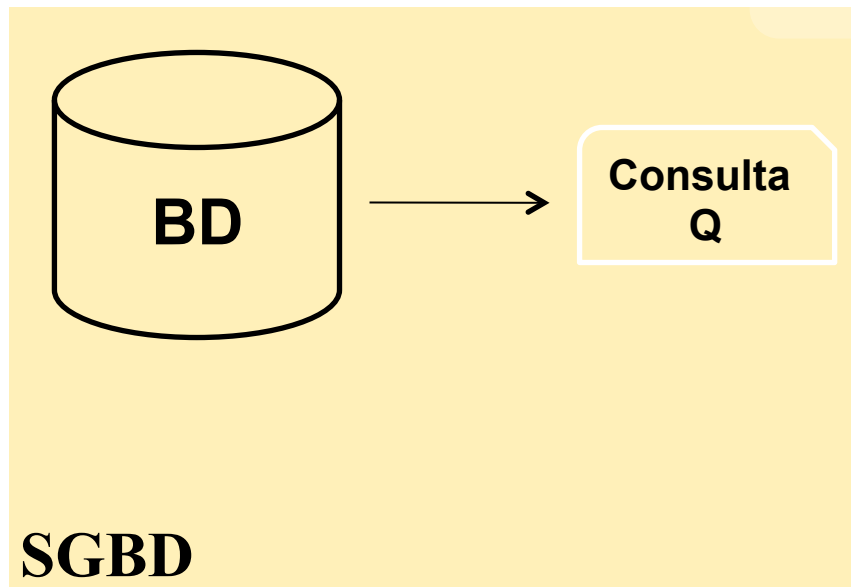
- Coleta é transparente para usuário
 - Sistema de arquivos
 - Procedência para criar, mover, remover, alterar arquivos
 - SGBD
 - Procedência das tuplas de um tabela
 - Aplicação
 - Procedência das transformações
 - Serviço
 - Fornece serviço de coleta de procedência às aplicações

- Procedência é “calculada” apenas quando requisitada



- Procedência é “calculada” apenas quando requisitada

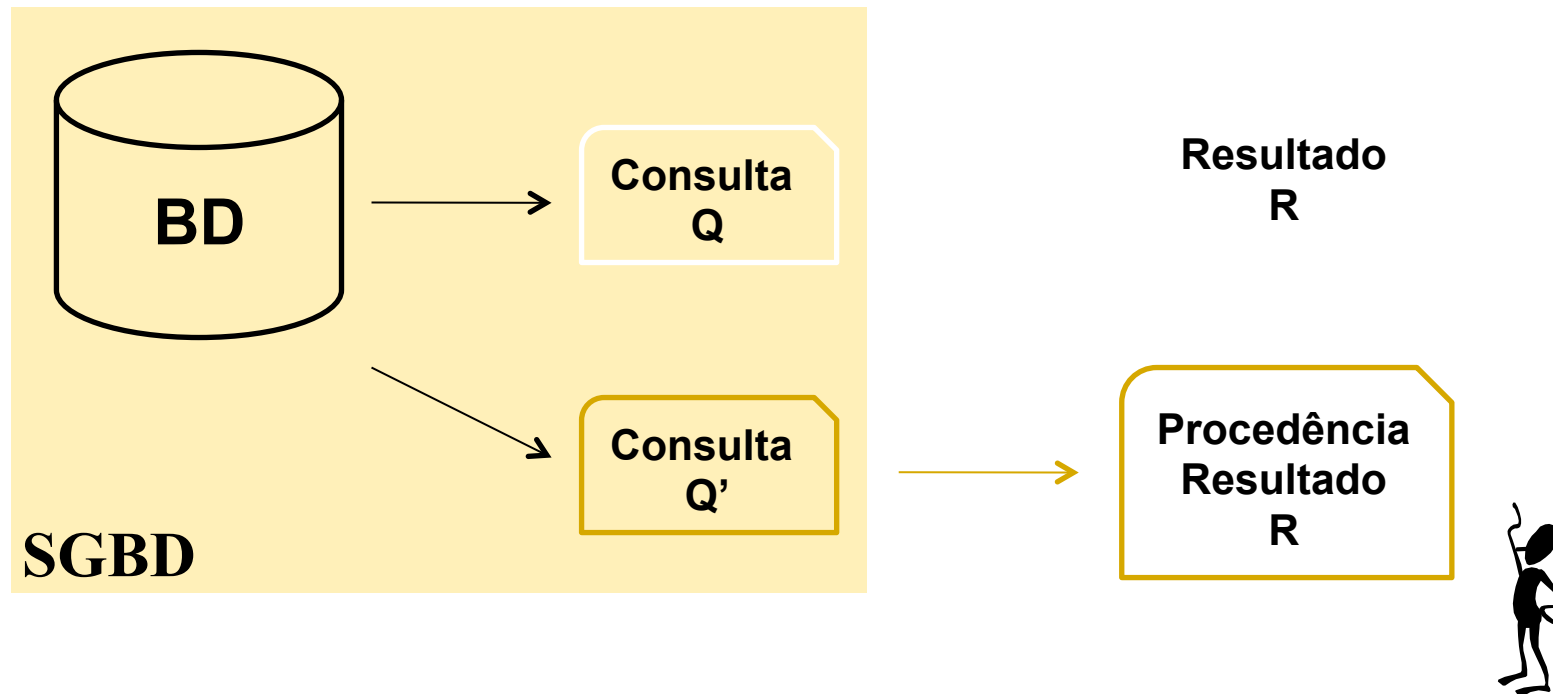
De onde veio esse resultado?



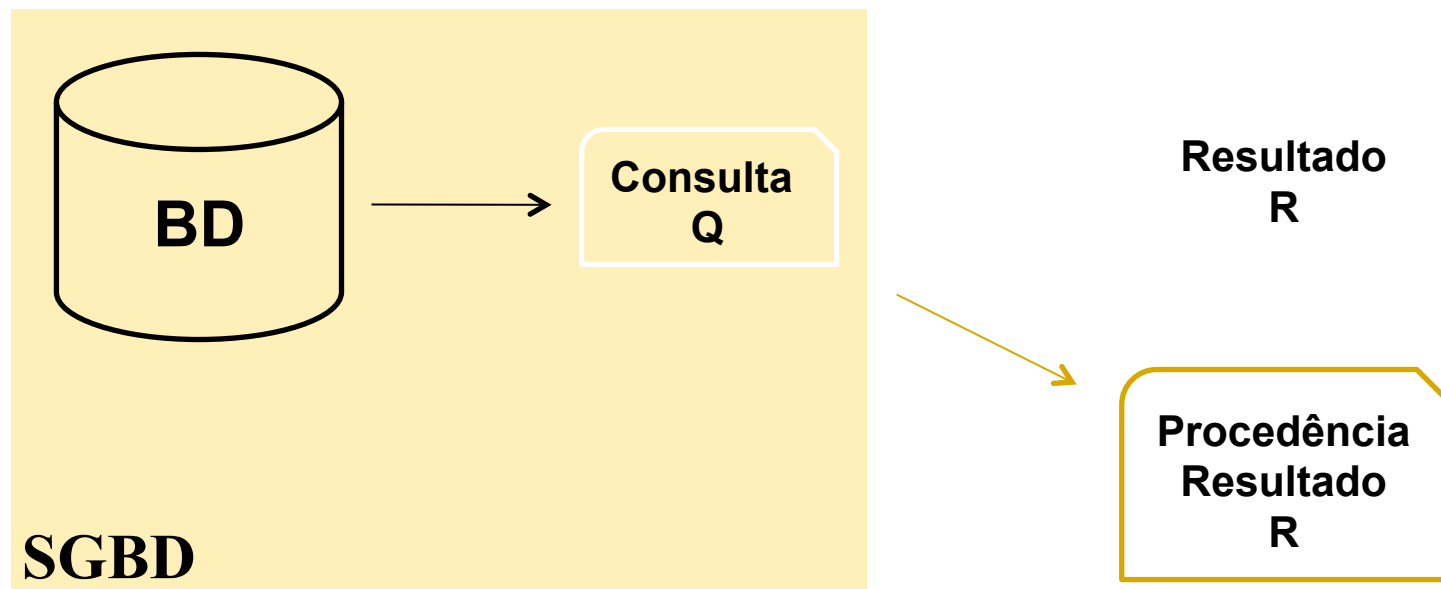
Resultado
R



- Procedência é coletada apenas quando requisitada



- Procedência é coletada conforme os dados são gerados



Como armazenar os dados coletados?

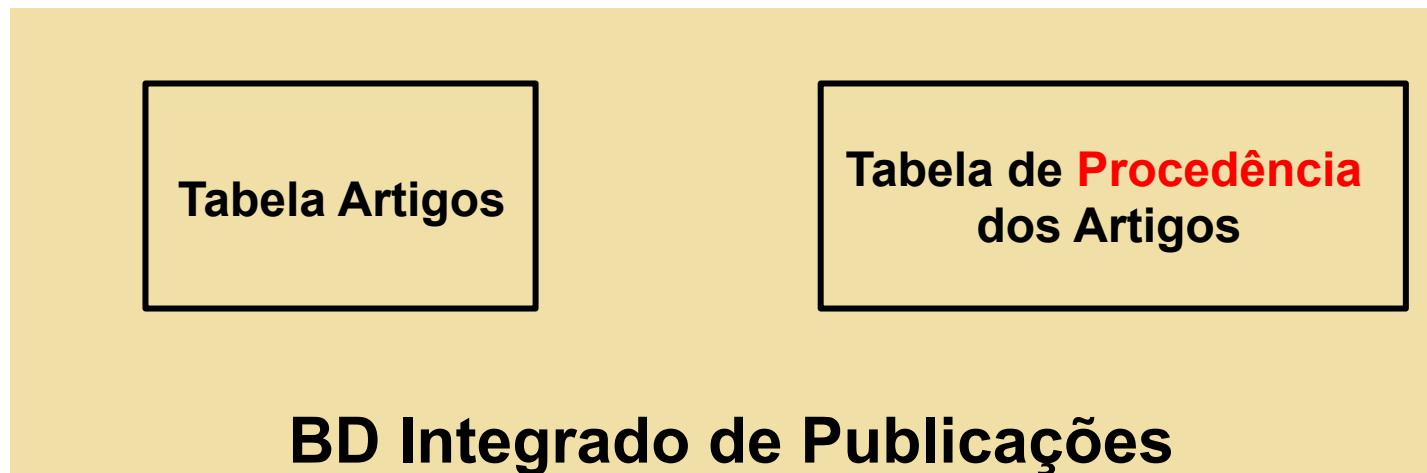
- Ligação entre dado e procedência





Ligação entre dado e procedência

- Procedência pode estar armazenada
 - Junto com o dado
 - Facilita a ligação entre o dado e sua procedência





Ligação entre dado e procedência

- Procedência pode estar armazenada
 - Separada do dado
 - Dificulta a ligação entre o dado e sua procedência

Tabela Artigos

BD Integrado de Publicações

Tabela de **Procedência**
dos Artigos

BD de Procedência

Como consultar os dados armazenados?

- Dois principais tipos de consulta:
 - Tipo rastreamento
 - Tipo filtro





Tipos de consulta

■ Tipo rastreamento

- ❑ Consultar os dados e verificar a procedência dos mesmos
- ❑ “Como esse relatório foi gerado?”

■ Tipo filtro

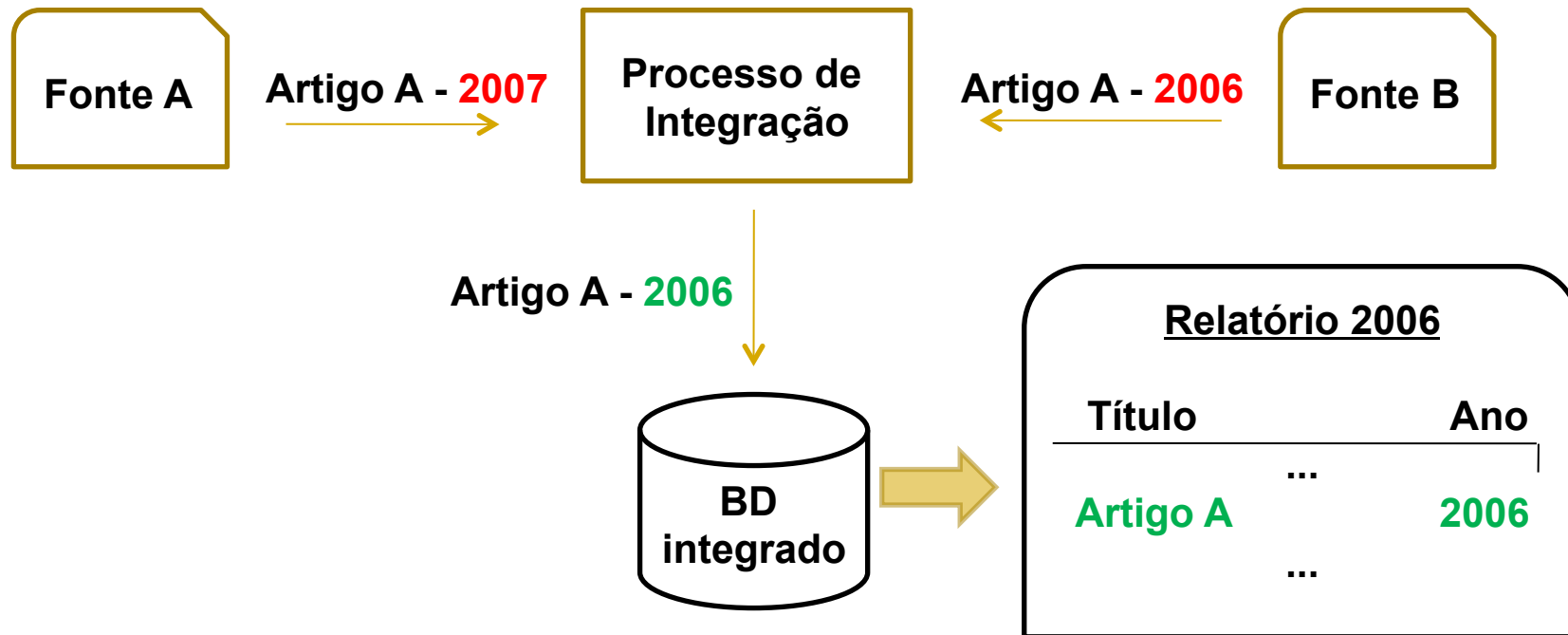
- ❑ Consultar os dados filtrando por um determinado critério de procedência
- ❑ “Gerar um relatório apenas com dados advindos do Lattes”

Motivações para Procedência dos Dados

- Verificar **histórico** dos dados
- Assegurar a **qualidade** dos dados
- Realizar processos de **auditoria** e **autoria** dos dados
- **Reenviar** dados para as fontes
- **Reproduzir decisões** de integração dos dados

Motivações para Procedência dos Dados

Verificar **histórico** dos dados
Ex: Identificar as **fontes** e as **versões** de um dado

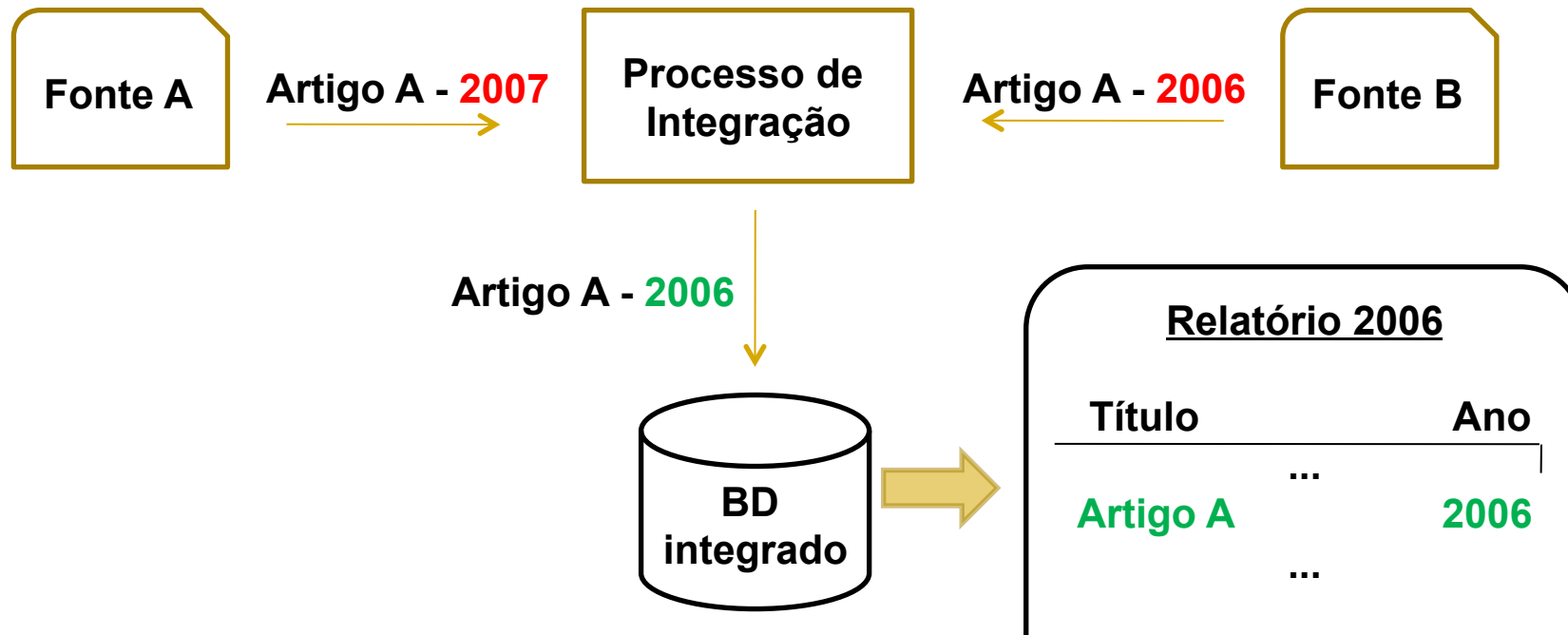


De onde veio esse artigo?



Motivações para Procedência dos Dados

Assegurar a **qualidade** dos dados
Ex: Fontes confiáveis x Fontes não-confiáveis

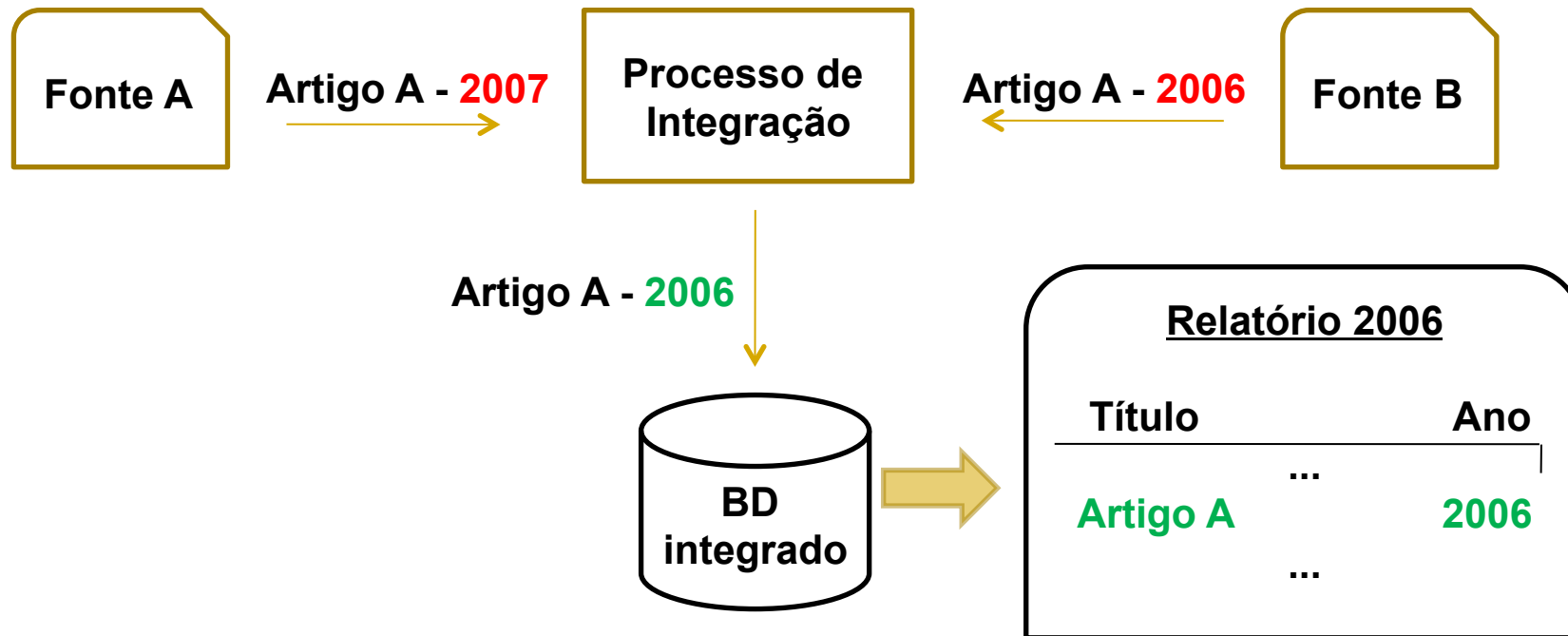


Esse relatório é confiável?



Motivações para Procedência dos Dados

Processos de auditoria
Ex: Verificar o processo de derivação de um dado

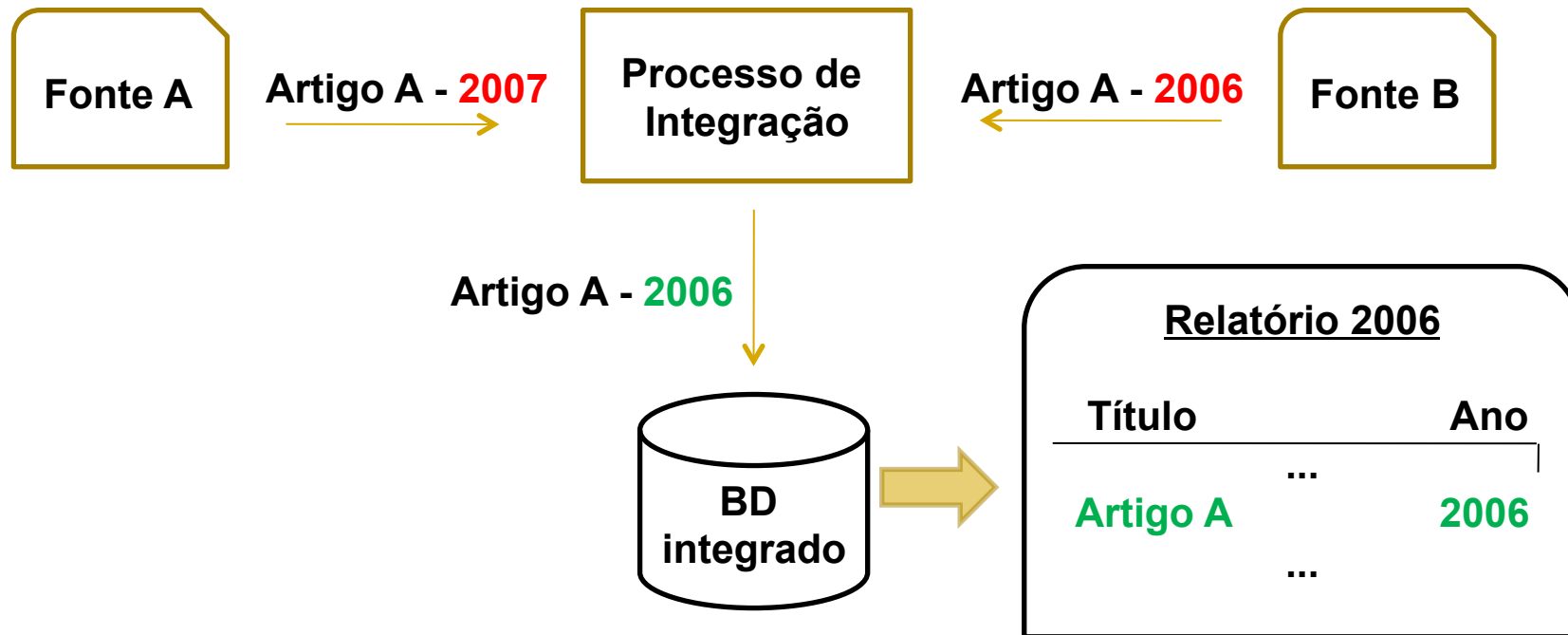


**Por que a Fonte B foi escolhida?
Quem tomou essa decisão?**



Motivações para Procedência dos Dados

Processos de **autoria**
Ex: Verificar o responsável por um dado

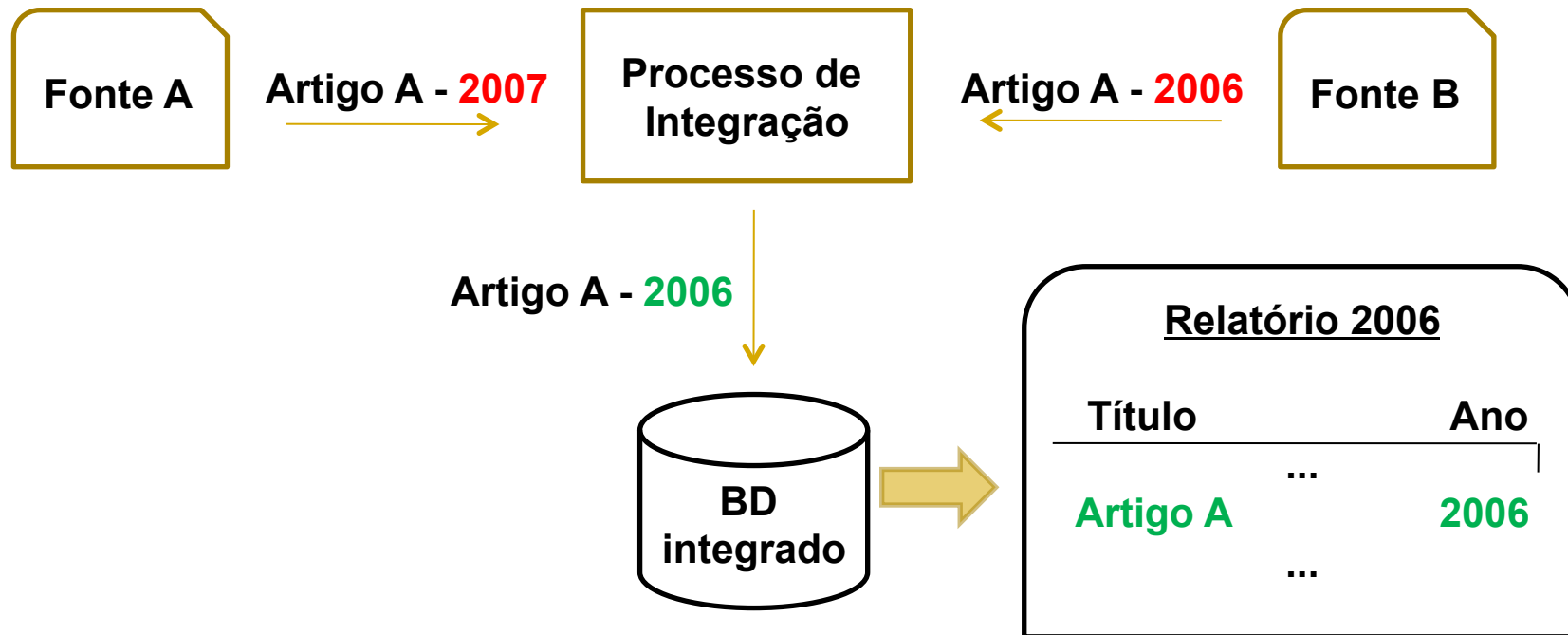


Quem é o responsável pelos dados desse artigo?



Motivações para Procedência dos Dados

Reenviar dados para as fontes
Ex: As fontes podem ser retificadas
com os dados integrados



Ok. Os dados do artigo A
estão corretos!



Motivações para Procedência dos Dados

Reproduzir decisões de integração dos dados
Ex: As fontes não podem ser retificadas com os dados integrados

- Fontes de Dados
 - são apenas para leitura (*read-only*)
 - sempre fornecem os mesmos dados inconsistentes
- Problema
 - necessidade de realizar a integração novamente, o que pode gerar diferentes decisões para um mesmo problema de integração e gasto adicional de tempo para prover o processo de integração

