## Análise de Dados com Base em Processamento Massivo em Paralelo

# Aula 4: Modelagem Conceitual de ETL/ELT

Cristina Dutra de Aguiar Ciferri
ICMC/USP
cdac@icmc.usp.br

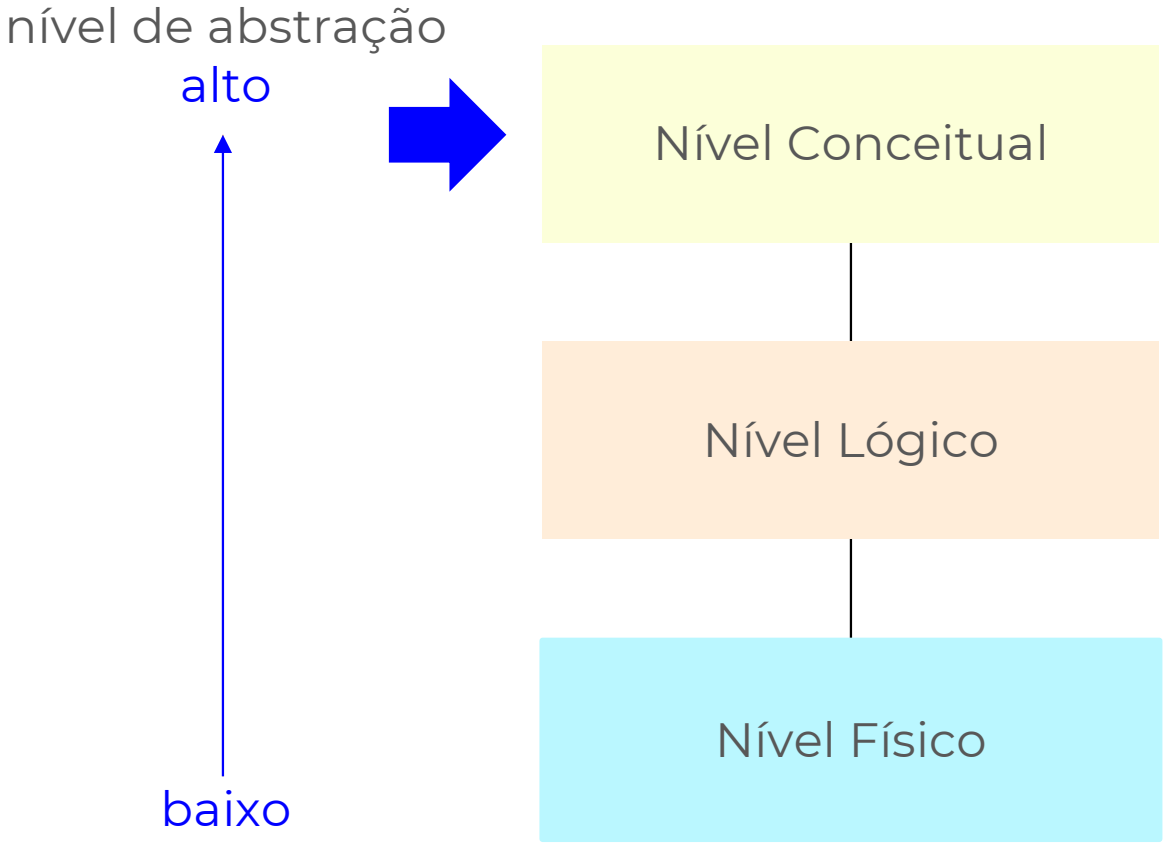
# Agenda

- Características
- Modelo Intuitive
- Exemplo para a BI Solutions

# Projeto de ETL/ELT

- Características desejadas para o sucesso
  - Robustez
  - Boa documentação
  - Facilidade de Manutenção
- Representado como um *workflow*
  - Cadeia de operações ou tarefas aplicadas aos dados
  - Representado por meio de um modelo

# Níveis de Abstração de um Modelo




complementa a análise de requisitos, facilitando o entendimento do processo

descreve os detalhes técnicos das tarefas envolvidas

incorpora aspectos de implementação e otimização

# Desafios e Motivação

- Desafios
  - Criticidade e complexidade do processo de ETL/ELT
  - Grande esforço despendido para a construção do processo
  - Propensão a falhas

- Motivação
  - Facilitar e padronizar a construção do processo de ETL/ELT
  - Melhorar a qualidade do processo de ETL/ELT e dos dados armazenados no DW

# Modelagem Conceitual

- Realizada na fase inicial do processo de ETL/ELT
  - Requisitos dos usuários de SSD
  - Entendimento do conteúdo e da estrutura das fontes de dados
  - Enfoque na estrutura proposta para o DW

- Produz um esquema conceitual
  - Representação gráfica e abstrata do processo de ETL/ELT

# Requisitos da Modelagem Conceitual

- Características desejadas
  - Simplicidade e completude
  - Clareza, consistência, não ambiguidade
- Diagrama produzido
  - Deve ser facilmente entendido pelos usuários finais que são conhecedores do negócio
    - muitas vezes não possuem conhecimento profundo de tecnologias
  - Deve contribuir para diminuir o esforço dos projetistas e desenvolvedores

# Funcionalidades Adicionais

- Documenta as decisões tomadas

- Possibilita a análise de impacto das alterações que ocorrem no ciclo de vida da aplicação de *data warehousing*
    - Alterações nas fontes de dados
    - Evolução dos requisitos ou das regras de negócio
    - Correção de erros cometidos durante a fase de projeto

- Facilita a exploração de cenários alternativos

# Modelagem Conceitual *versus* Ferramentas

- Modelagem Conceitual
  - Fornece alto nível de abstração, sendo independente de ferramentas específicas
- Ferramentas de ETL/ELT disponíveis
  - Relacionadas aos níveis lógico e físico
  - Exemplos
    - Pentaho Data Integration (Kettle)
    - Talend Open Studio
    - CloverETL
    - Oracle Warehouse Builder
    - IBM Infosphere
    - MSSQLServer Integration Services

# Agenda

- Características
- Modelo Intuitive
- Exemplo para a BI Solutions

# Características do Modelo

- Operadores
  - Entidades de alto nível
  - Representam as operações típicas de ETL/ELT
  - Possuem notação gráfica
- Combinação de operadores
  - Realizada por setas unidirecionais que indicam a propagação dos dados
  - Representa sequências que compõem o *workflow* de ETL/ELT

# Características do Workflow

- Início e final
  - Início: um ou mais repositórios de dados
  - Final: um ou mais repositórios, sendo o principal o *data warehouse* (ou *data mart*)

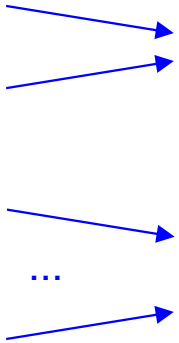
- Funcionalidades dos operadores
  - Manipulação de dados
  - Organização do fluxo de dados no *workflow*

Entrada
Parâmetro
Saída

# Especificação dos Operadores: Entrada

- Um ou mais conjuntos de dados

- Classificação
  - Unária: apenas um conjunto de dados
  - Binária: dois conjuntos de dados
  - N-ária: vários conjuntos de dados



# Especificação dos Operadores: Tipos de Parâmetro

- Lista de atributos
  - Nome de um atributo do conjunto de dados
  - Exemplo: funcNome, funcMatricula

- Condição
  - Expressão relacional
    - Exemplo: funcCidade = São Carlos
  - Expressão lógica
    - funcEstadoSigla = SP AND funcMatricula > 32879

Operações relacionais

= > < <> <= >=

Operadores lógicos

NOT, AND, OR

# Especificação dos Operadores: Tipos de Parâmetro

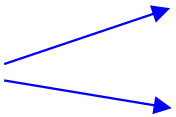
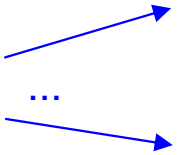
- Ordenação
  - Ordem crescente ou decrescente dos dados
  - Exemplo: asc e desc

- Precedência
  - Qual conjunto de dados deve ser analisado primeiro
  - Exemplo: dados do conjunto A — dados do conjunto B

- Lista de atribuições
  - Atributo <---  valor
  - Exemplo: funcMatricula <--- 234334, funcEstadoSigla <--- PE

# Especificação dos Operadores: Saída

- Um ou mais conjuntos de dados

- Classificação
  - Unária: apenas um conjunto de dados 
  - Binária: dois conjuntos de dados 
  - N-ária: vários conjuntos de dados 
    

# Categorias de Operadores

- Classificação baseada em
  - Características dos operadores
  - Efeitos que causam sobre os dados ou sobre a organização do processo

Armazenamento

Manipulação de Dados

Inicialização

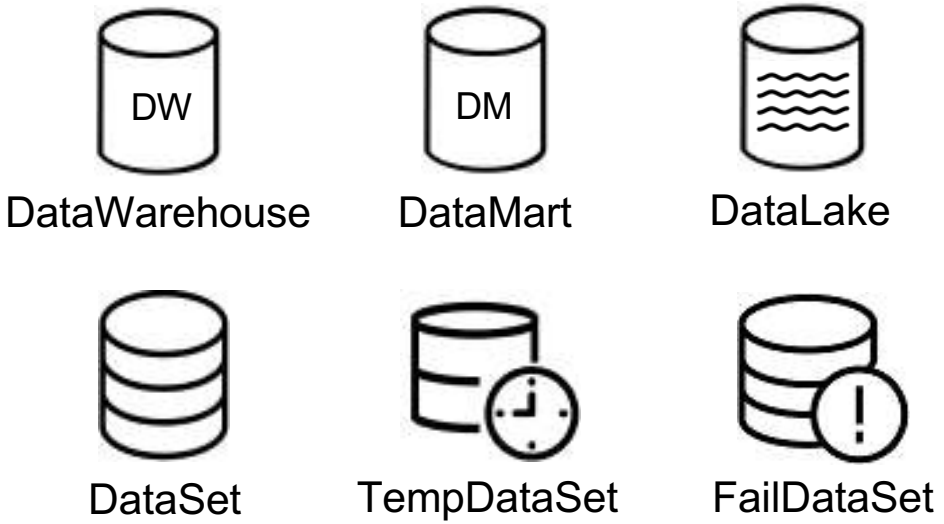
Agregação

Fluxo

Especiais

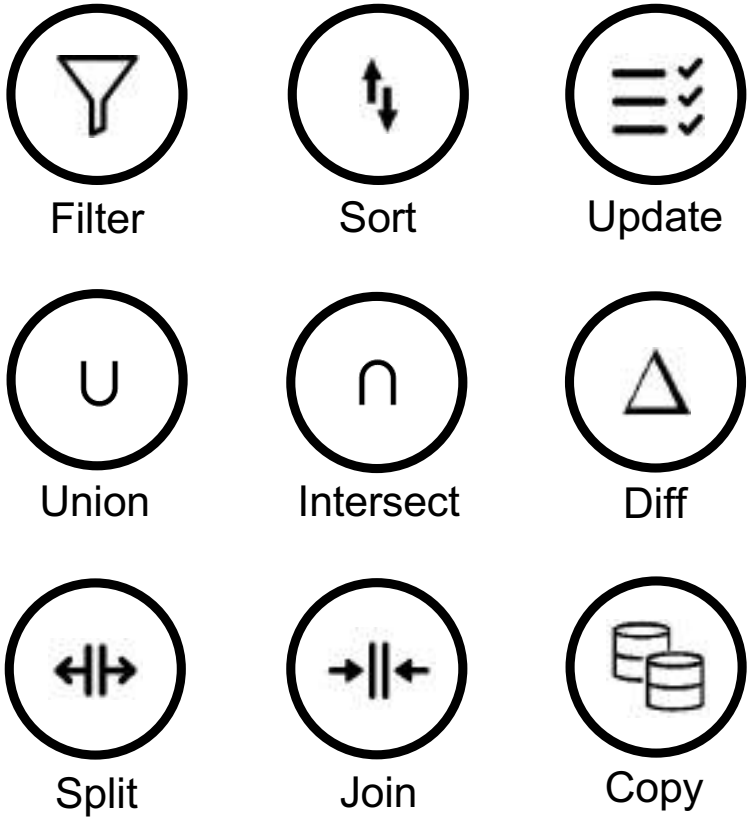
# Operadores de Armazenamento

Representam locais de armazenamento de dados, tais como repositórios, arquivos ou bancos de dados

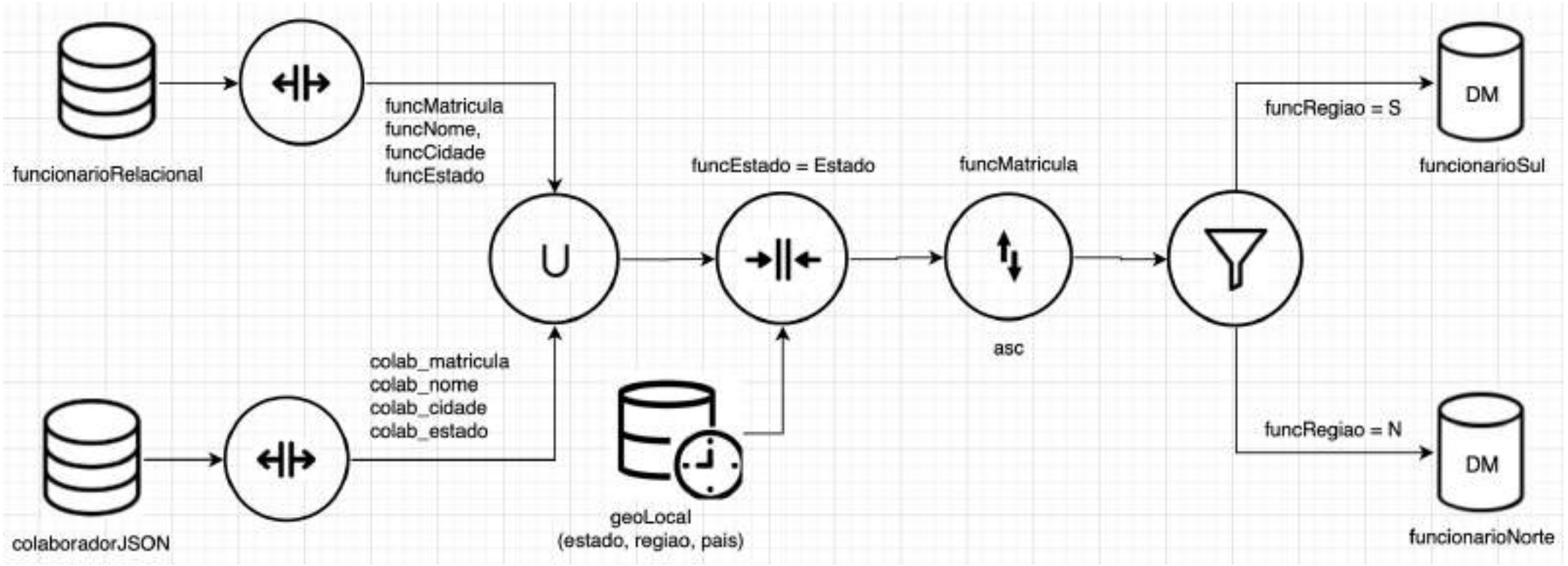


# Operadores de Manipulação de Dados

Representam operações de transformação e de limpeza dos dados

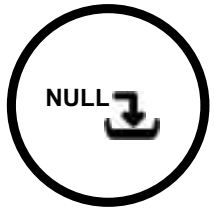


# Geração de Data Marts Regionais de Funcionários

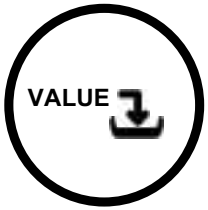


# Operadores de Inicialização

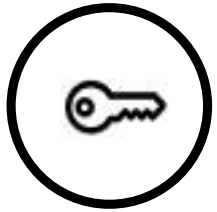
Representam a inicialização de um atributo com um valor específico



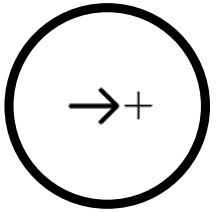

SetNullAsDefault




SetDefaultValue



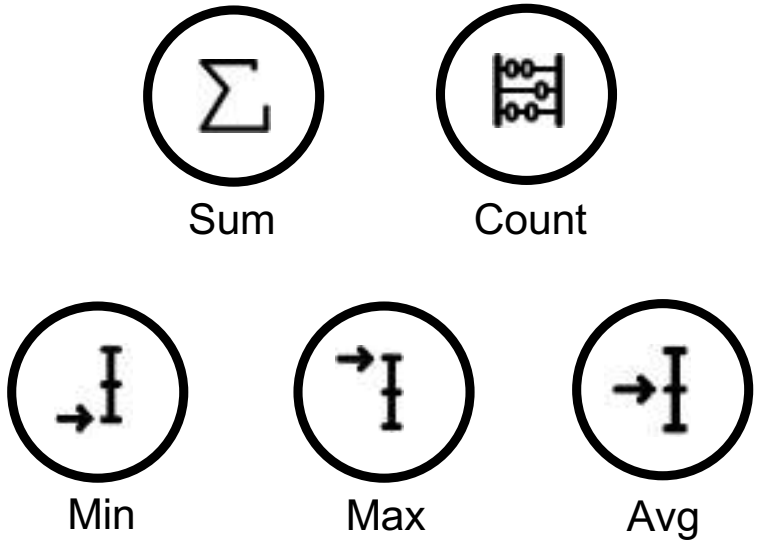
SurrogateKey



Sequence

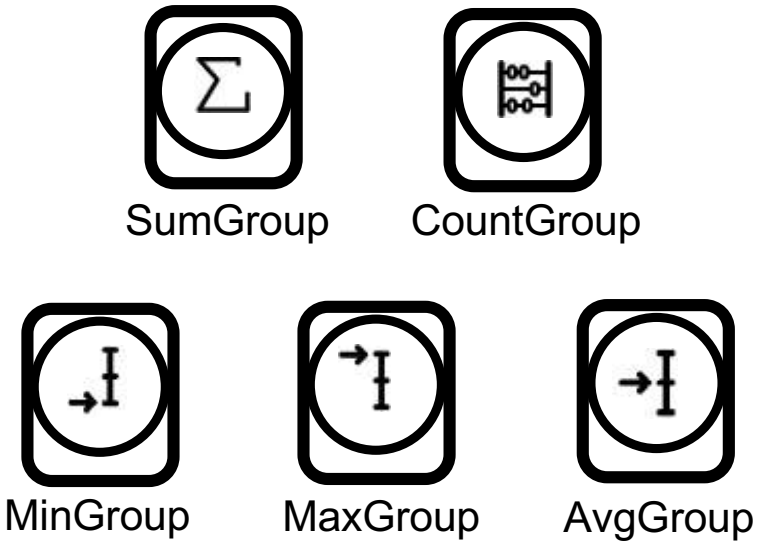
# Operadores de Agregação (1/2)

Representam funções que processam os valores de um atributo e retornam um único valor

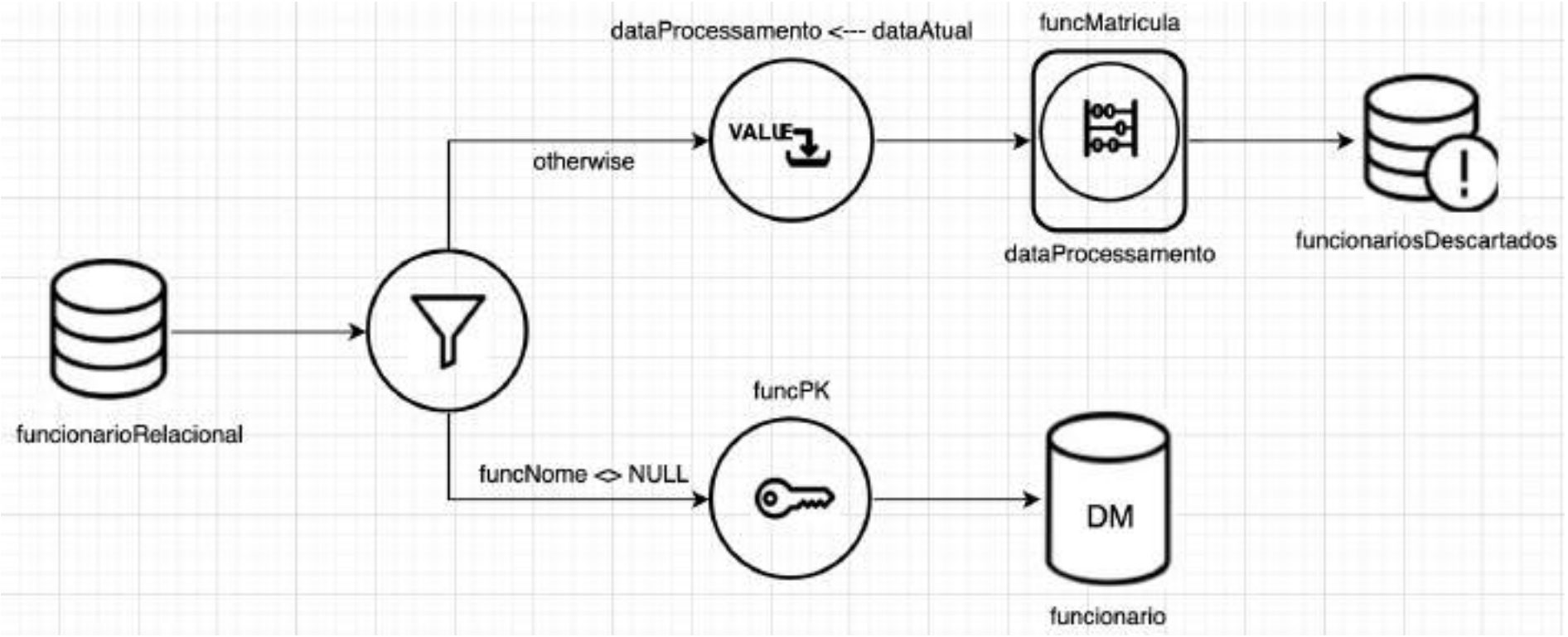


# Operadores de Agregação (2/2)

Representam funções que processam os valores de um atributo e retornam um único valor para cada atributo do agrupamento

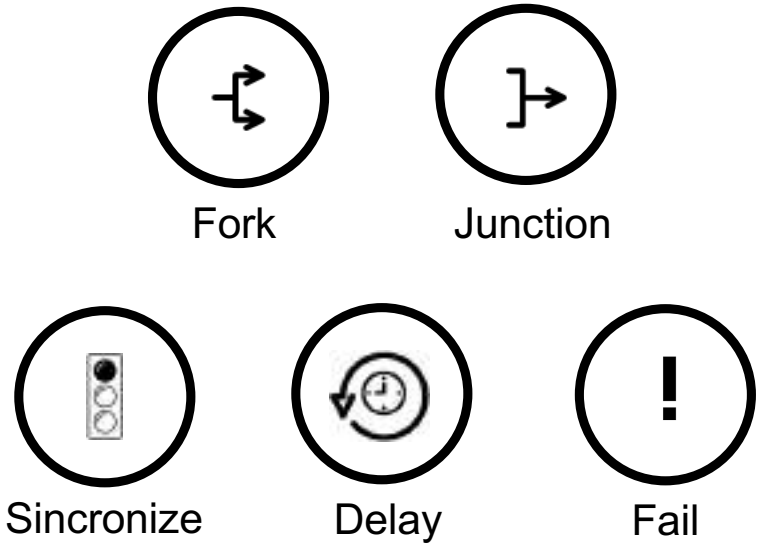


# Análise do Processamento de Funcionários

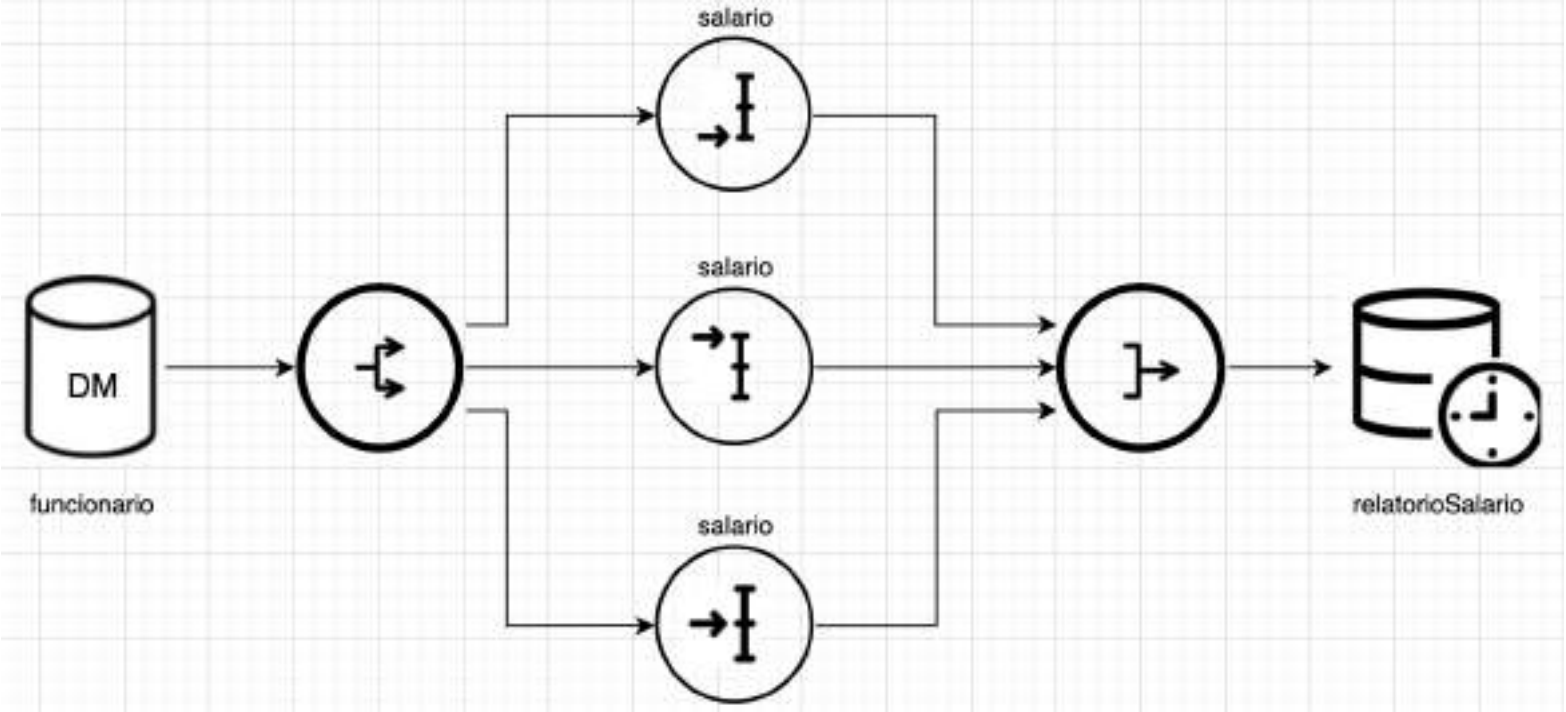


# Operadores de Fluxo

Representam uma alteração no fluxo dos dados, sem impactar esses dados

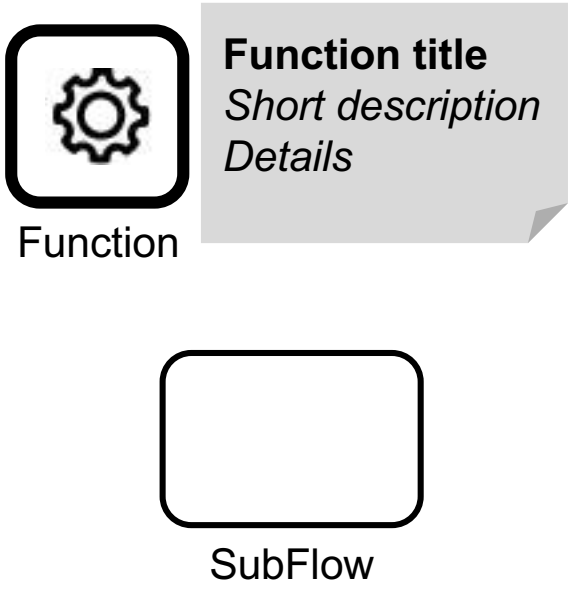


# Geração de Relatório de Salários



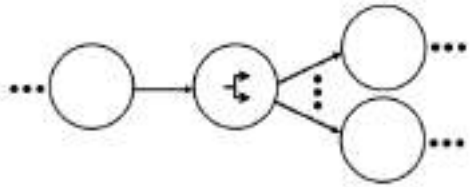
# Operadores Especiais

Representam operações que envolvem especificidades, complementando as funcionalidades dos demais operadores



# Material Suplementar

- Tabela descritiva dos operadores
  - Operador e sua funcionalidade
  - Representação gráfica, incluindo entrada, parâmetro e saída

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| **Fork** | direcionar um conjunto de dados para duas ou mais tarefas paralelas |  |

- Acompanha os slides no formato .pdf

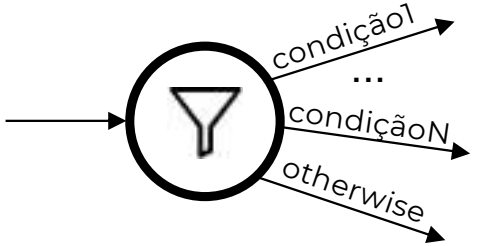
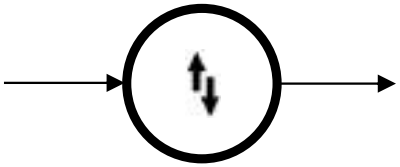
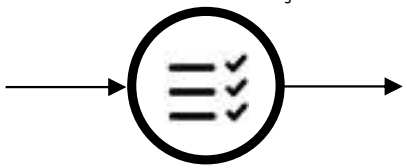
# Operadores de Armazenamento

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| DataWarehouse | armazena dados multidimensionais |   |
| DataMart | data warehouse com escopo limitado |   |
| DataLake | armazena dados brutos que ainda não foram transformados |   |

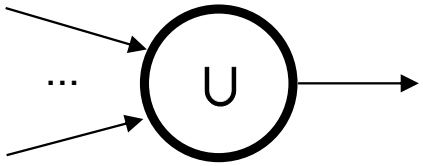
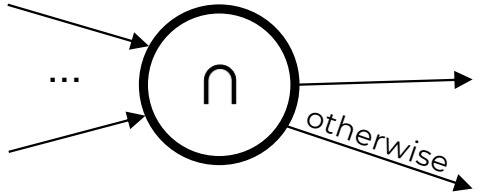
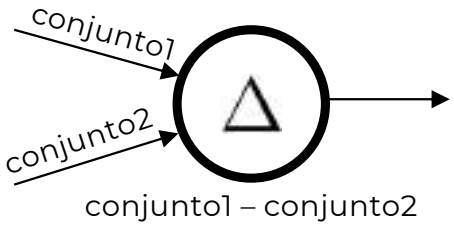
# Operadores de Armazenamento

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| DataSet | armazena dados |   |
| TempDataSet | área temporária de armazenamento de dados |   |
| FailDataSet | armazena dados rejeitados por uma operação ou para efeitos de log |   |

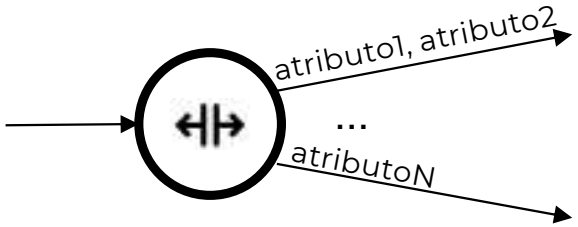
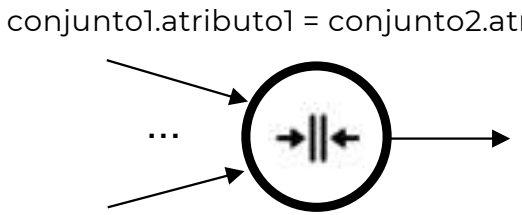
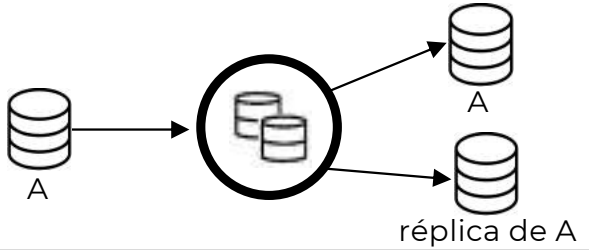
# Operadores de Manipulação de Dados

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Filter | seleciona subconjuntos de dados de acordo com condições definidas |   |
| Sort | ordena dados em ordem crescente ou decrescente, de acordo com atributos definidos |  atributo1 asc, ..., atributoN desc |
| Update | altera os valores dos dados, de acordo com condições definidas sobre atributos | lista de atribuições  condição |

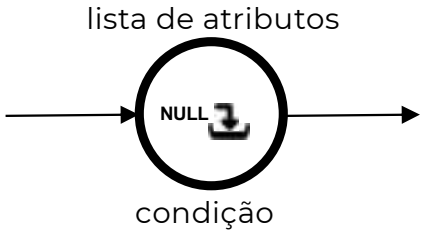
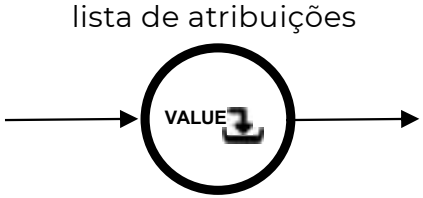
# Operadores de Manipulação de Dados

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Union | une conjuntos de dados, gerando um conjunto que contém todos os dados de entrada, sem repetição |   |
| Intersect | une conjuntos de dados, gerando um conjunto que contém apenas os dados em comum, sem repetição |   |
| Diff | gera os dados que estão presentes no primeiro conjunto de dados, mas não estão no segundo conjunto |   conjunto1 – conjunto2 |

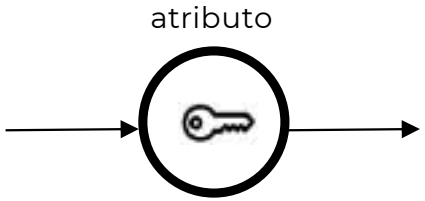
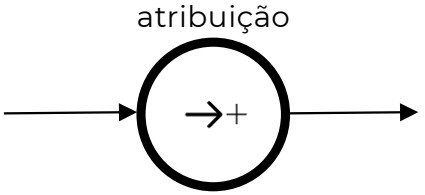
# Operadores de Manipulação de Dados

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Split | separa atributos de um conjunto de dados, direcionando-os para fluxos diferentes |   |
| Join | combina dois conjuntos de dados usando como base atributos em comum |   |
| Copy | a partir de um conjunto de dados de entrada, gera o próprio conjunto e uma réplica deste |   |

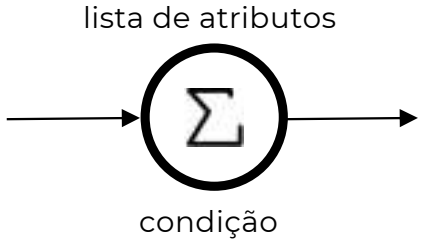
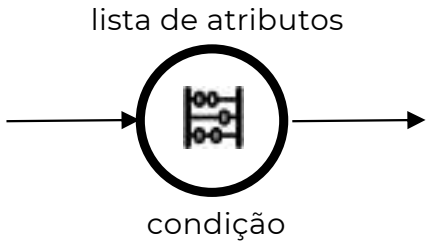
# Operadores de Inicialização

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| SetNullAsDefault | inicializa um atributo específico com o valor nulo, para todos os itens do conjunto de dados |   |
| SetDefaultValue | atribui um determinado valor para um atributo específico, para todos os todos os itens do conjunto de dados |   |

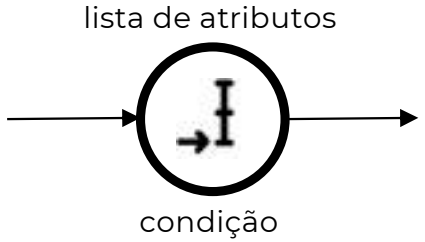
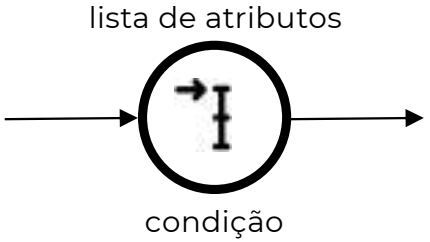
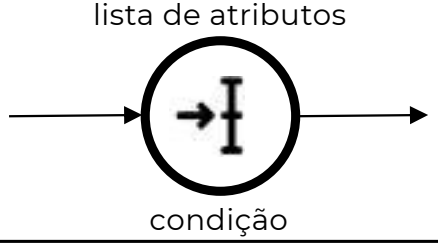
# Operadores de Inicialização

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| SurrogateKey | cria um atributo chave e atribui a ele um valor único para cada item do conjunto de dados |   |
| Sequence | cria um atributo não-chave e atribui a ele um valor único para cada item do conjunto de dados, o qual é gerado a partir de um valor inicial |   |

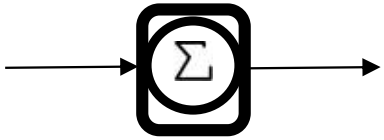
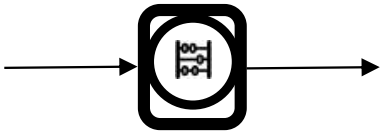
# Operadores de Agregação

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Sum | para cada atributo, soma seus valores e produz um único valor |   |
| Count | para cada atributo, conta seus valores e produz um único valor |   |

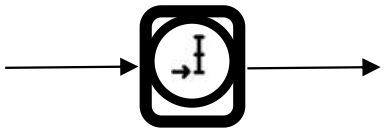
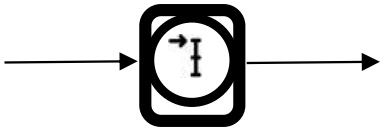
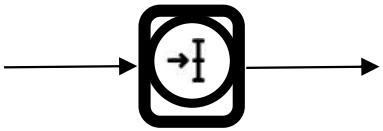
# Operadores de Agregação

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Min | para cada atributo, produz o menor valor |   |
| Max | para cada atributo, produz o maior valor |   |
| Avg | para cada atributo, produz o valor médio |   |

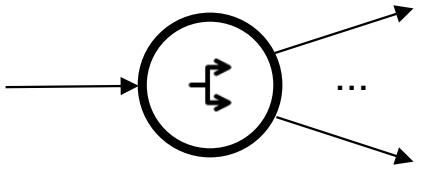
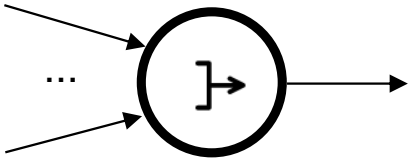
# Operadores de Agregação

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| SumGroup | para cada grupo, soma os valores dos dados de cada atributo | lista de atributos  lista de atributos de agrupamento condição |
| Count | para cada agrupamento, conta o número de dados de cada atributo | lista de atributos  lista de atributos de agrupamento condição |

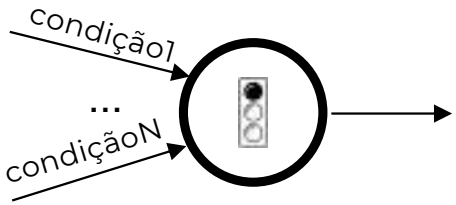
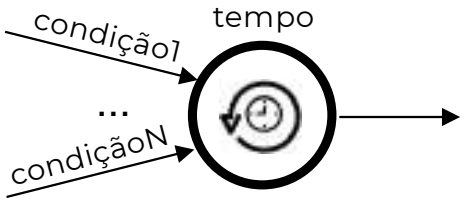
# Operadores de Agregação

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| MinGroup | para cada grupo, produz o menor valor de cada atributo | lista de atributos  lista de atributos de agrupamento condição |
| MaxGroup | para cada grupo, produz o maior valor de cada atributo | lista de atributos  lista de atributos de agrupamento condição |
| AvgGroup | para cada grupo, produz o valor médio de cada atributo | lista de atributos  lista de atributos de agrupamento condição |

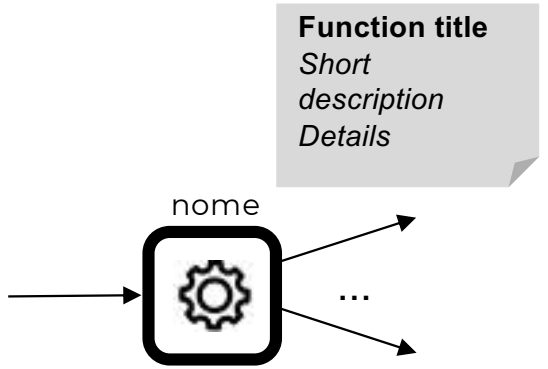
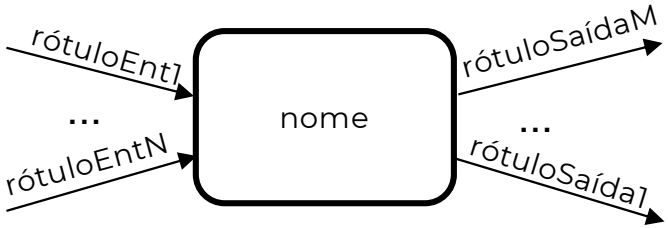
# Operadores de Fluxo

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Fork | direciona um conjunto de dados para dois ou mais fluxos executados em paralelo ou para um repositório e fluxos |  |
| Junction | junta dois ou mais fluxos executados em paralelo |  |

# Operadores de Fluxo

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Sincronize | sincroniza dois ou mais fluxos paralelos com base em uma condição de finalização |   |
| Delay | temporiza o tempo no qual será feita a análise de conjuntos de dados de fluxos paralelos |   |
| Fail | representa um fluxo alternativo para indicar falha |   |

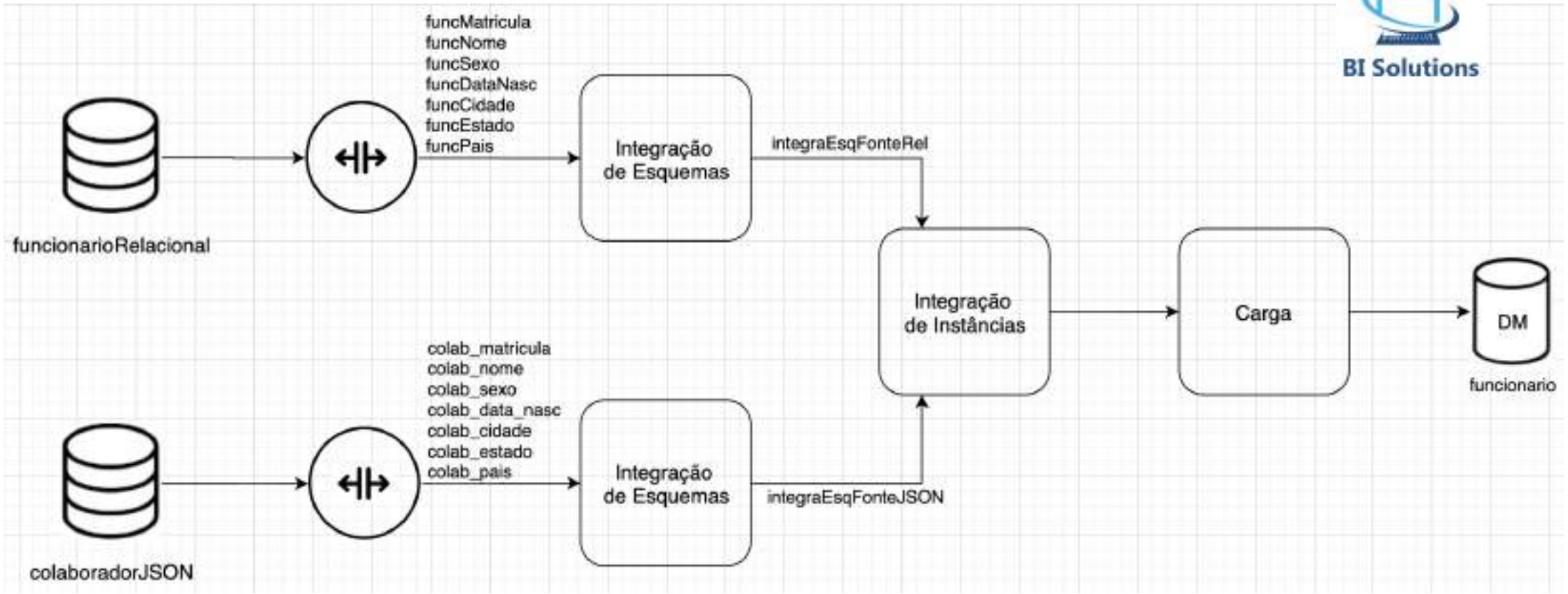
# Operadores Especiais

| Operador | Funcionalidade | Representação Gráfica |
|---|---|---|
| Function | representa operações ou atividades muito específicas que não podem ser representadas pelos outros operadores |  |
| SubFlow | encapsula subfluxos que envolvem conjuntos de tarefas específicas |  |

# Agenda

- Características
- Modelo Intuitive
- Exemplo para a BI Solutions

# Processo de ETL da BI Solutions



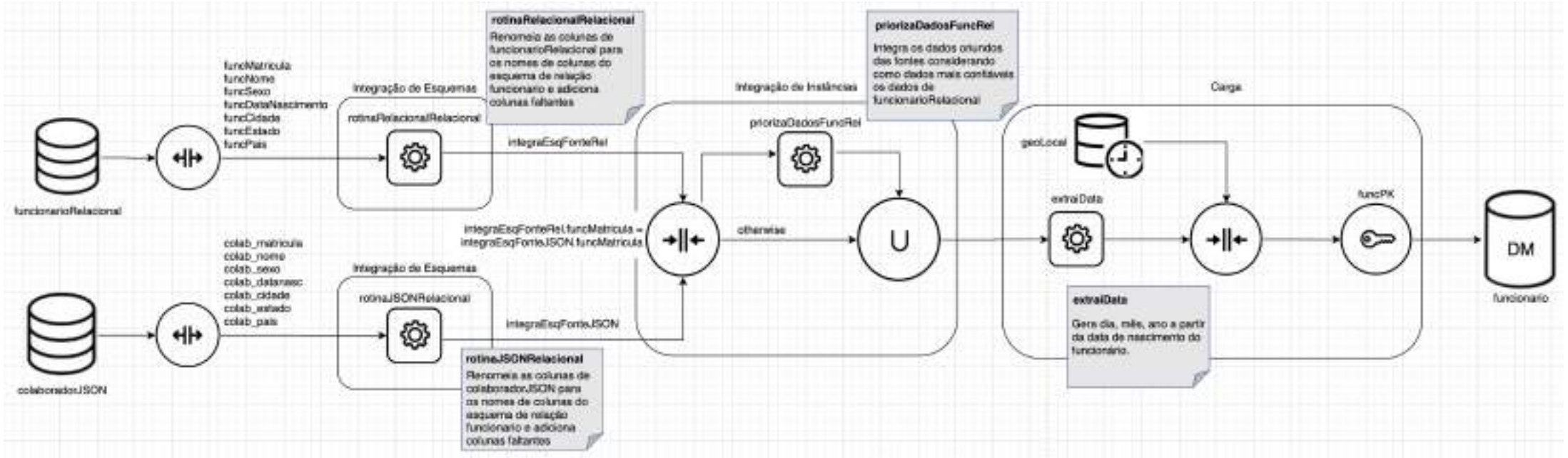
# Diagramas



- Exemplo do Processo de ETL
- Implementação em Pandas

# Diagrama Conceitual Completo

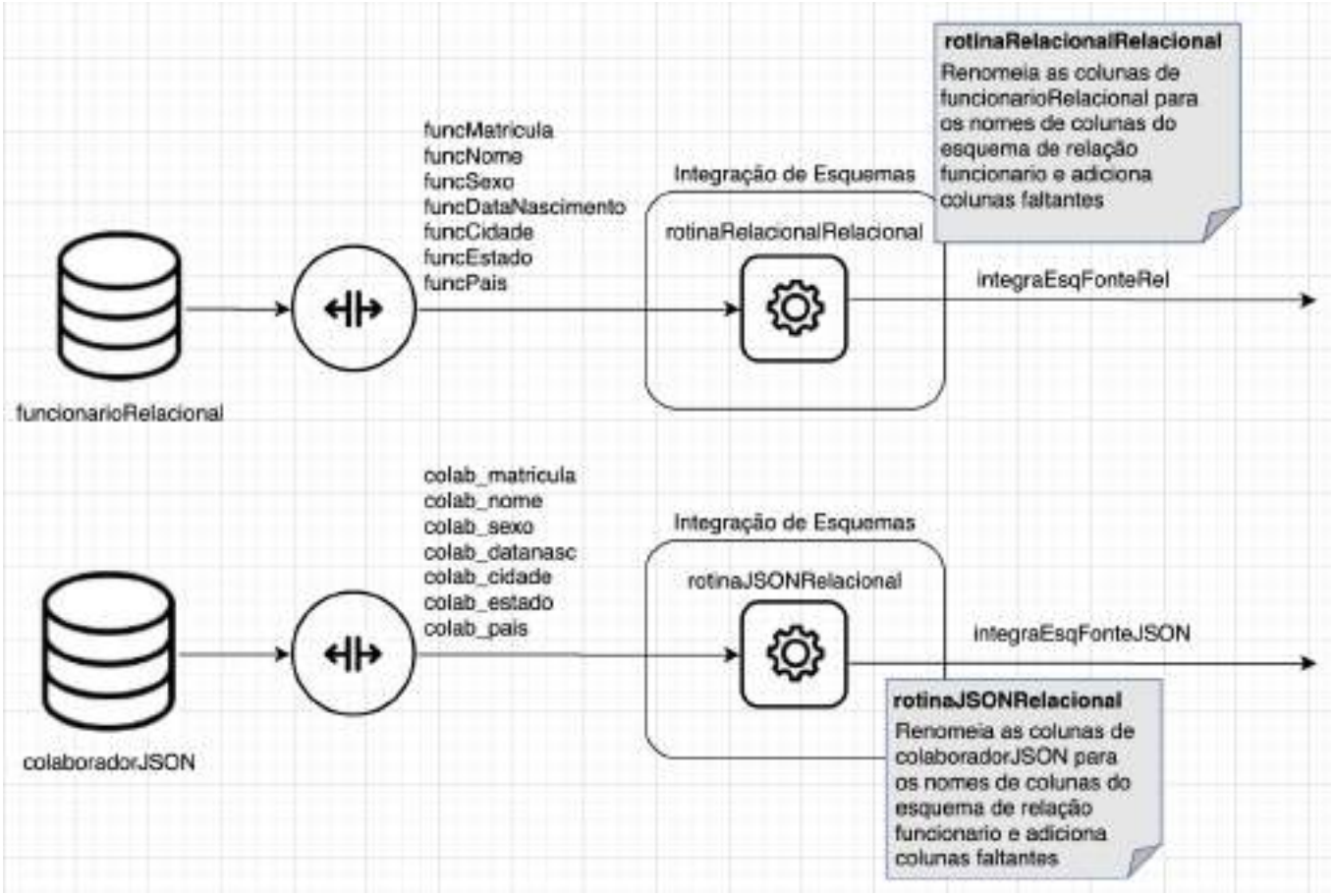
BI Solutions






USP ICMC CeMEAI

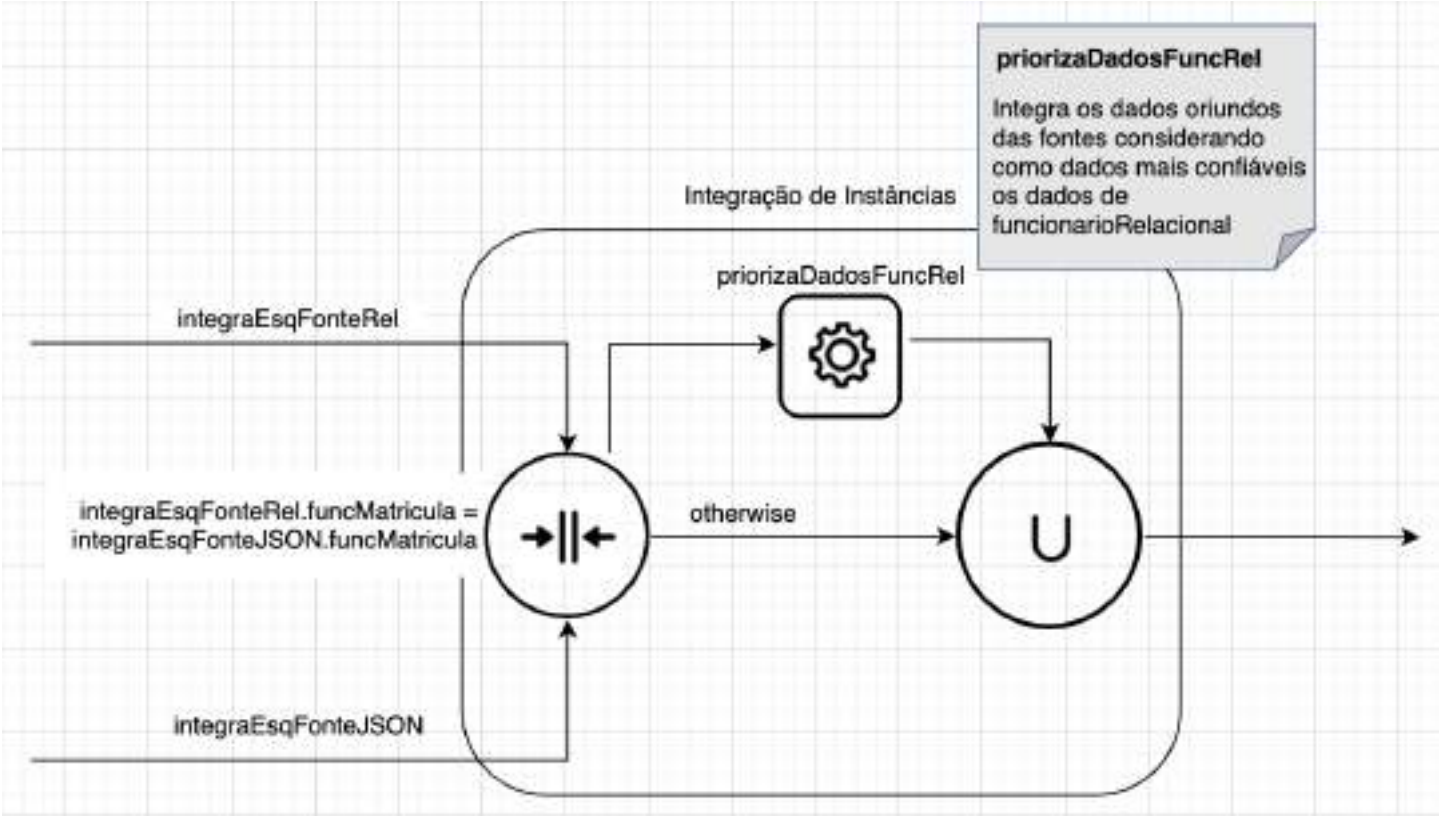
# Extração e Integração de Esquemas




BI Solutions

# Integração de Instâncias

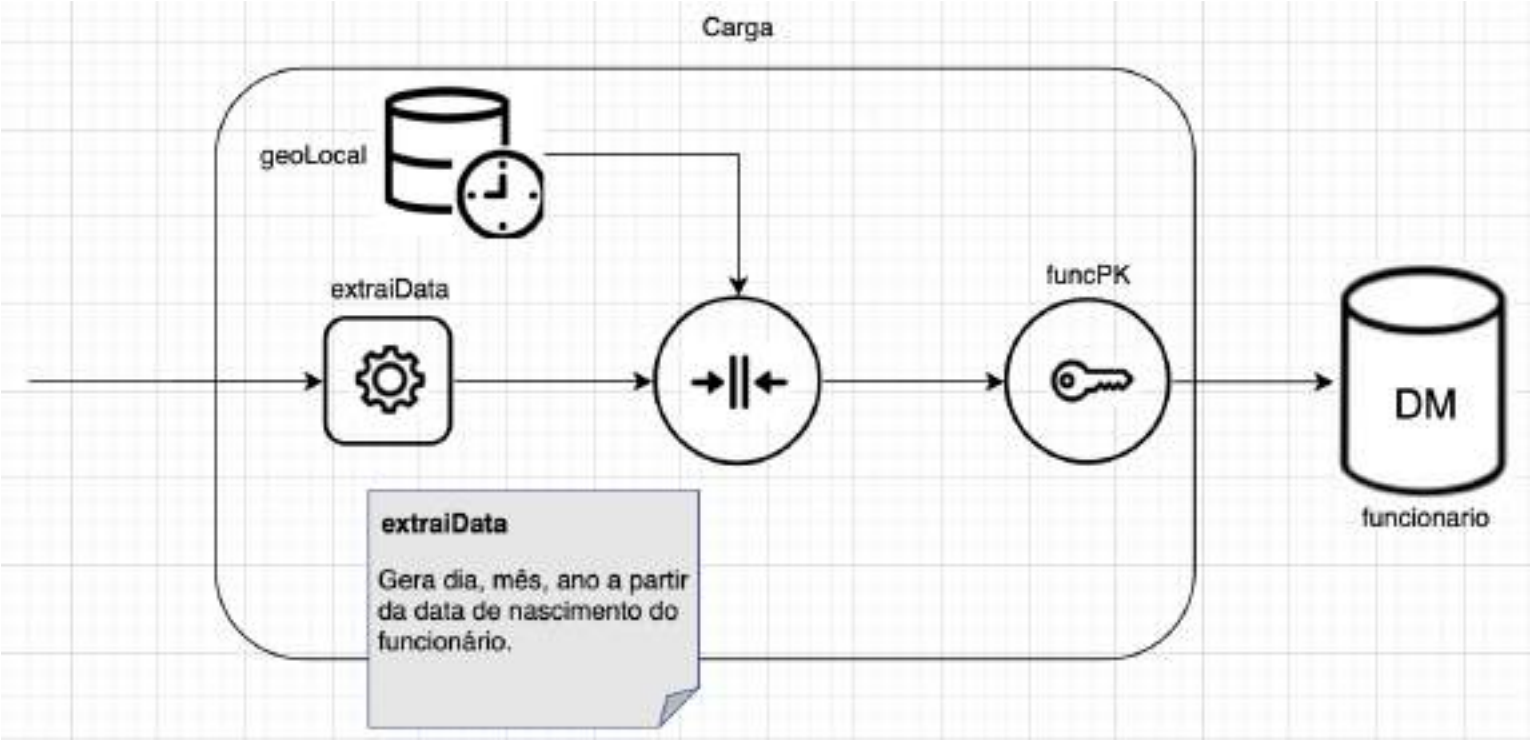
BI Solutions






USP ICMC CeMEAI

# Carga



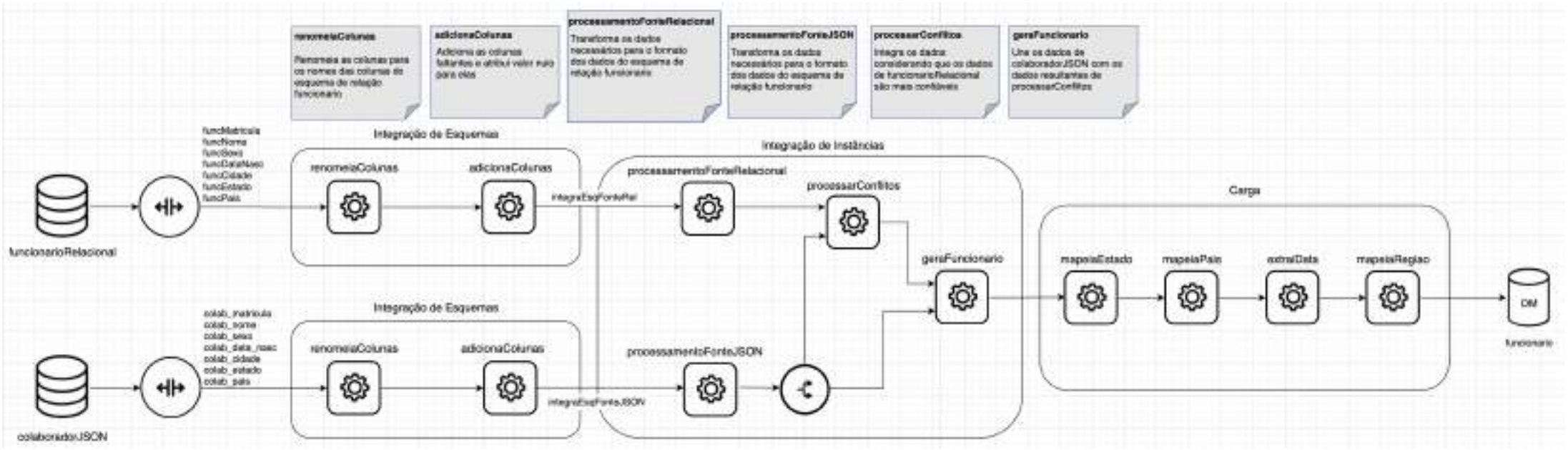
# Diagramas



- Exemplo do Processo de ETL
- Implementação em Pandas

# Visão Geral da Implementação em Pandas

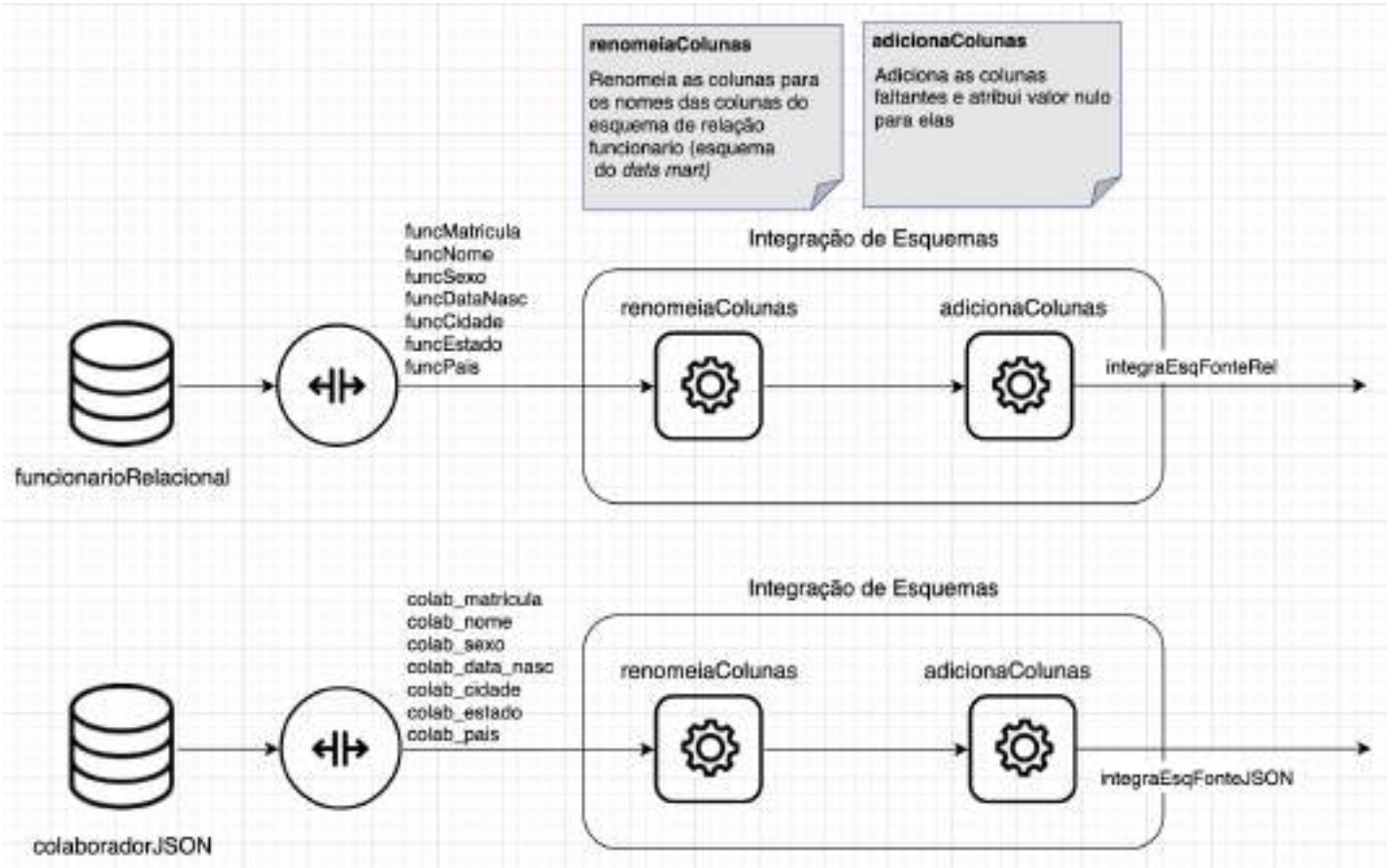
BI Solutions





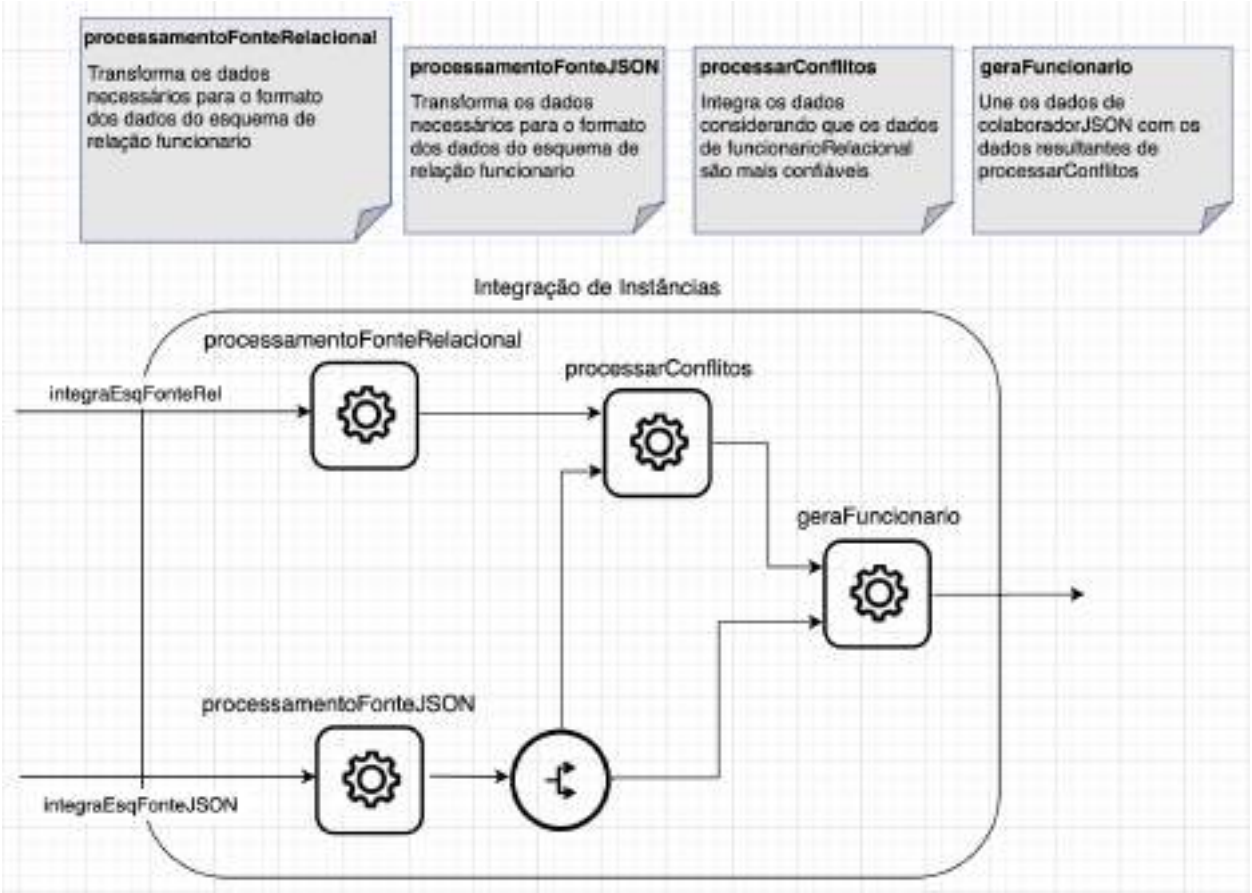

USP ICMC CeMEAI

# Extração e Integração de Esquemas




BI Solutions

# Integração de Instâncias




BI Solutions

# Carga

BI Solutions