UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS

Coleta e Pré-processamento de dados

Prof. Dr. Francisco Carlos Souza

Prof. Dr. Anderson Carniel

Sumário

- Fase 1 – Descoberta de Dados (Coleta de Dados)
  - Fontes de coleta
  - Formato dos dados
  - Tipo dos dados

- Fase 2 – Preparação dos Dados (Pré-processamento de Dados)
  - Limpeza dos Dados
  - Integração dos Dados
  - Redução dos Dados
  - Transformação dos Dados


06
Comunicação
dos Resultados
01
Descoberta
05
Operação
Processo de
Ciência de Dados
02
Preparação
dos Dados
04
Construção
de modelo
03
Planejamento
de modelo

Fase 1 - Descoberta

- Consiste na aquisição de dados de todas as fontes internas e externas identificadas, como:
  - Logs de servidores da web
  - Dados coletados das mídias sociais
  - Conjuntos de dados do censo
  - Dados transmitidos de fontes usando APIs

Fase 1 - Descoberta

COLETA DOS DADOS:

- Os dados podem ser obtidos por meio de diversas fontes podendo ser:

Interna

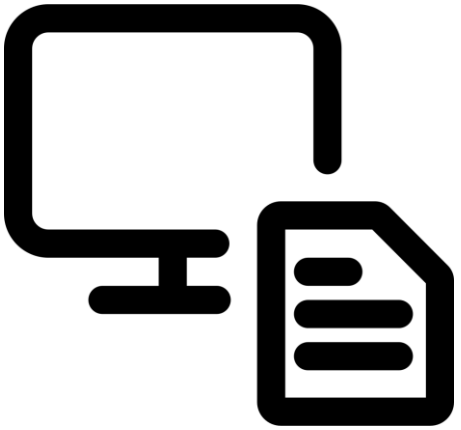
Externa

Fase 1 - Descoberta

COLETA DOS DADOS:

- Fontes internas

Banco de dados corporativos



Log do Sistema

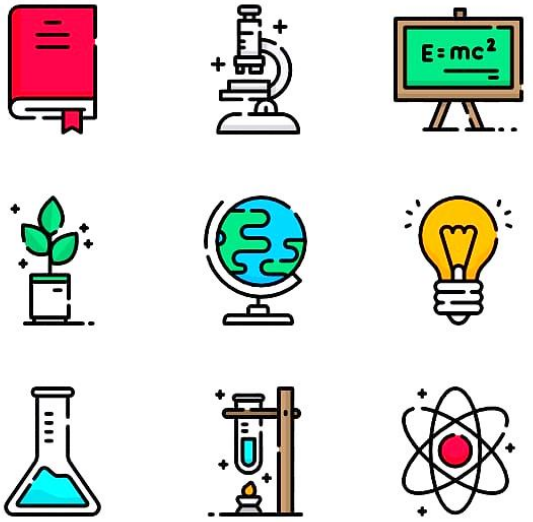
Documentos

Fase 1 - Descoberta

COLETA DOS DADOS:

- Fontes Externas:

Redes sociais



Dados acadêmicos

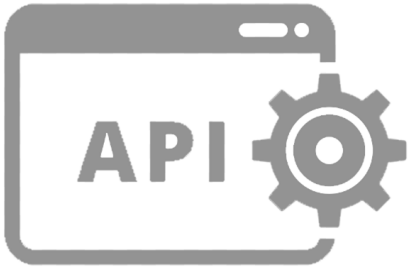
Banco de Dados Relacionais

Fase 1 - Descoberta

COLETA DOS DADOS:

- Fontes Externas:

Amigos



Web Scraping

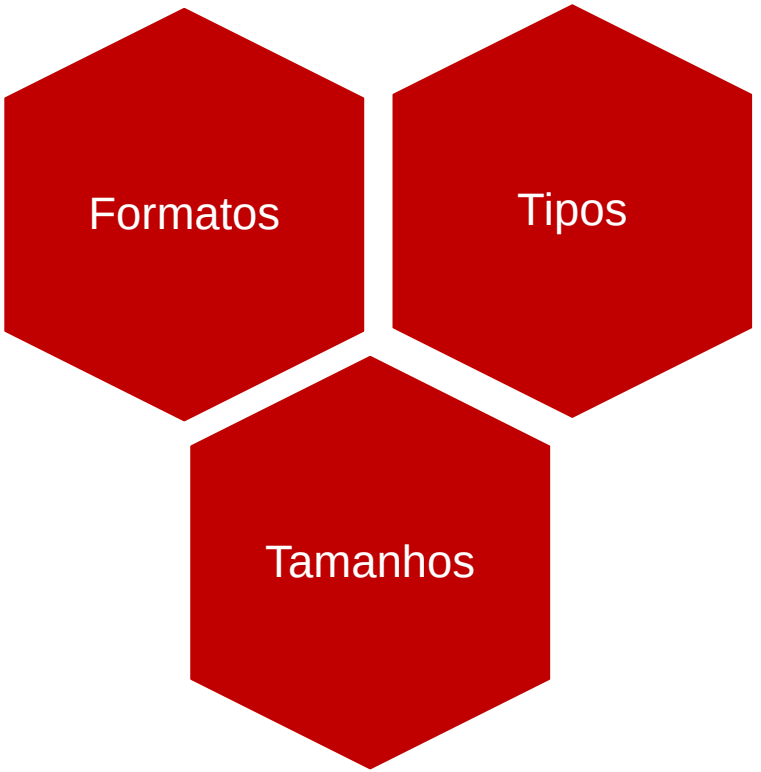
Web Crawler

Notícias

Fase 1 - Descoberta

COLETA DOS DADOS:

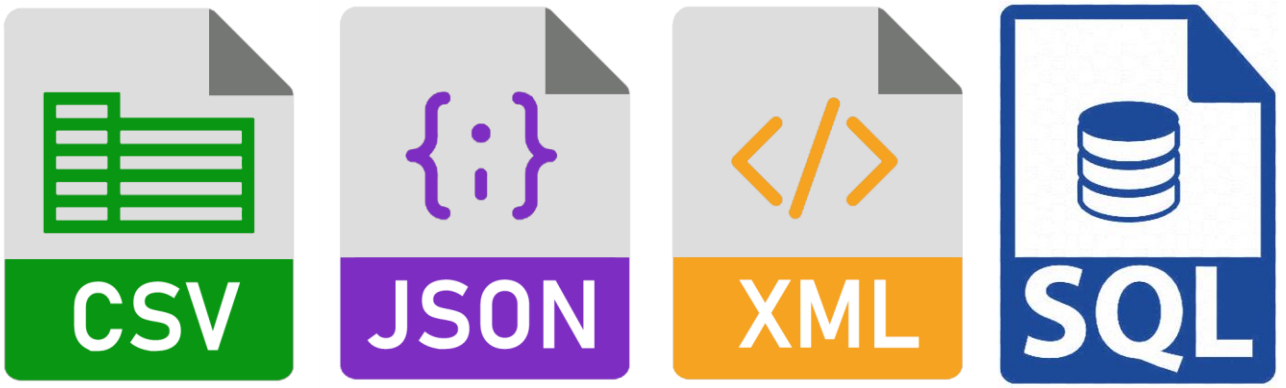
- Os dados podem ser apresentados em diferentes:



Fase 1 - Descoberta

COLETA DOS DADOS:

- Entre os **formatos** com maior frequência, têm-se:



Fase 1 - Descoberta

COLETA DOS DADOS:

- É importante entender que **tipo de dados** estão sendo utilizados e a propriedade dos dados coletados para resolver um determinado problema.
- Os tipos de dados podem ser:

Estruturados	Semi Estruturados	Não Estruturados

Fase 1 - Descoberta

COLETA DOS DADOS:

- **Dados estruturados**
  - Representam dados que são **armazenados**, **processados** e **manipulados** em sistemas tradicionais de bancos de dados relacionais;
  - Estrutura **rígida** e previamente planejada;
  - Representação **homogênea**;
  - Cada campo de dados tem um **formato bem definido**.

Fase 1 - Descoberta

COLETA DOS DADOS:

- **Dados estruturados**
  - Ex: Banco de Dados

ID	Nome	Idade	Titularidade
1	John	21	Bacharel
2	Davi	31	Doutor
3	Roberto	51	Doutor
4	Rick	26	Mestre
5	Michel	19	Mestre

Fase 1 - Descoberta

COLETA DOS DADOS:

- **Dados Semi estruturados**
  - Representam dados estruturados por **tags** que são úteis para criar **ordem** e **hierarquia** nos dados;
  - Estrutura **flexível**;
  - Representação **heterogênea**;
  - Cada **campo de dados** tem uma estrutura, mas **não existe** uma imposição de **formato**.

Fase 1 - Descoberta

COLETA DOS DADOS:

- **Dados Semi estruturados**
  - **Ex:** XML, JSON, RDF, OWL

```
<Universidade>
  <Estudante>
    <ID> 1 </ID>
    <Nome> John </Nome>
    <Idade> 21 </Idade>
    <Titularidade> Bacharel </Titularidade>
  </Estudante>
  <Estudante>
    <ID> 2 </ID>
    <Nome> Davi </Nome>
    <Idade> 31 </Idade>
    <Titularidade> Doutor </Titularidade>
  </Estudante>
  ...
</Universidade>
```

Fase 1 - Descoberta

COLETA DOS DADOS:

- **Não estruturados**
  - Representam dados **produzidos** a partir de **atividades humanas** e não se encaixam em um formato de banco de dados tradicional.;
  - **Sem estrutura pré-definida**;
  - Constituem a maioria dos dados corporativos;
  - Mais de 80% dos dados gerados no mundo é deste tipo.

Fase 1 - Descoberta

COLETA DOS DADOS:

- **Dados Não estruturados**
  - **Ex:**

Fase 1 - Descoberta

COLETA DOS DADOS:

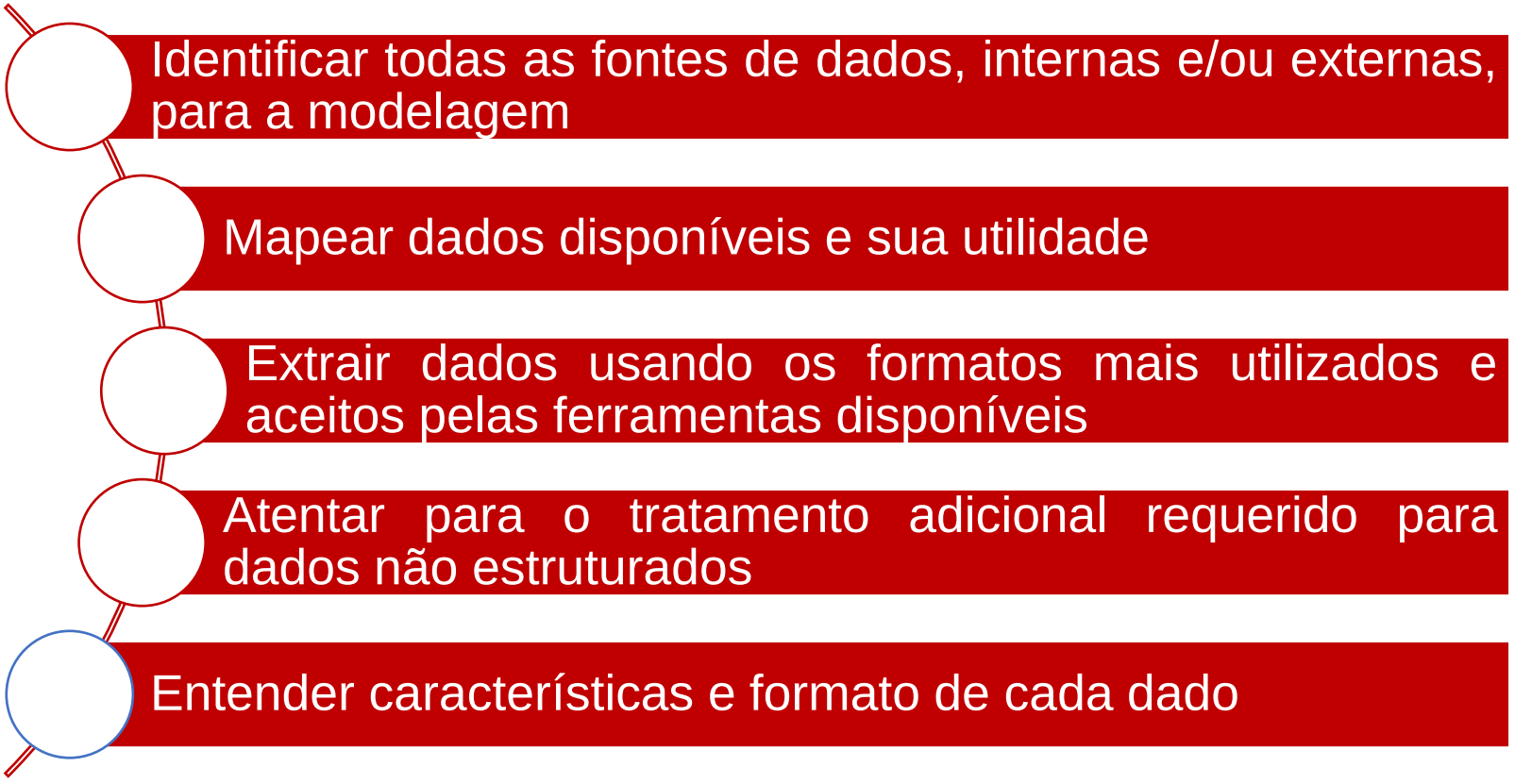
- **Dados Não estruturados**
  - **Ex:**

```
A Uiversidade possui 5600 estudantes. O ID de John é 1, ele tem 21
anos e é Bacharel. O ID de Davi é 2, ele tem 31 anos e é Doutor. O
ID de Roberto é 3, ele tem 51 anos e é Doutor.
```

Fase 1 - Descoberta

COLETA DOS DADOS:

- Nesta etapa é importante:



Fase 2 – Preparação dos Dados

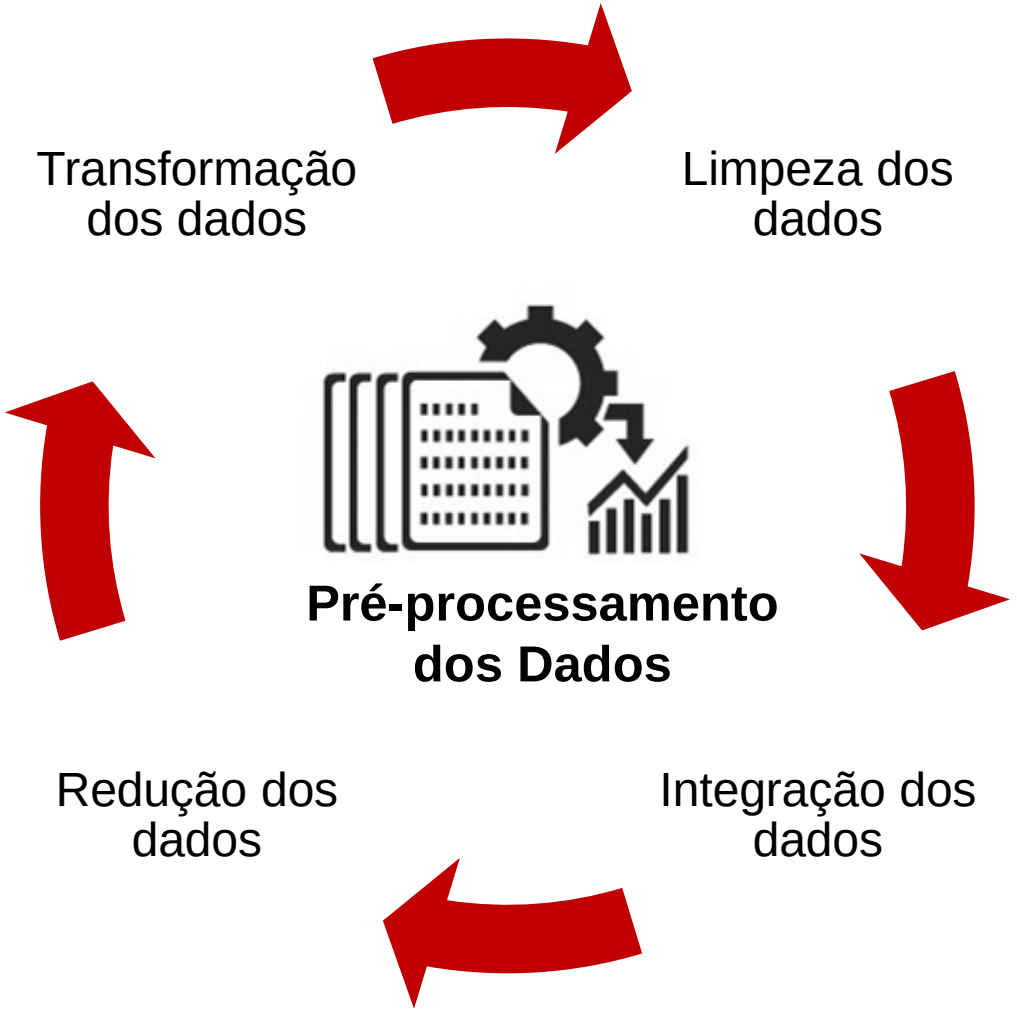
- Os dados podem ter muitas inconsistências, como:
  - valor ausente,
  - colunas em branco
  - formato de dados incorreto que precisa ser limpo.
- Processar, explorar e condicionar dados antes da modelagem.
- Quanto mais limpos os dados, melhores são as previsões.

Fase 2 – Preparação dos Dados

- Também chamada de **Pré-Processamento**
- É requerida para preparar dados para modelagem
- Dificilmente os dados coletados estarão prontos para análise imediata.
- Consiste no processo de **limpeza**, **transformação** (normalizar, combinar), **enriquecimento** e **estruturação** de **dados brutos** para utilizá-los nas análises, modelagens, reportes, visualização e no resultado.

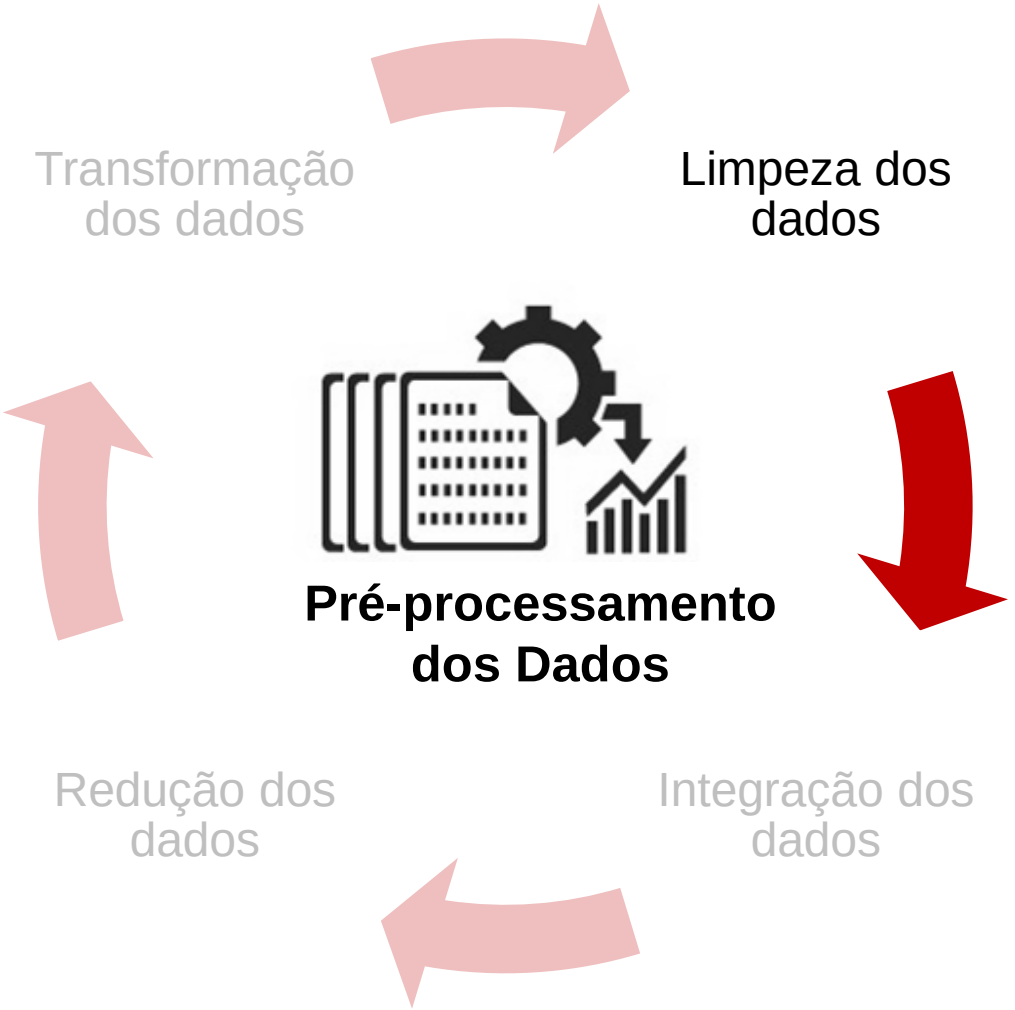
Fase 2 – Preparação dos Dados

- Etapas do pré-processamento dos dados:



Fase 2 – Preparação dos Dados

- Etapas do pré-processamento dos dados:



Fase 2 – Preparação dos Dados

1. LIMPEZA DOS DADOS:

- Refere-se as técnicas para "**limpar**" dados, removendo valores **discrepantes**, substituindo valores **ausentes**, suavizando dados **ruidosos** e corrigindo dados **inconsistentes**.

Fase 2 – Preparação dos Dados

1. LIMPEZA DOS DADOS:

- **Dados incompletos:**
  - Os dados podem conter alguns valores **ausentes** ou **nulos**
  - Existem diferentes métodos que auxiliam o preenchimento dos valores ausentes:
    - Ignorar a tupla;
    - Preencher valor ausente manualmente;
    - Usar um valor padrão para substituir o valor ausente;
    - Usar tendência central (média, mediana, modo) para atributo visando substituir o valor ausente;
    - Usando o valor mais provável para preencher o valor ausente.

Fase 2 – Preparação dos Dados

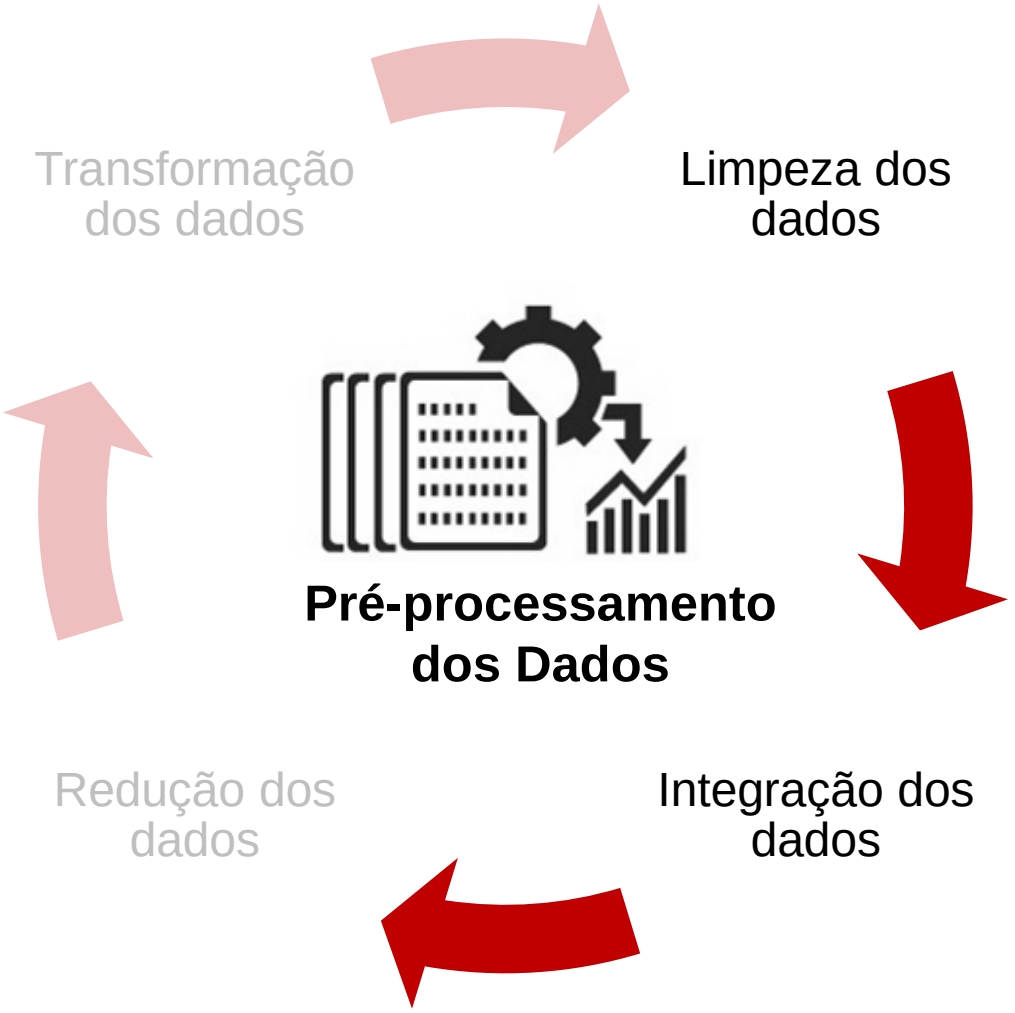
1. LIMPEZA DOS DADOS:

- **Dados com ruídos:**
  - Ruído é um **erro aleatório** ou **outlier** no atributo
  - Os dados podem ser suavizados usando as seguintes técnicas:

Método Binning	• Suaviza um valor de dados classificados consultando a vizinhança ou os valores ao redor.
Clustering	• Auxilia na detecção dos outliers, onde valores semelhantes são organizados em grupos ou cluster.
Regressão	• A regressão linear e a regressão linear múltipla podem ser usadas para suavizar os dados, onde os valores estão em conformidade com uma função.

Fase 2 – Preparação dos Dados

- Etapas do pré-processamento dos dados:



Fase 2 – Preparação dos Dados

2. INTEGRAÇÃO DOS DADOS:

- Como os dados são coletados de várias fontes, a integração de dados se tornou uma **parte vital** do processo.
- A integração pode levar a dados redundantes e inconsistentes, o que pode resultar em baixa precisão e velocidade do modelo de dados.

Fase 2 – Preparação dos Dados

2. INTEGRAÇÃO DOS DADOS:

- Abordagens mais comuns para integrar dados:

Consolidação dos dados	• Os dados são comprados fisicamente juntos em um armazenamento de dados. Isso geralmente envolve Data Warehousing.
Propagação dos dados	• Consiste em copiar dados de um local para outro usando aplicativos, podendo ser síncrono ou assíncrono e é orientado a evento.
Virtualização dos dados	• Uma interface é usada para fornecer uma visão unificada e em tempo real dos dados de várias fontes. Os dados podem ser visualizados a partir de um único ponto de acesso.

Fase 2 – Preparação dos Dados

- Etapas do pré-processamento dos dados:

Fase 2 – Preparação dos Dados

3. REDUÇÃO DOS DADOS:

- Visa ter uma **representação condensada** do conjunto de dados que seja menor em **volume**, mantendo a integridade do original.
- Resulta em resultados eficientes, mas similares.

Fase 2 – Preparação dos Dados

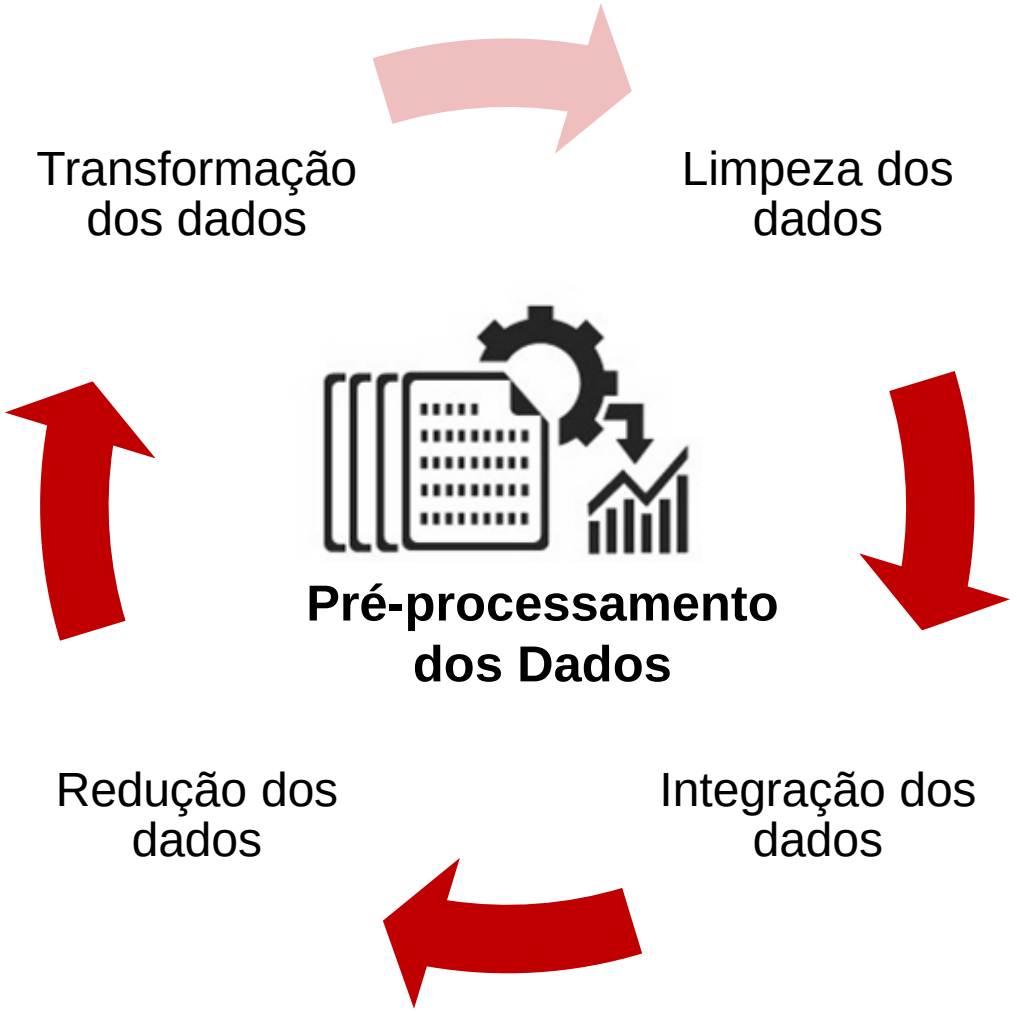
3. REDUÇÃO DOS DADOS:

- Métodos para reduzir o volume dos dados:

Relação de valores ausentes	• Os atributos que têm mais valores ausentes do que um limite são removidos.
Filtro de baixa variação	• Os os atributos normalizados que têm variação (distribuição) menor que um limite também são removidos.
Filtro de alta correlação	• Os atributos normalizados que têm um coeficiente de correlação maior que um limite também são removidos, pois tendências semelhantes significam que informações semelhantes são transportadas.
Análise de componentes principais	• É um método estatístico que reduz o número de atributos reunindo atributos altamente correlacionados. funciona apenas para recursos com valores numéricos.

Fase 2 – Preparação dos Dados

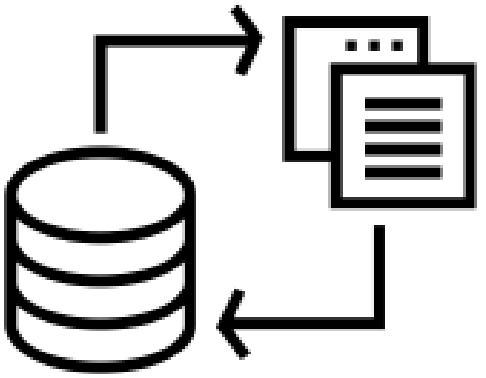
- Etapas do pré-processamento dos dados:



Fase 2 – Preparação dos Dados

4. TRANSFORMAÇÃO DOS DADOS:

- Visa transformar os dados em um formato apropriado para a Modelagem de Dados.



Fase 2 – Preparação dos Dados

4. TRANSFORMAÇÃO DOS DADOS:

- Estratégias que permitem a transformação de dados:

Suavização	• Aplicação dos métodos Bening, clusterização e regressão.
Construção de atributo / recurso	• Novos atributos são construídos a partir do conjunto de atributos fornecido.
Agregação	• São aplicadas no conjunto de atributos fornecido para criar novos atributos.
Normalização	• Os dados em cada atributo são redimensionados entre um intervalo menor, por exemplo 0 a 1 ou -1 a 1.
Discretização	• Os valores brutos dos atributos numéricos são substituídos por intervalos discretos ou conceituais.
Geração do conceito de hierarquia de para dados nominais	• Os valores para dados nominais são generalizados para conceitos de ordem superior

Pré-Processamento dos Dados

- Apesar da existência de várias abordagens para pré-processar dados, ainda é um campo pesquisado ativamente devido à **quantidade de dados incoerentes sendo gerados diariamente.**