Contexto e natureza dos dados em segurança pública

Dados Estruturados

BOs, sistemas policiais, bases jurídicas

Dados Semi-estruturados

Planilhas, PDFs, JSON, XML

Dados Não Estruturados

Imagens, vídeos, redes sociais

Dados Estruturados

Database

Dados armazenados em sistemas

📄 "BO" (Boletim de Ocorrência)

👮 "Registro Policial"

Região com mais ocorrências de...

SQL com GROUP BY e COUNT(*)

	KEY	DATA
0		
1		
2		
3		
4		

Formato fixo, ideais para consultas e estatísticas oficiais

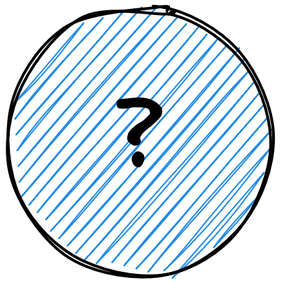
Como garantir que todos os registros de um mesmo tipo de crime estão padronizados?

Qualidade e integridade de dados

Dados Estruturados

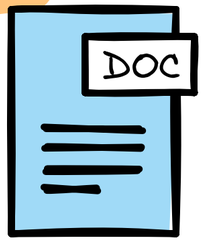
O número de ocorrências de um mesmo tipo cresce em períodos específicos?

Séries temporais e sazonalidade

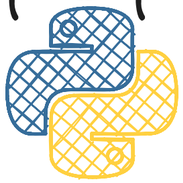


TA

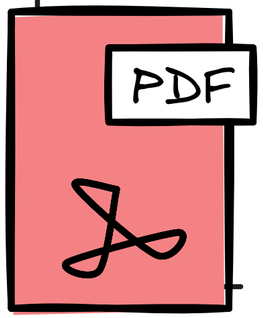
Dados Semi-estruturados



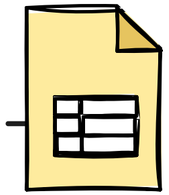
Exigem pré-processamento e parsing



Parte organizada, parte textual

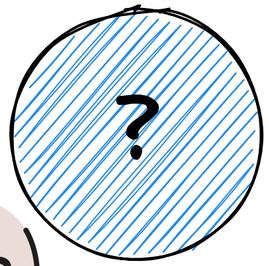


Complexidade intermediária, pois há dados disponíveis, mas não imediatamente utilizáveis.



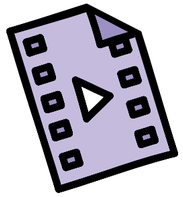
O que muda na análise quando o boletim está em PDF e não em um sistema?

Dados Semi-estruturados



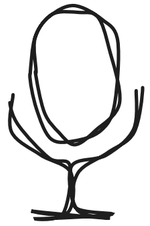
É possível cruzar os PDFs de laudos periciais com os dados de ocorrência?

Dados não estruturados

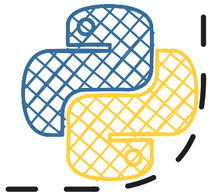


Oceano caótico

visão computacional, PLN, reconhecimento de padrões

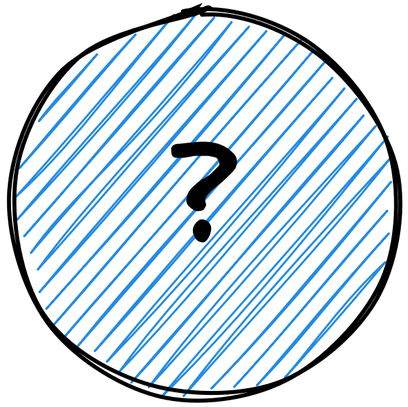


O valor dos dados não estruturados vem do processamento inteligente e do contexto ético



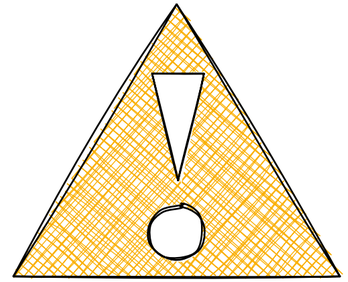
Como identificar padrões de veículos suspeitos em vídeos de câmeras públicas?

Dados não estruturados



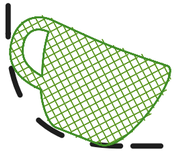
Micros Desafios

Estruturados: Campos mal padronizados inflacionam contagens.



Semi: OCR e parsing podem introduzir erros; registre versões.

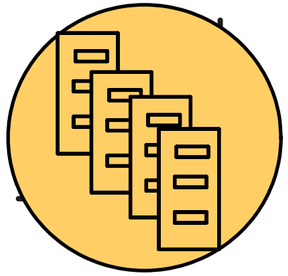
Não estruturados: Bias e ambiguidade — resultados dependem de contexto.



Exemplos básicos

```sql
SELECT regiao, COUNT(*) AS ocorrencias
FROM boletins
WHERE tipo='assalto' AND date(data_ocorrencia) >= date(?)
GROUP BY regiao
ORDER BY ocorrencias DESC;
```

```
$ ocrmypdf laudo.pdf laudo_ocr.pdf
```



Contexto e natureza dos dados em segurança pública

Dados Estruturados
BOs, sistemas policiais, bases jurídicas

Dados Semi-estruturados
Planilhas, PDFs, JSON, XML

Dados Não Estruturados
Imagens, vídeos, redes sociais

Dados Estruturados

Database

Dados armazenados em sistemas

"BO" (Boletim de Ocorrência)

"Registro Policial"

Região com mais ocorrências de...

SQL com GROUP BY e COUNT(*)

Formato fixo, ideais para consultas e estatísticas oficiais

KEY	DATA
0	
1	
2	
3	
N	

Como garantir que todos os registros de um mesmo tipo de crime estão padronizados?

Qualidade e integridade de dados

Dados Estruturados

O número de ocorrências de um mesmo tipo cresce em períodos específicos?

Séries temporais e sazonalidade

?

Dados não estruturados

Oceano caótico

Image

visão computacional, PLN, reconhecimento de padrões

O valor dos dados não estruturados vem do processamento inteligente e do contexto ético

Como identificar padrões de veículos suspeitos em vídeos de câmeras públicas?

Dados não estruturados

?

Dados Semi-estruturados

DOC

Exigem pré-processamento e parsing

Parte organizada, parte textual

PDF

Complexidade intermediária, pois há dados disponíveis, mas não imediatamente utilizáveis.

O que muda na análise quando o boletim está em PDF e não em um sistema?

Dados Semi-estruturados

É possível cruzar os PDFs de laudos periciais com os dados de ocorrência?

?

Micros Desafios

Estruturados: Campos mal padronizados inflacionam contagens.

Semi: OCR e parsing podem introduzir erros; registre versões.

Não estruturados: Bias e ambiguidade — resultados dependem de contexto.

Exemplos básicos

```sql
SELECT regiao, COUNT(*) AS ocorrencias
FROM boletins
WHERE tipo='assalto' AND date(data_ocorrencia) >= date(?)
GROUP BY regiao
ORDER BY ocorrencias DESC;
```

```
$ ocrmypdf laudo.pdf laudo_ocr.pdf
```

Fontes e tipos de coleta

Onde estão?

Natureza dos dados

Estruturados não e semi

Aquisição

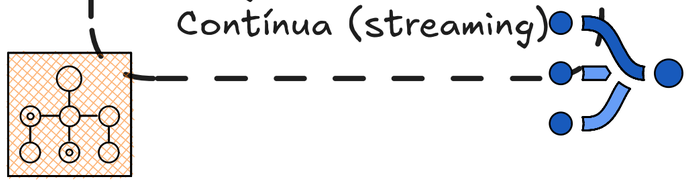
- Manual
- Automatizada
- Assistida por IA

Origem

- Coleta interna
- Coleta externa
- Coleta mista

Periodicidade

- Pontual (eventual)
- Programada (batch)
- Contínua (streaming)

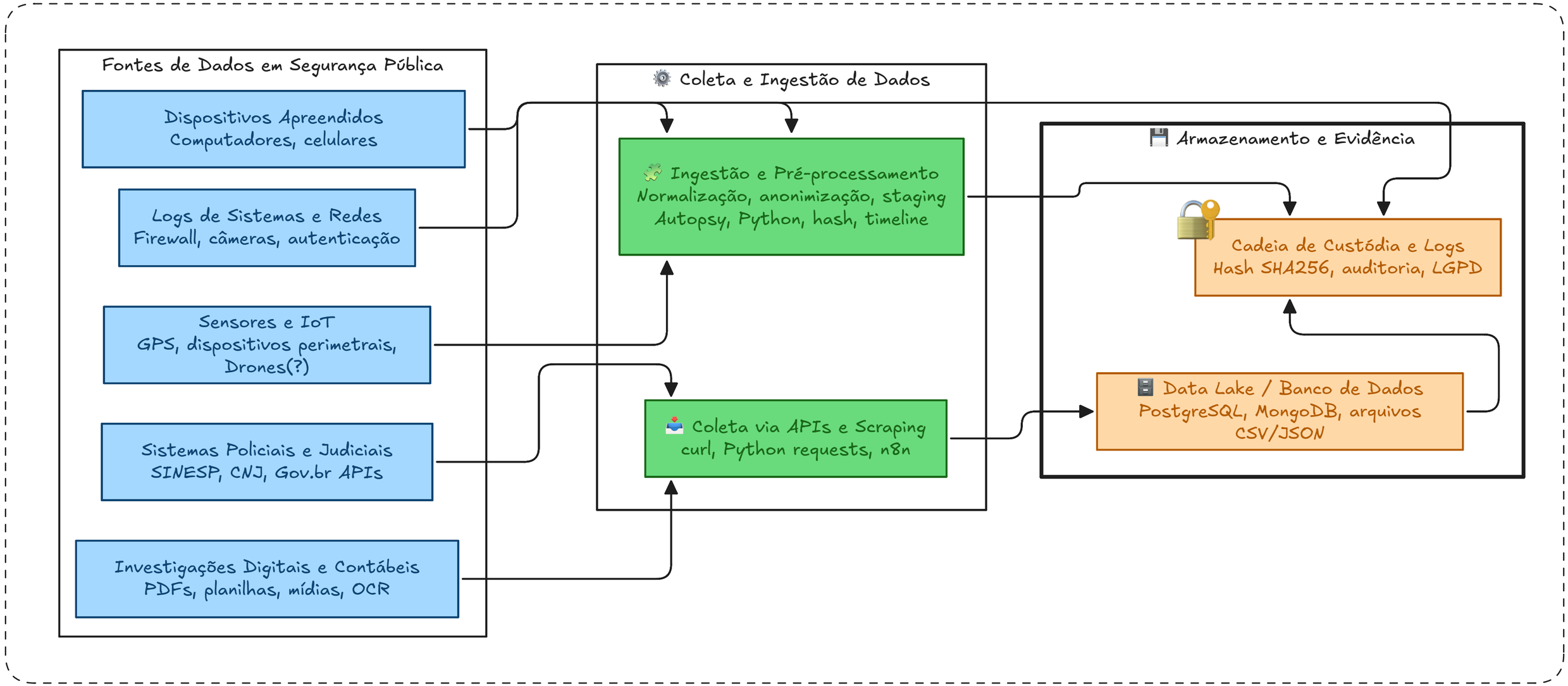


Nível de Intervenção

- Ativa, coletor solicita dado
- Passiva, sistema emite o dado

Logs

Acesso e Autorização


Fontes de Dados em Segurança Pública
Dispositivos Apreendidos
Computadores, celulares
Logs de Sistemas e Redes
Firewall, câmeras, autenticação
Sensores e IoT
GPS, dispositivos perimetrais, Drones(?)
Sistemas Policiais e Judiciais
SINESP, CNJ, Gov.br APIs
Investigações Digitais e Contábeis
PDFs, planilhas, mídias, OCR
Coleta e Ingestão de Dados
Ingestão e Pré-processamento
Normalização, anonimização, staging
Autopsy, Python, hash, timeline
Coleta via APIs e Scraping
curl, Python requests, n8n
Armazenamento e Evidência
Cadeia de Custódia e Logs
Hash SHA256, auditoria, LGPD
Data Lake / Banco de Dados
PostgreSQL, MongoDB, arquivos CSV/JSON

Atividade

Publique no fórum "Tipo de Coleta" um exemplo de sistema:
- Estruturado
- Não estruturado
- Semi estruturado

Tente caracterizar o tipo de coleta possível

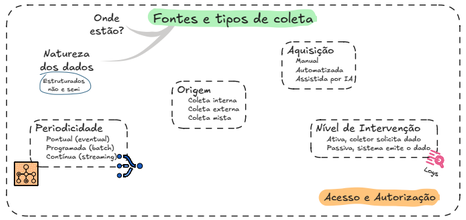
Aquisição
- Manual
- Automatizada
- Assistida por IA

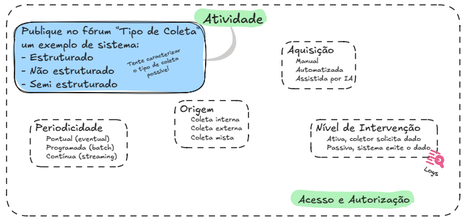
Origem
- Coleta interna
- Coleta externa
- Coleta mista

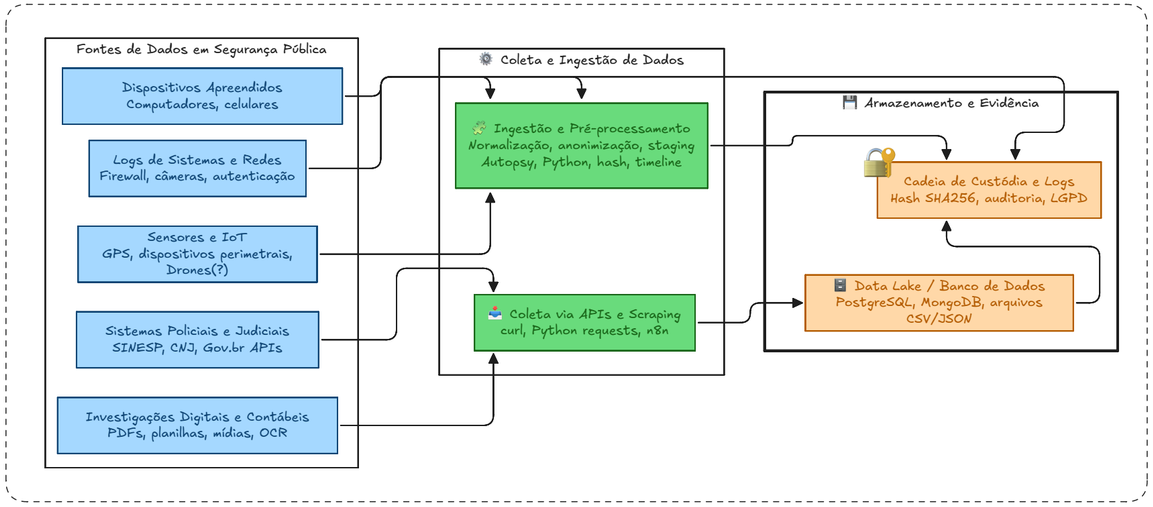
Periodicidade
- Pontual (eventual)
- Programada (batch)
- Contínua (streaming)

Nível de Intervenção
- Ativa, coletor solicita dado
- Passiva, sistema emite o dado

Acesso e Autorização


Onde estão?
Fontes e tipos de coleta
Natureza dos dados
Estruturados não e semi
Aquisição
Manual
Automatizada
Assistida por IA
Origem
Coleta interna
Coleta externa
Coleta mista
Periodicidade
Pontual (eventual)
Programada (batch)
Contínua (streaming)
Nível de Intervenção
Ativa, coletor solicita dado
Passiva, sistema emite o dado
Logs
Acesso e Autorização

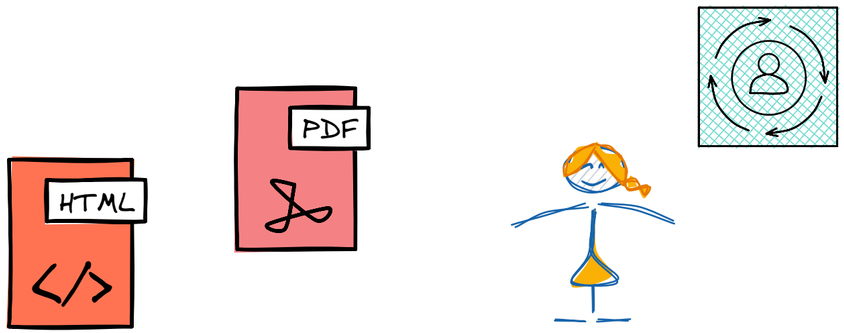

Atividade
Publique no fórum "Tipo de Coleta" um exemplo de sistema:
- Estruturado
- Não estruturado
- Semi estruturado
Tente se aproximar o tipo de coleta possível
Aquisição
Manual
Automatizada
Assistida por IA
Origem
Coleta interna
Coleta externa
Coleta mista
Periodicidade
Pontual (eventual)
Programada (batch)
Contínua (streaming)
Nível de Intervenção
Ativa, coletor solicita dado
Passiva, sistema emite o dado
Logs
Acesso e Autorização


Fontes de Dados em Segurança Pública
Dispositivos Apreendidos
Computadores, celulares
Logs de Sistemas e Redes
Firewall, câmeras, autenticação
Sensores e IoT
GPS, dispositivos perimetrais, Drones(?)
Sistemas Policiais e Judiciais
SINESP, CNJ, Gov.br APIs
Investigações Digitais e Contábeis
PDFs, planilhas, mídias, OCR
Coleta e Ingestão de Dados
Ingestão e Pré-processamento
Normalização, anonimização, staging
Autopsy, Python, hash, timeline
Coleta via APIs e Scraping
curl, Python requests, n8n
Armazenamento e Evidência
Cadeia de Custódia e Logs
Hash SHA256, auditoria, LGPD
Data Lake / Banco de Dados
PostgreSQL, MongoDB, arquivos
CSV/JSON

Raspagem e extração

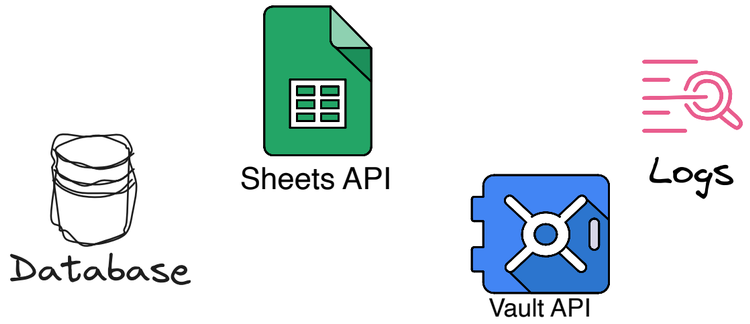
Raspagem (scraping)

Obter dados de interfaces
ou conteúdo publicado



Extração (extraction)

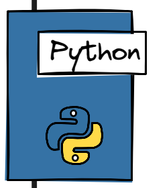
Obter dados diretamente
de uma fonte estruturada



Web Scraping

Ferramentas

requests
BeatifulSoup
Selenium
Scrapy



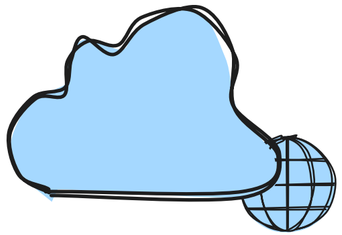
Extração de

Notícias
Portais de transparência
Redes sociais

Estabelecer limites de requisição

Evitar violar Terms of Service ou extrair dados pessoais sensíveis

Manter registro (hash e logs) como parte da cadeia de custódia

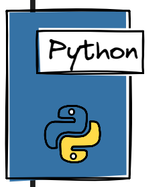


Exemplo: transparência

Extração via API e Integração

Tecnologias

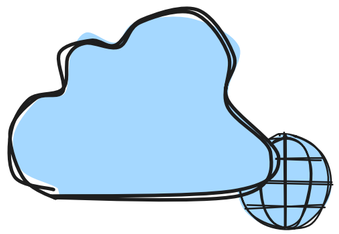
APIs REST/GraphQL
Paginação
Bancos de Dados
Change Data Capture



obtenção de dados de forma estruturada e autenticada

Prática

Integração interna
Portais públicos



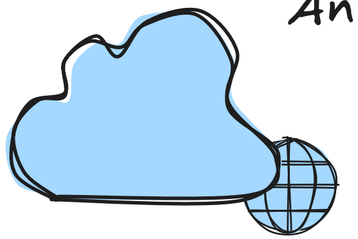
Cuidados éticos e legais

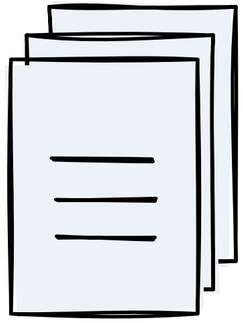
LGPD

Consentimento explícito
Tratamento de dados pessoais

Responsabilidade

Coletar apenas o necessário
Anonimizar dados pessoais





Cadeia de custódia

Registrar origem, data, hora, ferramentas e responsável pela coleta
Calcular hash de cada arquivo para garantir autenticidade

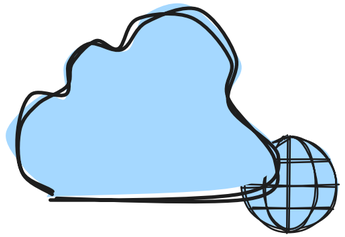
Hash para cadeia de custódia

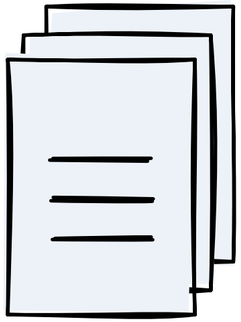
hash

é uma impressão digital de um arquivo

Algoritmo

SHA256: seguro e aceito





Usos principais

Verificar se o arquivo não foi alterado
Registrar no log de custódia
Comprovar autenticidade em auditorias

Exemplo

Raspagem e extração

Raspagem (scraping)

Obter dados de interfaces
ou conteúdo publicado

Extração (extraction)

Obter dados diretamente
de uma fonte estruturada

Web Scraping

Ferramentas

requests
BeatifulSoup
Selenium
Scrapy

Estabelecer limites de requisição

Evitar violar Terms of Service ou extrair dados pessoais sensíveis

Manter registro (hash e logs) como parte da cadeia de custódia

Extração de

Notícias
Portais de transparência
Redes sociais

Exemplo: transparência

Extração via API e Integração

Tecnologias

APIs REST/GraphQL
Paginação
Bancos de Dados
Change Data Capture

obtenção de dados de forma estruturada e autenticada

Prática

Integração interna
Portais públicos

Cuidados éticos e legais

LGPD

Consentimento explícito
Tratamento de dados pessoais

Responsabilidade

Coletar apenas o necessário
Anonimizar dados pessoais

Cadeia de custódia

Registrar origem, data, hora, ferramentas e responsável pela coleta
Calcular hash de cada arquivo para garantir autenticidade

Hash para cadeia de custódia

hash

é uma impressão digital de um arquivo

Algoritmo

SHA256: seguro e aceito

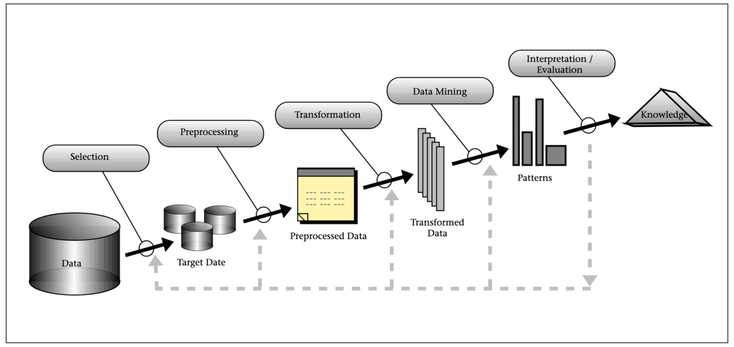
Usos principais

Verificar se o arquivo não foi alterado
Registrar no log de custódia
Comprovar autenticidade em auditorias

Exemplo

Do KDD Clássico aos Processos Modernos de Dados

Da Descoberta de Conhecimento à Engenharia de Dados:
a evolução do ciclo de valor do dado



Fayyad, 1996

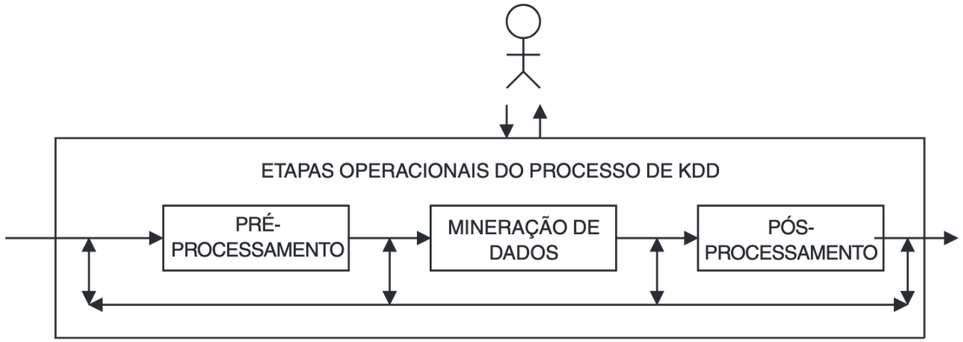


Figura 1.2. *Etapas Operacionais do Processo de KDD.*

Goldschmidt, 2015

Transformar bases estruturadas (bancos relacionais) em conhecimento útil

Contexto original: bases corporativas isoladas, foco em descoberta (insight científico/estatístico)

Transição para o Pipeline Moderno

Evolução tecnológica

Big Data, IoT, logs, streaming e dados não estruturados

Cloud, conteiners, APIs, ambiente distribuídos

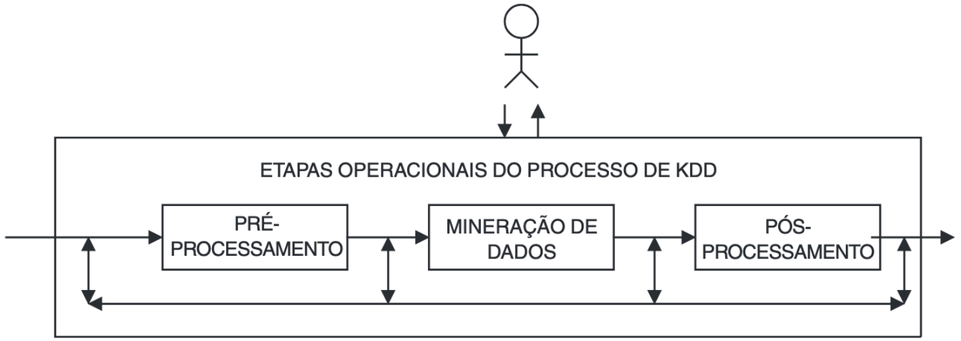


Figura 1.2. *Etapas Operacionais do Processo de KDD.*

Goldschmidt, 2015

Evolução conceitual

"Descobrir conhecimento" para "gerar e sustentar fluxos de dados contínuos e auditáveis"

Evolução ética

"Extrair informação" para "tratar dado como ativo sensível" sujeito à LGPD e à cadeia de custódia

Dado é o produto

Dado como produto

- Unidade de valor, não apenas o resultado de análise
- Deve ser produzido, versionado, validado e distribuído como um artefato
- Em segurança pública, cada registro é um produto com valor operacional e probatório
- Seu uso como prova depende da qualidade do pré-processamento e da confiança no processo de custódia

Implicações práticas

Fontes múltiplas: delegacias, tribunais, dispositivos pessoais, sensores e nuvem.

Necessidade de:

Data lineage: rastreabilidade de ponta a ponta.

Hash e cadeia de custódia digital

Staging areas e pipelines versionados

Governança ética e legal

Cada dado é uma evidência em potencial e, portanto, requer rastreabilidade, integridade e contexto.

KDD x Pipeline Moderno

Aspecto	KDD (anos 1990-2000)	Pipeline Moderno
Foco	Descoberta de Conhecimento	Produção e fluxo contínuo de dados
Estrutura	Sequência linear (offline)	Arquitetura distribuída e contínua
Entrada	Dados estruturados	Dados híbridos e em streaming
Papel humano	Cientista de dados	Engenheiro de dados/analista forense
Controle	Manual	Automatizado e versionado (ETL, DataOps)
Preocupação ética	Mínima	Central (LGPD, auditoria, custódia)
Segurança Pública	Estatístico ou preditivo	Probatório, evidência digital, perícia de dados

Limpeza

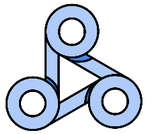
Data cleaning

Remover inconsistência, duplicações

Preencher valores ausentes

Corrigir erros de digitação, codificação e formatos

Converter campos textuais (datas, números)



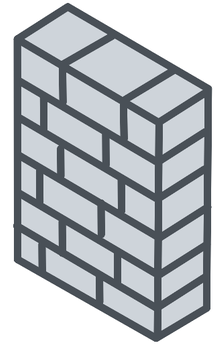
Anonimização e Pseudonimização

Dado

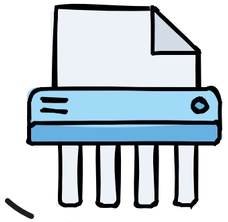
Anonimiado não é pessoal

Mantém a utilidade analítica e investigativa, sem expor dados sensíveis diretamente

Tornar impossível identificar o titular dos dados pessoais



Métodos: masking, hashing, noise addition, k-anonymity, truncamento



Aplicações: dashboards públicos, relatórios interinstitucionais, auditorias.

Normalização e Agregação

Intencionalidade

Ajuste de escala

Converter atributos categóricos para numéricos

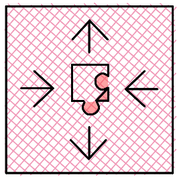
Unificar unidades (ex.: R$ → centavos, km/h → m/s)

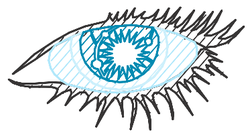
Campos categóricos (ex.: "PM", "Polícia Militar")

Agrupar registros por caso, período, local ou natureza do crime.

Usar funções de agregação (SUM, COUNT, AVG, MIN, MAX)

window functions para séries temporais (ocorrências por hora/dia/semana)

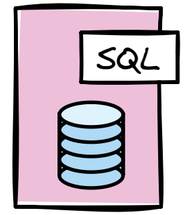




Data Staging Area

onde se verifica a qualidade
(completude, veracidade e integridade)

Finalidade



SQL

Zona de preparação antes da carga no data warehouse ou sistema analítico

Permite integração, limpeza, e auditoria sem afetar dados originais.

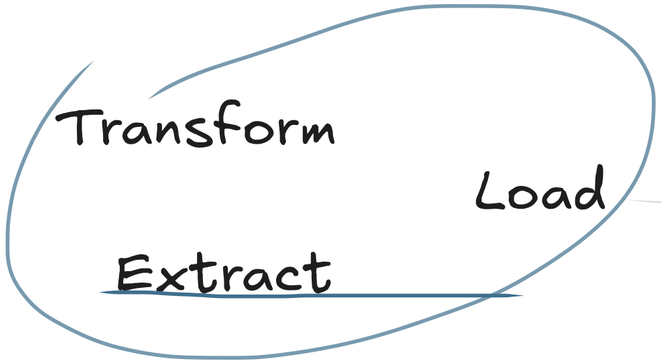
Requisito essencial em cadeias de custódia digitais.


S3 Bucket

Por que manter uma staging area separada do banco operacional?

Chambers & Zaharia (2018), Goldschmidt (2015)

ETL e ELT



elasticsearch

Data Lakes e Data Warehouses na nuvem

Repositório central

Hub de dados

ELT permite que as organizações trabalhem com dados brutos

ELT é ideal quando se deseja ter acesso a todos os dados originais (raw data)

kafka

Arquitetura Híbrida (Data Lakehouse)

Arquitetura	Finalidade na Segurança Pública	Natureza do dado
Data Warehouse (DW)	Relatórios Históricos e BI: Usado para consolidação, auditoria, estatísticas criminais e conformidade legal	Estruturado: Dados pré-processados e altamente transformados, com esquema rígido (schema-on-write)
Data Lake (DL)	Análise Preditiva e Investigativa: Usado para armazenar dados brutos e massivos, como vídeos, logs e mídias sociais, para Machine Learning e análise forense	Qualquer Formato: Não estruturado, semi-estruturado e estruturado, com esquema flexível (schema-on-read)

com dados brutos

seja ter acesso a todos os dados originais (raw data)

Arquitetura Híbrida (Data Lakehouse)

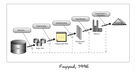
Arquitetura	Finalidade na Segurança Pública	Natureza do dado
Data Warehouse (DW)	Relatórios Históricos e BI: Usado para consolidação, auditoria, estatísticas criminais e conformidade legal	Estruturado: Dados pré-processados e altamente transformados, com esquema rígido (schema-on-write)
Data Lake (DL)	Análise Preditiva e Investigativa: Usado para armazenar dados brutos e massivos, como vídeos, logs e mídias sociais, para Machine Learning e análise forense	Qualquer Formato: Não estruturado, semi-estruturado e estruturado, com esquema flexível (schema-on-read)

Pré-processamento e organização em segurança pública

Evolução

Do KDD Clássico aos Processos Modernos de Dados

Da Descoberta de Conhecimento à Engenharia de Dados:
a evolução do ciclo de valor do dado



Fayyad, 1996

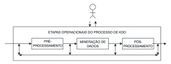


Figura 1.3. *Etapas Operacionais do Processo de KDD.*
Goldschmidt, 2015

Transformar bases estruturadas (bancos relacionais) em conhecimento útil.

Contexto original: bases corporativas isoladas, foco em descoberta (might cienTfico/estatístico)

Transição para o Pipeline Moderno

Evolução tecnológica

Big Data, IoT, logs, streaming e dados não estruturados.
Cloud, containers, APIs, ambiente distribuídos

Evolução conceitual

"Descobrir conhecimento" para "gerar e sustentar fluxos de dados contínuos e auditáveis".

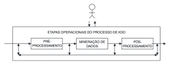


Figura 1.3. *Etapas Operacionais do Processo de KDD.*
Goldschmidt, 2015

Evolução ética

"Extrair informação" para "tratar dado como ativo sensível" sujeito à LGPD e à cadeia de custódia

Importância

Dado é o produto

Dado como produto

- Unidade de valor, não apenas o resultado da análise
- Deve ser produzido, versionado, validado e distribuído como um artefato
- Em segurança pública, cada registro é um produto com valor operacional e probatório
- Seu uso como prova depende da qualidade do préprocessamento e da confiança no processo de custódia

Implicações práticas

Fontes múltiplas: delegacias, tribunais, dispositivos pessoais, sensores e nuvem.

Necessidade de:

- Data lineage: rastreabilidade de ponta a ponta.
- Hash e cadeia de custódia digital
- Staging areas e pipelines versionados
- Governança ética e legal

Cada dado é uma evidência em potencial e, portanto, requer rastreabilidade, integridade e contexto.

KDD x Pipeline Moderno

Aspecto	KDD (anos 1990-2000)	Pipeline Moderno
Foco	Descoberta de Conhecimento	Produção e fluxo contínuo de dados
Estrutura	Sequência linear (offline)	Arquitetura distribuída e contínua
Entrada	Dados estruturados	Dados híbridos e em streaming
Papel humano	Cientista de dados	Engenheiro de dados/analista forense
Controle	Manual	Automatizado e versionado (ETL, DataOps)
Preocupação ética	Mínima	Central (LGPD, auditoria, custódia)
Segurança Pública	Estatístico ou preditivo	Probatório, evidência digital, perícia de dados

Pré-processamento

Limpeza

Data cleaning

Remover inconsistência, duplicações

Preencher valores ausentes

Corrigir erros de digitação, codificação e formatos

Converter campos textuais (datas, números)

Anonimização e Pseudonimização

Mantém a utilidade analítica e investigativa, sem expor dados sensíveis diretamente

Dado
Anonimiado não é pessoal

Tornar impossível identificar o titular dos dados pessoais

Métodos: masking, hashing, noise addition, k-anonymity, truncamento

Aplicações: dashboards públicos, relatórios interinstitucionais, auditorias.

Normalização e Agregação

Intencionalidade

Ajuste de escala

Converter atributos categóricos para numéricos

Agrupar registros por caso, período, local ou natureza do crime.

Unificar unidades (ex.: R$ → centavos, km/h → m/s)

Campos categóricos (ex.: "PM", "Polícia Militar")

Usar funções de agregação (SUM, COUNT, AVG, MIN, MAX)

window functions para séries temporais (ocorrências por hora/dia/semana)

Data Staging Area

onde se verifica a qualidade (completude, veracidade e integridade)

Finalidade

Zona de preparação antes da carga no data warehouse ou sistema analítico

SQL

Permite integração, limpeza, e auditoria sem afetar dados originais.

Requisito essencial em cadeias de custódia digitais.

S3 Bucket

Por que manter uma staging area separada do banco operacional?

Chambers & Zaharia (2018), Goldschmidt (2015)

Como

ETL e ELT

Transform
Load
Extract

elasticsearch

Data Lakes e Data Warehouses na nuvem

ELT permite que as organizações trabalhem com dados brutos

ELT é ideal quando se deseja ter acesso a todos os dados originais (raw data)

Arquitetura Híbrida (Data Lakehouse)

kafka