

82 REGISTROS DE ORIGEM



QUALIDADE E GOVERNANÇA DE DADOS

Uma pessoa, 82 registros

Como um programa estruturado de qualidade de dados — da padronização à governança de metadados — transforma bases fragmentadas em ativos confiáveis para análise e inteligência artificial.

A qualidade de dados deixou de ser um tema técnico. É a condição para que análises, decisões e IA funcionem.

SOBRE ESTE DOCUMENTO

Este white paper consolida aprendizados de projetos recentes da Vert Analytics em organizações de grande porte do setor público brasileiro, com bases que vão de dezenas de milhões a bilhões de registros. Nomes de clientes foram omitidos e dados pessoais foram substituídos por exemplos fictícios.

CONTATO

comercial@vert.com.br
vertanalytics.com.br

- **O custo é alto e já está no orçamento.** Segundo o Gartner, dados de baixa qualidade custam às organizações pelo menos US\$ 12,9 milhões por ano, em média — e, até 2026, 60% dos projetos de IA sem dados preparados devem ser abandonados.
- **O caos é invisível até alguém procurar uma pessoa.** Em uma base com 32,5 milhões de registros de pessoas, um mesmo indivíduo aparecia 82 vezes — com grafias, datas e documentos divergentes. Após padronização, deduplicação e gestão de dados mestres, a base convergiu para 11,9 milhões de pessoas únicas: 63% menos registros.
- **Sem coordenadas, não há análise territorial.** No ponto de partida, 35 milhões de endereços não tinham latitude e longitude — e 85% deles também não tinham CEP. Um pipeline de padronização e geocodificação georreferenciou esse acervo.
- **Nenhuma ferramenta resolve sozinha.** Resultados sustentáveis dependem de seis blocos integrados: padronização, deduplicação, dados mestres, regras de qualidade, monitoramento e governança de dados e metadados.
- **A qualidade decide o que a rede mostra.** Na análise de fraude fiscal por redes de relacionamento, entidades padronizadas revelam articuladores, endereços compartilhados e grupos de fato que cadastros fragmentados escondem.
- **O que não é medido volta a se degradar.** Propomos 15 indicadores — da taxa de duplicidade ao tempo de resolução de incidentes — e um roteiro em ondas, com resultados em semanas, não em anos.

CONTEÚDO

1	Dados ruins são um imposto invisível	3
2	Uma pessoa, 82 registros: a anatomia do caos	4
3	O mapa em branco: enriquecimento geográfico	6
4	Seis blocos de um programa estruturado	7
5	Redes de relacionamento contra a fraude fiscal	14
6	O painel de métricas	17
7	Como implantar: ondas, não big bang	18
8	Cinco lições para líderes	19
	Referências e sobre a Vert Analytics	20

Dados ruins são um imposto invisível

NO SETOR PÚBLICO

Bases de pessoas e endereços sustentam segurança, arrecadação, saúde e mobilidade. A Lei Geral de Proteção de Dados consagra o princípio da *qualidade dos dados*: exatidão, clareza, relevância e atualização.⁶ Um cadastro com dezenas de versões da mesma pessoa torna esse princípio difícil de cumprir.

Ninguém aprova orçamento para “dados duplicados”. Ainda assim, a conta é paga todos os dias. Toda organização que opera em escala acumula dados sobre as mesmas entidades — cidadãos, clientes, endereços, veículos — em sistemas que nasceram em épocas diferentes, com regras e finalidades diferentes. O resultado raramente aparece em um relatório de erros: aparece como um analista que passa a manhã conciliando planilhas, uma investigação que não conecta dois registros da mesma pessoa ou um indicador que conta o mesmo cidadão três vezes.

As evidências de mercado convergem. O Gartner estima que a baixa qualidade de dados custa, em média, pelo menos US\$ 12,9 milhões por ano a cada organização;¹ estimativas reunidas pela *MIT Sloan Management Review* situam esse custo entre 15% e 25% da receita da maioria das empresas.² Na pesquisa global da McKinsey, executivos relataram que, em média, 30% do tempo total das equipes era consumido por tarefas sem valor agregado causadas por dados ruins ou indisponíveis.³ E, em estudo publicado na *Harvard Business Review*, apenas 3% das avaliações de qualidade feitas por gestores sobre dados das próprias áreas foram consideradas aceitáveis.⁴

A inteligência artificial eleva a aposta. Em 2024, 63% das organizações disseram ao Gartner não ter — ou não saber se têm — práticas de gestão de dados adequadas para IA.⁵ Modelos aprendem com o que recebem: dados fragmentados produzem respostas fragmentadas. A qualidade de dados deixou de ser higiene operacional para se tornar condição da estratégia digital.

QUADRO 1

A má qualidade de dados é um custo sistêmico — e a inteligência artificial amplia a conta

US\$ 12,9 mi

custo médio anual da baixa qualidade de dados por organização

Gartner, 2020

30%

do tempo das equipes gasto em tarefas sem valor causadas por dados ruins ou indisponíveis

McKinsey, 2020

3%

das avaliações de qualidade de dados classificadas como aceitáveis

Harvard Business Review, 2017

47%

dos registros recém-criados com ao menos um erro crítico

Harvard Business Review, 2017

63%

das organizações sem práticas de dados adequadas para IA — ou sem saber se as têm

Gartner, 2025

60%

dos projetos de IA sem dados preparados devem ser abandonados até 2026

Gartner, 2025

Fonte: Gartner, pesquisa de 2020 e comunicado de 26 fev. 2025 (levantamento com 248 líderes de gestão de dados, 3º tri. 2024); McKinsey, “Designing data governance that delivers value”, 2020; T. Nagle, T. C. Redman e D. Sammon, *Harvard Business Review*, 2017

Uma pessoa, 82 registros: a anatomia do caos

O *golden record* — a versão única e confiável de cada entidade — não é um luxo de arquitetura. É o que separa uma base de dados de um arquivo morto.

32,5 mi

registros na base central de pessoas no ponto de partida

82

registros diferentes encontrados para um único indivíduo

Em um dos projetos que inspiram este documento, a base central de pessoas reunia 32,5 milhões de registros, alimentados ao longo de anos por sistemas operacionais, cargas legadas e integrações com outras instituições. Uma busca simples por um nome revelou a dimensão do problema: um mesmo indivíduo aparecia em 82 registros diferentes.

Não se tratava de fraude nem de erro isolado. Cada registro estava “correto” do ponto de vista do sistema que o criou; o problema só aparece quando alguém tenta enxergar a pessoa inteira (Quadro 2).

QUADRO 2

Um mesmo indivíduo aparecia em 82 registros, com grafias, datas e documentos divergentes

Amostra ilustrativa de 8 dos 82 registros vinculados a uma única pessoa e o *golden record* resultante

ID ORIGEM	NOME	NASCIMENTO	CPF	NOME DA MÃE	FONTE
301847562	JOSE CARLOS DOS SANTOS	28FEB1983:00:00:00	...482.117-...	MARIA APARECIDA DOS SANTOS	Sistema A
318204977	José Carlos dos Santos	28/02/1983	ausente	MARIA AP DOS SANTOS	Sistema A
296551038	JOSE C. DOS SANTOS	28/02/1983	...482.117-...	ausente	Sistema B
322970145	JOSE CARLOS SANTOS	ausente	...482117...	MARIA APARECIDA SANTOS	Carga legada
305118693	JOSE CARLOS DOS SANTO	1983-02-28	ausente	MARIA APARECIDA DOS SANTOS	Sistema C
327640219	J CARLOS DOS SANTOS	02/28/1983	...482.117-...	ausente	Integração ext.
299083416	JOSÉ CARLOS D. SANTOS	ausente	...428.117-...	MARIA A. DOS SANTOS	Sistema B
334712058	JOSECARLOS DOS SANTOS	28/02/1938	ausente	ausente	Carga legada

... mais 74 registros

M-184221	JOSÉ CARLOS DOS SANTOS	1983-02-28	...482.117-... ✓	MARIA APARECIDA DOS SANTOS	82 vínculos
----------	------------------------	------------	------------------	----------------------------	-------------

☐ Divergência em relação ao golden record

☐ Registro mestre consolidado

DIVERGÊNCIA SINTÁTICA

Acentos, caixa, abreviações, formatos

“José” e “JOSE”, “C.” e “CARLOS”, datas em quatro formatos. Resolvida pela padronização.

INCOMPLETUDE

CPF, data ou nome da mãe ausentes

Cada fonte conhece parte da pessoa. Resolvida pela **consolidação** de atributos entre fontes no golden record.

ERRO DE DIGITAÇÃO

Dígitos trocados, ano invertido, letras faltando

“SANTO”, “1938”, “428” no lugar de “482”. Resolvido por **similaridade probabilística** e curadoria.

Nota: dados fictícios e mascarados para preservar o sigilo; a estrutura e os tipos de divergência reproduzem o caso real. **Fonte:** projeto Vert Analytics

“Cada registro estava correto para o sistema que o criou. O problema só existe quando alguém tenta enxergar a pessoa inteira.”

2,7

registros, em média, para cada pessoa real antes do saneamento

O QUE MUDA

O objetivo não é o volume “removido”, e sim uma camada mestre que aponte, para cada pessoa real, todos os registros que a descrevem — com rastreabilidade até a origem, para desfazer vínculos equivocados e auditar decisões.

DE ONDE VEM A FRAGMENTAÇÃO

O caos é produto de cinco forças que atuam em quase toda organização de grande porte:

- **Ausência de chave comum.** O CPF nem sempre está presente; muitas vezes, a identificação depende só de nome e data de nascimento.
- **Captura sem validação.** Campos livres aceitam abreviações, erros de digitação, espaços extras e acentuação inconsistente.
- **Formatos legados.** Datas armazenadas como texto, documentos com e sem máscara, códigos de município de tabelas diferentes.
- **Integração por cópia.** Cada nova fonte é carregada como um novo lote, sem reconciliação com o que já existe.
- **Correções que não voltam.** Quando um analista corrige um dado, a correção fica no relatório — não na origem.

POR QUE ISSO IMPORTA

Uma base com, em média, 2,7 registros para cada pessoa real distorce tudo o que é construído sobre ela:

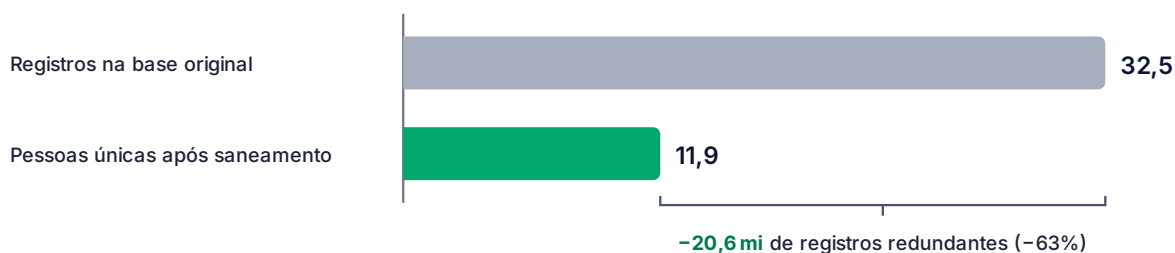
- **Análises incompletas.** Vínculos entre pessoas, endereços e eventos ficam espalhados e deixam de ser vistos.
- **Indicadores inflados.** Contagens de pessoas ou de cobertura de políticas passam a medir registros, não indivíduos.
- **Custo e desempenho.** Cada consulta e cada modelo processa quase três vezes mais linhas do que o necessário.
- **Risco regulatório.** É difícil garantir exatidão e os direitos do titular sobre um dado que existe em 82 versões.

A resposta combinou três blocos do programa descrito na seção 4 — padronização, deduplicação e gestão de dados mestres — e fez a base convergir de 32,5 milhões de registros para 11,9 milhões de pessoas únicas (Quadro 3). O mesmo raciocínio vale para o segundo domínio mais crítico de qualquer cadastro: o endereço.

QUADRO 3

Padronização, deduplicação e gestão de dados mestres reduziram a base de pessoas em 63%

Base central de pessoas, em milhões



Antes do saneamento: 2,7 registros, em média, para cada pessoa real

Nota: valores arredondados; “pessoas únicas” são entidades consolidadas em *golden records*. **Fonte:** projeto Vert Analytics

O mapa em branco: enriquecimento geográfico

Sem latitude e longitude, um registro não entra em mapas de calor nem em cruzamentos por território. Ele simplesmente não existe no mapa.

35 mi

de endereços sem coordenadas geográficas no ponto de partida

85%

deles também sem CEP — cerca de 29,8 milhões

No cenário de partida, 35 milhões de endereços não tinham coordenadas geográficas. O dado mais revelador: cerca de 29,8 milhões deles — 85% — também não tinham CEP, o que inviabiliza o atalho da geocodificação pelo código postal. O problema não era de enriquecimento, mas de qualidade do próprio endereço.

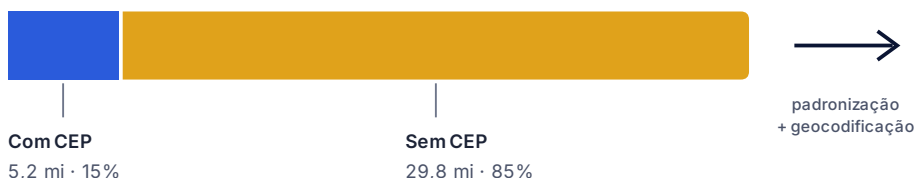
A resposta foi tratar o endereço como dado mestre e combinar padronização com geocodificação em larga escala, por meio da API da Google Maps Platform. O princípio é simples: padronizar antes de geocodificar e validar antes de gravar (Quadro 5). Aplicado sobre todo o acervo, o pipeline georreferenciou os 35 milhões de endereços.

QUADRO 4

85% dos endereços sem coordenadas também não tinham CEP; o enriquecimento georreferenciou os 35 milhões

Endereços sem latitude e longitude

Total: 35 milhões



Sem CEP, a geocodificação pelo código postal não é possível: o endereço precisa ser padronizado.

Nota: valores arredondados. Fonte: projeto Vert Analytics

QUADRO 5

Um pipeline de enriquecimento geográfico robusto tem cinco etapas

01

Normalizar

Separar logradouro, número, complemento, bairro, município e UF; expandir abreviações.

02

Completar

Inferir CEP e código IBGE do município a partir de bases de referência.

03

Geocodificar

Converter o endereço normalizado em latitude e longitude por API de geocodificação.

04

Validar

Conferir se o ponto cai no município declarado e classificar a precisão obtida.

05

Reusar

Guardar em cache o que já foi resolvido e processar só o incremental.

Fonte: Vert Analytics

Seis blocos de um programa estruturado

Uma limpeza pontual melhora a base por alguns meses. Sem regras, donos e monitoramento, a entropia volta.

“Projetos de qualidade de dados falham por dois motivos: começam pela ferramenta ou terminam na limpeza.”

O que funcionou nos projetos que analisamos foi tratar a qualidade como um programa de seis blocos interdependentes, organizados em três camadas. A primeira corrige o dado; a segunda consolida uma versão confiável de cada entidade; a terceira garante que essa confiança se sustente ao longo do tempo.

A ordem importa. Deduplicar sem padronizar produz falsos negativos; consolidar sem regras de sobrevivência explícitas produz um *golden record* que ninguém consegue auditar; e qualquer ganho se perde se ninguém for avisado quando uma carga falha ou um campo muda de formato. As páginas seguintes detalham cada bloco.

QUADRO 6

Um programa estruturado de qualidade de dados articula seis blocos em três camadas

Pergunta que cada bloco responde e principal entregável

CAMADA

Corrigir o dado

Reduz a variabilidade e identifica quem é quem

1 Padronização

O dado está escrito de um jeito só?

ENTREGÁVEL

Regras de normalização versionadas e chaves de comparação

2 Deduplicação

Quais registros descrevem a mesma entidade?

ENTREGÁVEL

Motor de resolução de entidades e clusters auditáveis

CAMADA

Consolidar a verdade

Cria a versão única e rastreável de cada entidade

3 Gestão de dados mestres

Qual é a versão confiável de cada pessoa, endereço ou organização?

ENTREGÁVEL

Golden record, referência cruzada com as origens e curadoria

CAMADA

Sustentar a confiança

Impede que a qualidade se degrade com o tempo

4 Qualidade de dados

O dado atende às regras de negócio?

ENTREGÁVEL

Regras por dimensão e DQ
Score por domínio

5 Monitoramento e observabilidade

Algo mudou e ninguém percebeu?

ENTREGÁVEL

Alertas, incidentes e acordos de nível de serviço

6 Governança e metadados

Quem é dono, o que significa e de onde vem?

ENTREGÁVEL

Catálogo, glossário, linhagem e papéis

Fonte: Vert Analytics, com base no DAMA-DMBOK e em projetos no setor público

Padronização: um jeito só de escrever cada coisa

EM SÍNTESE

Regras de normalização versionadas, aplicadas na entrada e na base histórica, com chaves derivadas (fonética, documento limpo) prontas para o matching.

Padronizar é reduzir a variabilidade sintática antes de comparar registros. Parece trivial, mas é o bloco que mais influencia a precisão da deduplicação: dois registros que diferem apenas por acento, abreviação ou formato de data são, para um algoritmo ingênuo, pessoas diferentes.

Três princípios orientam o trabalho. Primeiro, **regras como código**: cada transformação é documentada, versionada e testada, e pode ser reaplicada quando uma nova fonte entra. Segundo, **preservar o original**: o valor bruto é mantido ao lado do padronizado, para auditoria. Terceiro, **gerar chaves de comparação**: além da forma canônica, o bloco produz chaves derivadas — como a codificação fonética do nome, adaptada ao português — que alimentam a etapa seguinte.

QUADRO 7

A padronização converte variações de escrita em uma única forma canônica, sem perder o valor original

Exemplos de regras por atributo

ATRIBUTO	VALOR NA ORIGEM	FORMA CANÔNICA	REGRA APLICADA
Nome	" José Carlos dos Santos"	JOSE CARLOS DOS SANTOS	Caixa alta, remoção de acentos e espaços duplicados; geração de chave fonética
Data	28FEB1983:00:00:00 · 02/28/1983	1983-02-28	Conversão para ISO 8601; validação de faixa e de dia/mês invertidos
CPF / CNPJ	123.456.789-09 · 12345678909	12345678909	Somente dígitos; validação do dígito verificador; rejeição de sequências repetidas
Endereço	R. Sete de Setembro 1200 ap 3	RUA · SETE DE SETEMBRO · 1200 · AP 3	Separação em componentes; dicionário de abreviações de logradouro
Município	"S. Paulo" · "SAO PAULO-SP"	3550308 (IBGE)	Código IBGE como chave; nome oficial como descrição
Telefone	(61) 9 9999-0000	+5561999990000	Formato internacional E.164; validação de DDD

Nota: valores fictícios. Fonte: Vert Analytics

Na prática, a padronização acontece em dois momentos: uma vez, em lote, sobre o histórico; e continuamente, na entrada de cada nova carga. É a segunda aplicação que impede o problema de voltar — e é também onde a padronização encontra o monitoramento (bloco 5): uma regra que passa a rejeitar muitos valores é, em geral, o primeiro sinal de que um sistema de origem mudou.

Deduplicação: decidir, em escala, quem é quem

>500 tri

de comparações seriam necessárias para confrontar 32,5 milhões de registros todos contra todos

EM SÍNTESE

Motor de resolução de entidades com regras determinísticas e probabilísticas, limiares calibrados com amostras rotuladas e fila de curadoria para os casos ambíguos.

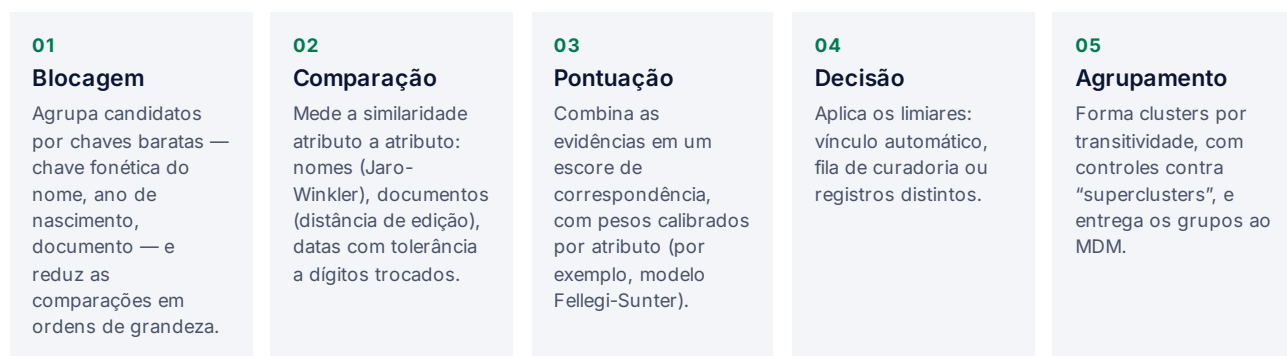
Deduplicar é decidir, par a par, se dois registros descrevem a mesma entidade — e fazer isso em escala. Confrontar 32,5 milhões de registros todos contra todos exigiria mais de 500 trilhões de comparações. Por isso a resolução de entidades é organizada em etapas que reduzem o espaço de busca antes de aplicar a lógica mais cara.

Regras **determinísticas** — por exemplo, CPF válido e data de nascimento idênticos — resolvem com rapidez os casos óbvios. O valor está nos registros sem chave comum, em que a comparação **probabilística** pondera evidências parciais: nomes parecidos, mesma mãe, datas com dígitos trocados. Os limiares de decisão são calibrados com amostras rotuladas por especialistas do negócio, medindo a **precisão** (dos vínculos propostos, quantos estão certos) e a **revocação** (dos vínculos existentes, quantos foram encontrados).

QUADRO 8

A resolução de entidades combina regras determinísticas, similaridade probabilística e revisão humana

Etapas do motor de deduplicação e zonas de decisão por pontuação de similaridade



Zonas de decisão por escore de similaridade (ilustrativo)



Nota: limiares ilustrativos; devem ser calibrados por domínio e revistos periodicamente. **Fonte:** Vert Analytics

Dois cuidados evitam os erros mais caros. O primeiro é controlar a **transitividade**: se A se parece com B e B com C, o agrupamento pode unir A e C mesmo sem evidência direta — e, sem limites, formar “superclusters” que fundem pessoas diferentes. O segundo é tratar **falsos positivos como mais graves que falsos negativos** em contextos sensíveis: unir duas pessoas distintas pode atribuir a alguém um histórico que não é seu.

Gestão de dados mestres: a versão que vale

EM SÍNTESE

Golden record por entidade, tabela de referência cruzada com as origens, regras de sobrevivência por atributo e um processo de curadoria com papéis definidos.

QUAL ESTILO DE MDM?

Do *registro*, que apenas indexa as origens, ao *centralizado*, em que o repositório mestre cria os dados. Para uso analítico, a **consolidação** — *golden record* a jusante, sem alterar as origens — equilibra valor e risco; a **coexistência**, que devolve as correções às origens, fecha o ciclo.

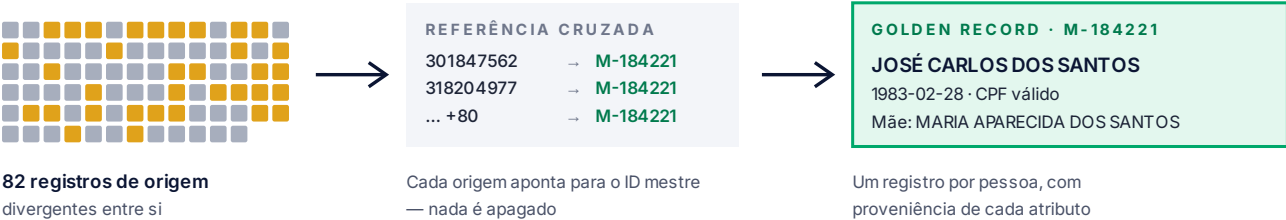
Encontrar duplicatas é metade do trabalho. A outra metade é decidir qual valor prevalece quando os registros discordam — e manter essa decisão rastreável. É o papel da gestão de dados mestres (MDM), que o Gartner define como a disciplina, apoiada por tecnologia, em que negócio e TI atuam juntos para garantir uniformidade, precisão, *stewardship*, governança, consistência semântica e responsabilização sobre os dados mestres oficiais e compartilhados da organização.⁷

Três princípios devem guiar o desenho. **Não apagar a origem:** o *golden record* é uma camada nova, e cada registro original recebe uma referência ao identificador mestre. **Sobrevivência por atributo, não por registro:** cada campo do registro mestre é escolhido por uma regra explícita, que registra de onde o valor veio. **Curadoria como processo:** casos ambíguos seguem para *stewards* de negócio, cujas decisões realimentam as regras.

QUADRO 9

Regras de sobrevivência definem, atributo a atributo, qual valor compõe o *golden record*

Da origem ao registro mestre, com proveniência de cada valor



ATRIBUTO	REGRA DE SOBREVIVÊNCIA	VALOR NO GOLDEN RECORD	PROVENIÊNCIA
Nome	Forma mais completa entre as fontes de maior confiabilidade	JOSÉ CARLOS DOS SANTOS	Sistema A
Nascimento	Valor válido mais frequente entre os registros vinculados	1983-02-28	Maioria das fontes
CPF	Dígito verificador válido e confirmado por mais de uma fonte	...482.117-...	Sistemas A e B
Nome da mãe	Forma mais completa, sem abreviações	MARIA APARECIDA DOS SANTOS	Sistema A
Endereço	Mais recente com georreferenciamento de alta precisão	Logradouro + lat/long	Integração externa
Telefone	Mais recente com formato válido	+55 610000	Sistema C

Nota: dados fictícios. Fonte: Vert Analytics

Qualidade de dados: regras de negócio, medidas por dimensão

EM SÍNTESE

Catálogo de regras por dimensão, associadas a tabelas e colunas, com histórico de execuções, evidências e um índice de qualidade (DQ Score) por domínio.

Com a base consolidada, o desafio muda: impedir que ela se degrade. Isso exige regras explícitas, escritas na linguagem do negócio e executadas automaticamente. Organizá-las por dimensão de qualidade permite comparar domínios, priorizar correções e comunicar a situação à liderança em um único índice.

O Gartner define qualidade de dados como a usabilidade e a aplicabilidade dos dados para os casos de uso prioritários da organização — incluindo IA — e recomenda uma abordagem guiada pelo negócio, em vez de iniciativas corporativas sem priorização.¹ Nos projetos que analisamos, as regras foram organizadas em dez dimensões (Quadro 10). Exemplos reais incluem a validação de CNPJ, a unicidade de inscrições cadastrais e a verificação de que códigos de situação pertencem ao domínio permitido.

QUADRO 10

Dez dimensões organizam as regras de qualidade — cada uma responde a uma pergunta de negócio

#	DIMENSÃO	PERGUNTA DE NEGÓCIO	EXEMPLO DE REGRA
01	Acurácia	O dado corresponde à realidade?	Endereço geocodificado cai no município declarado
02	Compleitude	Os campos obrigatórios estão preenchidos?	Percentual de pessoas com data de nascimento
03	Consistência	O mesmo fato tem o mesmo valor entre sistemas?	Nascimento igual no golden record e nas fontes vinculadas
04	Atualidade	O dado está atualizado para o uso?	Carga diária concluída no horário acordado
05	Validade	O valor respeita formato e domínio?	CNPJ com dígito verificador válido; situação cadastral em lista permitida
06	Unicidade	Cada entidade aparece uma única vez?	Inscrição cadastral sem valores repetidos
07	Conformidade	O dado segue o padrão definido?	Datas em ISO 8601; município em código IBGE
08	Integridade	As relações entre tabelas se sustentam?	Todo evento referencia uma pessoa existente
09	Precisão	O nível de detalhe é suficiente para o uso?	Coordenadas com precisão de número de porta
10	Acessibilidade	Quem precisa consegue acessar, com segurança?	Tabela crítica com responsável e política de acesso

COMO CALCULAR O DQ SCORE

Um índice por domínio, comparável ao longo do tempo

$$DQ\ Score = \Sigma (peso_d \times taxa\ de\ aprovação_d) \div \Sigma peso_d$$

Para cada dimensão d , a taxa de aprovação é a proporção de registros que passam nas regras daquela dimensão. Os pesos refletem o uso: em um cadastro de pessoas, unicidade e validade tendem a pesar mais que precisão. O índice é publicado por domínio e acompanhado em série histórica — o que importa é a tendência, não o valor isolado.

Fonte: Vert Analytics; o Gartner utiliza um conjunto semelhante de nove dimensões (acessibilidade, acurácia, completude, consistência, precisão, relevância, atualidade, unicidade e validade)

Monitoramento e observabilidade: saber antes do usuário

4 bi+

de linhas em uma única tabela perfilada continuamente, com 121 colunas monitoradas

EM SÍNTESE

Métricas de atualidade, volume e esquema habilitadas, detecção de anomalias, incidentes com responsável e visão executiva do estado dos dados.

Uma regra avaliada uma vez é uma fotografia; o valor está no filme. A observabilidade de dados acompanha continuamente os sinais vitais de cada tabela e pipeline, detecta desvios e aciona o responsável antes que o problema chegue a um painel, a um relatório oficial ou a um modelo.

Nos ambientes que analisamos, a camada de observabilidade perfila tabelas com mais de 4 bilhões de linhas e registra, a cada execução, contagens de nulos, distintos e valores únicos por coluna. Violações de limites geram incidentes e tarefas automáticas, com histórico de evidências e ações de correção. Alertas por severidade — por exemplo, a exclusão de uma tabela ou a alteração de uma *view* — notificam diretamente o *data owner* e os grupos que consomem o ativo, e podem abrir chamados no sistema de gestão de incidentes da organização por meio de APIs.

QUADRO 11

Quatro sinais que todo pipeline crítico deveria emitir — e o ciclo que transforma um alerta em correção

ATUALIDADE

A tabela foi atualizada quando deveria?

Horário da última carga comparado ao acordo de nível de serviço do dado.

VOLUME

O número de linhas está dentro do esperado?

Quedas ou saltos em relação à série histórica indicam carga parcial ou duplicada.

ESQUEMA

Alguma coluna mudou de nome, tipo ou sumiu?

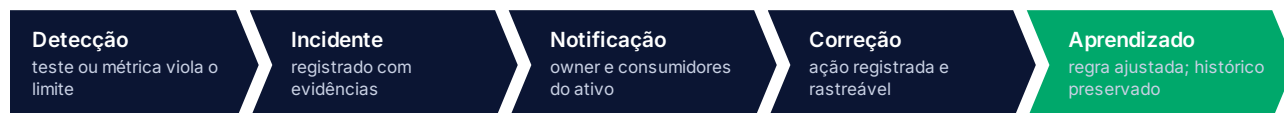
Mudanças estruturais quebram pipelines a jusante sem gerar erro na origem.

DISTRIBUIÇÃO

Os valores continuam com o perfil de sempre?

Proporção de nulos, distintos e valores fora do domínio, coluna a coluna.

Ciclo de um incidente de dados



Fonte: Vert Analytics

O monitoramento também alimenta a própria melhoria do programa. Em um dos projetos, a qualidade foi aplicada primeiro a um conjunto limitado de tabelas cadastrais, justamente para entender o consumo de infraestrutura, o comportamento das métricas em atualizações incrementais e os ajustes necessários antes de expandir para outras fontes.

NA PRÁTICA: PERFIL DE UMA COLUNA IDENTIFICADORA EM UMA TABELA MONITORADA

4,08 bi

linhas na tabela

121

colunas perfiladas

539 mi

valores distintos na coluna

0

valores nulos

100%

de amostra no perfil

Série diária acompanhada em janela de sete dias; desvios nessas contagens podem disparar alertas. Fonte: projeto Vert Analytics

Governança de dados e metadados: quem é dono, o que significa, de onde vem

130+

conectores nativos na plataforma de metadados de código aberto utilizada

8

pipelines prioritários no escopo inicial de linhagem inferida por IA, a um ritmo de dois por mês

O sexto bloco sustenta os demais. Sem saber quem responde por uma tabela, o que cada campo significa e de onde o dado vem, nenhuma regra de qualidade sobrevive à primeira reestruturação de sistemas. O DAMA-DMBOK, referência internacional da disciplina, trata governança e gestão de metadados como funções centrais da gestão de dados.⁹

Na prática, o bloco combina cinco capacidades: um **catálogo** pesquisável de bancos, esquemas, tabelas, colunas e pipelines, atualizado por agentes de ingestão automáticos; **linhagem de ponta a ponta**, em nível de tabela e de coluna, que apoia análise de impacto e resposta a incidentes; **glossário e classificação**, com termos de negócio e marcações como dado pessoal ou sensível, que apoiam a LGPD; **papéis** — *owners*, *stewards* e domínios — com políticas de governança operacional; e **colaboração** em torno de cada ativo, com solicitações de descrição e anúncios de mudança.

A plataforma adotada foi o OpenMetadata, de código aberto, que reúne descoberta, linhagem, qualidade, observabilidade e governança em um único ambiente.⁸ Um obstáculo comum é o legado: nem todo banco tem conector capaz de extrair linhagem. A solução foi usar modelos de linguagem executados em um **ambiente de IA privada** — cluster Kubernetes com GPUs dedicadas, dentro do perímetro da organização — para inferir a linhagem a partir do código dos pipelines, sem expor dados sensíveis.

QUADRO 12

Uma arquitetura de referência integra catálogo, qualidade, observabilidade e IA privada em um só ambiente



Nota: no projeto de referência, a plataforma de metadados somou 7 nós, 22 vCPUs e 104 GB de memória, com busca em cluster de três nós. **Fonte:** Vert Analytics

Redes de relacionamento: onde a qualidade de dados vira inteligência fiscal

Na administração tributária, a fraude raramente está em um cadastro. Está nos vínculos — e vínculos só se formam entre entidades bem identificadas.

VOCABULÁRIO DE REDES

Nó: uma entidade — pessoa, empresa, endereço, contato.

Vínculo: uma relação societária, cadastral ou transacional.

Grau: quantos vínculos um nó tem.

Intermediação: quanto um nó está no caminho entre os demais — as “pontes”.

Comunidade: grupo densamente conectado.

Esquemas de fraude estruturada contra o fisco — empresas “noteiras” que emitem documentos sem operação real, interpostas pessoas no quadro societário, grupos econômicos de fato que abrem e fecham empresas para escapar de débitos, circulação simulada de mercadorias para gerar créditos indevidos de ICMS — são, por natureza, fenômenos de rede. Nenhum registro isolado parece suspeito; o padrão só aparece quando se conectam sócios, endereços, contadores, telefones e fluxos de notas fiscais.

A análise de redes de relacionamento transforma essas conexões em evidência. Métricas de grafo como o grau, a intermediação¹⁰ e a detecção de comunidades¹¹ apontam articuladores, pontes entre grupos e núcleos densos que merecem auditoria (Quadro 13). Mas a rede é tão boa quanto a resolução das entidades que a compõem. Sobre cadastros não padronizados, ela erra de duas maneiras opostas:

- **Fragmentação: o articulador desaparece.** Se o mesmo sócio aparece como três nós, sua centralidade é dividida e ele se perde no ruído. Endereços escritos de formas diferentes nunca formam o vínculo “mesmo endereço” (Quadro 14).
- **Fusão indevida: surgem vínculos que não existem.** Se homônimos são unidos sem critério, um contribuinte regular passa a parecer parte de um esquema — erro que compromete a sustentação técnica de uma autuação.

QUADRO 13

Cinco métricas de rede — e como a má qualidade de dados distorce cada uma

GRAU

Concentração

Quantos vínculos um nó tem. Duplicatas dividem o grau: o sócio de 12 empresas parece sócio de 4.

INTERMEDIAÇÃO

Pontes

Quanto um nó está no caminho entre grupos. Fragmentos cortam os caminhos e a ponte some.

COMUNIDADES

Grupos de fato

Núcleos densamente conectados. Vínculos faltantes partem um grupo em vários menores.

CICLOS

Circularidade

Notas que voltam à origem. CNPJs não normalizados quebram o ciclo e escondem o carrossel.

COMPONENTES

Fragmentação

Partes desconectadas. Muitos componentes pequenos são, em si, sinal de baixa qualidade.

Fonte: Vert Analytics

“A qualidade de dados não é um pré-requisito burocrático da análise de fraude. É o que determina se a rede mostra o esquema ou o esconde.”

O MESMO ESQUEMA, DUAS REDES

O Quadro 14 reproduz, com dados fictícios, um padrão recorrente em investigações fiscais: cinco empresas, uma interposta pessoa, um contador e um sócio que aparece com grafias diferentes em cada cadastro. Sobre os registros brutos, a rede mostra dois grupos sem relação e nenhum nó que se destaque. Depois da padronização, da deduplicação e do georreferenciamento, os três fragmentos do sócio viram um único nó, as três grafias do endereço viram um único imóvel e as duas variações do contador viram um só escritório.

Nenhum vínculo novo foi inventado: são os mesmos 13 relacionamentos. O que muda é que eles passam a apontar para as entidades certas — e o articulador, antes invisível, torna-se o nó de maior intermediação da rede.

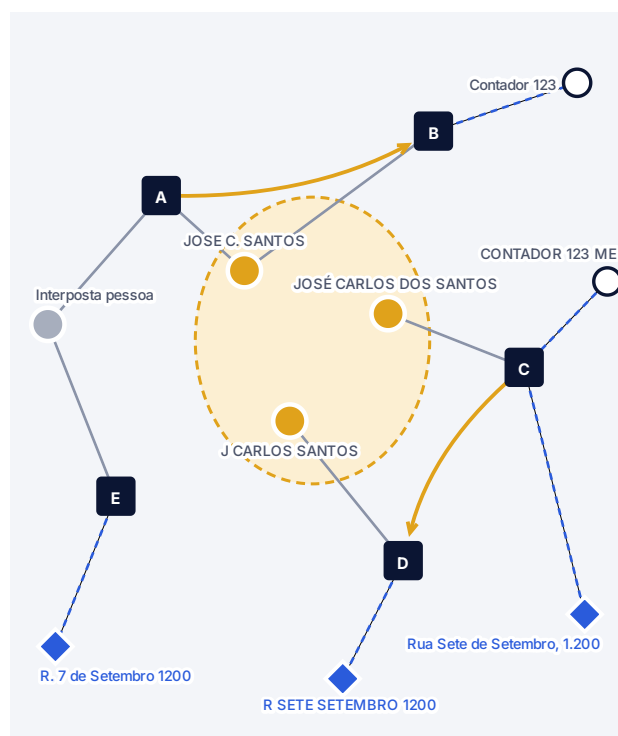
QUADRO 14

Sem resolução de entidades, a rede esconde o articulador; com entidades padronizadas, o esquema aparece

Os mesmos vínculos, antes e depois da padronização, da deduplicação e do georreferenciamento — exemplo ilustrativo

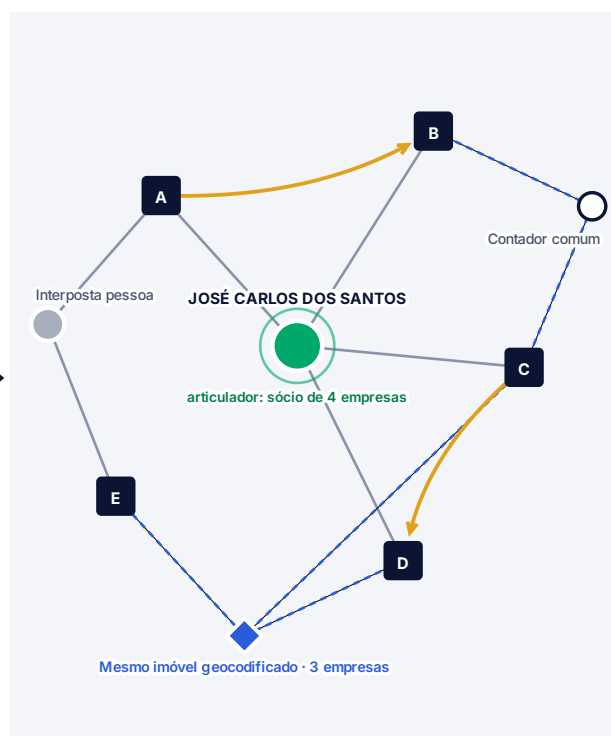
ANTES: CADASTROS BRUTOS

14 nós · 2 grupos desconectados · o sócio aparece em 3 nós



DEPOIS: ENTIDADES RESOLVIDAS

9 nós · 1 rede conectada



● Pessoa ● Fragmento da mesma pessoa ● Pessoa resolvida ■ Empresa ◆ Endereço ○ Contador — Vínculo societário
- - - Vínculo cadastral → Nota fiscal emitida

14 → 9

nós na rede, com os mesmos 13 vínculos

2 → 1

componente: os dois grupos eram um só

2 → 4

empresas ligadas ao sócio (maior fragmento x entidade)

0 → 0,21

intermediação do sócio — a maior da rede

3 → 1

endereço, agora compartilhado por 3 empresas

Nota: dados fictícios; métricas calculadas sobre o próprio exemplo (intermediação normalizada). Fonte: Vert Analytics

NA PRÁTICA

Os vínculos são configurados sobre identificadores mestres, nunca sobre registros de origem. Assim, cada nova carga atualiza a rede sem retrabalho, e cada decisão de curadoria no MDM se propaga automaticamente para o grafo.

POR QUE A QUALIDADE FACILITA A CONFIGURAÇÃO DOS VÍNCULOS

Configurar uma rede é escrever regras: que tipos de entidade existem, que relação forma um vínculo, com que peso e por quanto tempo ela vale. Com entidades padronizadas e um identificador mestre para cada uma, cada regra se reduz a uma junção exata sobre chaves canônicas — determinística, auditável e reaplicável a cada nova carga. Sem isso, cada vínculo depende de comparações aproximadas feitas caso a caso, a cada investigação, por cada analista.

Na prática, os blocos do programa se traduzem diretamente em tipos de vínculo (Quadro 15): o *golden record* garante um nó por pessoa; o endereço padronizado e geocodificado cria o vínculo “mesmo imóvel”; documentos validados ligam sócios, contadores e empresas; e datas padronizadas dão à rede uma dimensão temporal.

QUADRO 15

Cada tipo de vínculo depende de um atributo — e da qualidade dele

VÍNCULO	O QUE PODE REVELAR	SEM QUALIDADE DE DADOS	COM QUALIDADE DE DADOS
Societário pessoa ↔ empresa	Interpostas pessoas, sócios ocultos, grupos de fato	CPF ausente ou mal digitado; o mesmo sócio vira vários nós	CPF validado e <i>golden record</i> : um nó por pessoa
Endereço comum empresa ↔ imóvel	Empresas de fachada no mesmo imóvel; endereços de aluguel	“R. 7 de Setembro 1200” e “Rua Sete de Setembro, 1.200” não se encontram	Endereço padronizado e geocodificado: mesmo ponto, mesmo nó
Contato comum telefone, e-mail	O mesmo operador por trás de várias empresas	Formatos distintos, DDD ausente, e-mails com variações	Telefone em E.164 e e-mail normalizado
Responsável contábil empresa ↔ contador	Escritórios que concentram empresas de risco	Nome do escritório com variações e sem identificador	Contador identificado por documento e registro profissional
Transacional emitente → destinatário	Circularidade de notas; “carrossel” de créditos	Matriz e filiais tratadas como entidades sem relação	CNPJ normalizado, com a raiz ligando matriz e filiais
Temporal vigência do vínculo	Sucessão de empresas com os mesmos sócios e endereço	Datas de entrada e saída ausentes ou em formatos diversos	Vínculos com início e fim: a rede reconstituída em qualquer data

Fonte: Vert Analytics

EXEMPLO: TRÊS REGRAS DE VÍNCULO ESCRITAS SOBRE ENTIDADES PADRONIZADAS

mesmo_socio	Pessoa (ID mestre) ↔ empresa (CNPJ raiz) , quando o CPF validado consta do quadro societário. Vigência: datas de entrada e saída do sócio.
mesmo_imovel	Empresa ↔ empresa , quando o ponto geocodificado é o mesmo, com precisão de número e mesmo complemento. Peso maior se os períodos de atividade se sobrepõem.
mesmo_contato	Empresa ↔ empresa , quando o telefone em E.164 ou o e-mail normalizado coincidem — excluídos números de centrais e de escritórios contábeis já conhecidos, para evitar falsos vínculos.

Cada regra é uma junção exata sobre chaves canônicas; sem padronização, exigiria comparação aproximada, com resultados que mudam de analista para analista.

Há, ainda, uma exigência que só a governança atende: cada vínculo precisa carregar sua proveniência — a fonte, a regra que o formou e a data. Catálogo e linhagem (bloco 6) tornam o grafo auditável, condição para que uma evidência de rede sustente uma ação fiscal.

O painel de métricas

O que não é medido volta a se degradar. O painel precisa ser simples o bastante para a liderança acompanhar e detalhado o bastante para orientar a operação. Propomos 15 indicadores em quatro famílias, alinhados aos seis blocos do programa. Os de consolidação e qualidade mostram o estado do dado; os de monitoramento e governança mostram se a organização é capaz de mantê-lo. Onde o caso descrito neste documento oferece um resultado, ele aparece na última coluna; nos demais, indicamos metas de partida, a serem calibradas por domínio.

QUADRO 16

Quinze indicadores acompanham o estado do dado e a capacidade de mantê-lo

INDICADOR	COMO CALCULAR	RESULTADO NO CASO · META DE PARTIDA
CONSOLIDAÇÃO DE ENTIDADES		
Taxa de duplicidade	Registros de origem ÷ entidades únicas	2,7 → 1,0 registro por pessoa
Redução de redundância	(Registros – entidades únicas) ÷ registros	63% (32,5 mi → 11,9 mi)
Precisão e revocação do matching	Vínculos corretos ÷ propostos; vínculos encontrados ÷ existentes, em amostra auditada	Meta: ≥ 95% de precisão em domínios sensíveis
Fila de curadoria	Casos pendentes e tempo médio até a decisão do <i>steward</i>	Meta: fila estável; decisão em até 5 dias úteis
QUALIDADE DO DADO		
DQ Score	Média ponderada da taxa de aprovação das regras, por dimensão e domínio	Meta: tendência de alta, com piso por domínio
Completeness de atributos críticos	% de registros com CPF, nascimento e nome da mãe preenchidos	Meta definida por atributo e por uso
Validade de documentos	% de CPF e CNPJ com dígito verificador válido	Meta: 100% dos preenchidos
Cobertura geográfica	% de endereços com latitude e longitude e precisão mínima de logradouro	35 mi de endereços georreferenciados
MONITORAMENTO E OBSERVABILIDADE		
Cobertura de testes	% de tabelas críticas com regras de qualidade ativas	Meta: 100% das tabelas críticas
Taxa de sucesso dos testes	Testes aprovados ÷ testes executados, por período	Acompanhar tendência e reincidência
SLA de atualidade	% de cargas concluídas dentro do horário acordado	Meta: ≥ 99% em pipelines críticos
MTTD e MTTR	Tempo médio até detectar e até resolver incidentes de dados	Meta: detecção antes do consumidor
GOVERNANÇA E METADADOS		
Cobertura do catálogo	% de ativos críticos catalogados com descrição de negócio	Meta: 100% dos ativos críticos
Cobertura de responsáveis	% de ativos críticos com <i>owner</i> e <i>steward</i> nomeados	Meta: 100% dos ativos críticos
Cobertura de linhagem	% de pipelines prioritários com linhagem de ponta a ponta	8 pipelines no escopo inicial

Nota: metas de partida sugeridas pela Vert Analytics; devem ser calibradas conforme criticidade e uso de cada domínio. Fonte: Vert Analytics

Como implantar: ondas, não big bang

10–20%

dos dados costumam ser elementos críticos, segundo a McKinsey — é por eles que se começa

Programas de dados fracassam quando tentam abraçar a organização inteira de uma vez. A McKinsey recomenda concentrar o esforço inicial em dois ou três domínios prioritários e identificar os elementos de dados críticos — tipicamente de 10% a 20% do total.³ Nossa experiência confirma: começar por pessoas e endereços, os domínios que mais alavancam análises, gera resultados visíveis rapidamente e forma a coalizão necessária para as ondas seguintes.

O Quadro 17 mostra o plano de implantação da plataforma de metadados em um dos projetos: 31 semanas, com frentes sobrepostas. A maturidade evolui em níveis (Quadro 18); o primeiro ciclo mira sair do estágio inicial e chegar a processos definidos.

QUADRO 17

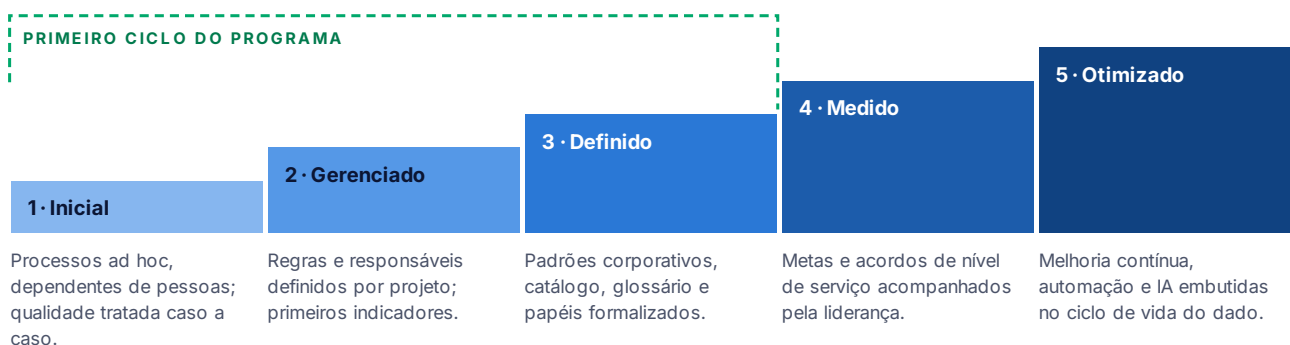
A implantação da plataforma de metadados foi planejada em 31 semanas, com frentes sobrepostas



Fonte: plano de projeto Vert Analytics

QUADRO 18

A maturidade em governança de dados evolui em cinco níveis; o primeiro ciclo leva a organização do nível 1 ao 3



Nota: o DAMA-DMBOK não prescreve uma curva de maturidade; a escala é uma forma prática de estruturar o programa. Fonte: Vert Analytics

Cinco lições para líderes

O maior desafio dos projetos que analisamos não foi técnico: foi transformar definições de negócio em regras executáveis.

1 Comece pelo uso, não pela ferramenta.

Escolha dois ou três casos de uso com dono claro e valor mensurável — uma investigação mais rápida, um indicador confiável, um modelo em produção. A tecnologia vem depois.

2 Regras de qualidade são decisões de negócio.

Quem define se um registro está “certo” é quem usa o dado. Um piloto com tabelas escolhidas junto às áreas de negócio destrava o processo.

3 Não apague: vincule.

O *golden record* deve ser uma camada rastreável sobre as origens — condição para auditoria, correção de vínculos e conformidade com a LGPD.

4 Automatize para replicar.

Templates de ingestão, padrões de regras e estruturas de governança reutilizáveis incorporam novas fontes com previsibilidade, em boa parte pela interface, sem código.

5 Use a IA como acelerador — dentro de casa.

Modelos de linguagem em ambiente privado preenchem lacunas de conectores, sugerem descrições e aceleram a curadoria, sem que dados sensíveis deixem a organização.

PERGUNTAS QUE A LIDERANÇA DEVERIA FAZER

- 1 Quantos registros, em média, temos para cada cidadão ou cliente real?
- 2 Que porcentagem dos nossos endereços pode ser colocada em um mapa hoje?
- 3 Quem responde, nominalmente, por cada uma das nossas 20 tabelas mais críticas?
- 4 Se um sistema de origem mudar um campo hoje, quando saberemos — e quem será avisado?
- 5 Nossos dados estão prontos para os casos de uso de IA que planejamos?

Qualidade de dados não é um projeto com data de término, mas uma capacidade organizacional — como segurança da informação ou gestão financeira. O caminho é conhecido, os componentes são maduros — muitos de código aberto — e os primeiros resultados aparecem em semanas. A pessoa com 82 registros existe em quase toda grande base de dados. A diferença está em quem decide procurá-la.

REFERÊNCIAS

- 1 Gartner. "Data quality", página temática de Data & Analytics, com base em pesquisa Gartner de 2020. gartner.com/en/data-analytics/topics/data-quality.
- 2 Redman, T. C. "Seizing opportunity in data quality". *MIT Sloan Management Review*, 2017.
- 3 Petzold, B.; Roggendorf, M.; Rowshankish, K.; Sporleder, C. "Designing data governance that delivers value". McKinsey & Company, 26 jun. 2020. Dados da McKinsey Global Data Transformation Survey 2019.
- 4 Nagle, T.; Redman, T. C.; Sammon, D. "Only 3% of companies' data meets basic quality standards". *Harvard Business Review*, 11 set. 2017.
- 5 Gartner. "Lack of AI-ready data puts AI projects at risk". Comunicado à imprensa, 26 fev. 2025.
- 6 Brasil. Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais), art. 6º, inciso V.
- 7 Gartner. "Master data management (MDM)". *Gartner Information Technology Glossary*.
- 8 OpenMetadata. Documentação oficial. docs.open-metadata.org.
- 9 DAMA International. *DAMA-DMBOK: Data Management Body of Knowledge*. 2ª ed., revisão de 2024. Disponível em português do Brasil.
- 10 Freeman, L. C. "A set of measures of centrality based on betweenness". *Sociometry*, v. 40, n. 1, p. 35-41, 1977.
- 11 Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment*, P10008, 2008.

METODOLOGIA

Os números de casos apresentados neste documento vêm de projetos recentes conduzidos pela Vert Analytics em organizações de grande porte do setor público brasileiro. Para preservar o sigilo, clientes não são identificados, dados pessoais foram substituídos por exemplos fictícios e valores foram arredondados. Estatísticas de mercado são atribuídas às fontes originais listadas acima.

SOBRE A VERT



A Vert Analytics é uma empresa brasileira especializada em dados, analytics e inteligência artificial, com atuação em projetos de grande escala nos setores público e privado — de plataformas analíticas e governança de dados a soluções de IA aplicada.

Contato: comercial@vert.com.br

vertanalytics.com.br



vertanalytics.com.br
comercial@vert.com.br