Unitins – Sede Administrativa – Qd. 108 Sul, Alameda 11, lote 03 – CEP 77020-122 | www.unitins.br

UNITINS

Ingestão e Coleta de Dados em Segurança Pública

Fundamentos, Arquiteturas, Governança e Ética

Marco Antonio Firmino de Sousa
marco.af@unitins.br

Outubro de 2025

Conteúdo

21 de outubro de 2025

Resumo

Este texto básico aborda os desafios técnicos, éticos e legais envolvidos na coleta, ingestão e processamento de dados aplicados à segurança pública. Explora a natureza dos dados estruturados e não estruturados, destacando suas particularidades e as exigências para garantir qualidade, integridade e conformidade com a Lei Geral de Proteção de Dados (LGPD). São discutidas as fontes principais de dados, como sistemas policiais, judiciais, redes sociais e plataformas digitais, evidenciando a necessidade de arquiteturas flexíveis que integrem diferentes formatos e volumes informacionais. Apresenta-se uma análise detalhada dos modos de ingestão, incluindo processamento em lotes e contínuo, enfatizando a importância da automação e do uso de agentes inteligentes para otimizar pipelines, assegurar governança e observabilidade. As arquiteturas Lambda e Kappa são examinadas quanto à sua aplicabilidade e limitações no contexto da segurança pública, assim como ferramentas como Apache Kafka e Apache Spark, que suportam o processamento distribuído e em tempo real. Por fim, destaca-se a relevância das políticas internas de gestão de dados e da transparência no uso das informações públicas, ressaltando a necessidade de equilibrar eficiência operacional com a proteção dos direitos individuais e a construção da confiança social.

1 Introdução

A coleta e ingestão de dados voltados para a segurança pública carregam uma série de desafios técnicos, éticos e legais que não podem ser ignorados. Logo no momento inicial do ciclo de vida dos dados, a natureza bruta dessas informações se evidencia: registros oriundos de aplicações, fluxos de cliques, sensores ao vivo ou mesmo bases heterogêneas acumulam ruídos e inconsistências que exigem estratégias adequadas de filtragem e organização (Moses, 2023). Essa etapa inicial se caracteriza por reunir conteúdo estruturalmente variado (estruturado e não estruturado), o que aumenta a probabilidade de conflitos semânticos e redundâncias. Em contextos de segurança pública, onde o impacto social das decisões baseadas em dados é elevado, esse cuidado com a qualidade da informação é inevitável. O armazenamento intermediário dos dados entre as fontes operacionais e um Data Warehouse aparece como prática recorrente para mitigar problemas originados durante a extração. Esse estágio envolve tanto sincronização quanto integração dos conjuntos, abrindo espaço para atividades de limpeza e consolidação. A ideia é que a preparação prévia refina os insumos antes da análise avançada, evitando que incoerências técnicas ou semânticas interfiram nos modelos analíticos posteriores. Do ponto de vista de um sistema destinado à segurança pública, isso pode representar desde a padronização de códigos numéricos até a harmonização de formatos textuais utilizados por diferentes órgãos. A qualidade do dado também depende fortemente do contexto da aplicação e das rotinas estabelecidas nas interfaces de entrada. Quando inexistem mecanismos automáticos para crítica dos valores inseridos, aumenta o risco de poluição na base, seja por erros humanos ou falhas sistêmicas (Goldschmidt et al., 2015). Em áreas como segurança pública, esse tipo de poluição pode levar não só a conclusões incorretas mas também afetar diretamente operações estratégicas. Assim, definir domínios aceitáveis para atributos sensíveis (exemplo: intervalos plausíveis para datas ou coordenadas geográficas) torna-se medida preventiva essencial. Paralelamente aos aspectos técnicos, as obrigações legais moldam profundamente a abordagem sobre ingestão e uso desses dados. No caso brasileiro, a Lei Geral de Proteção de Dados (LGPD) estabelece princípios como o livre acesso aos dados pelo titular, com consulta gratuita sobre forma, duração do tratamento e integralidade das informações mantidas. Isso implica que qualquer sistema em segurança pública deve estar preparado não só para processar grandes volumes em tempo hábil, mas também para oferecer transparência integral ao cidadão afetado pelo tratamento dessas informações. Há ainda delimitações normativas claras quanto ao tipo de tratamento excluído da LGPD. Processamentos para fins particulares não econômicos, jornalísticos, artísticos ou acadêmicos, bem como atividades específicas do Estado voltadas à segurança pública ou defesa nacional ficam fora do

escopo direto da lei. No entanto, tal exclusão não significa ausência absoluta de regras: outras leis complementares regulam o manuseio desses conjuntos dentro da esfera estatal. Essa estrutura legislativa sugere uma necessidade constante de alinhamento entre arquitetura tecnológica e normativa vigente. No âmbito internacional, padrões técnicos e jurídicos são impulsionados por autoridades autônomas que atuam na fiscalização e cooperação com outros países (Lima, 2019). Essa cooperação é particularmente sensível na transferência transfronteiriça de dados pessoais envolvidos em investigações digitais. Para sistemas integrados à segurança pública isso significa pensar na arquitetura desde cedo com salvaguardas robustas contra acesso indevido por autoridades externas quando essas não cumpram requisitos mínimos previstos pelas normas. Do ponto de vista arquitetural contemporâneo, padrões como ETL (Extract, Transform, Load) ou ELT (Extract, Load, Transform) oferecem caminhos distintos para incorporar dados às plataformas analíticas (Rucco et al., 2025). Operações clássicas convivem com modelos mais amplos como Data Mesh ou Data Vault que priorizam escalabilidade e resiliência dentro ambientes distribuídos. Esses conceitos ganham importância quando se deseja cruzar bases heterogêneas, por exemplo, registros criminais com dados socioeconômicos, mantendo governança centralizada mas execução descentralizada. Ferramentas capazes de acessar múltiplas fontes heterogêneas revelam-se fundamentais frente ao ecossistema variado encontrado nos órgãos públicos (Goldschmidt et al., 2015). No caso da segurança pública isso pode envolver desde bancos relacionais legados até planilhas provenientes do trabalho administrativo interno. A verdadeira força vem quando essa diversidade é acompanhada pela capacidade efetiva de integrar os conjuntos relacionados em estruturas coesas que favoreçam análises multidimensionais. Para assegurar eficiência nessa jornada tecnológica-legal é preciso conjugar automação das rotinas de ingestão com práticas claras sobre anonimização e ética. Anonimizar dados sensíveis preserva identidades individuais sem inviabilizar padrões estatísticos relevantes para decisões estratégicas em investigação criminal ou prevenção (Lima, 2019). Por outro lado, negligenciar esses cuidados abre margem para usos indevidos que podem comprometer tanto direitos fundamentais quanto credibilidade institucional. Assim, arquiteturas voltadas à ingestão segura em segurança pública devem prevê-la desde seu projeto inicial como uma operação multifásica envolvendo captura heterogênea (Moses, 2023), preparação intermediária (Goldschmidt et al., 2015), conformidade legal rígida (Lima, 2019) e integração técnica adaptativa (Rucco et al., 2025). Esta combinação atua como sustentação prática para alcançar modelos analíticos confiáveis e socialmente responsáveis dentro deste domínio altamente sensível.

2 Fundamentos da Coleta de Dados em Segurança Pública

2.1 Natureza e Tipologia dos Dados

2.1.1 Dados Estruturados

Dados estruturados, no contexto de aplicações voltadas à segurança pública, consistem em informações que seguem um formato rígido, previamente definido por esquemas formais. Geralmente, são armazenados em bases relacionais ou sistemas compatíveis com modelos fixos de definição de campos e tipos. Estruturas desse tipo impõem coerência na entrada e permitem que consultas derivem resultados consistentes usando linguagens como SQL. Por exemplo, registros criminais podem conter colunas específicas para data do ocorrido, classificação da ocorrência, localização georreferenciada e identificação do responsável pela coleta. A padronização facilita auditorias e integração posterior com sistemas analíticos mais amplos. Embora a previsibilidade seja um ponto positivo, ela também gera dependência do esquema preexistente. Alterações na organização dos dados demandam mudanças de modelagem e podem impactar múltiplas aplicações conectadas. Esse cenário exige atenção especial na segurança pública, pois alterações não planejadas no esquema, como mudar o formato de códigos de classificação de crimes, podem prejudicar a comparabilidade histórica das estatísticas (Reis e Housley, 2022). É por isso que estratégias de governança devem considerar tanto o controle rigoroso da estrutura quanto uma documentação acessível sobre seus campos obrigatórios e permissões. Uma característica relevante dos dados estruturados é a sua previsibilidade quanto ao relacionamento entre as tabelas. Modelos relacionais frequentemente usam chaves primárias e estrangeiras que asseguram integridade referencial. Em bases policiais ou judiciais, isto significa que um registro de ocorrência só pode existir vinculado a um identificador válido da unidade policial responsável ou ao número oficial do inquérito. Essa relação controlada reduz inconsistências internas mas não dispensa cuidados no pré-processamento para eliminar registros duplicados ou incompletos (Moses, 2023). Erros se acumulam facilmente se existirem

inconsistências na fonte primária, especialmente em operações que envolvem diferentes órgãos trocando dados via integrações automatizadas. A integração com Data Warehouses costuma representar um objetivo comum para consolidar dados estruturados oriundos de múltiplos sistemas operacionais distintos (Goldschmidt et al., 2015). Essa etapa envolve extração periódica seguida da transformação dos valores para adequação ao modelo analítico centralizado. No caso do policiamento ostensivo, por exemplo, os dados capturados diariamente nas viaturas podem ser processados em lotes e enviados ao repositório central apenas após validação automática da completude dos campos exigidos. Isso permite manter a consistência global sem comprometer o andamento das operações locais. A busca permanente pela qualidade nesses depósitos requer rotinas sistemáticas como limpeza e padronização (Moses, 2023), sem as quais discrepâncias numéricas ou textuais dificultam comparações entre períodos. Do ponto de vista legal, os dados estruturados carregam responsabilidades específicas sob o escopo da Lei Geral de Proteção de Dados quando incluírem informações pessoais. Mesmo contendo formato rígido e bem definido, eles precisam seguir princípios como finalidade específica, necessidade mínima e livre acesso pelo titular (Lima, 2019). Um desafio frequente é assegurar anonimização adequada sem comprometer a utilidade estatística necessária para ações estratégicas. Por exemplo, nomes completos podem ser substituídos por identificadores aleatórios preservando vínculos lógicos internos e permitindo cálculos agregados sobre perfis criminais ou distribuições espaciais. A gestão desses dados demanda processos claros de governança que assegurem acesso controlado e rastreável (Reis e Housley, 2022). Há riscos tanto técnicos quanto sociais quando estruturas bem definidas são mal protegidas: consultas excessivas ou mal formuladas por usuários internos podem expor indevidamente informações sensíveis; já falhas no controle externo abrem portas para vazamentos com impacto direto na confiança da população nos órgãos públicos. Assim, mecanismos como trilhas de auditoria extensivas, permissões granulares por perfil funcional e monitoramento contínuo do uso são medidas preventivas fundamentais. Outro aspecto técnico diz respeito ao processo pós-ingestão em arquiteturas modernas como Data Mesh ou Data Vault aplicadas mesmo a fontes rigidamente estruturadas. Essas abordagens permitem descentralizar a responsabilidade sobre conjuntos específicos garantindo ainda conformidade com padrões globais da organização. No caso da segurança pública isso pode significar que cada departamento policial mantém seu próprio conjunto devidamente organizado segundo um esquema padrão acordado nacionalmente, enquanto uma camada superior cuida da interoperabilidade entre todos esses domínios. Apesar das vantagens óbvias associadas aos dados estruturados, facilidade analítica, simplicidade no cruzamento com outras fontes igualmente padronizadas e alta eficiência nas consultas, é preciso reconhecer limitações intrínsecas. Em cenários dinâmicos como operações emergenciais ou análise criminal preditiva com sinais não convencionais (por exemplo fluxos massivos de vídeo ou interações em redes sociais), o excesso de rigidez pode atrasar inclusão dessas novas variáveis nos modelos decisórios. A adaptabilidade deve ser ponderada junto à longevidade dos esquemas existentes para que as estruturas não se tornem obstáculos à inovação operacional. Por fim, práticas automatizadas como validação em tempo real durante a inserção ajudam a garantir qualidade desde a origem (Moses, 2023). Scripts concebidos para verificar formatos esperados, datas dentro de intervalos plausíveis ou coordenadas geográficas compatíveis, reduzem custos posteriores com retrabalho no pré-processamento (Goldschmidt et al., 2015). Ferramentas contemporâneas já possibilitam executar checagens mais complexas baseando-se em regras predefinidas e alertar operadores imediatamente sobre desvios críticos antes que causem impactos administrativos ou legais irreversíveis no tratamento dos dados estruturados utilizados na segurança pública.

2.1.2 Dados Não Estruturados

Diferentemente das estruturas rígidas e previsíveis, os dados não estruturados apresentam um formato livre, sem esquemas predefinidos para sua disposição. Eles podem surgir em textos livres de boletins de ocorrência, registros audiovisuais de câmeras de monitoramento urbano, conversações telefônicas interceptadas mediante ordem judicial ou mesmo publicações em redes sociais coletadas para fins investigativos. Uma particularidade é que esses conteúdos muitas vezes coexistem em meio a informações contextuais e metadados que auxiliam na interpretação, mas cujo armazenamento e indexação exigem soluções específicas. No contexto operacional da segurança pública, esse tipo de dado frequentemente provém de fontes múltiplas e heterogêneas, com níveis distintos de qualidade e confiabilidade, dificultando uma integração direta com sistemas convencionais baseados em modelos relacionais (Foidl et al., 2023). Há também questões práticas ligadas ao volume e à velocidade com que tais dados são gerados. Vídeos contínuos em alta definição a partir de uma rede urbana de câmeras podem produzir terabytes

por dia; transcrições automáticas de conversas capturadas durante investigações operam na mesma lógica volumétrica intensa. Nesses casos, arquiteturas de ingestão precisam lidar tanto com fluxos contínuos quanto com cargas eventuais em lote. A utilização concomitante de processamento em tempo real e processamento batch permite não apenas armazenar o acervo integral para tratamento posterior como também agir sobre informações críticas imediatamente, como no reconhecimento facial ao vivo ou detecção automática de sons específicos associados a disparos. A diversidade estrutural apresenta outro obstáculo: formatos totalmente diferentes precisam ser tratados conforme suas especificidades técnicas. Um áudio captado por microfones ambientais exige extração e transformação distintas das necessárias para um conjunto massivo de imagens fixas ou documentos em formato PDF. Essa variedade obriga as equipes a adotarem mecanismos robustos para extração, transformação e carga (ETL ou ELT), garantindo a integridade do conteúdo mesmo após múltiplas etapas (Rucco et al., 2025). Além disso, tarefas como limpeza e eliminação de redundâncias são menos triviais quando o próprio conceito de duplicação pode não ser puramente sintático; dois vídeos distintos podem retratar o mesmo evento sob ângulos diferentes e isso requer estratégias analíticas mais sofisticadas do que meras comparações byte a byte (Goldschmidt et al., 2015). Do ponto de vista legal e ético, tratar dados não estruturados relacionados à segurança pública impõe desafios adicionais. Conteúdos audiovisuais ou textuais frequentemente contêm dados pessoais sensíveis cuja exposição pode afetar diretamente a privacidade dos indivíduos retratados. A Lei Geral de Proteção de Dados traz requisitos claros sobre finalidade e necessidade mínima na coleta, mesmo quando se trata de atividades estatais isentas do seu escopo direto, como certas operações voltadas à segurança. O alinhamento com esses princípios implica no uso extensivo de técnicas como anonimização ou pseudonimização para remover identificadores diretos antes do compartilhamento interno ou externo (Lima, 2019). Analisando-se ainda pela perspectiva da governança, a ausência de estrutura definida dificulta a aplicação automática das regras tradicionais adotadas para dados tabulares. É comum que iniciativas baseadas em catálogos manuais falhem nesse cenário devido ao alto custo humano para classificar ativos informacionais desse tipo. Automatizar metadados gerados via ferramentas dotadas de aprendizado de máquina passa a ser quase obrigatório para manter rastreabilidade, descobrir rapidamente conteúdos específicos e cumprir exigências normativas quanto à auditabilidade dos acessos. Assim como nos dados estruturados descritos na Seção 2.1.1, é imperativo estabelecer políticas claras sobre quem pode acessar cada tipo específico desse conteúdo e sob quais condições tal acesso é concedido. A questão da observabilidade ganha relevância quando se pretende monitorar continuamente a "saúde" desses ecossistemas informacionais não estruturados (Moses, 2023). Sem métricas adequadas, por exemplo, sobre completude das legendas geradas automaticamente ou taxa média de erro nos reconhecimentos faciais, torna-se difícil manter decisões táticas confiáveis no dia a dia operacional. Sistemas modernos já incorporam camadas especializadas capazes de detectar anomalias nos pipelines que manipulam fotos, áudios e vídeos; essas camadas identificam falhas na ingestão ou transformações indevidas logo no momento em que ocorrem. Outro fator prático: enquanto operações sobre dados tabulares permitem filtragens prévias relativamente simples utilizando SQL, nos dados não estruturados é comum depender fortemente de técnicas avançadas como processamento automático de linguagem natural (NLP) ou análise preditiva aplicada a imagens mediante redes neurais convolucionais treinadas. Isso abre espaço para integrações diretas com pipelines orientados a aprendizagem profunda que processem lotes históricos paralelamente às transmissões ao vivo (Rucco et al., 2025). No entanto essa abordagem precisa vir acompanhada por verificações rigorosas da acurácia alcançada, pois vieses algorítmicos podem provocar desde detecções incorretas até implicações discriminatórias nas atuações policiais. Finalmente há o impacto social da manipulação indevida desses insumos. O armazenamento prolongado sem controles efetivos multiplica riscos: vídeos sensíveis podem vazar comprometendo investigações; transcrições parciais fora do contexto original podem gerar interpretações distorcidas pela mídia ou pelo público generalista. A proteção contra esses cenários inclui práticas como criptografia forte em repouso e em trânsito combinada ao fracionamento seguro dos arquivos mais sensíveis entre diferentes domínios administrativos; além disso recomenda-se auditoria regular sobre quem acessou quais fragmentos desses conteúdos e para qual finalidade registrada no sistema central (Reis e Housley, 2022). Ao integrar esses mecanismos dentro das arquiteturas modernas orientadas ao tratamento heterogêneo é possível diminuir substancialmente o espaço para incidentes, mantendo simultaneamente aderência normativa, eficiência operacional e respeito absoluto às garantias individuais previstas pela legislação vigente na área da segurança pública brasileira.

2.2 Fontes de Dados

2.2.1 Sistemas Policiais e Judiciais

Sistemas policiais e judiciais constituem a espinha dorsal da coleta de dados formais na segurança pública. Esses sistemas têm natureza híbrida: armazenam, processam e compartilham informações que vão desde registros de ocorrências, mandados, inquéritos policiais até decisões judiciais consolidadas. Por essa razão, operam com dados predominantemente estruturados, mas também incluem segmentos não estruturados, como anexos textuais livres em boletins ou vídeos de audiências registradas para fins probatórios (Reis e Housley, 2022). A integração técnica desses diversos formatos exige arquiteturas capazes de harmonizar padrões rígidos com conteúdos mais livres sem perder integridade nem valor jurídico. Do ponto de vista funcional, os sistemas policiais precisam estabelecer relações claras entre identificadores únicos dos cidadãos envolvidos e ações registradas ao longo do tempo. Isso implica modelagem relacional rigorosa sustentada por chaves primárias e estrangeiras que assegurem consistência na entidade "ocorrência" ou "processo" (Moses, 2023). Ao mesmo tempo, há necessidade de acompanhar metadados relevantes sobre o fluxo das informações, como evolução do esquema de banco de dados ou modificações nas regras processuais internas (Reis e Housley, 2022). Esse mapeamento auxilia na interoperabilidade entre unidades geograficamente dispersas ou administrativamente distintas. No âmbito legal, a Lei Geral de Proteção de Dados (LGPD) impõe atenção especial à manipulação dos dados pessoais sensíveis que transitam nesses sistemas. Ainda que atividades voltadas à segurança pública possuam prerrogativas específicas fora do escopo direto da LGPD em alguns casos, subsiste a obrigação ética e técnica de preservar confidencialidade e limitar acesso apenas ao pessoal autorizado. Medidas como pseudonimização aparecem como possibilidade concreta para proteger identidades proporcionando que análises estatísticas sejam realizadas sem exposição direta dos indivíduos. Contudo, esse procedimento deve ser cuidadosamente implementado pois a literatura aponta fragilidades na anonimização pura e possibilidade de reidentificação quando diferentes bases são cruzadas (Lima, 2019). Automação no processamento desses dados surge como elemento vital. Operações diárias em sistemas policiais englobam entrada contínua e massiva, desde notificações automáticas geradas por leitura de placas veiculares até atualizações manuais oriundas de agentes em campo. A utilização de pipelines adaptativos capazes de alternar entre modos de ingestão completos (full ingestion) para sincronizações periódicas e incrementais para atualizações pontuais proporciona maior eficácia operacional (Rucco et al., 2025). Essa flexibilidade é crucial para evitar gargalos nos momentos críticos ou sobrecarga desnecessária durante baixas demandas. A observabilidade amplia o controle sobre a "saúde" dos sistemas policiais e judiciais. Métricas relacionadas à frescor dos dados (tempo decorrido entre evento e registro), volume diário processado, distribuição por tipo de ocorrência e linhagem completa desde a origem até os relatórios finais tornam-se fatores determinantes para manutenção da confiabilidade institucional (Moses, 2023). Ferramentas modernas conectadas diretamente à pilha tecnológica existente conseguem monitorar fluxos internos sem extrair dados brutos desnecessariamente, reduzindo risco de vazamento ao mesmo tempo em que simplificam auditorias posteriormente necessárias diante de investigações internas ou externas. Governança ganha relevo nessas plataformas ao disciplinar quem pode acessar quais partes do sistema e sob quais circunstâncias esse acesso acontece. Autenticação forte aliada a políticas definidas por perfis (IAM – Identity and Access Management) delimita escopo funcional evitando que operadores realizem consultas fora da sua competência técnica ou administrativa (Reis e Housley, 2022). O histórico detalhado das ações é essencial tanto para responsabilização indivíduo-institucional quanto para eventual defesa judicial do próprio órgão frente a alegações externas sobre uso indevido das informações armazenadas. Há também um aspecto estratégico relacionado à integração entre sistemas policiais e judiciais: cruzar automaticamente informação registrada pela polícia com despachos emitidos pelo poder judiciário oferece uma visão mais completa das dinâmicas criminais e procedimentais. Esse acoplamento favorece análises preditivas dentro de programas específicos, como acompanhamento eletrônico em liberdade condicional ou padrões temporais no surgimento de mandados vinculados a determinados perfis criminosos (Goldschmidt et al., 2015). Entretanto, integrar essas bases demanda protocolos formais sobre armazenamento seguro em diferentes localizações geográficas conforme requisitos regulatórios, evitando exposição indevida especialmente quando sistemas se encontram hospedados em nuvens públicas multinacionais nas quais legislações divergentes podem influenciar garantias jurídicas aplicáveis. Segurança técnica aplicada aqui engloba criptografia tanto no armazenamento quanto no trânsito das informações sensíveis capturadas destes sistemas; mecanismos adicionais incluem mascaramento seletivo em interfaces situacionais (exemplos: ocultar parcialmente

nome completo ou endereço enquanto exibir dados operacionais essenciais) e tokenização para campos críticos garantindo que qualquer exportação indevida mantenha o conteúdo inaproveitável fora do contexto original (Reis e Housley, 2022). O monitoramento constante contra tentativas internas ou externas não autorizadas complementa esta camada protetiva, falhas nesta área podem ter impacto social direto minando credibilidade pública nos órgãos responsáveis pela segurança coletiva. Finalmente existe o componente ético implicado no uso intensivo desses dados: cada extração ou correlação realizada tem potencial de afetar direitos individuais, inclusive quando motivada por boas intenções institucionais. Assim práticas transparentes apoiadas em documentação clara sobre os motivos da coleta, métodos empregados e resultados obtidos devem ser incorporadas às rotinas administrativas desses sistemas policiais e judiciais. Isso ajuda não apenas na conformidade com normas vigentes mas também na construção progressiva da confiança social tão necessária à efetividade das ações estatais relacionadas à segurança pública (Lima, 2019).

2.2.2 Redes Sociais e Plataformas Digitais

O uso de redes sociais e plataformas digitais como fonte de dados para segurança pública traz um grau elevado de complexidade tanto técnica quanto ética. Essas fontes geram um volume expressivo de informações dinâmicas, variando desde textos curtos, imagens e vídeos, até transmissões em tempo real via *streaming*, e frequentemente com metadados associados que indicam localização, hora ou interações entre usuários (Foidl et al., 2023). Essa combinação de elementos heterogêneos pode resultar em cenários ricos para análise preditiva e monitoramento situacional, mas também exige arquiteturas capazes de processar conteúdo contínuo com ingestão adaptativa. No aspecto técnico, a ingestão desses dados demanda que pipelines sejam flexíveis o suficiente para lidar com formatos e origens distintas, por exemplo, coletar dados estruturados provenientes de APIs públicas ao mesmo tempo em que se indexam conteúdos audiovisuais extraídos diretamente da superfície de redes sociais. Em casos nos quais o objetivo é monitorar eventos críticos ou identificar padrões comportamentais suspeitos, é comum combinar processamento *batch* com sistemas orientados a eventos para capturar instantaneamente peças relevantes. Contudo, essa integração simultânea é desafiadora e pode comprometer desempenho ou consistência se não houver um mapeamento claro das fontes e seus fluxos (Rucco et al., 2025). Ferramentas modernas permitem ajustar dinamicamente esquemas e tipos de ingestão sem alterar significativamente a estrutura implantada, viabilizando atualizações rápidas frente às alterações contínuas nas APIs dessas plataformas. Outra dimensão crítica está na confiabilidade dos dados coletados. Informações obtidas em redes sociais nem sempre têm veracidade garantida; há risco elevado de ruído, duplicação ou até falsificação deliberada. Estratégias robustas de limpeza tornam-se indispensáveis aqui: aplicar filtros semânticos para eliminar conteúdos irrelevantes ou detectar duplicações baseadas em contexto ajuda a manter acurácia no resultado final (Goldschmidt et al., 2015). No entanto, tal processo precisa considerar que dois conteúdos aparentemente distintos (por exemplo, fotos feitas em momentos próximos) podem representar exatamente o mesmo evento relevante. Do ponto de vista legal e ético, existe sensibilidade extrema na utilização desses dados para fins estatais. Quando se tratam de publicações contendo informações pessoais identificáveis, mesmo que essas estejam acessíveis publicamente, devem ser observados princípios da LGPD no Brasil, como finalidade explícita e respeito ao direito do titular (Lima, 2019). Embora algumas atividades ligadas à segurança pública possam ter prerrogativas específicas fora do escopo direto da lei, isso não elimina a responsabilidade em preservar confidencialidade e adotar medidas apropriadas para anonimização ou pseudonimização antes da análise massiva. Por exemplo, extrair apenas elementos estatísticos agregados sobre padrões temporais ou geográficos evita exposição direta dos indivíduos envolvidos. A governança também exerce papel central nesse contexto: controlar quem tem acesso aos dados brutos capturados nas plataformas digitais minimiza riscos associados a uso indevido interno ou vazamentos externos. Políticas rígidas de autorização integradas ao gerenciamento de identidades (IAM) asseguram limite funcional adequado ao perfil do operador da informação (Reis e Housley, 2022). Além disso, trilhas completas de auditoria são indispensáveis para posterior verificação tanto da origem quanto do uso das informações extraídas dessas fontes. Essas práticas precisam coexistir com mecanismos avançados de observabilidade que permitam avaliar continuamente qualidade e frescor dos conteúdos coletados (Moses, 2023). Métricas específicas podem incluir taxa média diária de ingestão por tipo de mídia (texto, imagem, vídeo), latência entre publicação original e registro no sistema central ou precisão dos algoritmos utilizados na classificação automática destes dados. Acompanhar esses indicadores permite reagir rapidamente a falhas, como quedas na ingestão devido à alteração nas políticas das APIs externas, mantendo es-

tabilidade operacional mesmo sob alta demanda. O tratamento desse ecossistema informacional exige ainda atenção à questão do viés algorítmico quando se aplicam técnicas automáticas sobre os conjuntos não estruturados coletados (Rucco et al., 2025). Algoritmos responsáveis por reconhecimento facial ou categorização temática podem apresentar resultados enviesados caso os dados originais reflitam desigualdades pré-existentes nas representações das redes sociais. No âmbito policial isso pode levar a inferências incorretas ou discriminação indireta contra grupos específicos. Para mitigar esse risco é necessário adotar métodos sistemáticos de avaliação da acurácia durante todo o ciclo dos modelos aplicados nessas análises. Há também desafios operacionais ligados à escala temporal e geográfica das informações disponíveis nas plataformas digitais. Algumas ocorrências acontecem em lapsos mínimos cujos registros desaparecem rapidamente devido às rotinas internas das redes sociais; outras são espalhadas globalmente exigindo integração com sistemas externos localizados em diferentes jurisdições legais. Nesses casos entram novamente exigências normativas relacionadas à transferência transfronteiriça de dados pessoais (Lima, 2019), exigindo salvaguardas específicas para tráfego seguro e conforme regulações locais vigentes nos países envolvidos. Por fim, o impacto social dessa coleta deve ser ponderado cuidadosamente: cidadãos tendem a perceber negativamente ações estatais que pareçam invasivas na esfera pessoal digital; portanto transparência sobre métodos usados, incluindo explicitação clara da motivação por trás da coleta, serve como importante medida preventiva contra erosão da confiança pública. Documentar processos desde o momento inicial até o descarte seguro do dado processado impede interpretações equivocadas externas sobre objetivos adotados pelas autoridades responsáveis pela segurança coletiva (Reis e Housley, 2022). Dessa forma é possível alinhar efetividade técnica à preservação dos direitos fundamentais enquanto se aproveita todo potencial analítico oferecido pelo vasto universo informacional presente nas redes sociais e plataformas digitais atuais.

3 Aspectos Éticos, Legais e Pré-processamento

3.1 Conformidade com a LGPD

A conformidade com a Lei Geral de Proteção de Dados (LGPD) em contextos de segurança pública exige um alinhamento entre o aparato tecnológico, as práticas operacionais e os princípios jurídicos estabelecidos. Embora determinadas atividades estatais voltadas à segurança pública possam estar fora do escopo direto da LGPD, há obrigações éticas e técnicas que não se diluem pelo amparo legal à exceção. Isso significa que órgãos responsáveis pela coleta e processamento de informações precisam adotar salvaguardas compatíveis com padrões normativos mesmo em situações nas quais a lei não impõe todos os requisitos formais. A primeira dimensão prática dessa conformidade se relaciona com a transparência no tratamento dos dados pessoais. Em ambientes nos quais a coleta é massiva, como sistemas policiais integrando múltiplas fontes heterogêneas, torna-se necessário disponibilizar mecanismos que permitam ao titular consultar quais dados foram coletados, por quanto tempo serão mantidos e para qual finalidade estão sendo tratados. Essa clareza não apenas atende ao princípio da publicidade, mas também atua como elemento preventivo contra abusos internos e externos, ajudando a fortalecer a confiança pública na operação estatal. Outro ponto central envolve o princípio da necessidade mínima: apenas informações diretamente relacionadas à execução da atividade de segurança devem ser processadas. Faz sentido limitar atributos sensíveis ou reduzir temporalmente a guarda desses dados quando não houver justificativa operacional robusta para mantê-los além do estritamente necessário. A adoção de políticas automatizadas de expurgo periódico pode evitar sobrecarga nos sistemas e reduzir riscos associados ao armazenamento prolongado. Essa gestão inteligente depende de integração estreita entre módulos técnicos e regras administrativas predeterminadas, garantindo que as exclusões sejam consistentes em todos os ambientes onde o dado transita. Do ponto de vista arquitetural, práticas como anonimização ou pseudonimização são essenciais na conformidade com a LGPD, ainda mais considerando que nem toda anonimização é realmente definitiva (Lima, 2019). Estudos indicam possibilidade concreta de reidentificação quando diferentes bases são cruzadas inadvertidamente; portanto, aplicar máscaras ou tokens isolados para campos críticos é uma etapa importante, mas precisa vir acompanhada de restrições sobre o acesso aos conjuntos originais. Aqui entram diretrizes rígidas sob responsabilidade da governança: delimitar perfis funcionais conforme regras institucionais evita que operadores manipulem grandes volumes sem autorização formal ou motivação legítima. Além disso, mecanismos tecnológicos como trilhas completas de auditoria ajudam tanto na prevenção quanto na remediação de incidentes. Cada ação executada sobre os dados, consulta, exportação ou alteração,

deve deixar registro vinculando operador, horário e justificativa operacional. Essas trilhas possibilitam investigações posteriores caso ocorram suspeitas de uso indevido e reforçam responsabilidade individual dentro dos órgãos públicos. A proteção contra vazamentos é outro pilar relevante: criptografia forte tanto em repouso quanto em trânsito deve ser padrão adotado para todas as camadas onde circulem informações pessoais sensíveis (Reis e Housley, 2022). Essa medida dificulta acesso não autorizado mesmo em cenários extremos como comprometimento físico do hardware ou interceptação das comunicações entre sistemas distribuídos. O fracionamento seguro entre diferentes domínios administrativos pode reduzir ainda mais o risco, separando partes críticas do conjunto total e exigindo múltiplos fatores para recomposição do conteúdo integral. Nas integrações internacionais que envolvem transferência transfronteiriça de dados pessoais usados para fins investigativos digitais, exigem-se cuidados adicionais previstos pela LGPD ao reconhecer somente países ou organizações cujas leis e práticas ofereçam grau adequado de proteção (Lima, 2019). Isso impacta diretamente projetos que dependem da cooperação técnica com autoridades estrangeiras; nesses casos é recomendável incluir cláusulas contratuais específicas limitando escopo e duração desse compartilhamento. A conformidade também passa por medidas proativas na detecção e correção imediata de problemas no pipeline dos dados (Moses, 2023). Rotinas automáticas capazes de identificar anomalias, seja na completude das informações cadastradas ou na estrutura esperada das variáveis, permitem agir rapidamente antes que falhas se propaguem pelo sistema operacional. Quando aplicadas em grande escala dentro dos serviços estatais, essas verificações reduzem probabilidade de decisões estratégicas serem afetadas por erros acumulados desde a base original (Rucco et al., 2025). Na perspectiva social da conformidade com a LGPD, há um fator não técnico igualmente relevante: comunicar claramente à população os métodos e objetivos relacionados à coleta e tratamento das informações gera percepção positiva quanto ao respeito aos direitos fundamentais. Em tempos nos quais cidadãos são cada vez mais conscientes acerca da sua privacidade digital, omitir detalhamento dos processos pode enfraquecer apoio comunitário às ações promovidas pelo Estado. Por outro lado, publicações regulares sobre políticas internas e resultados alcançados demonstram compromisso ativo com transparência administrativa (Reis e Housley, 2022). Também aparece nesse contexto a importância da acurácia algorítmica aplicada sobre os dados coletados. Sistemas automatizados destinados à segurança pública podem introduzir vieses discriminatórios quando treinados sobre amostras desbalanceadas; isso não apenas viola princípios éticos como pode colidir diretamente com diretrizes legais voltadas à igualdade no tratamento dos cidadãos (Rucco et al., 2025). Avaliações sistemáticas sobre performance desses modelos devem fazer parte integrante do ciclo operacional garantindo que qualquer prática analítica mantenha equidade sem prejuízos indevidos para indivíduos ou grupos específicos. Assim, alinhar tecnologia, processos administrativos e marcos normativos sob a ótica da LGPD significa construir um ecossistema informacional seguro e auditável dentro da estrutura estatal responsável pela segurança pública. Ao atender simultaneamente às exigências legais explícitas, aos padrões internacionais reconhecidos e às melhores práticas técnicas descritas anteriormente (Lima, 2019), estabelece-se uma base sólida tanto para efetividade operacional quanto para preservação duradoura dos direitos individuais frente às demandas crescentes dessa área estratégica para a sociedade brasileira.

3.2 Anonimização e Limpeza de Dados

A anonimização e a limpeza de dados representam processos complementares que sustentam a integridade e a conformidade legal em sistemas de segurança pública. Partindo do cenário apresentado na Seção 3.1, fica evidente que não basta apenas coletar informações; é necessário garantir que elas possam ser usadas de forma responsável, protegendo as identidades individuais e assegurando a qualidade necessária para análises confiáveis. A anonimização procura remover ou transformar elementos capazes de identificar diretamente uma pessoa, reduzindo drasticamente o risco de violação de privacidade mesmo quando datasets são compartilhados ou analisados em larga escala (Lima, 2019). Entretanto, há uma nuance importante: certas técnicas comuns como a substituição por pseudônimos podem não ser suficientes para evitar reidentificação se bases distintas forem combinadas inadvertidamente. Isso exige aplicação criteriosa de métodos avançados, incluindo randomização difícil de inverter, supressão seletiva de atributos e generalização de valores críticos. Essas abordagens precisam ser escolhidas segundo avaliação prévia da sensibilidade do conjunto e dos riscos operacionais envolvidos. Dentro de fluxos automatizados modernos, a anonimização tende a ocorrer ainda nos primeiros estágios da ingestão para que qualquer processamento subsequente já opere sobre dados com menor grau de sensibilidade (Rucco et al., 2025). Configurar esse passo logo no início diminui superfícies de ataque contra informações

críticas e simplifica requisitos posteriores de governança. Arquiteturas adaptativas possibilitam inserir módulos dedicados à anonimização nos pipelines sem interromper rotinas já consolidadas, permitindo inclusive aplicar diferentes esquemas conforme categoria ou origem da informação, por exemplo, aplicar supressão mais intensa em transcrições interceptadas judicialmente do que em estatísticas agregadas sobre patrulhamento. A limpeza de dados, por sua vez, é focada em corrigir ou eliminar registros incorretos, incompletos ou inconsistentes antes que sejam enviados aos sistemas analíticos. Trata-se de um componente essencial quando volumes massivos vindos de sensores urbanos, sistemas policiais ou redes sociais carregam ruídos inevitáveis. Entre as estratégias mais usuais estão o tratamento de valores ausentes (imputação controlada, exclusão), remoção ou ajuste de outliers anômalos, temperatura marcada como 99 999 em sensores seria um exemplo claro (Moses, 2023), além da normalização e harmonização dos formatos para garantir compatibilidade futura entre diferentes bases. Falhas nessas operações comprometem diretamente previsões e diagnósticos estratégicos; um campo geográfico com coordenadas fora do intervalo esperado pode distorcer mapas criminais usados no planejamento tático. Integrar anonimizacão e limpeza requer atenção especial à ordem das operações. Em algumas situações, aplicar anonimização primeiro pode atrapalhar limpezas posteriores se identificadores originais eram essenciais para deteção de duplicatas; em outras, limpar previamente evita perder dados válidos durante transformações irreversíveis. Essa decisão deve ser embasada tanto no desenho arquitetural do pipeline quanto nas exigências legais aplicáveis ao contexto específico. Outro aspecto central envolve a automação desses processos para reduzir custos humanos e manter consistência (Jindal et al., 2017). Rotinas automáticas conseguem detectar violações simples das regras empresariais, registros fora dos padrões definidos, e aplicar reparos básicos sem intervenção manual. Porém nem todo caso comporta correções automáticas seguras; situações complexas pedem validação humana sobre o que deve ser mantido ou descartado. Esse equilíbrio entre automação e supervisão é delicado na segurança pública porque o custo potencial de uma decisão errada pode significar liberar ou excluir informação vital para investigação criminal. No contexto da LGPD, etapas sistemáticas de limpeza ajudam no cumprimento do princípio da necessidade mínima ao eliminar informações desnecessárias (Lima, 2019). Manter campos irrelevantes aumenta superfície para vazamentos e abre margem para uso indevido interno. Limpeza adequada funciona portanto como filtragem consciente das variáveis cujo valor agregado à análise final não compensa risco associado à sua retenção prolongada. Do ponto de vista social, adotar anonimização robusta combinada a processos claros e documentados de limpeza reforça percepções positivas sobre respeito aos direitos individuais. Cidadãos que têm certeza de que suas informações serão tratadas com rigor técnico desde a captura até o descarte tendem a apoiar políticas públicas baseadas em evidências extraídas dessas mesmas fontes (Reis e Housley, 2022). Já falhas nesse cuidado abrem espaço para controvérsia pública, judicializações e perda da confiança institucional. Os desafios técnicos também não são triviais: erros na classificação automática dos tipos de dados foram identificados como causa recorrente em problemas operacionais (33% dos casos investigados) (Foidl et al., 2023). Tipos incorretos dificultam tanto limpezas quanto anonimizações porque funções aplicadas presumem formatos esperados; ao receber conteúdos fora desse padrão muitas ferramentas simplesmente falham silenciosamente ou descartam material importante sem aviso. Correção precoce desses aspectos amplia estabilidade geral do pipeline. Por fim, cabe destacar que mesmo processos eficientes não são imunes à necessidade constante de auditoria. Toda anonimização deve ter parâmetros revisados periodicamente considerando avanços tecnológicos capazes de reverter técnicas antes tidas como seguras; toda rotina automática precisa ser recalibrada frente a mudanças nas características originais dos dados recebidos (Rucco et al., 2025). Incorporar observabilidade plena às etapas facilita identificar pontos onde acurácia caiu ou onde novas anomalias surgiram devido à alteração na fonte primária (Moses, 2023). Aliar práticas sofisticadas desses dois domínios constrói um ambiente informacional mais resiliente dentro da segurança pública: limpa-se o excesso desnecessário que compromete análises enquanto protege-se identidades com barreiras técnicas sólidas contra exposição indevida. Tal combinação fortalece tanto eficácia operacional quanto credibilidade institucional frente aos cidadãos e organismos reguladores nacionais e internacionais (Lima, 2019; Reis e Housley, 2022).

4 Ingestão e Integração de Dados

4.1 Modos de Ingestão

4.1.1 Processamento em Lotes (Batch)

O processamento em lotes, ou *batch processing*, caracteriza-se por agrupar dados recebidos ao longo de um período definido e processá-los de forma unificada em momentos pré-determinados. Diferente de fluxos contínuos, nos quais cada registro é tratado assim que chega, esse modo pressupõe uma periodicidade ou gatilhos específicos para iniciar o ciclo. Essa estratégia é particularmente útil em sistemas de segurança pública quando existe a necessidade de consolidar grandes volumes de informações antes de submetê-las a análises mais complexas (Foidl et al., 2023). Exemplos clássicos incluem a extração diária de registros policiais gerados em unidades distintas ou a compilação semanal de dados forenses digitais recuperados em investigações. A operação em lotes amplia a eficiência quando há ganho em acumular dados e aplicar transformações coletivas, como conversões de formatos, deduplicações ou integração com outros conjuntos, sem sobrecarregar recursos computacionais em tempo real (Moses, 2023). Em contextos operacionais distribuídos, rotinas desse tipo permitem padronizar conteúdos oriundos de várias fontes antes que entrem no ambiente analítico centralizado, garantindo consistência entre bases geograficamente dispersas. Na prática, pode-se configurar gatilhos baseados no tamanho acumulado da carga ou no horário programado para executar pipelines completos durante períodos de menor uso do sistema. Por outro lado, esse modelo impõe um atraso natural na disponibilidade da informação mais recente. No caso da segurança pública, isso significa que decisões táticas baseadas nesses lotes estarão sempre um passo atrás do cenário atual. Porém, nem todas as aplicações demandam resposta imediata; estatísticas históricas, relatórios periódicos e modelagens estratégicas podem se beneficiar da densidade informacional obtida ao trabalhar com pacotes mais amplos. Uma rede municipal de câmeras, por exemplo, pode processar diariamente os arquivos captados para treinar algoritmos de reconhecimento ou extrair métricas agregadas sobre movimentação urbana sem precisar acionar recursos intensivos 24 horas por dia. Do ponto de vista técnico, arquiteturas clássicas como ETL (Extract, Transform, Load) se alinham bem ao conceito batch (Rucco et al., 2025). Nesse fluxo, dados são extraídos das fontes originais em janelas predefinidas, recebem tratamento adequado (limpeza, padronização e possivelmente anonimização) e só então carregados no repositório final. Esse processo facilita auditorias pelos logs detalhados gerados durante cada execução, permitindo identificar rapidamente falhas ou inconsistências localizadas no lote processado (Moses, 2023). Mais recentemente, combinações híbridas como a arquitetura Lambda implementam camadas batch junto a componentes em tempo real para balancear frescor e profundidade histórica dos dados. No aspecto legal e ético, o processamento em lotes oferece algumas vantagens ligadas à conformidade com normas como a Lei Geral de Proteção de Dados (LGPD). Isso porque as rotinas periódicas permitem inserir estágios controlados para aplicação sistemática de anonimização e pseudonimização logo na transição entre ambientes (Lima, 2019). Ao operar com cargas consolidadas há oportunidade para revisar detalhadamente cada pacote antes da disponibilização analítica, removendo atributos desnecessários segundo o princípio da necessidade mínima ou mascarando identificadores diretos que possam comprometer privacidade. Além disso, trilhas completas sobre quem aprovou cada lote e quais critérios foram aplicados favorecem transparência e responsabilização institucional (Reis e Housley, 2022). A governança nesse contexto exige coordenação estreita entre equipes técnicas responsáveis pelos pipelines e gestores administrativos que definem regras sobre prazos máximos para retenção dos dados brutos processados. Como cargas antigas podem conter informações sensíveis irrelevantes às novas análises estratégicas, é recomendável implementar políticas automáticas que descartem parcelas não essenciais logo após validações conclusivas (Lima, 2019). Essa prática reduz superfície para incidentes e minimiza custo com armazenamento seguro. Um cuidado específico relacionado à segurança pública envolve manter alta qualidade dentro dos pacotes processados: valores inválidos introduzidos na origem se propagam pelo lote inteiro caso não sejam corretamente filtrados na etapa inicial (Goldschmidt et al., 2015). Estratégias robustas de limpeza, como detecção automática de outliers absurdos (exemplo: latitude fora do intervalo permitido) ou harmonização semântica, devem estar integradas ao roteiro batch. Desconsiderar essa etapa arrisca gerar interpretações distorcidas nos relatórios consolidados usados para definir políticas públicas. Arquiteturas modernas também permitem distribuir o processamento batch entre múltiplos nós computacionais escaláveis horizontalmente. Tal abordagem é útil quando se trabalha com datasets massivos típicos da segurança pública contemporânea: desde logs integrados de rádios comunicadores até coleções mul-

timídia extensas provenientes das perícias criminais digitais. Esse paralelismo otimiza tempo total sem abrir mão do controle centralizado sobre resultados finais. Outro ponto relevante diz respeito à observabilidade desses processos: monitorar métricas como duração média por execução, volume total ingerido por ciclo e número médio de correções aplicadas aumenta visibilidade sobre performance geral do pipeline. Ocorrências anômalas, como aumento repentino no número de registros excluídos por inconsistências, podem sinalizar mudanças na qualidade dos insumos ou problemas técnicos nas fontes originais. O impacto social desse modo também precisa ser previsto nas estratégias institucionais. Embora batch preserve eficiência operacional para muitos casos administrativos ou investigativos prolongados, sua temporalidade pode não responder adequadamente a crises emergenciais; para estas cobra-se integração complementar com mecanismos que capturem eventos críticos em intervalos menores (Moses, 2023). Decidir por uma configuração exclusivamente batch deve vir acompanhada dessa avaliação pragmática alinhada aos objetivos funcionais esperados. Em síntese técnica-operacional contextualizada pela ética pública: processamento em lotes permanece componente central no mosaico tecnológico aplicado à segurança coletiva desde que calibrado tanto quanto às restrições normativas quanto às demandas táticas locais. Integrando etapas obrigatórias ligadas à LGPD como anonimização adequada e expurgo criterioso (Lima, 2019), aliando limpeza sólida (Goldschmidt et al., 2015) e governança clara (Reis e Housley, 2022), essa modalidade sustenta soluções consistentes onde profundidade histórica é mais valiosa que instantaneidade absoluta, desde que não se perca a vigilância frente às implicações sociais resultantes dessa escolha arquitetural.

4.1.2 Processamento Contínuo (Streaming)

O processamento contínuo, ou *streaming processing*, diferencia-se das operações em lotes ao tratar cada evento ou registro praticamente no momento de sua chegada ao sistema. Essa abordagem permite que sistemas de segurança pública respondam com agilidade a ocorrências em andamento, integrando capacidades de ingestão e análise quase instantâneas (Reis e Housley, 2022). Enquanto no modelo batch a latência entre coleta e processamento pode ser de horas ou dias, o paradigma streaming busca minimizar essa lacuna para segundos ou milissegundos, dependendo da aplicação. Em cenários como monitoramento urbano por câmeras, leitura automática de placas veiculares ou acompanhamento em tempo real de comunicações interceptadas judicialmente, essa capacidade de resposta é decisiva para prevenir crimes ou acelerar ações operacionais. O aspecto técnico central do streaming é a ingestão contínua de dados gerados em fluxo pelos sistemas de captura. No contexto da segurança pública, esse fluxo pode vir de sensores distribuídos pela cidade, aplicativos móveis usados por agentes em campo ou APIs conectadas diretamente às redes sociais. Arquiteturas modernas tendem a integrar soluções especializadas como Kafka, que atua não apenas como transporte eficiente desses eventos mas como plataforma orientada a fluxos contínuos capazes de alimentar múltiplos consumidores simultaneamente (Shapira et al., 2022). Essa multiplicidade facilita correlacionar diferentes fontes em tempo real, por exemplo, associar um alerta disparado por reconhecimento facial automatizado à verificação simultânea nos bancos policiais e judiciais já discutidos na seção anterior 4.1.1. Do ponto de vista operacional, o streaming impõe exigências específicas sobre escalabilidade e tolerância a falhas. Como os fluxos são incessantes e potencialmente volumosos, o sistema precisa absorver variações abruptas na taxa de eventos sem perder registros importantes. Técnicas como particionamento dinâmico e uso de *thread pools* otimizados ampliam a capacidade do pipeline lidar com cargas súbitas (Jindal et al., 2017). No entanto, esse paralelismo exige coordenação cuidadosa para garantir que cada evento seja processado uma vez e apenas uma vez; duplicações ou omissões podem distorcer análises táticas sensíveis. Os benefícios práticos dessa modalidade se multiplicam quando analisados sob ótica da prevenção. De acordo com o princípio da segurança previsto na LGPD, aplicar medidas técnicas aptas a proteger dados pessoais inclui não apenas proteção contra acessos indevidos mas também ações preventivas para evitar danos decorrentes do tratamento incorreto. No streaming, isso se traduz em protocolos que validam atributos críticos imediatamente antes de encaminhar o evento para análise. Por exemplo: ao capturar coordenadas geográficas é possível verificar instantaneamente se estão dentro dos limites plausíveis; casos fora desse padrão são descartados ou sinalizados sem atrasar o restante do fluxo. O componente legal e ético é mais delicado aqui porque a rapidez do processamento contínuo deixa pouco espaço para revisões humanas antes da utilização dos dados coletados. Isso torna imperativa a implementação prévia de anonimização automática sobre campos sensíveis logo no momento da ingestão (Lima, 2019). Ferramentas analíticas capazes de respeitar essas restrições asseguram que respostas rápidas não comprometam garantias fundamentais dos cidadãos. Além disso, procedimentos sólidos

de governança devem regular quem pode acessar camadas brutas do fluxo versus quem trabalha com informações já processadas e anonimizadas (Reis e Housley, 2022). A observabilidade neste contexto adquire relevância extrema para manutenção da qualidade global do sistema (Moses, 2023). Monitorar métricas como taxa média de eventos processados por segundo, latência fim-a-fim entre captura e ação tomada, além da frequência relativa de erros detectados, ajuda a identificar gargalos nas rotinas contínuas. Indicadores anômalos, aumento repentino na rejeição por inconsistência estrutural, podem sinalizar problemas nas fontes originais ou mudanças no comportamento operativo das equipes. Uma característica marcante do streaming na segurança pública é combinar dados estruturados e não estruturados sem esperar ciclos completos. Informações textuais vindas do despacho policial podem ser confrontadas imediatamente com imagens adquiridas pela rede urbana; áudios interceptados são transcritos automaticamente usando NLP e correlacionados com ocorrências registradas minutos antes. Essa integração direta possibilita detectar padrões emergentes que só se tornam perceptíveis quando diferentes formatos informacionais convergem rapidamente no ambiente analítico. Ainda assim existe necessidade clara de calibragem constante dos algoritmos envolvidos para evitar vieses discriminatórios durante análises instantâneas (Rucco et al., 2025). Modelos aplicados sobre fluxos contínuos tendem a operar sobre conjuntos menores que os usados em treinamento offline; nesses casos podem surgir distorções devido à representatividade limitada no sample imediato. Avaliações sistemáticas devem ser inseridas ciclicamente como parte das rotinas automáticas garantindo acurácia mínima aceitável antes das decisões derivadas desse processamento. No plano arquitetural mais amplo, soluções híbridas como a arquitetura Lambda combinam camadas contínuas e batch dentro do mesmo pipeline permitindo balancear frescor informacional com profundidade histórica (Reis e Housley, 2022). Assim integra-se o melhor dos dois mundos: análises imediatas aliadas à validação mais profunda feita posteriormente sobre o conjunto consolidado. Para operações policiais isso significa agir prontamente sobre alertas críticos enquanto cálculos posteriores refinam os modelos preditivos utilizados na prevenção. A transferência transfronteiriça surge novamente como questão sensível quando fluxos recebem contribuições internacionais, por exemplo compartilhamentos entre autoridades estrangeiras durante investigações conjuntas, exigindo cumprimento rigoroso às salvaguardas previstas pela LGPD sobre reconhecimento apenas aos países com proteção equivalente à brasileira (Lima, 2019). Nesses casos é recomendável segmentar logicamente os pipelines para que apenas subconjuntos anonimizados sejam exportados fora da jurisdição local. Do ponto social, utilizar processamento contínuo pode aumentar aceitação pública quando aliado à transparência institucional sobre métodos adotados. Relatórios regulares indicando tempos médios de resposta e explicando quais medidas protegem privacidade ajudam a mitigar percepções negativas relacionadas à vigilância permanente. Por outro lado exageros na coleta sem justificativa clara podem gerar forte resistência comunitária exigindo revisão política das práticas vigentes. Oportunidades adicionais aparecem nas aplicações móveis direcionadas ao cidadão comum, como botões virtuais integrados diretamente às forças policiais enviando geolocalização imediata durante incidentes emergenciais. Esses eventos entram no pipeline streaming recebendo prioridade máxima até retornarem verificação e resposta apropriada pelas equipes competentes. Em última análise técnica-legal-ética, o streaming processa rápido demais para permitir controles manuais extensivos; este fato impõe total dependência das rotinas automatizadas pré-configuradas quanto à limpeza (Goldschmidt et al., 2015), anonimização e validação semântica dos insumos ingeridos. Construir esse alicerce técnico robusto garante que velocidade não venha acompanhada de imprudência institucional nem comprometa conformidade normativa exigida pela LGPD na esfera da segurança pública brasileira.

4.2 Automação de Pipelines

4.2.1 Ferramentas de Automação

A automação de pipelines em segurança pública representa um avanço técnico que liga diretamente práticas tradicionais de ingestão e integração às arquiteturas modernas discutidas anteriormente. Ferramentas de automação permitem converter tarefas extensas, repetitivas ou sujeitas a erro humano em processos controlados por código, melhorando confiabilidade e consistência (Reis e Housley, 2022). A lógica "pipelines como código" viabiliza que o desenho das etapas, desde captura até transformação, seja declarativo, geralmente utilizando linguagens como Python para configurar dependências, condicionais e recursos necessários. Em termos práticos, isso significa que rotinas de ingestão contínua descritas na abordagem *streaming* podem ser orquestradas por sistemas interpretadores que acionam cada tarefa apenas quando as condições especificadas estão atendidas, otimizando uso computacio-

nal e evitando gargalos. No contexto de segurança pública, onde dados sensíveis fluem por múltiplos sistemas simultaneamente, essa automação torna-se quase mandatória para manter auditoria plena e governança efetiva. Ferramentas como Airflow, Prefect ou Apache NiFi oferecem funcionalidades integradas de agendamento dinâmico, monitoramento dos estados das tarefas e reprocessamento automático em caso de falha (Moses, 2023). Ao contrário da execução manual em blocos isolados, essas plataformas trabalham seguindo grafos de dependências: se uma etapa de limpeza disruptiva falhar, impede-se automaticamente que o conjunto seja carregado na camada analítica até a correção. Essa abordagem previne o risco de decisões operacionais serem tomadas com base em registros incompletos ou incorretos. O nível global de automação também reforça conformidade com a Lei Geral de Proteção de Dados (LGPD) ao permitir inserir estágios obrigatórios no fluxo sem depender exclusivamente da intervenção humana. Por exemplo, a anonimização pode ser definida como função permanente logo após as rotinas iniciais de extração; se tal função não for executada corretamente, o próprio sistema interrompe o pipeline evitando que campos identificáveis cheguem à análise (Lima, 2019). Com isso, protege-se dados pessoais sensíveis mesmo em situações onde a rapidez é vital. Em arquitetura distribuída ou híbrida (como DataOps inspirada em DevOps), a automação também engloba versionamento do código e do ambiente. Isso assegura rastreabilidade das alterações aplicadas nas rotinas e facilita auditorias internas ou externas sobre modificações estruturais (Reis e Housley, 2022). Um benefício adicional da automação bem estruturada é sua capacidade de trabalhar com múltiplos domínios informacionais dentro do mesmo ciclo operacional. Uma pipelina pode orquestrar ingestão simultânea de dados textuais oriundos dos sistemas judiciais enquanto processa fluxos audiovisuais capturados pelas redes urbanas. Cada tipo recebe tratamento específico conforme os requisitos técnicos, NLP para textos; reconhecimento facial via modelos treinados para imagens, mas todos passam pelas mesmas camadas centrais exigidas pela governança: validação estrutural, anonimização mínima requerida e logging detalhado (Rucco et al., 2025). No caso da segurança pública essa uniformidade garante que diferentes fontes mantenham coerência regulatória independentemente do formato bruto dos insumos. Além disso, ferramentas modernas já incorporam mecanismos proativos voltados para observabilidade integral dos pipelines. Esses mecanismos incluem coleta automática de métricas de desempenho (tempo médio por tarefa), taxa diária de eventos processados e alertas imediatos em caso de anomalias detectadas (Moses, 2023). Essa vigilância constante torna possível reagir rapidamente a alterações inesperadas nas fontes originais, como mudanças no esquema da base policial ou quebra temporária no fornecimento de feed das câmeras urbanas, antes que tais interrupções prejudiquem o panorama analítico central. Claro que há desafios relacionados à complexidade crescente na automação aplicada à segurança pública. Integrar diferentes ferramentas externas exige garantir compatibilidade tanto técnica quanto legal entre elas: conectar bases hospedadas no exterior impõe salvaguardas adicionais previstas pela LGPD sobre transferências transfronteiriças (Lima, 2019). A solução prática pode envolver segmentar logicamente os pipelines para exportar apenas subconjuntos anonimizados fora da jurisdição nacional ou configurar bloqueios rígidos contra execução automática desses módulos quando determinadas variáveis legais estiverem em conflito. Do ponto de vista operacional local, automatizar scripts destinados a identificar erros no input inicial também reduz custos posteriores com retrabalho (Goldschmidt et al., 2015). Na segurança pública isso significa que valores geográficos incoerentes inseridos manualmente são rejeitados imediatamente sem impactar o restante do fluxo; notas forenses ausentes recebem marcação específica para revisão humana antes do avanço na análise. Com esse mecanismo preventivo diminui-se substancialmente a propagação do erro pelas demais camadas. A adoção dessas ferramentas demanda ainda atenção ao aspecto ético: quanto mais intrusivo for o alcance técnico da automação sobre múltiplas fontes públicas e privadas maior será a responsabilidade institucional sobre seu uso adequado. Orquestrar correlações instantâneas entre diferentes conjuntos informacionais pode potencializar investigações legítimas mas também criar cenários propensos ao abuso caso não haja controle estrito sobre acessos e finalidade dos cruzamentos realizados. Portanto é inevitável incluir autenticação forte associada ao perfil funcional autorizado para acionar ou alterar rotinas presentes nos pipelines automatizados. Por fim é relevante observar tendências globais destacadas pelo conceito moderno de “DataOps”, que transporta práticas maduras da área DevOps para engenharia e operação dos dados (Reis e Housley, 2022). A automatização nesse escopo amplia agilidade tanto na entrega rápida das análises quanto na implementação segura das melhorias sugeridas por incidentes anteriores. Esse ciclo contínuo beneficia sistemas públicos ao aumentar confiabilidade sem abdicar da flexibilidade necessária para incorporar novas fontes ou métodos investigativos em curto prazo. Combinando controle programático rigoroso das etapas técnicas

com vigilância normativa ativa e responsabilidade ética explícita consegue-se um equilíbrio saudável entre velocidade operacional e preservação dos direitos fundamentais no tratamento digital complexo característico da segurança pública contemporânea.

4.2.2 Agentes Inteligentes

O texto que você forneceu já está integralmente em português, portanto, seguindo suas instruções, aqui está o mesmo conteúdo sem alterações: Agentes inteligentes aplicados à automação de pipelines em segurança pública representam uma evolução natural da integração entre processamento de dados, observabilidade e governança discutida anteriormente. A principal característica desses agentes é a capacidade de tomar decisões autônomas dentro de um conjunto de parâmetros pré-estabelecidos, ajustando o fluxo segundo as condições circunstanciais detectadas em tempo real. Diferenciam-se de scripts estáticos tradicionais por incorporarem modelos analíticos e mecanismos adaptativos, capazes de identificar padrões, anomalias e oportunidades de otimização diretamente no ambiente operacional (Moses, 2023). Do ponto de vista técnico, esses agentes podem estar embutidos em plataformas orquestradoras como módulos especializados que não apenas executam tarefas agendadas, mas também monitoram continuamente a qualidade dos insumos e os estados intermediários do pipeline. Ao encontrar registros inconsistentes ou sinais de degradação de performance, como aumento da latência na ingestão contínua, eles são aptos a tomar ações corretivas automatizadas. Tais procedimentos incluem reprocessamento seletivo de lotes críticos, roteamento condicional para fontes secundárias ou mesmo ativação imediata de protocolos de contingência previstos para sistemas sob alta demanda (Reis e Housley, 2022). No contexto legal e ético, inserir capacidade decisória nos agentes traz implicações importantes: qualquer ação automatizada que envolva dados pessoais sensíveis deve obedecer irrestritamente aos princípios da Lei Geral de Proteção de Dados (LGPD), com destaque para necessidade mínima e finalidade específica (Lima, 2019). Por isso, muitos desses agentes incorporam rotinas internas obrigatórias que validam conformidade antes de liberar determinada operação. Se detectarem campos contento identificadores não tratados por anonimização prévia, barram o avanço do dado bruto e sinalizam para intervenção humana. Esse tipo de comportamento programado evita violação inadvertida da privacidade individual, mesmo em cenários nos quais decisões precisam ser tomadas em frações de segundo. Outra função crítica dos agentes inteligentes é otimizar a utilização combinada dos modos batch e streaming já descritos (Rucco et al., 2025). Em operações policiais, um agente pode redirecionar automaticamente determinadas cargas do fluxo contínuo para processamento diferido quando identifica que não há urgência operacional; o inverso também ocorre quando detecta eventos cujas características coincidam com padrões preditivos associados a incidentes iminentes. Essa alternância dinâmica exige que o agente seja capaz tanto de ler metadados contextuais quanto compreender aspectos semânticos dos conteúdos recebidos (Foidl et al., 2023). A utilização desses módulos pode ainda enriquecer diagnósticos mediante integração transparente com catálogos de dados institucionais (Moses, 2023). Ao conhecer taxonomias previamente definidas sobre tipos, formatos e políticas associadas aos datasets disponíveis, o agente inteligente mapeia automaticamente ativos informacionais adequados à análise desejada sem necessidade de consulta manual extensiva. Isso reduz tempo na correlação entre diferentes fontes, por exemplo cruzar geolocalizações oriundas das redes sociais tratadas conforme discutido antes com registros judiciais relevantes, mantendo uniformidade na governança aplicada aos dois domínios. No entanto, operar com autonomia requer abordagens rigorosas sobre segurança da informação e controle organizacional. Permitir que um agente ajuste rotinas críticas implica conceder algum grau de permissão privilegiada; portanto deve-se aplicar políticas estritas como autenticação multifator interna antes da execução de ações delicadas (Reis e Housley, 2022). Além disso, trilhas completas das decisões tomadas pelos agentes tornam-se indispensáveis tanto para auditoria técnica quanto para eventual revisão jurídica em caso de questionamento posterior sobre a legitimidade das operações realizadas. Casos práticos mostram que agentes equipados com capacidades avançadas, como algoritmos especializados em detecção automática baseada em aprendizado supervisionado ou não supervisionado, conseguem reduzir falsos positivos na identificação de eventos relevantes ao mesmo tempo em que preservam agilidade operacional (Rucco et al., 2025). Porém esse benefício vem acompanhado do risco potencial derivado do viés algorítmico: se os dados usados no treinamento forem desbalanceados ou apresentarem distorções históricas ligadas a perfis demográficos específicos, as recomendações resultantes podem reforçar desigualdades pré-existentes sob aparência técnica neutra. Mitigar esse risco envolve avaliações repetidas do desempenho preditivo frente a amostras representativas atualizadas periodicamente (Lima, 2019). Os agentes inteligentes também podem desempenhar

papel central na manutenção preventiva da integridade estrutural dos pipelines (Goldschmidt et al., 2015). Por exemplo, monitorando alterações inesperadas no esquema de uma base policial integrada ao sistema central: se uma coluna obrigatória desaparecer ou mudar formato sem aviso prévio, o agente dispara alerta automático e suspende temporariamente operações dependentes daquela estrutura até validação pela equipe responsável. Tal medida evita propagação silenciosa de erros estruturais pelas camadas subsequentes do pipeline analítico. Outro aspecto relevante é a capacidade desses módulos anteciparem gargalos ou indisponibilidades através da análise preditiva aplicada aos indicadores coletados pela camada supervisora. Tendências como crescimento acelerado no volume diário processado ou variação abrupta nas taxas médias por tarefa podem sinalizar necessidade iminente de escalonamento do ambiente computacional, ajuste automático na distribuição das cargas entre nós processadores ou contratação temporária de recursos adicionais nas nuvens públicas compatíveis com as restrições legais vigentes (Reis e Housley, 2022). A decisão autônoma sobre expandir infraestrutura deve sempre respeitar acordos sobre onde dados sensíveis podem ser hospedados para não ferir dispositivos normativos sobre transferência internacional (Lima, 2019). No plano social e político-institucional cabe reconhecer que agentes capazes de agir sem supervisão direta humana despertam inevitavelmente debate público sobre legitimidade e transparência dessas práticas na esfera estatal. A aceitação desse tipo tecnológico depende fortemente da existência simultânea de mecanismos claros que possibilitem auditoria independente das regras implementadas nos modelos decisórios e análise externa periódica dos impactos gerados nas populações-alvo das operações conduzidas pelo Estado. Descrever abertamente metodologias usadas pelos agentes, resguardando aspectos sigilosos imprescindíveis à segurança, favorece construção gradativa da confiança pública nesse recurso tecnológico aplicado ao tratamento e integração avançada dos dados sensíveis coletados para fins legítimos da segurança coletiva.

5 Arquiteturas e Ferramentas de Processamento

5.1 Arquiteturas de Ingestão

5.1.1 Lambda

A arquitetura Lambda tem como propósito integrar, de forma harmônica, duas abordagens distintas de processamento de dados: a análise histórica e consolidada típica do modelo em lotes (*batch*) e o tratamento rápido e contínuo presente no processamento em fluxo (*streaming*). Essa união atende especialmente ambientes que não podem abrir mão nem da profundidade analítica obtida com grandes blocos consolidados, nem da agilidade operacional derivada de respostas imediatas (Reis e Housley, 2022). Em aplicações voltadas à segurança pública, essa flexibilidade se torna estratégica. Com a Lambda, por exemplo, é possível cruzar registros criminais acumulados durante meses com alertas gerados em tempo real por câmeras urbanas ou sistemas automáticos de leitura de placas, maximizando tanto a eficácia preventiva quanto a capacidade investigativa. Na prática, arquiteturas Lambda segregam o pipeline em duas camadas que operam paralelamente. A camada batch absorve entradas massivas vindas de sistemas como bancos relacionais policiais ou históricos judiciais, processando-as segundo rotinas mais pesadas e demoradas que envolvem extração, limpeza sistemática, anonimização controlada e integração com Data Warehouses para posterior exploração (Rucco et al., 2025). Nessas rotinas há espaço para aplicar procedimentos regulatórios previstos pela Lei Geral de Proteção de Dados (LGPD), como expurgo de atributos irrelevantes conforme o princípio da necessidade mínima e pseudonimização para proteger sujeitos envolvidos nos registros (Lima, 2019). Já a camada streaming recebe eventos instantâneos oriundos de sensores urbanos, aplicativos táticos usados pelos agentes ou fluxos provenientes das redes sociais monitoradas para fins autorizados. Esses eventos são tratados quase simultaneamente à captura, permitindo resposta policial imediata ou acionamento automático de protocolos preventivos. A coordenação entre essas duas camadas impõe desafios técnicos importantes: reconciliar formatos distintos, estruturados nas fontes batch e frequentemente não estruturados no streaming, exige mecanismos robustos de transformação e harmonização (Goldschmidt et al., 2015). Ferramentas modernas já permitem que pipelines híbridos processem texto livre com NLP ao mesmo tempo que executam consultas SQL complexas sobre tabelas consolidadas, mas a orquestração precisa considerar regras estritas sobre acesso e circulação dos dados. Governança nesse contexto requer políticas explícitas sobre quais perfis podem interagir com cada camada: operadores táticos podem consumir saídas anonimizadas do streaming para ações imediatas; analistas estratégicos ficam responsáveis por correlacionar os resultados históricos do batch com movimentações detectadas na camada contínua

(Reis e Housley, 2022). Do ponto de vista legal e ético existe um aspecto sensível: a camada streaming processa informação bruta em alta velocidade, deixando pouca margem para retificação manual antes da análise. Isso exige configuração obrigatória para que qualquer campo sensível seja anonimizado automaticamente na ingestão (Lima, 2019). Já na camada batch há maior latitude temporal para revisões humanas e aplicação meticulosa das normas regulatórias antes da dispensa dos dados ao ambiente analítico. No entanto falhas na sincronização entre as camadas podem gerar inconsistências ou duplicação de eventos; agentes inteligentes integrados ao pipeline conseguem detectar discrepâncias entre registros históricos e fluxos contínuos, atuando proativamente na correção. A observabilidade torna-se crítica nesse arranjo híbrido. Monitorar indicadores como latência fim-a-fim na camada contínua, duração média das execuções batch e taxas relativas de erros estruturais fornece diagnóstico preciso da "saúde" operacional (Moses, 2023). Sinais inesperados, aumento drástico na rejeição de eventos por inconsistências geográficas no streaming ou queda abrupta no volume processado pelo batch, demandam respostas rápidas para evitar prejuízo às decisões estratégicas baseadas nesses insumos. Um exemplo concreto aplicado à segurança pública: durante grandes eventos urbanos um sistema basado em Lambda pode correlacionar padrões históricos apreendidos via batch (como ocorrências recorrentes mapeadas geograficamente) com entradas imediatas capturadas pelo streaming (fluxo detectado por sensores na área do evento). Ao identificar correspondências perigosas ou situações potencialmente críticas, aciona alertas simultâneos aos agentes em campo enquanto registra todos os elementos contextuais necessários para investigações posteriores. Isso mantém coerência regulatória porque as saídas utilitárias do streaming chegam aos operadores já sem identificadores diretos graças às rotinas automáticas configuradas conforme LGPD. Outra vantagem prática dessa arquitetura é a compatibilidade com diferentes ferramentas já empregadas nos órgãos públicos. Soluções como Apache Kafka alimentam o lado streaming, enquanto sistemas ETL robustos cuidam do lado batch usando infraestrutura escalável horizontalmente para lidar com volumes massivos acumulados nas bases históricas policiais. A integração desses ambientes pode ser coordenada por orquestradores capazes de aplicar políticas específicas a cada tipo de carga e garantir resiliência frente a falhas parciais no pipeline híbrido. Ao mesmo tempo é importante reconhecer riscos inerentes: vieses oriundos dos modelos aplicados sobre o fluxo possam reforçar interpretações enviesadas se não forem recalibrados periodicamente (Rucco et al., 2025). Avaliações constantes ajudam a impedir que padrões históricos incorretos sejam perpetuados automaticamente pela combinação das camadas. Nesse sentido a governança deve incluir revisão cíclica tanto do código operador nos pipelines quanto dos critérios semânticos usados para classificar dados sensíveis. Por fim o impacto social da adoção desse tipo arquitetural demanda comunicação transparente perante a população. Explicar publicamente, dentro dos limites exigidos pela segurança institucional, como duas frentes distintas trabalham juntas na detecção e prevenção favorece aceitação comunitária dessa prática tecnológica (Reis e Housley, 2022). Documentar claramente metodologias aplicadas, protocolos regulamentares seguidos e resultados obtidos reduz percepções equivocadas sobre invasão indevida da esfera privada. A arquitetura Lambda, assim implementada dentro das balizas técnicas, legais e éticas adequadas, oferece um equilíbrio sofisticado entre profundidade histórica e velocidade reativa capaz de atender demandas complexas típicas da segurança pública contemporânea enquanto preserva direitos fundamentais reconhecidos pela legislação vigente.

5.1.2 Kappa

A arquitetura Kappa surge como alternativa conceitual e prática ao modelo Lambda, questionando a necessidade da duplicação de caminhos de código e ambientes separados para processamento em lote e em fluxo. Seu princípio central é utilizar um único sistema de processamento orientado a eventos para todo o ciclo de ingestão, armazenamento e consumo de dados. Em vez de separar dados históricos em pipelines batch e dados recentes em pipelines streaming, a Kappa processa fluxos contínuos e recorre à reprocessamento do histórico diretamente a partir do mesmo stream, seja para análises tardias ou correções. Essa abordagem simplifica alguns aspectos operacionais ao eliminar a complexidade de coordenar duas arquiteturas distintas, mas também introduz desafios próprios, especialmente quando o volume acumulado é massivo e as exigências legais sobre retenção e uso restrito se fazem presentes. Na segurança pública, essa proposta pode ser atraente justamente pela homogeneidade dos pipelines: sensores urbanos, sistemas policiais internos e captações de redes sociais integrariam o mesmo backbone tecnológico orientado a eventos. Se um algoritmo precisa ser atualizado, por exemplo, um classificador que detecta padrões suspeitos em vídeos, basta reprocessar o fluxo armazenado na mesma infraestrutura para que todos os dados antigos sejam reinterpretados pelo novo modelo. Isso oferece agilidade

na correção ou evolução das análises sem necessidade de duplicações complicadas entre ambiente batch e streaming (Reis e Housley, 2022). Contudo, essa unificação amplia pressão sobre escalabilidade: fluxos contínuos com cargas históricas exigem recursos computacionais capazes de manter latências aceitáveis tanto no processamento instantâneo quanto no replay potencialmente pesado. Do ponto de vista técnico-operacional, adotar Kappa implica investir numa camada robusta de armazenamento que conserve os eventos originais por tempo suficiente para atender demandas legais e institucionais sem contrariar a Lei Geral de Proteção de Dados (LGPD) (Lima, 2019). Na prática, isso significa conciliar interesse investigativo em revisitar eventos passados com obrigação legal de descartar ou anonimizar dados pessoais cujo prazo legítimo tenha expirado. Como não há separação natural entre dados antigos e novos na arquitetura Kappa, mecanismos automáticos precisam rastrear constantemente metadados temporais e aplicar políticas de expurgo ou transformação inline durante reprocessamentos previstos. Uma característica essencial da Kappa está na capacidade nativa de combinar operações sobre dados estruturados e não estruturados sem alternância arquitetural. Aplicações específicas na segurança pública incluem desde correlacionar transcrições textuais obtidas por sistemas de interceptação autorizada até examinar imagens oriundas das câmeras urbanas usando redes neurais convolucionais no mesmo pipeline (Rucco et al., 2025). Essa convergência cria oportunidades para detectar padrões emergentes cruzando diferentes formatos informacionais quase simultaneamente à ingestão; porém exige módulos dedicados para padronização dos insumos antes que estes avancem para as camadas analíticas compartilhadas. A governança em um cenário Kappa assume papel complexo porque todos os acessos ocorrem sobre um mesmo conjunto lógico. Diferente da segmentação natural existente na Lambda, onde perfis distintos atuam sobre conjuntos diferenciados no batch ou streaming, aqui é necessário controlar permissões granulares dentro do próprio stream persistido (Reis e Housley, 2022). Operadores táticos podem ver apenas campos anonimizados; analistas estratégicos recebem acesso expandido aos identificadores conforme autorizações expressas. Implementar isso com consistência demanda integração forte com sistemas IAM que assegurem logs detalhados sobre qualquer interação, leitura, exportação ou modificação, realizada nos registros. Outro ponto importante é a observabilidade. Monitorar métricas como throughput médio actual dos eventos processados em tempo real versus taxas alcançadas durante replays ajuda a identificar gargalos latentes (Moses, 2023). O risco aqui é dobrado: problemas podem afetar tanto respostas imediatas quanto revisões históricas se não houver alertas diferenciados para cada contexto. Ferramentas modernas podem emitir sinais preventivos quando estimativas projetadas indicam degradação iminente sob carga acumulada alta. Em termos éticos e sociais, usar Kappa na segurança pública exige cautela redobrada: manter fluxos completos por períodos prolongados aumenta superfície potencial para abuso interno ou externo das informações sensíveis coletadas (Lima, 2019). Estratégias como anonimização automática logo na ingestão ajudam a reduzir riscos iniciais, mas não substituem auditorias regulares sobre quem acessa quais partes do histórico persistente. Essas auditorias devem avaliar inclusive se o uso efetivo respeitou finalidade declarada original; qualquer desvio pode comprometer confiança institucional já que o modelo facilita acesso similar aos eventos capturados ontem e há meses atrás sem distinção clara. A automação interage fortemente com esse desenho arquitetural: agentes inteligentes previamente discutidos podem agir em tempo real dentro da Kappa ajustando filtros conforme padrões descobertos ou interrompendo replays ao detectar presença inadvertida de campos sensíveis não tratados (Moses, 2023). Essa capacidade autônoma diminui dependência da supervisão humana contínua sem abrir mão da conformidade normativa, desde que programada sob algoritmos eticamente calibrados e revistos periodicamente contra vieses discriminatórios (Rucco et al., 2025). Por fim, convém reconhecer limitações observadas na adoção global da arquitetura Kappa: relatos apontam custo elevado operacionalmente associado à escalabilidade necessária para grandes datasets históricos aliados ao streaming ativo. No ambiente estatal brasileiro essas restrições poderiam impactar desde orçamentos limitados até conformidade com diretrizes ligadas à soberania digital sobre dados sensíveis nacionais. Uma mitigação possível seria adotar variações parciais da filosofia Kappa aplicando seu conceito unificado apenas a subconjuntos críticos nos quais simplicidade operacional supere vantagem trazida pela separação tradicional batch/streaming. Implementada sob salvaguardas técnicas sólidas, criptografia forte tanto no armazenamento quanto no trânsito (Reis e Housley, 2022), governança minuciosa sobre permissões e observabilidade constante, aliada às garantias jurídicas exigidas pela LGPD (Lima, 2019), a arquitetura Kappa pode fornecer ganhos importantes em agilidade analítica e redução da duplicidade operacional nas aplicações voltadas à segurança pública. Ainda assim precisa ser cuidadosamente avaliada frente ao impacto social potencial do seu modelo homogêneo sobre o acúmulo prolongado dos eventos capturados e sua utilização futura pelas autoridades

competentes.

5.2 Ferramentas de Integração

5.2.1 Apache Kafka

Apache Kafka é uma plataforma de *streaming* de eventos de código aberto voltada para ingestão, transporte e processamento contínuo de dados, com aplicação direta em cenários de segurança pública que exigem baixa latência e alta confiabilidade na comunicação entre sistemas. A utilização do Kafka nesse contexto tem como benefício central a capacidade de receber fluxos heterogêneos, registros estruturados oriundos dos sistemas policiais, dados não estruturados provenientes de câmeras urbanas ou mensagens interceptadas mediante ordem judicial, e encaminhá-los a múltiplos consumidores simultaneamente (Shapira et al., 2022). A arquitetura interna do Kafka organiza os dados em *topics*, cada um representando uma sequência ordenada de eventos. Esse modelo facilita criar "canais" dedicados a diferentes tipos de informações, permitindo que apenas serviços autorizados consumam aqueles tópicos específicos. Assim, no uso estatal voltado à segurança pública, é possível segmentar fluxos sensíveis aplicando governança granular conforme perfil funcional definido para cada unidade operacional (Reis e Housley, 2022). Do ponto de vista técnico-operacional, o Kafka se destaca pela capacidade de entregar latências extremamente baixas (reportadas como próximas a 2 milissegundos), o que possibilita respostas quase instantâneas a eventos críticos (Moses, 2023). Essa característica o torna peça fundamental quando integrado a arquiteturas descritas anteriormente como Lambda ou Kappa (5.1.2), fornecendo à camada contínua um meio robusto para transportar eventos diretamente dos dispositivos captadores até as ferramentas analíticas e sistemas de decisão. Essa eficiência depende entretanto da correta configuração dos parâmetros relacionados ao *throughput* e à persistência dos dados; ajustar fatores como número de partições e replicação entre nós é essencial para comportar variações bruscas na taxa de ingestão sem perdas ou duplicações. A natureza distribuída do Kafka favorece escalabilidade horizontal: novos nós podem ser adicionados ao cluster sem interrupção significativa do serviço, ampliando o processamento conforme a demanda, um requisito típico em grandes operações policiais que necessitem aumentar instantaneamente a capacidade durante eventos extraordinários. No entanto essa flexibilidade carrega obrigações legais quando envolve dados pessoais sensíveis: se parte da infraestrutura estiver hospedada fora da jurisdição nacional, entram em cena as salvaguardas previstas pela Lei Geral de Proteção de Dados (LGPD), incluindo restrição ao envio internacional sem garantias equivalentes às nacionais. Nesses casos recomenda-se configurar *brokers* locais para filtrar ou anonimizar dados antes do envio a clusters externos. Quanto à conformidade normativa e ética, configurar pipelines no Kafka deve incluir estágios automáticos que anonimizem atributos identificáveis logo na entrada do tópico. Isso previne que consumidores internos tenham acesso direto a informações brutas capazes de identificar indivíduos em contextos delicados. Implementar tal prática pode utilizar processadores intermediários (*stream processors*) ligados ao próprio cluster que executam rotinas pré-definidas para substituição por pseudônimos ou agregação estatística (Lima, 2019). Além disso, trilhas detalhadas (logs) sobre quem publicou e quem consumiu determinado evento precisam ser mantidas para auditoria posterior e responsabilização institucional em casos suspeitos (Reis e Housley, 2022). No campo da automação descrito anteriormente (4.2.1), agentes inteligentes podem ser integrados diretamente ao ecossistema Kafka usando APIs compatíveis com sua arquitetura distribuída. Esses módulos analisam padrões nos fluxos em tempo real, identificam anomalias, como mensagens corrompidas ou volumes fora do padrão esperado, e reconfiguram topologias dinamicamente para contê-las antes que causem impacto operacional mais amplo. Tais ações automatizadas devem seguir parâmetros rigorosos previamente acordados com autoridades competentes para evitar intervenções indevidas no fluxo legítimo das informações capturadas. O aspecto da observabilidade é especialmente relevante: o Kafka permite coletar métricas como atraso médio por consumidor, taxa atualizada de publicação por produtor e tamanho acumulado das filas em cada tópico. Esses indicadores ajudam equipes técnicas a diagnosticar gargalos e planejar expansões ou otimizações específicas (Moses, 2023). Em operações críticas da segurança pública, quedas abruptas na taxa média podem indicar falha nas fontes originais, p. ex., interrupção no sistema de câmeras, exigindo mobilização imediata das equipes responsáveis por garantir continuidade informacional. Embora possua vantagens evidentes sobre ferramentas tradicionais de integração (*ETL*), o uso do Kafka também apresenta riscos caso falte calibragem nos modelos analíticos acoplados. Aplicações que utilizem aprendizado supervisionado ou não supervisionado dentro dos processadores podem herdar vieses discriminatórios enraizados nos próprios datasets ingeridos;

portanto é necessário revisar continuamente tais modelos com amostras equilibradas e representativas (Rucco et al., 2025). Esse cuidado evita inferências injustas contra determinados grupos sociais, protegendo direitos fundamentais enquanto mantém qualidade técnica das predições operacionais. Outro ponto crítico envolve retenção prolongada dos tópicos. A configuração padrão pode manter eventos armazenados por períodos definidos; prolongar esse tempo traz benefícios investigativos mas amplia risco potencial de abuso interno ou externo caso controles não sejam suficientemente restritivos. É recomendável delimitar retenção segundo princípio da necessidade mínima previsto pela LGPD (Lima, 2019), descartando conteúdos cujo valor operacional tenha expirado ou transformando-os irreversivelmente antes da retenção ampliada. Em integrações com outras plataformas tecnológicas internas aos órgãos públicos, como Data Warehouses ou sistemas judiciais eletrônicos, o Kafka atua como camada intermediária desacoplada capaz de encaminhar eventos aos destinos corretos sem expor diretamente estruturas internas sensíveis. Isso mantém isolamento apropriado entre bases distintas preservando coerência regulatória mesmo sob altos índices transacionais diários (Reis e Housley, 2022). No entanto carece-se aqui novamente garantir que tais roteamentos associados estejam documentados formalmente para passar por auditoria governamental quando necessário. No plano social, usar Apache Kafka em larga escala exige comunicação clara sobre seu papel nas operações estatais: explicar à população que trata-se de uma ferramenta técnica voltada à eficiência operacional e respeito às normas vigentes ajuda a mitigar resistências ligadas à percepção negativa sobre "monitoramento constante". Fornecer relatórios públicos consolidados sobre desempenho agregado sem expor indivíduos contribui para reforçar confiança institucional nos órgãos responsáveis pela segurança coletiva utilizando essa tecnologia. Quando esses elementos técnicos, legais e éticos são harmonizados dentro das práticas adotadas no Kafka, anonimização preventiva, governança rígida sobre consumo e produção, observabilidade contínua e mitigação ativa contra vieses algorítmicos, obtém-se uma base sólida para transporte seguro y eficiente das informações críticas tratadas pela segurança pública nacional (Shapira et al., 2022; Moses, 2023).

5.2.2 Apache Spark

Apache Spark é um mecanismo unificado de computação distribuída, amplamente utilizado em cenários que exigem processamento intensivo de grandes volumes de dados, e mostra-se particularmente adequado para aplicações sensíveis como as relacionadas à segurança pública. Ele combina um núcleo capaz de executar tarefas paralelas em clusters com bibliotecas especializadas para diferentes finalidades, desde consultas SQL sobre dados estruturados até aprendizado de máquina e processamento de fluxos contínuos. Essa versatilidade permite integrar em uma mesma plataforma múltiplos tipos de dados, como registros tabulares oriundos de sistemas administrativos e sinais multimídia capturados por sensores urbanos, sem necessidade de alternar infraestrutura ou lógica operacional. No contexto técnico, Spark oferece APIs consistentes em linguagens como Python, Scala, Java e R, fato que facilita a interoperabilidade entre equipes diversas e a integração com outros componentes já utilizados nas arquiteturas discutidas anteriormente. A filosofia unificada do framework significa que operações analíticas heterogêneas podem compartilhar o mesmo motor de execução, beneficiando-se da otimização cruzada entre etapas. Por exemplo, ao carregar informações consolidadas de ocorrências criminais via comando SQL e aplicar sobre ellas um modelo preditivo do MLlib, o Spark consegue fundir leituras e cálculos em varreduras únicas sobre os dados, reduzindo custo computacional e tempo de resposta. Essa característica é valiosa na segurança pública quando se deseja reagir rapidamente a padrões identificados durante eventos críticos. Um diferencial relevante do Spark está na gama de bibliotecas nativas: Spark SQL lida com dados estruturados usando sintaxe declarativa; Spark Streaming e Structured Streaming oferecem suporte a ingestão contínua; MLlib traz algoritmos robustos que podem ser aplicados desde segmentações até detecção de anomalias; GraphX viabiliza análises mais complexas envolvendo redes relacionais (Chambers e Zaharia, 2018). Além disso há centenas de bibliotecas externas desenvolvidas pela comunidade que expandem funcionalidades, incluindo conectores especializados para bases forenses ou sistemas geoespaciais relevantes para policiamento ostensivo. Essa extensibilidade abre espaço para customizar pipelines conforme exigências normativas específicas da Lei Geral de Proteção de Dados (LGPD), por exemplo aplicando transformações anonimizadoras logo nos estágios iniciais da análise. Do ponto de vista arquitetural-operacional voltado à segurança pública, a habilidade do Spark em escalar horizontalmente, executando desde um único nó até milhares, favorece iniciativas que precisam atender tanto demandas cotidianas quanto picos extraordinários gerados por eventos imprevistos. Em operações rotineiras é possível processar lotes diários dos sistemas policiais

centrais; já em situações emergenciais pode-se configurar Structured Streaming para consumir tópicos transmitidos por um sistema intermediário como Apache Kafka (5.2.1), garantindo latência reduzida no fluxo até análises cruciais. Este tipo de integração se beneficia das capacidades distribuídas do Spark combinadas à durabilidade e ordenação assegurada pelo Kafka. A governança na manipulação dos dados dentro do ecossistema Spark exige atenção direcionada: dado seu potencial para unificar múltiplas fontes no mesmo pipeline lógico, controles granulares sobre permissões são indispensáveis para evitar acessos indevidos. Modelos IAM integrados aos jobs executados pelo Spark podem restringir quais operadores submetem códigos capazes de acessar colunas sensíveis; logs detalhados devem registrar cada ação realizada no cluster (Reis e Housley, 2022). Adicionalmente, incorporar verificações automáticas nos scripts ajuda a assegurar conformidade prévia com princípios da LGPD antes que qualquer carga seja persistida ou compartilhada externamente (Lima, 2019). Outra aplicação prática é o pré-processamento inteligente realizado ainda no cluster principal. À medida que dados são ingeridos, sejam streams contínuos ou lotes consolidados, o framework pode aplicar funções definidas pelo usuário (UDFs) para higienização imediata: correções em campos inválidos, remoção seletiva baseada em regras legais ou pseudonimização automática (Moses, 2023). Com isto garante-se qualidade suficiente sem criar etapas adicionais fora do fluxo centralizado, reforçando também observabilidade através das métricas internas expostas pelo próprio motor sobre tempo médio por tarefa, volume processado e taxa de falhas detectadas. A questão da escalabilidade acompanha implicações éticas: quanto mais robusta for a capacidade técnica para correlar diferentes conjuntos tratados pelo Spark, maior o potencial risco associado à reidentificação indireta. Por exemplo, cruzar registros anonimizados mas temporalmente próximos oriundos das redes sociais monitoradas com logs oficiais pode revelar padrões inadvertidamente pessoais. Esse cenário reforça a necessidade não apenas da anonimização formal mas também da avaliação contínua dos modelos analíticos contra vieses discriminatórios ou falhas sistêmicas (Rucco et al., 2025). Dentro dessa ótica crítica, usar Spark para aprendizado supervisionado em segurança pública requer práticas metodológicas sólidas: separar amostras representativas; realizar validação cruzada frequente; monitorar deriva dos modelos implantados. Ao aparecer sinal significativo de degradação preditiva ou enviesamento contra determinados grupos demográficos, as rotinas precisam ser revistas e corrigidas, idealmente com acompanhamento por auditorias independentes compostas por especialistas técnicos e jurídicos. No plano social-operacional é recomendável comunicar à população papéis concretos desempenhados por tecnologias como o Spark nos programas estatais. Relatar resultados agregados obtidos com base na análise desses dado pode auxiliar na construção gradual da confiança pública e neutralizar discursos infundados sobre uso indiscriminado da informação sensível. Evidenciar publicamente etapas como anonimização sistemática aplicada antes do processamento paralelo ou explicitar salvaguardas normativas empregadas contribui tanto para transparência quanto para aceitação comunitária dessas práticas. Por fim vale considerar que o ecossistema do Spark pode operar tanto isoladamente on-premises quanto em nuvens públicas/privadas compatíveis com exigências regulatórias impostas pela LGPD sobre localização física dos dados pessoais (Reis e Housley, 2022). A escolha entre essas modalidades deve avaliar simultaneamente custo-benefício operacional e grau aceitável de exposição jurídica na transferência internacional, aplicando criptografia forte end-to-end sempre que fluxos trafegarem entre jurisdições distintas. Sob esse conjunto integrado de aspectos técnicos, legais e éticos, Apache Spark configura-se como peça central capaz de sustentar operações complexas da segurança pública contemporânea combinando processamento massivo distribuído com governança rigorosa e respeito efetivo aos direitos fundamentais previstos pela legislação vigente.

6 Governança e Observabilidade

6.1 Políticas Internas de Gestão de Dados

As políticas internas de gestão de dados, aplicadas no contexto de segurança pública, constituem o eixo de sustentação organizacional para garantir que práticas técnicas e operacionais estejam alinhadas com princípios éticos e regulatórios. Elas não se restringem apenas a definir regras abstratas sobre uso da informação, mas também estabelecem protocolos concretos para coleta, armazenamento, processamento e descarte dos dados. A efetividade dessas políticas depende da integração direta com os mecanismos técnicos descritos anteriormente em arquiteturas como Lambda, Kappa e ferramentas de ingestão como Apache Kafka ou Spark, uma vez que essas tecnologias só oferecerão resultados confiáveis se operarem em conformidade com diretrizes institucionais claras (Reis e Housley, 2022). No

plano estrutural, uma política bem delineada descreve papéis e responsabilidades individuais e coletivos no ciclo de vida dos dados. Isso envolve determinar quem pode acessar informações brutas vindas de sistemas policiais ou redes sociais monitoradas, sob quais motivos documentados e quais limites temporais regem essa permissão. Perfis funcionais distintos, como operadores táticos em campo versus analistas estratégicos, devem ter escopos específicos calibrados de acordo com a sensibilidade do conteúdo que manipulam. Essas delimitações reduzem a probabilidade de uso indevido e sustentam conformidade com exigências da Lei Geral de Proteção de Dados (LGPD) (Lima, 2019). A governança participativa também é componente-chave dessas políticas, pois a tomada de decisão sobre regras internas deve incluir tanto especialistas técnicos quanto representantes jurídicos e administrativos. Tal integração propicia que definições sobre retenção máxima, anonimização obrigatória ou padrões mínimos de qualidade sejam realistas frente às capacidades tecnológicas existentes e aos marcos normativos vigentes. É recomendável que cada política interna seja periodicamente revisada considerando avanços em automação, como agentes inteligentes capazes de agir autonomamente na rotina dos pipelines, sem abrir mão da supervisão humana nos pontos sensíveis (Moses, 2023). Outro elemento essencial é a formalização das práticas relacionadas à anonimização e pseudonimização no fluxo institucional. Definir claramente métodos aceitos para remover identificadores diretos antes que dados sejam utilizados em análises ou transferidos entre unidades previne violações inadvertidas à privacidade individual (Lima, 2019). Essa etapa deve estar acoplada a gatilhos automáticos configurados nas ferramentas de ingestão; por exemplo, processadores intermediários no Kafka podem executar anonimização prévia antes do evento entrar num tópico acessível por múltiplos consumidores (Shapira et al., 2022). Em paralelismo, rotinas batch consolidadas no Spark podem aplicar função padronizada para supressão seletiva baseada em regras legais durante execução periódica (Rucco et al., 2025). A observabilidade operacional precisa figurar no centro das políticas internas para permitir acompanhamento constante da "saúde" dos pipelines e bases gerenciadas. Determinar métricas obrigatórias, tais como latência média entre captura e disponibilização analítica, volume tratado por janela temporal ou porcentagem de registros rejeitados por inconsistência estrutural, auxilia na detecção imediata de falhas técnicas ou alterações inesperadas na qualidade dos insumos (Moses, 2023). Esses indicadores não devem ser apenas coletados automaticamente; é necessário explicitar, na política interna, o fluxo decisório que será disparado quando se observar um desvio crítico nesses parâmetros. Do ponto social, as políticas internas abrangem procedimentos transparentes para comunicação pública quando operações envolvem coleta massiva ou uso intensivo de dados potencialmente sensíveis. Embora nem todos os detalhes possam ser divulgados por questões operacionais, apresentar relatórios agregados sobre cumprimento das diretrizes normativas reforça a confiança comunitária nas instituições responsáveis pela segurança coletiva (Reis e Housley, 2022). Essa transparência deve ser construída evitando termos genéricos excessivos; ao invés disso, demonstrar exemplos reais (anonimizados) das salvaguardas aplicadas aumenta compreensão popular sobre empenho institucional em proteger direitos fundamentais. Os mecanismos internos também têm papel fundamental na gestão da retenção e descarte dos dados armazenados. A LGPD preconiza o princípio da necessidade mínima; portanto arquivos cuja utilidade operacional cessa devem ser eliminados ou transformados irreversivelmente nos prazos definidos pela política (Lima, 2019). Integrar scripts automatizados capazes de realizar expurgo controlado ajuda evitar acumulação desnecessária que poderia se converter em risco legal ou reputacional. Não menos importante é contemplar nas políticas internas estratégias robustas contra ameaças internas e externas à integridade informacional. Plano detalhado sobre controle de acesso físico às salas onde servidores críticos estão instalados deve coexistir com autenticação multifator nos sistemas virtuais que processam os dados mais sensíveis (Reis e Housley, 2022). Cada ação executada nesses ambientes precisa gerar registro indivisível vinculando operador responsável, momento exato e justificativa, trilhas essenciais para auditoria técnica ou jurídica posterior. Integrar medidas proativas relacionadas ao viés algorítmico reforça ainda mais essas diretrizes organizacionais. Dados usados para treinar modelos implantados diretamente na segurança pública devem ser analisados periodicamente por equipes mistas (técnicas/jurídicas) para detectar distorções discriminatórias (Rucco et al., 2025). Se identificadas desigualdades nas inferências geradas por esses algoritmos, fruto do balanceamento inadequado nas amostras históricas processadas pelo Spark ou fluxo contínuo via Kafka, a política interna deve exigir ajuste imediato antes que tais modelos voltem à produção. Por fim cabe enfatizar que essas políticas não são estáticas: precisam evoluir sincronizadas aos avanços tecnológicos e mudanças no cenário social-legal. Ferramentas como agentes inteligentes podem servir não só para atuar dentro dos pipelines mas também para monitorar continuamente aderência às regras definidas nas próprias políticas institucionais (Moses, 2023). As-

sim constrói-se um ciclo autorregulatório onde tecnologia serve simultaneamente como meio executivo das diretrizes internas e instrumento fiscalizador do seu cumprimento efetivo. Combinando clareza documental, integração automática das salvaguardas nos sistemas operacionais e ampla fiscalização intrainstitucional obtém-se um ecossistema informacional mais seguro, transparente e eficiente na gestão dos dados aplicados à segurança pública brasileira.

6.2 Transparência no Uso de Dados Públicos

A transparência no uso de dados públicos, especialmente no campo da segurança pública, implica um equilíbrio delicado entre a prestação de contas à sociedade e a manutenção das salvaguardas necessárias para proteger direitos individuais. Esse princípio pressupõe que cidadãos, organizações da sociedade civil e órgãos de controle possam compreender, com clareza suficiente, como as informações sob gestão estatal são coletadas, processadas, armazenadas e aplicadas na formulação ou execução de políticas públicas. No entanto, essa abertura não pode se dar de forma indiscriminada: no contexto brasileiro, a Lei Geral de Proteção de Dados (LGPD) impõe limites claros sobre a divulgação de dados pessoais e sensíveis, inclusive quando esses dados são tratados por entes públicos com atribuições relacionadas à segurança. No plano operacional, mecanismos efetivos de transparência dependem fortemente de processos técnicos consolidados que promovam anonimização ou pseudonimização antes da eventual disponibilização dos conjuntos informacionais (Lima, 2019). Isso significa que qualquer dado individualmente identificável precisa passar por transformação segura que impossibilite, ou torne extremamente difícil, a reidentificação posterior. Ferramentas modernas integradas aos pipelines já permitem implementar essas operações logo nas etapas iniciais do fluxo (Rucco et al., 2025), diminuindo o risco de exposição inadvertida. Por exemplo, ao processar registros georreferenciados provenientes de rondas policiais, é possível agregar coordenadas por áreas amplas em vez de exibir localizações exatas vinculadas a ocorrências específicas. A criação e manutenção de portais públicos para divulgação desses dados devem observar padrões técnicos mínimos para assegurar que as informações exibidas sejam compreensíveis e contextualizadas. Publicar estatísticas brutas sem notas explicativas ou metadados pode induzir interpretações distorcidas com impacto social direto. É recomendável adotar um modelo em camadas: disponibilizar ao grande público relatórios agregados e análises interpretativas; reservar datasets mais detalhados apenas a pesquisadores ou instituições legalmente autorizadas sob acordos específicos de confidencialidade (Reis e Housley, 2022). Do ponto de vista da governança interna, torna-se essencial manter trilhas completas das interações realizadas com os dados liberados, desde quem solicitou acesso até quais manipulações ocorreram até sua publicação. Assim, além de cumprir exigências normativas relacionadas à rastreabilidade, cria-se um mecanismo robusto para investigar desvios ou vazamentos eventualmente associados à cadeia interna de processamento. Esta abordagem alinha-se às práticas já descritas anteriormente quanto ao monitoramento constante da saúde dos pipelines e à definição rigorosa das permissões por perfil funcional. A tecnologia desempenha papel decisivo nesse processo. Soluções como Apache Kafka (5.2.1) e Apache Spark (5.2.2), amplamente usadas para integração e análise em larga escala na segurança pública, podem ser configuradas para gerar automaticamente indicadores públicos sobre desempenho do processamento sem expor conteúdo sensível (Moses, 2023). Essa automação favorece transparência ativa porque reduz o tempo entre a compilação das métricas internas e sua publicação controlada nos canais institucionais. Outro ponto relevante diz respeito às metodologias adotadas nos modelos analíticos aplicados sobre tais dados. Se esses modelos influenciam diretamente decisões estatais, como alocação policial em determinadas áreas ou definição de prioridades investigativas, é fundamental documentar publicamente seus princípios gerais, critérios utilizados e processos contínuos de avaliação anti-vieses (Rucco et al., 2025). A ausência dessa camada explicativa enfraquece o valor democrático da transparência porque impede o escrutínio técnico-científico independente sobre a justiça e acurácia dessas ferramentas. Os desafios aumentam quando se trata de integrações internacionais envolvendo transferência transfronteiriça dos dados públicos tratados com finalidades securitárias. Nesses casos deve-se respeitar integralmente as salvaguardas previstas pela LGPD quanto ao reconhecimento apenas de países com nível adequado de proteção, documentando as garantias adicionais providas em contratos ou protocolos multilaterais. Tornar essa documentação acessível em formato resumido fortalece a percepção pública sobre controle estatal responsável frente à cooperação internacional. Do ponto social, há um aspecto crítico frequentemente negligenciado: formatos e linguagens adotados nas divulgações precisam ser inclusivos, possibilitando compreensão por diferentes segmentos populacionais sem exigir conhecimento técnico avançado. Transparência que se limita a publicar arquivos massivos sem estrutura amigável equivale

a restringir acesso prático à informação. Nesse sentido, boas práticas incluem visualizações interativas que apresentem séries temporais sobre indicadores criminais anonimizados ou mapas agregados por região administrativa que permitam ao cidadão acompanhar tendências sem violar privacidade individual. Ainda assim persiste o dilema entre maximizar abertura e prevenir usos secundários indevidos dos dados lançados ao domínio público. Mesmo conjuntos aparentemente inócuos podem ser correlacionados por terceiros com outras fontes externas para gerar inferências invasivas sobre indivíduos ou grupos específicos. Implementar licenças claras definindo finalidades lícitas do reuso ajuda mitigar tais riscos; no entanto é inevitável adotar técnicas técnicas resistentes à engenharia reversa nos processos prévios à publicação (Lima, 2019). Por fim é importante ressaltar que a transparência não deve ser evento pontual mas sim processo contínuo acompanhado pela atualização periódica das informações oferecidas ao público (Reis e Housley, 2022). Atualizações regulares mostram compromisso institucional ativo com prestação constante de contas e mantêm relevância analítica dos conteúdos divulgados. Somente ao integrar governança clara, anonimização efetiva, observabilidade plena dos sistemas envolvidos e canais abertos para comunicação compreensível com a sociedade será possível consolidar um ecossistema transparente capaz de servir simultaneamente à eficácia operacional da segurança pública e à preservação inegociável dos direitos individuais reconhecidos na legislação brasileira vigente.

7 Conclusão

A integração eficiente e segura de dados em segurança pública demanda uma abordagem que combine rigor técnico, conformidade legal e sensibilidade ética. A diversidade das fontes, que vão desde sistemas policiais e judiciais até redes sociais e plataformas digitais, impõe desafios que exigem arquiteturas flexíveis capazes de lidar com dados estruturados e não estruturados, garantindo qualidade, integridade e proteção das informações. A adoção de modelos híbridos de ingestão, como as arquiteturas lambda e kappa, permite equilibrar a necessidade de análises históricas detalhadas com a agilidade requerida para respostas em tempo real, assegurando que decisões operacionais sejam fundamentadas em dados confiáveis e atualizados.

A conformidade com a Lei Geral de Proteção de Dados (LGPD) emerge como elemento fundamental, orientando práticas de anonimização, pseudonimização e controle de acesso que preservam direitos individuais sem comprometer a eficácia das ações estatais. A automação dos pipelines, por meio de ferramentas especializadas e agentes inteligentes, contribui para a manutenção da qualidade dos dados e para a observabilidade contínua dos processos, possibilitando a detecção precoce de anomalias e a mitigação de riscos operacionais e legais. Além disso, a governança institucional, estruturada em políticas internas claras e integradas aos sistemas tecnológicos, assegura a responsabilização, a transparência e o respeito aos princípios éticos que regem o tratamento de informações sensíveis.

A transparência no uso dos dados públicos, quando articulada com mecanismos técnicos adequados, fortalece a confiança da sociedade nas instituições responsáveis pela segurança coletiva, ao mesmo tempo em que protege a privacidade dos cidadãos. A disponibilização de informações agregadas e contextualizadas, acompanhada de relatórios sobre os métodos analíticos e as salvaguardas adotadas, contribui para o equilíbrio entre prestação de contas e segurança operacional. Por fim, a constante atualização das práticas e a adaptação às mudanças tecnológicas e normativas são indispensáveis para garantir que os sistemas de segurança pública permaneçam eficazes, éticos e alinhados às expectativas sociais, promovendo um ambiente informacional que respeite os direitos fundamentais e apoie a construção de políticas públicas baseadas em evidências confiáveis.

Referências

Chambers, Bill e Matei Zaharia (2018). *Spark: The Definitive Guide*. O'Reilly Media. ISBN: 978-1-4919-1221-8.

Foidl, Harald et al. (set. de 2023). *Data Pipeline Quality: Influencing Factors, Root Causes of Data-related Issues, and Processing Problem Areas for Developers*. DOI: 10.48550/arXiv.2309.07067. arXiv: 2309.07067 [cs]. URL: http://arxiv.org/abs/2309.07067 (acedido em 19/10/2025).

Goldschmidt, Ronaldo, Emmanuel Passos e Eduardo Bezerra (jan. de 2015). *Data Mining: Conceitos, Tecnicas, Algoritmos, Orientacoes e Aplicacoes*. Elsevier Editora Ltda. ISBN: 978-85-352-7822-4. URL: https://app.minhabiblioteca.com.br/reader/books/9788595156395/.

Jindal, Alekh, Jorge-Arnulfo Quiane-Ruiz e Samuel Madden (jan. de 2017). *INGESTBASE: A Declarative Data Ingestion System*. DOI: 10.48550/arXiv.1701.06093. arXiv: 1701.06093 [cs]. URL: http://arxiv.org/abs/1701.06093.

Lima, Cíntia Rosa Pereira de (nov. de 2019). *Comentários à lei geral de proteção de dados: lei n. 13.709/2018, com alteração da lei n. 13.853/2019*. Almedina. ISBN: 978-85-8493-579-6. URL: https://app.minhabiblioteca.com.br/reader/books/9788584935796/.

Moses, Barr (jan. de 2023). *Fundamentos da Qualidade de Dados: Guia Prático Para Criar Pipelines De Dados Confiáveis*. Rio de Janeiro, RJ: Alta Books. ISBN: 978-85-508-2122-1. URL: https://app.minhabiblioteca.com.br/reader/books/9788550821221/.

Reis, Joe e Matt Housley (2022). *Fundamentals of Data Engineering*. First. O'Reilly Media, Inc. ISBN: 978-1-0981-0830-4.

Rucco, Chiara, Antonella Longo e Motaz Saad (abr. de 2025). *Efficient Data Ingestion in Cloud-based Architecture: A Data Engineering Design Pattern Proposal*. DOI: 10.48550/arXiv.2503.16079. arXiv: 2503.16079 [cs]. URL: http://arxiv.org/abs/2503.16079 (acedido em 19/10/2025).

Shapira, Gwen et al. (2022). *Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale*. Second edition. Bejing Boston Farnham Sebastopol Tokyo: O'Reilly. ISBN: 978-1-4920-4308-9.