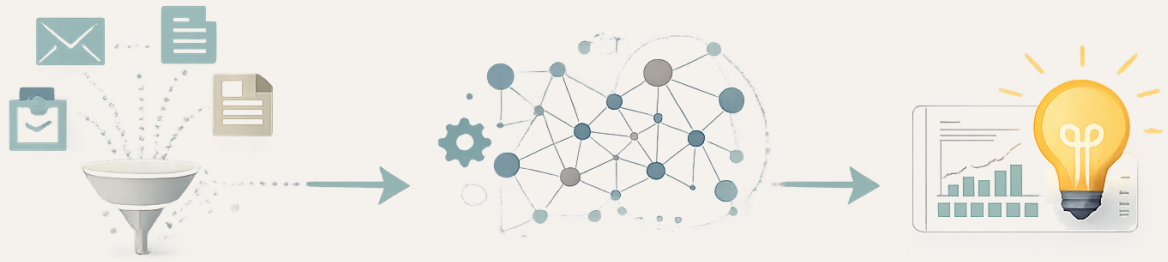
Data Pipeline Induction Guide

Minimum data set for governance analysis. Standard fields from systems already in use.



Data Harvesting

Email · Ticketing Systems

Incident Logs · Operational Metadata

Pipeline Processing

- Graph Construction
- Behavioral Scoring
- Structural Risk Propagation

Gain Insight

- Governance Evidence
- Facility Stratification
- Capital Decisions

I.

How the Pipeline Consumes Data

BBCO builds a directed communication graph from operational metadata: who communicated with whom, when, through which system, and how the conversation threaded. That metadata is already produced by the systems that support day-to-day operations. The upstream source may be an email server, an EHR, an incident management platform, a ticketing system, or a messaging application. The specific system matters less than the presence of a consistent set of structural fields: a timestamp, a sender, one or more recipients, and a thread, ticket, or encounter-linked identifier that groups related communications into a chain.

- The pipeline brings disparate sources into a unified graph. An escalation that begins in an EHR or incident platform, continues as an email thread between the DON and the Regional Director, and results in a follow-up ticket in ServiceNow can appear as a single connected path when the exports contain enough metadata to support linkage.
- Email is often the richest single source. Governance-relevant communications tend to follow a compliance gradient: as stakes rise, the conversation is more likely to remain in or return to auditable channels. Corporate counsel, regulatory reporting, and board-level escalation commonly route through email. In many environments, email alone is enough to begin.
- EHR, ticketing, and incident management systems are usually the next most valuable sources. They provide structured severity fields, assignment changes, and state transitions that align directly with the behavioral markers used in the analysis: decision signals, priority shifts, and escalation routing.
- This guide is intended as a field reference rather than a set of step-by-step extraction instructions. Interfaces vary by vendor, though the information needed is largely the same across platforms. The practical task is simply to map available exports to the schemas below.

II.
Quick Start

A simple starting point is three CSV files, reviewed before handoff.

- **File 1a** is the incident or event header extract from the RMIS, EHR, or incident platform: one row per incident or event with category, severity, and facility.
- **File 1b** is the line-level extract from the same system: one row per escalation action, status change, or follow-up step, with timestamps and roles.
- **File 2** comes from one or more communication sources: email (File 2a), ticketing systems (File 2b), or both. Email alone is sufficient to start.

Standard export tools already available in those systems are usually sufficient. Where the source system stores escalation history as a flat field rather than a separate table, Files 1a and 1b can be combined into a single export.

- **Format**: For incident data (Files 1a and 1b), CSV with column headers, UTF-8 encoding, one row per record. Excel (.xlsx) is also accepted. For email data (File 2), the pipeline can ingest raw mailbox archives directly: PST (Outlook/Exchange), MBOX, or EML files. No manual export or reformatting is required for these formats. If your environment uses Microsoft 365 or Google Workspace, an admin-level message trace or log export (CSV) also works. The pipeline handles header extraction and thread reconstruction from any of these sources.
- **Date range**: The most recent 12-24 months is a practical starting range. A longer history can improve baseline calibration, though it is not required for initial analysis.
- **Delivery**: Encrypted upload portal (link provided separately) or encrypted email attachment.
- **What is never requested**: Resident names, diagnoses, medications, room numbers, care plans, or any of the 18 HIPAA identifiers. Message body text may be used to identify structural governance markers such as decision language, escalation signals, or priority shifts, but it is not retained after scoring.

III.
File 1a: Incident Header Extract

One row per incident or event. This is the summary record: what happened, where, and how it was classified. Source: any RMIS, EHR workflow module, or incident management platform. PointClickCare, RLDatix, American Data Network, MatrixCare, and most enterprise RMIS platforms support CSV export with configurable field selection.

#	Field	Req	Format	Example	Purpose
1	**incident_id**	Yes	Text	INC-2025-04821	Primary key; links header to line-level detail
2	**facility_id**	Yes	Text	SPRINGS-BOZEMAN-01	Facility-level governance stratification
3	**incident_date**	Yes	Datetime	2025-06-14 08:30	Temporal position for rolling-window analysis
4	**incident_category**	Yes	Text	Fall, Elopement, Med Error	Escalation shape class assignment
5	**severity**	Yes	Text	Minor, Moderate, Major, Sentinel	Maps to escalation depth expectations
6	**department**	Opt	Text	Nursing, Dietary, Maintenance	Cross-functional boundary detection
7	**resolution_date**	Opt	Datetime	2025-06-16 14:00	Containment duration measurement

IV.

File 1b: Incident Line-Level Extract

Multiple rows per incident, one per escalation action. This is the audit trail: who was notified, when, and what they did. Linked to the header record by incident_id. Most RMIS, EHR workflow, and incident platforms store this as a separate activity or history table.

#	Field	Req	Format	Example	Purpose
1	**incident_id**	Yes	Text	INC-2025-04821	Foreign key to the header record
2	**action_date**	Yes	Datetime	2025-06-14 09:10	When this escalation step occurred
3	**action_by_role**	Yes	Text	Charge Nurse, DON, Administrator	Who performed this action; node in the graph
4	**action_type**	Yes	Text	Report, Notify, Escalate, Resolve	Distinguishes routine steps from governance events
5	**escalated_to_role**	Opt	Text	DON, Regional Dir, Risk Manager	Recipient of this escalation step
6	**notes**	Opt	Text	DON notified, family contacted	Supplementary routing detail
7	**follow_up_actions**	Opt	Text	RCA completed; state report filed	Governance forum activation at path tail

The line-level extract preserves the temporal ordering of escalation steps within each incident. Knowing that the DON was notified at 09:10 and escalated to the Regional Director at 14:30 is a different governance signal than knowing both were involved at some unspecified point. This sequencing allows the pipeline to compute route termination depth, meaning the organizational level at which an escalation path ends, and to measure containment consistency per facility. Where the source system stores escalation history as a flat field rather than a separate table (e.g., a single "escalated_to" column with semicolon-separated roles), a combined File 1 export is acceptable. The pipeline can work with either structure. Exclude from both exports: resident names, room numbers, diagnoses, medications, care plan notes, or any field containing PHI. If the system cannot exclude PHI columns directly, those columns can be redacted before processing.

V.

File 2a: Email Header Extract

Email provides the strongest native threading of any communication source. RFC 5322 headers carry globally unique message identifiers, reply chains, and a full chain of custody through the References header. For many organizations, the email export is the single most valuable input to the pipeline. One row per email message. Source: Microsoft 365 message trace, Google Workspace email log, Exchange admin export, or MBOX/EML/PST archive parsing.

#	Field	Req	Format	Example	Purpose
1	**date**	Yes	Datetime	2025-06-14 09:15	Temporal ordering within escalation threads

#	Field	Req	Format	Example	Purpose
2	**from**	Yes	Text	DON_BOZEMAN or Person_042	Sender node in the communication graph
3	**to**	Yes	Text	ADMIN_BOZEMAN; RD_WEST	Directed edges; semicolon-separated for multiple
4	**cc**	Opt	Text	Risk_Mgr; Compliance_Officer	Governance breadth at each escalation step
5	**subject**	Yes	Text	RE: Fall incident Unit 3	Thread grouping via subject-line fallback
6	**body**	Yes	Text	(message content)	Scored for structural governance markers; not retained after scoring
7	**message_id**	Yes	Text	abc123@corp.com	Globally unique ID per RFC 5322; primary linkage key
8	**in_reply_to**	Yes	Text	def456@corp.com	Parent message ID; reconstructs reply chains
9	**references**	Opt	Text	def456@corp.com abc123@corp.com	Full ancestor chain; gold standard for thread reconstruction
10	**x_priority**	Opt	Text	1 (Highest) to 5 (Lowest)	RFC 2156 priority header; maps to priority-shift scoring
11	**importance**	Opt	Text	High, Normal, Low	RFC 4021 importance header; alternative priority signal

- **Threading**: Message-ID uniquely identifies each message. In-Reply-To links a reply to its parent. References preserves the full thread lineage even when intermediate messages are missing from the corpus. The pipeline parses References first, falling back to In-Reply-To when References is absent.
- **Mailbox scope**: The export should cover mailboxes for team leads, department heads, facility administrators, regional leadership, and executive staff involved in operational oversight. The pipeline identifies structurally significant participants through centrality analysis after ingestion, so a broader initial scope is preferable to a narrow one.
- **Message body**: Body text is included so structural governance markers can be scored. It is not retained after scoring.
- **PHI in subject lines**: Where subject lines contain resident names or other PHI, redaction can be handled during initial data processing. Practical support during extraction and anonymization is available for pilot engagements.
- **Encryption**: All data is encrypted at rest and in transit.

VI.

File 2b: Ticketing and Incident Communication Extract

When governance escalation routes through a ticketing system such as ServiceNow or through internal messaging within the RMIS, EHR, or incident platform, those records provide structured escalation metadata that complements or substitutes for the email export. One row per ticket action, such as a comment, assignment change, status transition, or escalation. Source: standard ticket export with history.

#	Field	Req	Format	Example	Purpose
1	**date_time**	Yes	Datetime	2025-06-14 09:30	Temporal ordering of escalation actions
2	**ticket_id**	Yes	Text	INC0012345 or OPS-1234	Thread identifier; groups all actions on one issue
3	**action_by**	Yes	Text	DON_BOZEMAN or Person_042	Who performed the action; sender-equivalent node
4	**action_type**	Yes	Text	Comment, Assign, Escalate, Resolve	Distinguishes routine updates from governance events
5	**assigned_to**	Opt	Text	RD_REGION_WEST	Current assignee after this action; recipient-equivalent
6	**status**	Opt	Text	Open, In Progress, Escalated, Resolved	State transitions map to containment depth
7	**priority**	Opt	Text	P1, P2, P3, P4	Priority changes map to priority-shift scoring
8	**linked_tickets**	Opt	Text	INC0012340; CHG0004521	Cross-ticket linkage for multi-system escalation paths

Ticketing systems are particularly useful because they record structured severity fields, assignment changes, and state transitions that correspond directly to behavioral markers: decision signals when a ticket is escalated or reassigned, priority shifts when severity is upgraded, and route depth when a ticket resolves at a specific organizational layer.

VII.

Linking Across Systems

An escalation that starts in an EHR or incident platform, triggers an email thread, and generates a ServiceNow ticket can appear as three disconnected fragments unless the exports contain enough metadata to support linkage. The pipeline resolves cross-system linkage through three mechanisms:

- **Identity resolution**: The same person may appear as jsmith@corp.com in email, John Smith in the RMIS or EHR, and jsmith in ServiceNow. The anonymization step described below is intended to produce a single consistent identifier, such as DON_BOZEMAN, across all exports so the graph resolves to one node.
- **Temporal proximity**: When an incident record and an email thread share the same date, facility, and incident category, linkage can often be inferred even without an explicit cross-reference.
- **Explicit cross-references**: Some systems embed ticket IDs in email subjects or link related tickets. These explicit references provide the strongest linkage signal.

For an initial deployment, email alone is often sufficient. Cross-system integration can be added later as an incremental enhancement that deepens the governance graph over time.

VIII.
Anonymization

The operator does not need to anonymize the data before handoff. Send it as-is, encrypted. The pipeline handles identity resolution and anonymization as part of initial processing. An organization chart or other mapping matrix may be needed to uniquely identify each node in the correspondence. The table below illustrates what the pipeline produces internally. Personal names are replaced with role-based or hash-based identifiers so that the governance graph tracks organizational function rather than individual identity.

#	Raw Input	Pipeline Output (Role-Based)	Pipeline Output (Hash-Based)
1	Jane Smith (DON, Bozeman)	DON_BOZEMAN	Person_042
2	Robert Chen (Admin, Bozeman)	ADMIN_BOZEMAN	Person_017
3	Maria Lopez (Regional Dir, West)	RD_REGION_WEST	Person_003
4	Sarah Park (CNA, Bozeman, Day)	CNA1_BOZEMAN	Person_088
5	Kim Nguyen (CNA, Bozeman, Night)	CNA2_BOZEMAN	Person_091

- **Role-based identifiers** are the default output. The pipeline maps each person to their organizational role and facility, producing identifiers like DON_BOZEMAN that preserve governance function without personal identity.
- **Hash-based identifiers** are used when role mapping is not available from the source data. Each person receives a consistent opaque identifier (Person_001, Person_002) that persists across all exports so the graph resolves correctly.
- **Multiple people in the same role** receive a numeric suffix automatically (CNA1_BOZEMAN, CNA2_BOZEMAN) to distinguish individuals within the graph.
- **If the operator prefers to pre-anonymize**, we support that. A role-mapping reference table can be provided to align identifiers across systems before handoff. This is optional, not required.

IX.
Data Handling and HIPAA

- BBCO does not ingest, process, or store Protected Health Information (PHI). Incident category and severity are operational metadata. The 18 HIPAA identifiers are never requested.
- Files are transmitted through encrypted channels and processed in an isolated environment with no internet egress. Source files are deleted after the governance graph is constructed.
- As an additional safeguard, automated data loss prevention (DLP) redaction can be activated as a pipeline processing stage.