WHITEPAPER

keyrus
make data matter

Responsible AI starts with responsible data engineering

This paper explores the **regulatory context for AI**, defines principles of **ethical data engineering** and illustrates **best practices.**

TL; DR

Responsible AI doesn't begin with algorithms, it begins with **responsible data engineering**. As regulations like the EU AI Act and South Africa's POPIA demand transparency, accountability, and governance, organisations must embed ethical principles into their data pipelines before AI models are ever trained.

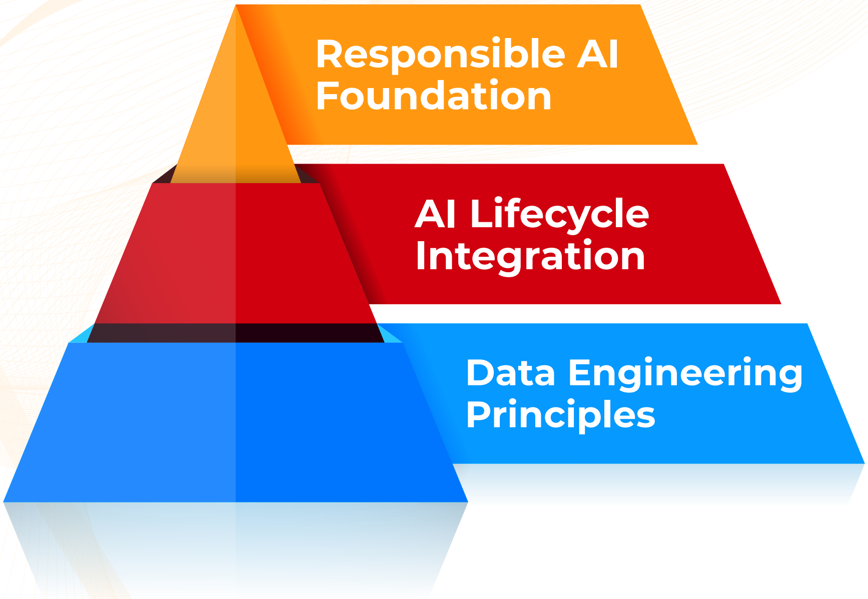
Five key principles guide this approach: transparency, quality, governance, fairness, and oversight. Modern tools (e.g., dbt, Snowflake, Soda) can help enforce lineage tracking, automated testing, access control and bias monitoring.

Ignoring these safeguards creates risks, technical debt, biased outcomes, regulatory penalties, and reputational harm. Success requires cross-functional collaboration across engineering, legal, compliance, and domain experts.

Practical actions include: auditing pipelines, documenting lineage, assigning governance roles, running automated data tests, engaging diverse teams early, and aligning with ethical frameworks.

Bottom line: Responsible AI is not a future aspiration but a present necessity. By embedding ethical data practices now, organisations can build AI systems that are not only compliant and trustworthy but also deliver long-term strategic advantage.

As artificial intelligence (AI) becomes embedded in more business-critical systems, there is growing concern about how these technologies are governed. Questions about fairness, bias, explainability, and accountability are no longer theoretical. They are urgent, especially as AI begins to shape decision-making across industries. The common narrative focuses on building responsible AI by improving algorithms or introducing regulations. While important, this approach overlooks a fundamental truth: **responsible AI starts with responsible data engineering.**



The regulatory and industry context for responsible AI

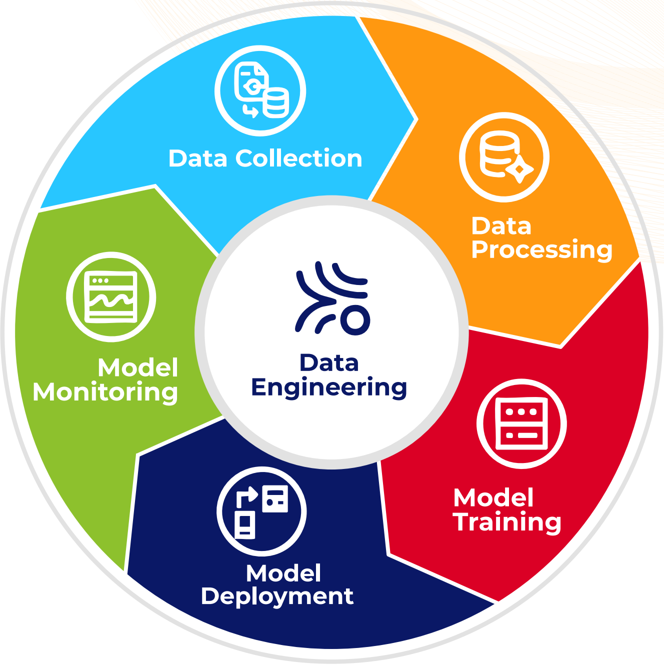
The growing pressure for responsible AI is not occurring in a vacuum. Governments, regulators and industry bodies around the world are introducing frameworks that make transparency, accountability and data governance mandatory components of AI practice. The European Union's Artificial Intelligence Act (European Commission, 2021) sets out risk-based compliance obligations, including documentation, traceability and human oversight. Similarly, South Africa's Protection of Personal Information Act (POPIA) places legal duties on how data is collected, processed and secured.

These regulatory developments highlight a critical point: organisations will increasingly be required to demonstrate responsible handling of data, especially when it informs automated or AI-driven decisions. This reinforces the need to embed ethical principles directly into data engineering workflows, from sourcing and transformation through to access and auditability. In this context, data engineers are not only technical contributors; they are **frontline participants in building AI systems** that meet both ethical and legal expectations.

Understanding the link between AI and data engineering

AI systems are only as reliable as the data that feeds them. Before a model is trained, before predictions are made, data must be collected, cleaned, transformed, and validated. These tasks fall squarely within the domain of data engineering. When done irresponsibly, this stage introduces hidden risks into AI pipelines. Poor data quality, undocumented transformations, lack of testing, and unrestricted access can all lead to opaque and untrustworthy outcomes. **Ethical AI is not just about what a model does; it is about how the data got there in the first place.**

Data Engineering's Role in the AI Lifecycle



Key principles of ethical data engineering (where AI gets involved)

To build AI systems that are transparent, fair and robust, organisations must first establish ethical data engineering practices. We outline **five key principles** and demonstrate how recent developments in modern data tooling can support these goals.

Transparency
Ensuring clarity and openness in data processes

Quality
Maintaining high standards of data accuracy and reliability

Governance
Establishing clear rules and responsibilities for data management

Fairness
Promoting equitable treatment and avoiding bias in data use

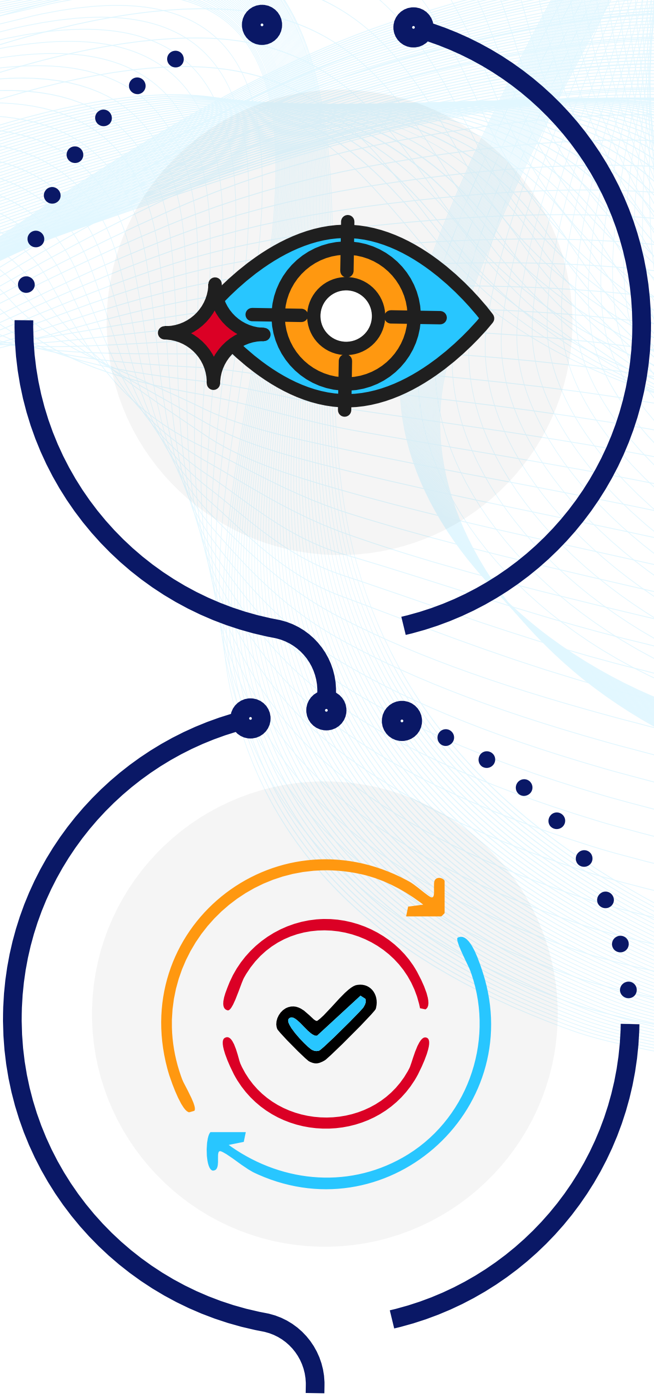
Oversight
Implementing mechanisms for monitoring and accountability

To ground this discussion, we refer to examples from dbt and Snowflake. The recent release of the dbt Model Context Protocol (MCP) server and Snowflake's existing governance structure provide a useful frame for showing how **responsible data engineering can be embedded into real-world workflows.** These are not endorsements. Rather, they are illustrative technologies that demonstrate what is possible when responsibility is designed into the data layer.

1 Transparency and lineage

Stakeholders must be able to trace how data has been sourced, transformed, and delivered. Without visibility, it is impossible to understand or explain AI behaviour.

- dbt enforces transformation logic as code, maintaining documented lineage and version control.
- dbt MCP enables AI agents to access lineage context programmatically, bringing traceability into AI workflows.
- Snowflake integrates with dbt to extend lineage visibility across the analytics stack.



2 Data quality and testing

AI is highly sensitive to the quality of its inputs. Biased, incomplete, or incorrect data can have damaging effects downstream.

- dbt allows for automated tests (e.g. uniqueness, null checks) that validate datasets before they enter the model pipeline.
- dbt MCP exposes metadata about test coverage and results, helping AI agents reason about data reliability.
- Snowflake, when paired with tools like Soda or Monte Carlo, can monitor live data quality at scale.

3 Governanace and access control

Access to data must be regulated and documented to avoid misuse and protect privacy.

- Snowflake enforces fine-grained permissions, ensuring that AI systems only interact with approved data.
- dbt's Semantic Layer, now accessible via MCP, ensures AI agents operate using governed definitions, not raw tables.

4 Fairness and representativeness

Training data must reflect the diversity of real-world populations to avoid reinforcing discrimination.

- dbt models can be audited for skewed distributions or missing categories.
- dbt MCP allows AI agents to detect upstream filters or joins that might introduce bias.
- Snowflake's Data Marketplace offers access to diverse datasets for enrichment.

5 Automation with oversights

While automation improves efficiency, it must be controlled. AI systems that manipulate data or run models should be auditable.

- dbt MCP enables AI agents to trigger dbt tasks within the bounds of project governance.
- Snowflake supports execution auditing and compute controls to monitor AI-driven actions.

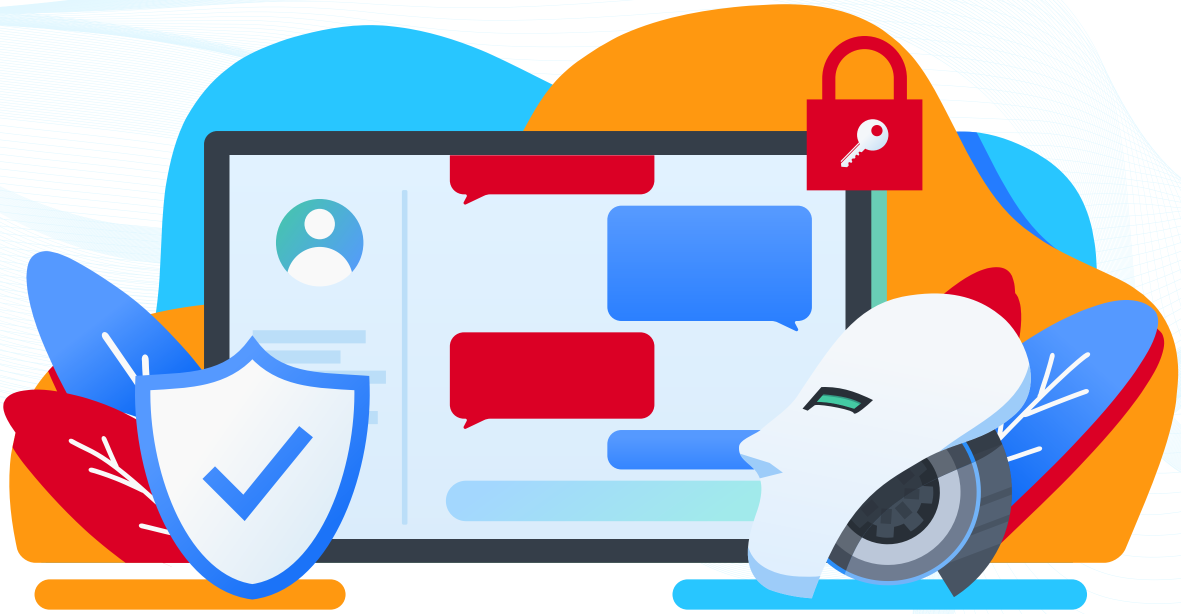
These principles are applicable to any organisation, regardless of tooling. The goal is to ensure that **ethical safeguards are embedded within the data engineering process** before AI ever interacts with the data.

The risks of ignoring ethical data engineering

Neglecting ethical data engineering principles does not only limit the effectiveness of AI systems—it actively introduces risk. These risks are wide-ranging and can manifest across technical, legal, reputational, and societal dimensions. Organisations that fail to ensure data quality and lineage may find their AI outputs difficult to audit or defend. A lack of governance can lead to data breaches, privacy violations, or unintended algorithmic bias. In sectors like finance, healthcare, or public services, these failures can result in **regulatory penalties, litigation, and loss of public trust.**

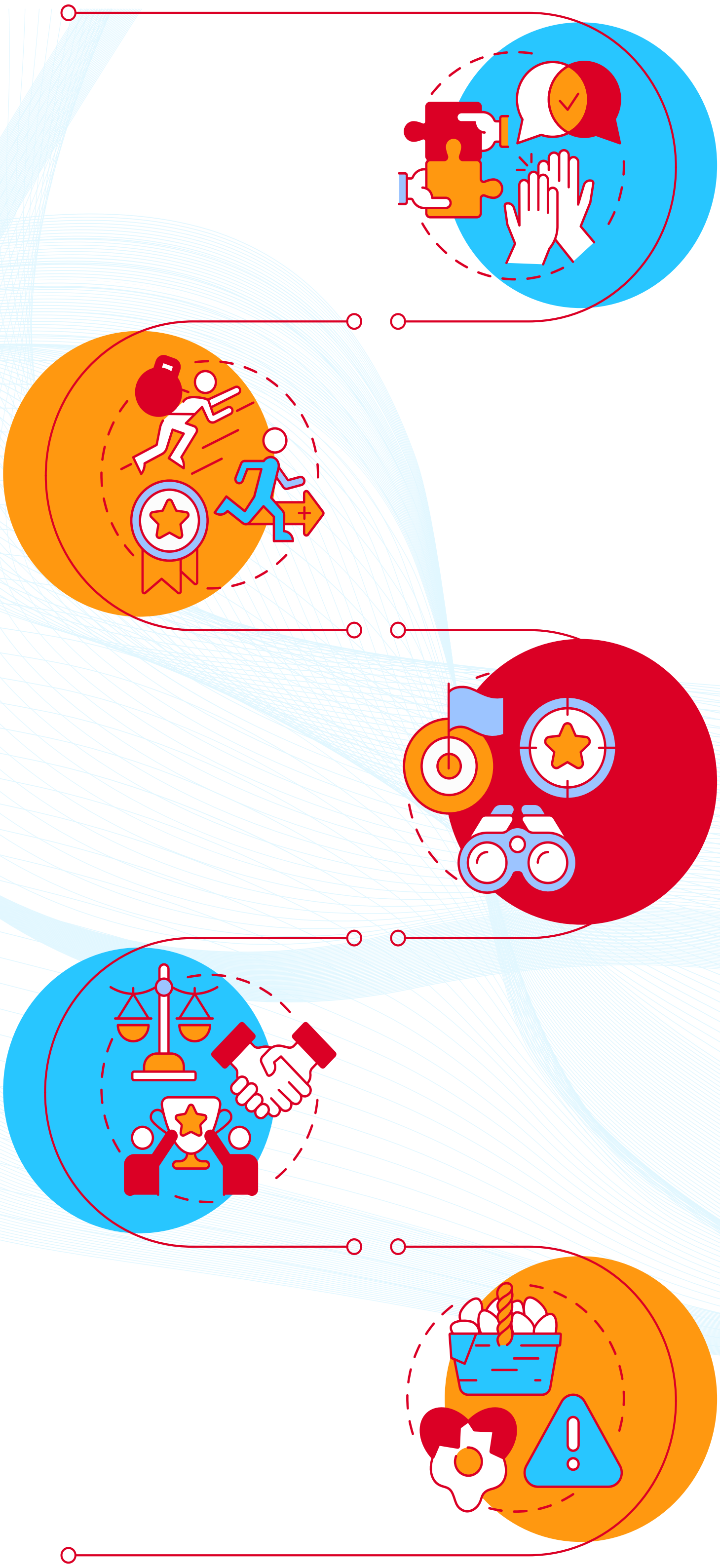
One common failure mode is the use of historical data without proper assessment of its suitability or bias. For example, training models on skewed datasets can reinforce discrimination in hiring, lending, or insurance decisions. Another example includes opaque transformations in data pipelines, which can make it impossible to explain how an AI system arrived at a particular outcome, undermining trust and compliance.

Ethical data engineering is therefore not just a good practice; it is a strategic imperative. It helps organisations avoid technical debt, safeguard their reputation, and ensure they can meet the growing demands of AI regulation.



Cross-functional collaboration is essential

Ethical AI is not solely the responsibility of data engineers or data scientists. It requires a collaborative effort across technical and non-technical roles. Legal, compliance, product, and domain experts all bring critical perspectives to how data is collected, transformed, and used in AI systems. Without this collaboration, ethical blind spots can emerge, particularly in understanding user impact, regulatory obligations, or unintended consequences.



Responsible data engineering must therefore be framed as a cross-functional initiative.

Teams must jointly define what 'good' looks like in terms of fairness, auditability, and accountability. When governance frameworks are developed in isolation or retrofitted after deployment, AI systems are more likely to produce outcomes that lack transparency or legitimacy. Embedding these conversations early in the data lifecycle helps ensure that AI operates in alignment with business values and social expectations.

Responsible AI begins before the model

Although this article presents benefits and best practices, it is important to acknowledge some challenges. Embedding ethical data engineering requires cultural change, investment in tools, and cross-team alignment. Smaller organisations or legacy environments may face constraints in achieving full transparency and governance. However, these limitations reinforce the importance of incremental progress, starting with attainable changes such as pipeline audits, data testing, and documentation.

The push for responsible AI must expand its scope. Data scientists and ML engineers cannot be the only stewards of ethics. **Responsibility must be a shared commitment that begins much earlier in the data workflow.** It starts with how data is sourced, how transformations are documented, how tests are run, and how access is controlled.

Practical action points

To support the development of responsible AI, organisations can begin by evaluating their current data engineering practices and identifying areas for improvement. We recommend the following actions:

Audit your data pipelines

Check for undocumented transformations, missing test coverage and unclear ownership.

Establish data lineage

Ensure every dataset used in AI systems can be traced back to its source and transformation history.

Define governance roles

Clarify who owns each dataset and who is accountable for its quality and accessibility.

Implement automated testing

Use tools that enforce data validation at every step of the pipeline.

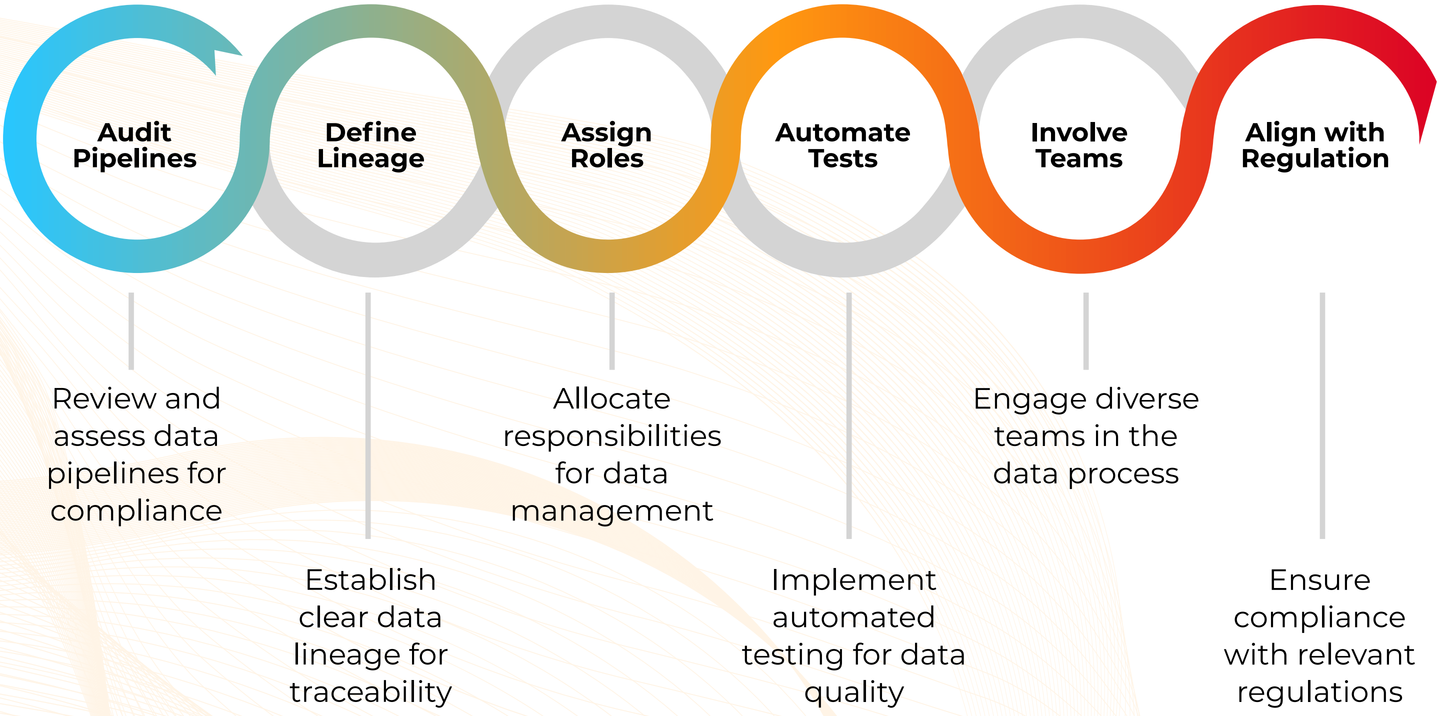
Involve diverse teams early

Engage legal, compliance and domain experts from the start of any AI project.

Aiign with ethical frameworks

Familiarise your teams with emerging standards like the EU AI Act and POPIA.

Responsible AI is not a future aspiration. It is a present necessity that begins with the choices made at the data engineering level. Organisations that act now will not only mitigate risk, but also gain a strategic advantage in building trustworthy, future-ready AI systems.



Review and assess data pipelines for compliance

Establish clear data lineage for traceability

Allocate responsibilities for data management

Implement automated testing for data quality

Engage diverse teams in the data process

Ensure compliance with relevant regulations

References

dbt Labs (2025) Build reliable AI agents with the dbt MCP server [webinar], 30 & 31 July 2025. Available at: https://www.getdbt.com/resources/webinars/build-reliable-ai-agents-with-the-dbt-mcp-server (Accessed: 31 July 2025).

dbt Labs (2025) Introducing the dbt MCP Server. Available at: https://docs.getdbt.com/blog/introducing-dbt-mcp-server (Accessed: 31 July 2025).

Snowflake Inc. (2023) Snowpark for Python – Empowering Secure and Scalable ML. Available at: https://www.snowflake.com/blog/snowpark-for-python-now-generally-available (Accessed: 31 July 2025).

Snowflake Inc. (2024) Data Cloud Security and Governance Overview. Available at: https://www.snowflake.com/guides/data-security-and-governance-overview (Accessed: 31 July 2025).

Gebru, T. et al. (2018) Datasheets for Datasets. Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning. Available at: https://arxiv.org/abs/1803.09010 (Accessed: 31 July 2025).

European Commission (2021) Proposal for a Regulation on Artificial Intelligence (AI Act). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (Accessed: 31 July 2025).

How Keyrus can assist your responsible AI journey

Keyrus draws on deep expertise in both data engineering and AI to help organisations implement the practical steps outlined above. Our multidisciplinary team partners with clients to assess current data workflows, design robust governance frameworks, and embed automated testing, documentation, and lineage tracking into your existing pipelines. Leveraging leading tools such as dbt, Snowflake, and others, we tailor solutions to ensure your data infrastructure fully supports ethical, auditable, and compliant AI systems.

Craig Andrew
Head of Data Analytics

Scan here

Contact us today

www.keyrus.com/za

Whether you are modernising legacy environments or building new AI capabilities, Keyrus provides hands-on support with pipeline audits, tool implementation, team alignment, and regulatory readiness. By working alongside your technical and business stakeholders, we help turn **responsible data engineering into a source of competitive advantage**, making your AI initiatives both trustworthy and future-ready.