WHITEPAPER

keyrus
make data matter

(R)EVOLUTION OF DATA ENGINEERING

The journey toward auto-engineering

Data at all costs?

It's worth pausing to ask: are we truly getting value from our data? Despite being positioned as a performance lever, data often remains trapped in brittle, highly manual, and expertise-dependent processes. The promises of agility and insight frequently collide with the burden of legacy architectures—costly to maintain, slow to adapt, and hard to scale.

At Keyrus, we believe the time has come to rethink how data pipelines are delivered. This eBook outlines a tested and pragmatic approach: automation applied at the system level—not to eliminate engineering, but to industrialise and accelerate it. By formalising business logic, codifying best practices, and leveraging composable frameworks, we can scale engineering efforts without sacrificing governance or quality.

"Automated data engineering shifts the focus from building pipelines to designing systems. It's a transformation that improves scalability, governance, and collaboration between data and business teams."

~ Nkosinathi Xulu
Head of Data Engineering

The rise of auto-engineering

Data has become a strategic asset for companies, fuelling innovation and competitive advantage. To harness this potential, organisations have long built data pipelines to collect, transform, and operationalise increasingly large and diverse datasets. Traditionally, building and maintaining these pipelines involved hands-on, custom-built solutions tailored to each use case. These approaches, while effective at the time, often required significant technical effort, lacked scalability, and were difficult to adapt as business needs evolved.

Today, the rise of automation and generative AI is transforming this model. Tasks once handled manually are now being streamlined by sophisticated tools and frameworks. For instance, DBT automates data transformations using SQL and Python, while Infrastructure-as-Code accelerates reproducible, environment-agnostic deployments.

Innovations such as business manifests help translate functional needs into technical specifications. Text2SQL tools can generate queries from natural language inputs. Generative AI can now propose entire models and data workflows—suggesting a future where many components of the data lifecycle may be automatically generated and governed. While this direction offers clear benefits in terms of speed, cost, and agility, it also introduces new responsibilities. How do we ensure the reliability of automatically generated pipelines? What governance frameworks must evolve to support these tools? And how do we ensure business needs remain at the center of increasingly automated systems?

Data engineering in the age of automation

The challenge for modern data systems is no longer just volume, but diversity - encompassing text, images, sensor outputs, and geospatial data. These varied data types often demand processing that extends beyond the capabilities of traditional SQL-based approaches.

In the past, data pipelines were largely handcrafted — each use case often requiring custom development, rigorous testing, and manual maintenance to ensure performance and reliability. This approach, while functional, often resulted in brittle systems that were difficult to scale or adapt quickly.

Today, **a combination of automation, modern frameworks, and AI-driven tooling** is transforming how data pipelines are built and managed:

Technological Evolution: Frameworks like dbt introduce software engineering principles — version control, modularity, and testing — into the analytics workflow. Infrastructure-as-Code enables consistent, scalable pipeline deployment.

Generative AI: AI now assists in generating code, surfacing documentation, and proactively identifying anomalies — supporting more intelligent, adaptive systems.

Self-Service Enablement: Business users increasingly seek direct access to data insights. While full autonomy may not be feasible, modern platforms offer controlled self-service options with built-in governance and data quality enforcement.

As automation reshapes the landscape, data engineering evolves into a discipline focused on designing resilient systems, embedding best practices, and enabling scalable collaboration — rather than writing every line of code from scratch.

> “Automated data engineering isn’t a future concept — it’s already changing how we deliver value. By combining modular tooling like dbt with business-aligned manifests and CI/CD practices, we’re able to standardise, scale, and govern data pipelines faster and more reliably. It’s about making data engineering more predictable, not less human.”

~ Nkosinathi Xulu
Head of Data Engineering

Key concepts in modern data engineering

As the data landscape evolves, several foundational concepts are reshaping how data systems are built, maintained, and consumed:

Business Manifests

Structured documents , typically in YAML or JSON, that encode business logic, rules, and requirements in a machine-readable format. These serve as the blueprint for automated data pipelines and often require input from data teams to define elements like primary keys, relationships, and transformation logic accurately.

Modern Frameworks

Technologies such as dbt and Python bring software engineering discipline into the data domain. By incorporating practices like version control, automated testing, and CI/CD pipelines, these tools improve reliability, scalability, and collaboration in data engineering workflows.

Self-Service Platforms

Modern data platforms aim to empower business users to access, analyse, and visualise data independently. However, this empowerment must be balanced with governance controls, data quality standards, and lineage tracking to ensure responsible and compliant usage.

Total Cost of Ownership (TCO)

TCO includes all visible and hidden costs , spanning software licenses, cloud infrastructure, personnel training, support, and inefficiencies. Reducing TCO through automation, standardisation, and scalable design not only lowers expenses but also accelerates time to value and operational resilience.

Building a pipeline factory: A systemic approach

The traditional approach of crafting bespoke pipelines for each data need is no longer sustainable. As data demands grow in scale and complexity, a more industrialised model is required - one that enables repeatability, governance, and speed.

A pipeline factory achieves this by automating the generation of data workflows from well-defined business needs. Rather than building pipelines manually, teams define the "what" and let the system generate the "how."

This shift relies on **three foundational pillars:**

1

Ideation Workshops: Collaborative sessions that capture the full context of each use case, including source systems, update frequency, data transformations, formats, quality rules, and documentation requirements. These workshops align technical output with business goals from the outset.

2

Business Manifests: Structured, machine-readable specifications (e.g., YAML or JSON) that translate business logic into input for the pipeline factory. These manifests act as living contracts between business and engineering, and serve as the single source of truth for automation.

3

Automated Orchestration: The heart of the factory, combining tools like dbt, Python, Text2SQL, APIs, and Infrastructure as Code to auto -generate & deploy pipelines. This framework ensures modularity, scalability & governance, turning manual development into systemised delivery.

By adopting this model, organisations replace ad-hoc, artisanal practices with a systematised, engineering-driven approach, accelerating delivery while maintaining consistency, compliance and quality.

Tangible, measurable & immediate benefits

The pipeline factory model isn't just a technical shift, it delivers practical advantages across speed, cost, governance, and performance.

Organisations adopting this approach realise value quickly and at scale:

Accelerated Delivery: Business users can articulate their data needs through guided interfaces that generate manifests and trigger automated pipelines, reducing turnaround time from weeks to hours.

Reduced Total Cost of Ownership (TCO): Pipelines are no longer permanent assets. They are spun up, executed, and torn down on demand, minimising infrastructure costs and operational overhead.

Stronger Governance: Standardised manifests automatically populate lineage, monitoring, and documentation tools, improving auditability and regulatory compliance by design.

Optimised Performance: Seamless integration with high-performance engines such as Indexima enables sub-second response times and a frictionless user experience, even at scale.

Evolved Engineering Roles: Rather than being confined to repetitive tasks, data engineers take on a more impactful role: guiding the structure of manifests, contributing reusable components, and ensuring systems are robust, secure, and scalable. Their expertise becomes critical in shaping the automation framework, not just maintaining it.

Human expertise remains critical

What automation can't replace:

Functional Interpretation: Clarifying business needs and constraints.

Schema Design: Defining keys, relationships, and models.

Continuous Supervision: Auditing, monitoring, and evolving systems.

Auto-engineering aims not to replace experts but to empower them with the right tools.

Always start from the business need

In a world cluttered with tools and complex architectures, Keyrus advocates a product mindset over project mindset, and systemic automation over manual development. Keyrus builds platforms like:

K-Observe – for data quality
K-Convert – for automated code migration

These leverage collaborative AI chains that work together methodically.

By focusing on use cases first, and designing from right to left (starting from outcomes), Keyrus repositions tech to serve business needs, not its own complexity.

Avoid the pitfalls of tech hype

With number of AI projects completed in a year, Keyrus champions useful, scalable transformation—reducing friction between business and tech teams.
Auto-engineering is not a trend. It's a method.

At Keyrus, we are your partner in crafting a cutting-edge data architecture and building a robust data platform. We bring together expertise in data strategy, technology, and innovation to help your organisation design and implement a modern data ecosystem that drives insights, agility, and competitive advantage.

Scan here

Craig Andrew
Head of Data Analytics

About Keyrus

Keyrus is a leading global consultancy in data intelligence and digital solutions. We combine business and technical expertise to unlock the maximum value from our customers' data.

Contact us today

www.keyrus.com/za