INFOSYS DE.AI - THE AGENTIC LEAP FOR DATA ENGINEERING
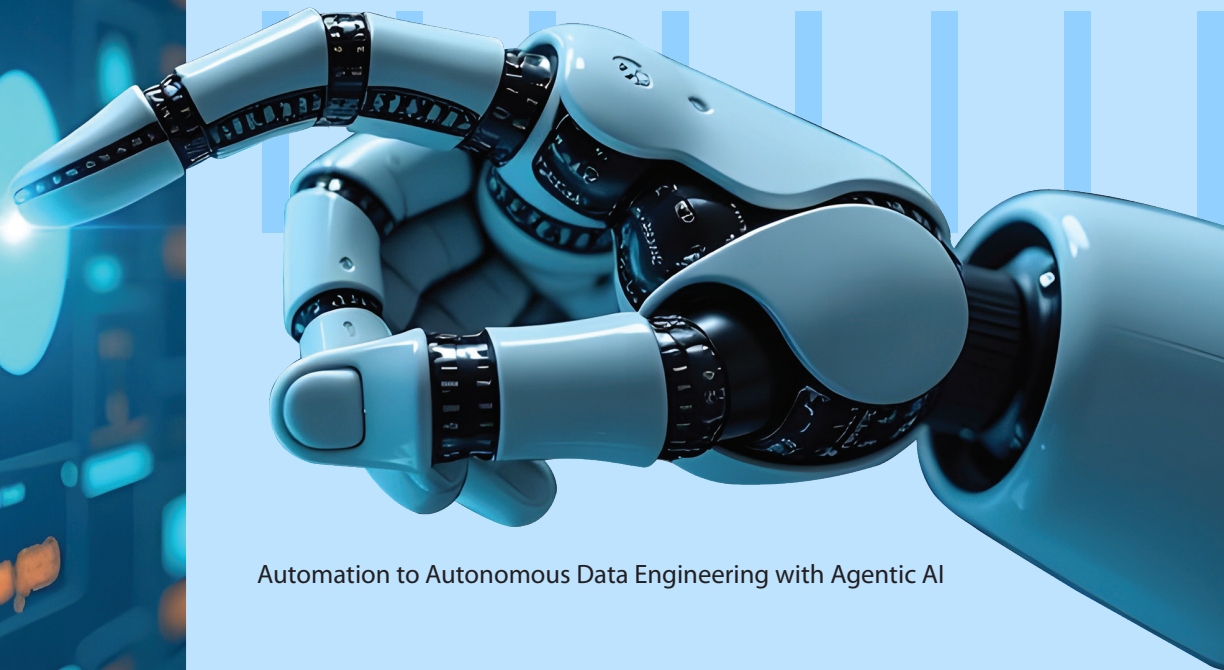
Automation to Autonomous Data Engineering with Agentic AI

Infosys
topaz

Infosys®
Navigate your next

An Evolution in Autonomy

Just as cars evolved from manual operation to cruise control (automation), driver-assist(assistance), and finally to fully autonomous vehicles (like Waymo), data engineering ison the same journey. We are moving from manually scripting pipelines to a futurewhere autonomous agents manage the entire data lifecycle, from intent to production.

The Evolution of Data Engineering

Data engineering is evolving beyond manual scripting. Agentic AI introduces a newparadigm, moving from simple automation to intelligent, autonomous assistants thatmanage the entire data lifecycle.

Stage 1: Automation Driven
Rule-based and schedule-driven. Orchestrators (Airflow, cron) run pre-defined scripts. This is"automation," but it's brittle and 100% human-defined.

Stage 2: Assisted Automation
AI "co-pilots" help engineers write code, fix bugs, and draft documentation. This is an "assistant"model, reducing tasks but requiring human oversight.

Stage 3: Agentic for Personas
Specialized agents assist specific roles: an "Architect Agent" suggests designs, a "Test Agent"generates test cases, a "Dev Agent" writes pipeline code.

Stage 4: Completely Autonomous
A system of agents manages the full lifecycle based on human "intent." It can self-heal, self-optimize, and handle new requests from discovery to publishing.
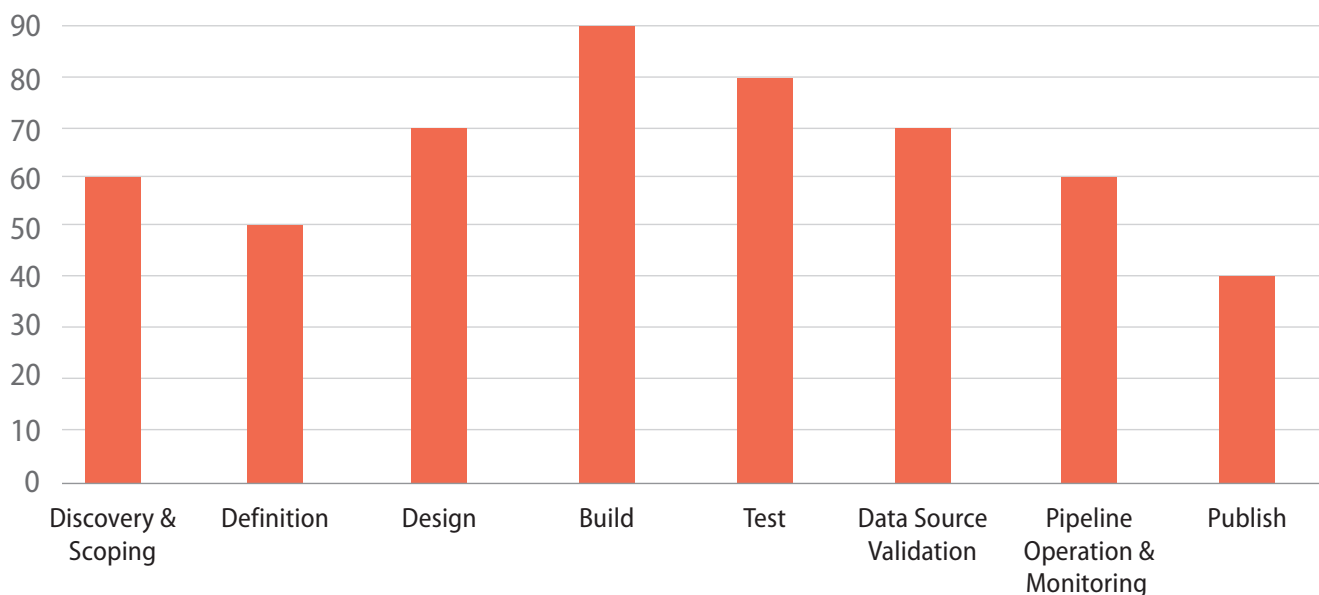
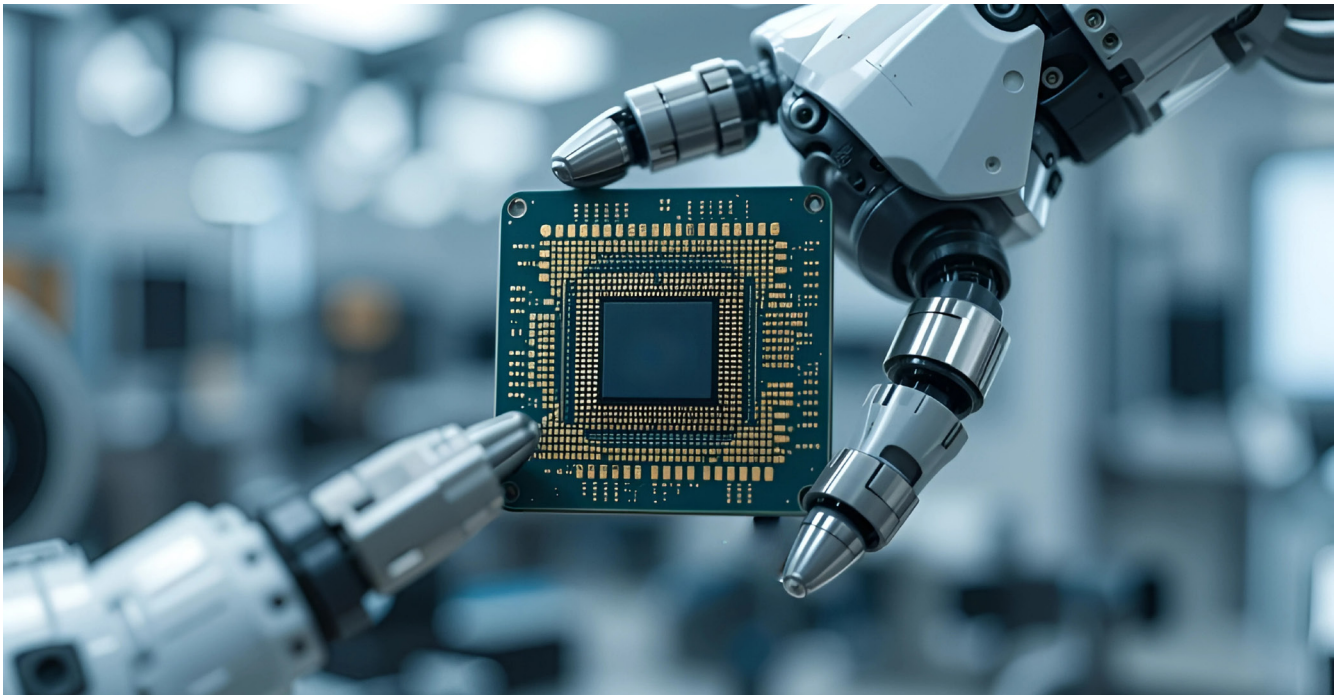# The Bottleneck: Disconnected Data Engineering andTechnology Evolution

The conventional data engineering lifecycle is linear and labor-intensive. While data *platforms*have continuously evolved—from warehouse appliances to cloud platforms, data platforms for AI,and now unified platforms—the core *process* of engineering data has remained a manualbottleneck. This disconnect is where value is lost.

- **High Effort:** 'Build' and 'Test' phases are notoriously time-consuming.
- **Reactive Operations:** 'Operate' is often reactive, fixing breaks rather than pre-empting them.
- **Rigid Design:** 'Design' is a static blueprint that's difficult to change.

Traditional Lifecycle: Manual Effort

| Phase	Effort
Discovery & Scoping	60
Definition	50
Design	70
Build	90
Test	80
Data Source Validation	70
Pipeline Operation & Monitoring	60
Publish	40

The Solution: The Agentic DE Lifecycle

Agentic AI transforms the lifecycle by offloading the majority of effort to intelligent agents. Humanengineers transition to a strategic role, defining high-level goals and validating outcomes.

Theagent handles the asynchronous execution of discovery, design, building, and validation.
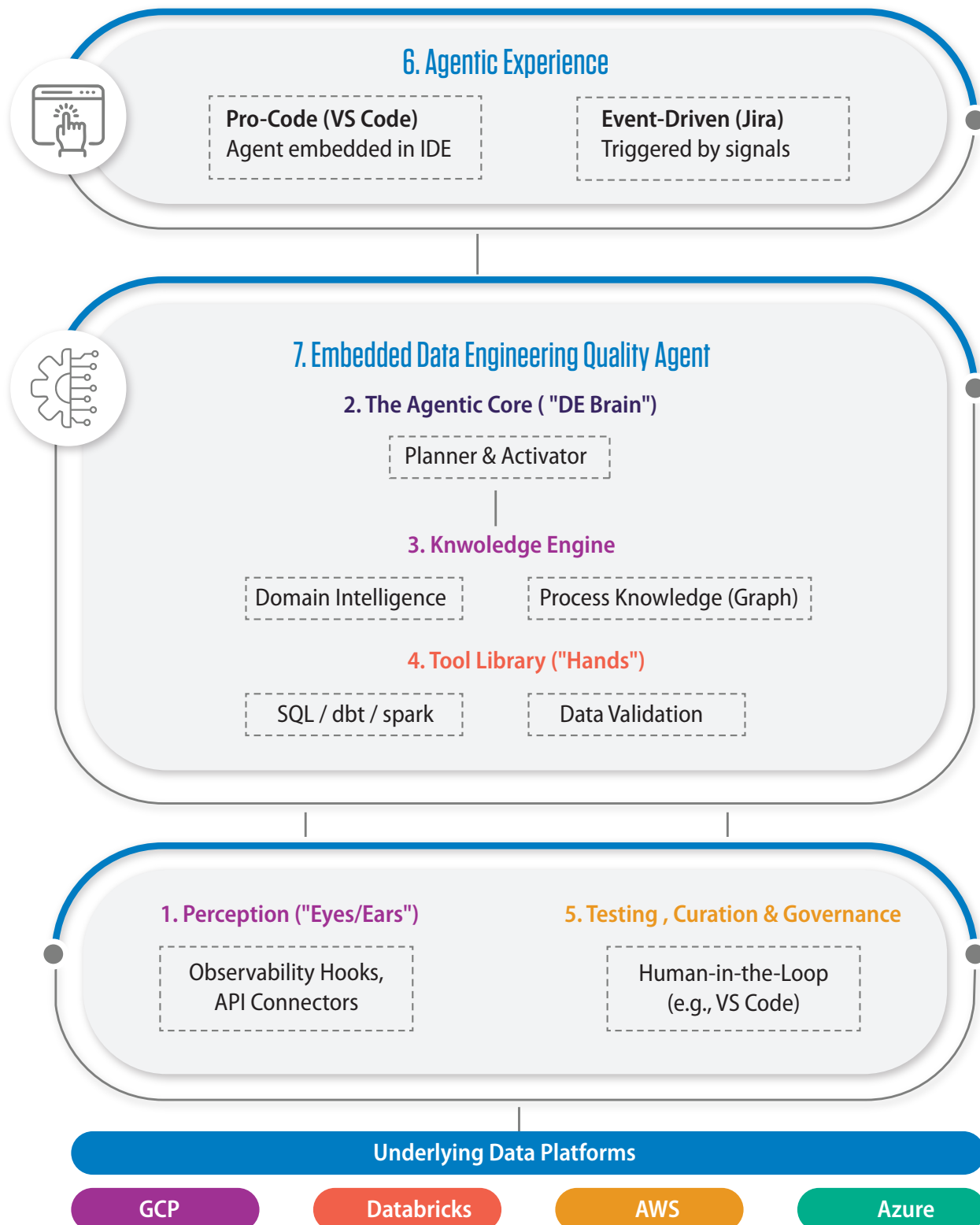
- **Human-in-the-Loop:** Manual effort shifts to goal-setting and validation.
- **Autonomous Execution:** Agents autonomously perform the heavy lifting.
- **Continuous Adaptation:** The system can self-heal and adapt to schema drift or sourcechanges.

Agentic Lifecycle: Effort Rebalanced
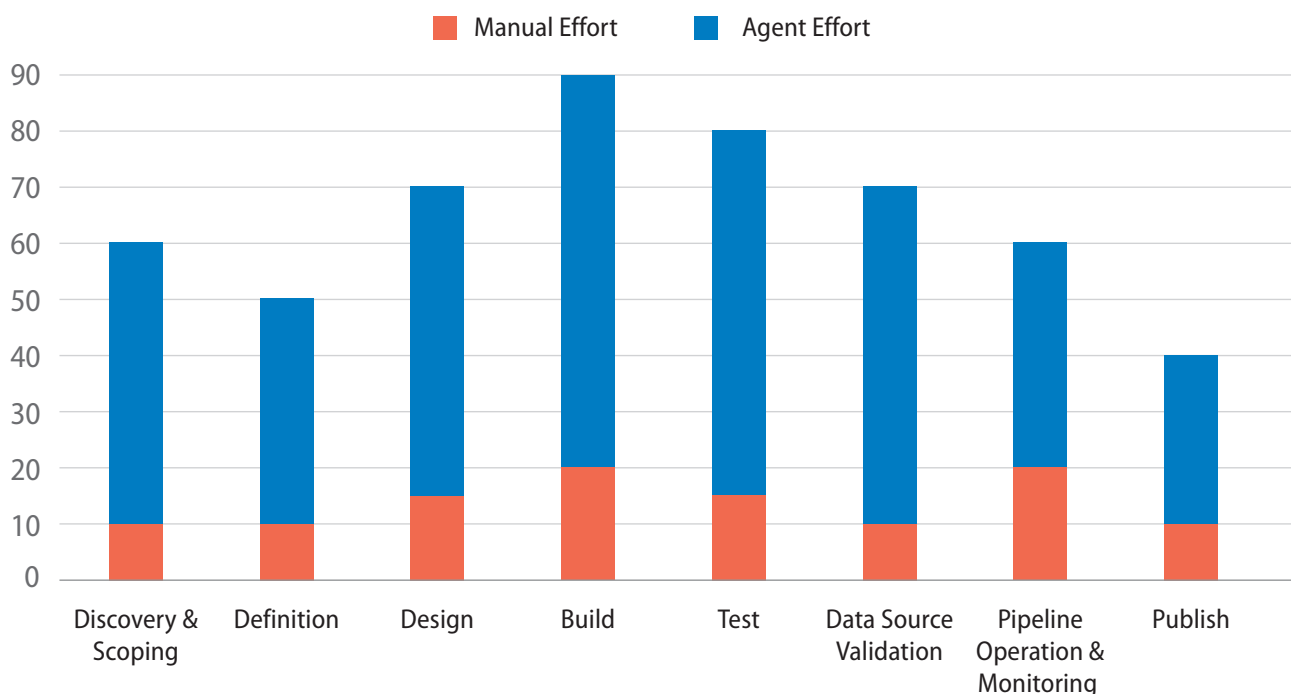
Capability Blueprint: 'Lego Blocks' for an Agentic Data Engineering System

Agentic Data Engineering Architecture

6. Agentic Experience

Pro-Code (VS Code)
Agent embedded in IDE

Event-Driven (Jira)
Triggered by signals

7. Embedded Data Engineering Quality Agent

2. The Agentic Core ("DE Brain")

Planner & Activator

3. Knwoledge Engine

Domain Intelligence

Process Knowledge (Graph)

4. Tool Library ("Hands")

SQL / dbt / spark

Data Validation

1. Perception ("Eyes/Ears")

Observability Hooks,
API Connectors

5. Testing , Curation & Governance

Human-in-the-Loop
(e.g., VS Code)

Underlying Data Platforms

GCP

Databricks

AWS

Azure

1. Perception & Interface

The "eyes and ears" of the agent. This includes API connectors to data sources and observabilityhooks to monitor data quality and pipeline status.

2. The Agentic Core (The "Brain")

The central planning and reasoning engine . It decomposes complex requests (e.g., "Build a salesdashboard") into a series of executable steps.

3. Knowledge Engine

Provides context. Includes Domain Data Intelligence (metadata, business rules) and ProcessIntelligence (pipeline logs, best practices).

4. Tool Library (The "Hands")

A curated set of functions the agent can call: schema discovery, data validation, running SQLqueries, code generation, and pipeline deployment (dbt, Airflow).

5. Testing, Curation & Governance

The human-in-the-loop component, often integrated into a known IDE like VS Code, for validatingagent-generated code, approving data publications, and providing feedback.
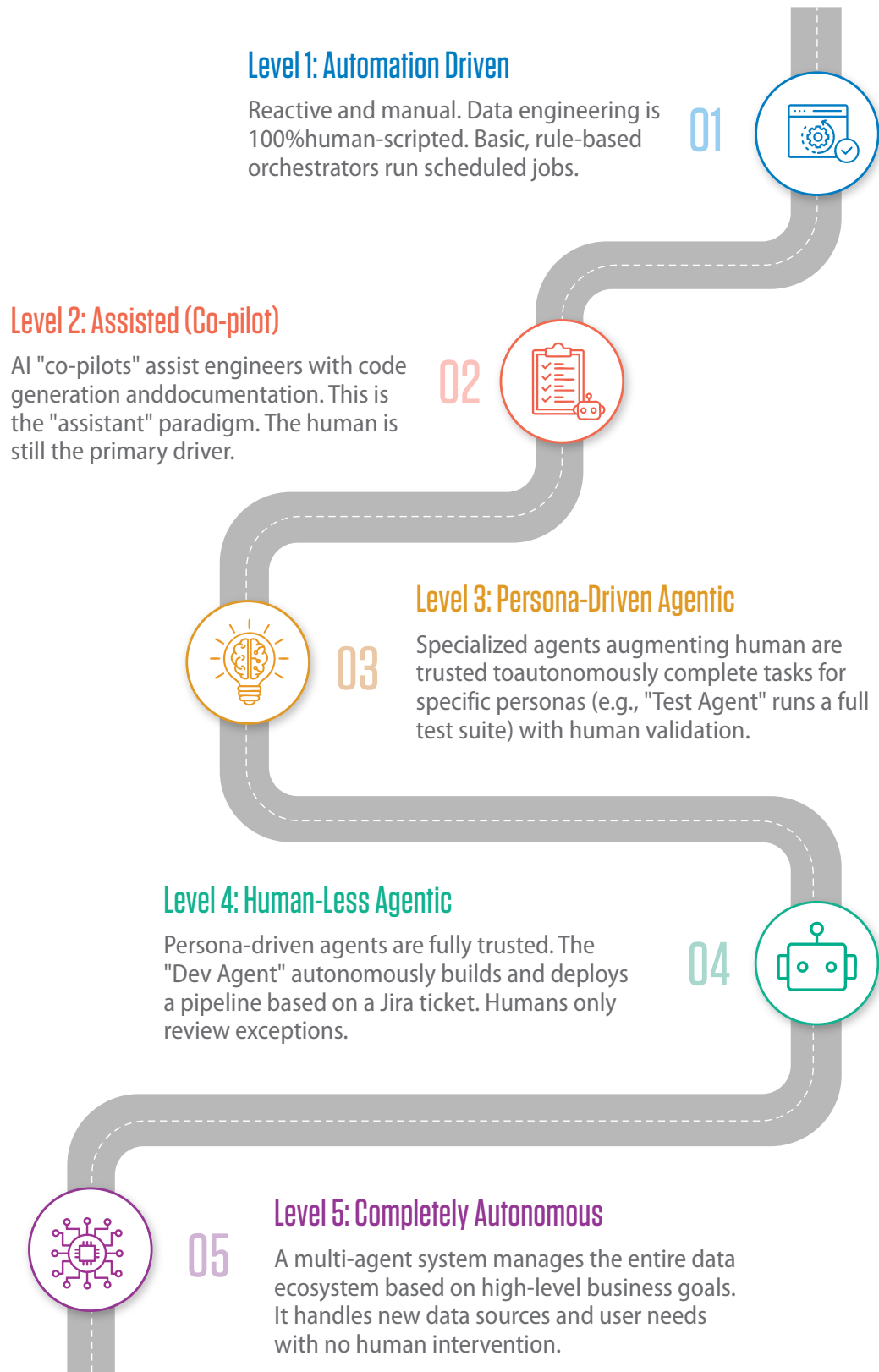
6. Agentic Experience

How humans interact with the agent. Includes "pro-code" experiences (embedded in VS Code) and event-driven triggers (a change in a Jira story triggers an agent).

7. Embedded Data Engineering Quality

An overarching layer of non-negotiable standards enforced by and adjudication agent. Thisincludes Code Completeness, Correctness, Security by Design, Compliance by Design, andadherence to engineering Best Practices.

# Roadmap to Organizational Maturity

Adopting agentic AI is a journey, not a single step. Organizations typically progress through five stages of maturity, building capability and trust at each level.

Level 1: Automation Driven
01

Reactive and manual. Data engineering is 100%human-scripted. Basic, rule-based orchestrators run scheduled jobs.

Level 2: Assisted (Co-pilot)
02

AI "co-pilots" assist engineers with code generation anddocumentation. This is the "assistant" paradigm. The human is still the primary driver.

Level 3: Persona-Driven Agentic
03

Specialized agents augmenting human are trusted toautonomously complete tasks for specific personas (e.g., "Test Agent" runs a full test suite) with human validation.

Level 4: Human-Less Agentic
04

Persona-driven agents are fully trusted. The "Dev Agent" autonomously builds and deploys a pipeline based on a Jira ticket. Humans only review exceptions.

Level 5: Completely Autonomous
05

A multi-agent system manages the entire data ecosystem based on high-level business goals. It handles new data sources and user needs with no human intervention.

The Transformative Impact

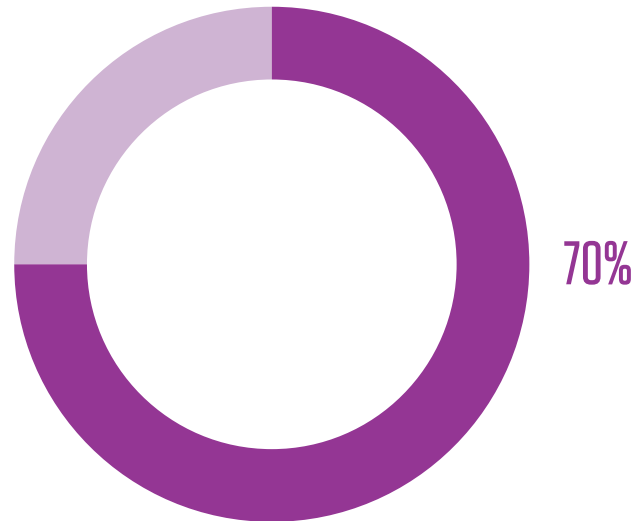The shift to an agentic model unlocks profound business value by freeing your most skilledengineers from low-level maintenance to focus on high-impact strategic work.
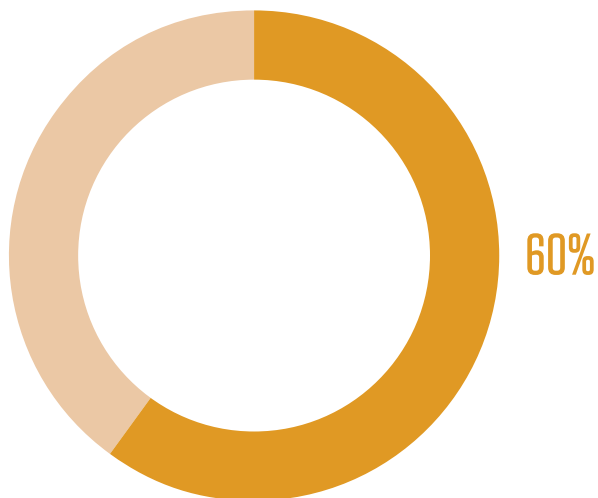
Faster Development

75%
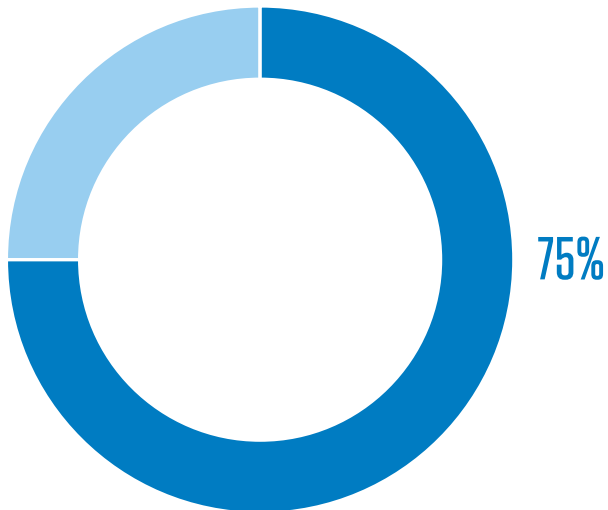
Reduction in pipeline development and deployment time by automating design, build, and test phases.

Fewer Errors

70%

Reduction in data quality issues and pipeline failures through autonomous validation and self-healing.

Cost to Value

60%

Improvement in cost-to-value ratio by automating high-effort, low-value tasks and focusingresources on innovation.

Authors

Rajan Padmanabhan

Unit Technology Officer, Data Analytics and AI, Infosys

Vishnu Sankar

Principal Technology Architect, Data Analytics and AI, Infosys

For more information, contact askus@infosys.com

Infosys®
Navigate your next

Infosys.com | NYSE: INFY

Stay Connected