A recent MIT Tech Review Report shows that 88% of surveyed organizations are either investing in, adopting or experimenting with generative AI (GenAI) and 71% intend to build their own GenAI models. This increased interest in AI is fueling major investments as AI becomes a differentiating competitive advantage in every industry. As more organizations work to leverage their proprietary data for this purpose, many encounter the same hard truth:

The best GenAI models in the world will not succeed without good data.

This reality emphasizes the importance of building reliable data pipelines that can ingest or stream vast amounts of data efficiently and ensure high data quality. In other words, good data engineering is an essential component of success in every data and AI initiative and especially for GenAI.

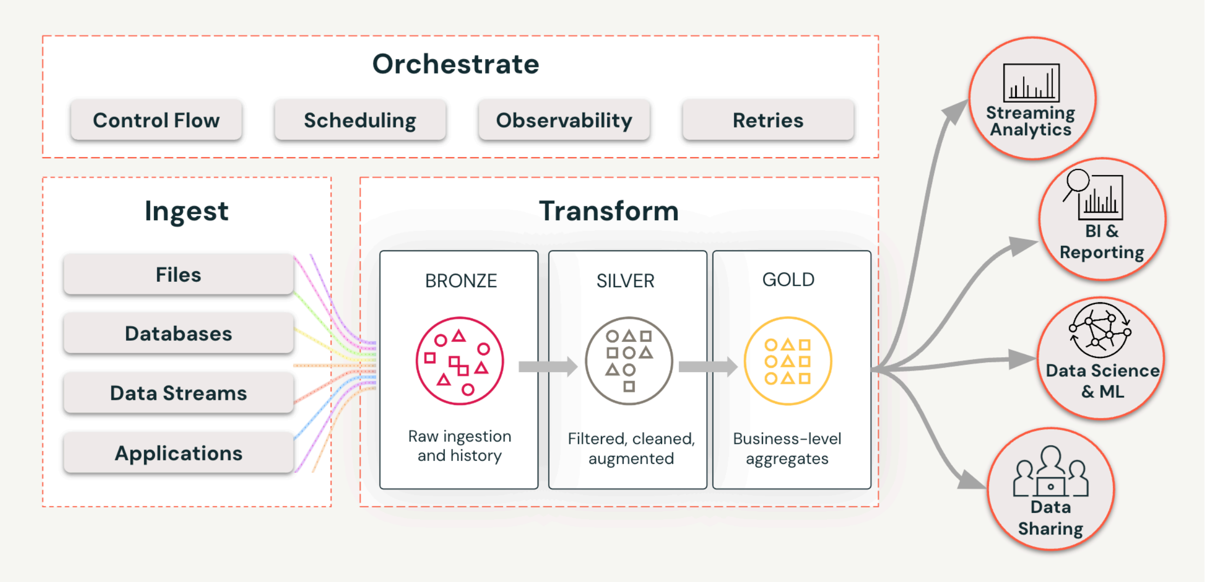
Using practical guidance, useful patterns, best practices and real-world examples, this book will provide you with an understanding of how the Databricks Data Intelligence Platform helps data engineers meet the challenges of this new era.

What is data engineering?

Data engineering is the practice of taking raw data from a data source and processing it so it's stored and organized for a downstream use case such as data analytics, business intelligence (BI) or machine learning (ML) model training. In other words, it's the process of preparing data so value can be extracted from it.

A useful way of thinking about data engineering is by using the following framework, which includes three main parts:

1. **Ingest**
   Data ingestion is the process of bringing data from one or more data sources into a data platform. These data sources can be files stored on-premises or on cloud storage services, databases, applications and increasingly — data streams that produce real-time events.
2. **Transform**
   Data transformation takes raw ingested data and uses a series of steps (referred to as "transformations") to filter, standardize, clean and finally aggregate it so it's stored in a usable way. A popular pattern is the medallion architecture, which defines three stages in the process — Bronze, Silver and Gold.
3. **Orchestrate**
   Data orchestration refers to the way a data pipeline that performs ingestion and transformation is scheduled and monitored as well as the control of the various pipeline steps and handling failures (e.g., by executing a retry run).



Challenges of data engineering in the AI era

As previously mentioned, data engineering is key to ensuring reliable data for AI initiatives. Data engineers who build and maintain ETL pipelines and the data infrastructure that underpins analytics and AI workloads face specific challenges in this fast-moving landscape.

- **Handling real-time data:** From mobile applications to sensor data on factory floors, more and more data is created and streamed in real time and requires low-latency processing so it can be used in real-time decision-making.
- **Scaling data pipelines reliably:** With data coming in large quantities and often in real time, scaling the compute infrastructure that runs data pipelines is challenging, especially when trying to keep costs low and performance high. Running data pipelines reliably, monitoring data pipelines and troubleshooting when failures occur are some of the most important responsibilities of data engineers.
- **Data quality:** "Garbage in, garbage out." High data quality is essential to training high-quality models and gaining actionable insights from data. Ensuring data quality is a key challenge for data engineers.
- **Governance and security:** Data governance is becoming a key challenge for organizations who find their data spread across multiple systems with increasingly larger numbers of internal teams looking to access and utilize it for different purposes. Securing and governing data is also an important regulatory concern many organizations face, especially in highly regulated industries.

These challenges stress the importance of choosing the right data platform for navigating new waters in the age of AI. But a data platform in this new age can also go beyond addressing just the challenges of building AI solutions. The right platform can improve the experience and productivity of data practitioners, including data engineers, by infusing intelligence and using AI to assist with daily engineering tasks.

In other words, the new data platform is a data *intelligence* platform.