BARC

# The Ultimate Guide to Data Pipelines for Generative AI: Five Criteria to Evaluate Tools

Author: Kevin Petrie

Publication: June 2024

## Abstract

This ultimate guide defines five product evaluation criteria for GenAI data pipeline tools: functional breadth, ease of use, governance, performance & scale, and cost.

# Inhalt

# The Ultimate Guide to Data Pipelines for Generative AI: Five Criteria to Evaluate Tools

As companies apply Generative AI language models to their own domain-specific data, they create both promise and peril. The promise: to boost productivity and gain competitive advantage by enriching business functions such as customer service, document processing, and content development. But the peril ranges from broken workflows to angry customers and inquisitive regulators. To realize the promise and avoid the peril, companies need to prepare GenAI inputs that accurately describe business reality.

Achieving this requires a new class of data pipelines – and new tools to manage them. This ultimate guide defines five product evaluation criteria for GenAI data pipeline tools: functional breadth, ease of use, governance, performance & scale, and cost. For each criterion it recommends key questions for data engineering leaders to pose to vendors.

***Companies need new pipelines to transform their unstructured data into usable, trustworthy GenAI inputs***

We start by defining the market and product category.

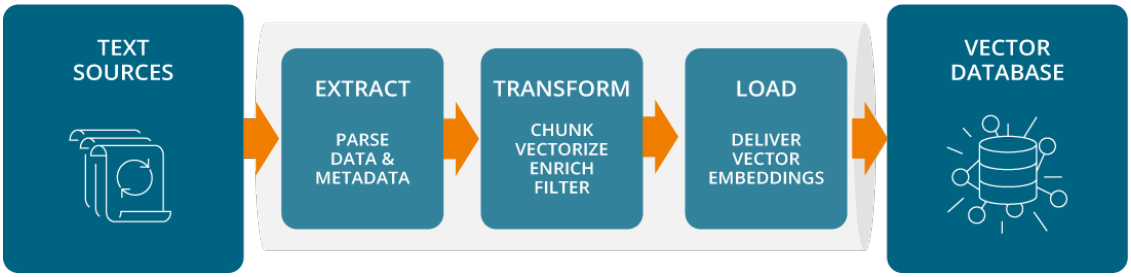
## Market

Generative AI (GenAI) is a type of neural network that data scientists train to interpret and create text, images, audio, or even video. Much of the current GenAI excitement centers on language models (LMs) such as ChatGPT from OpenAI, Gemini from Google, or Llama from META. These LMs generate content, often strings of words, after studying how words or objects relate to one another in an existing corpus of similar content. Companies now are customizing LMs to understand their own content. But first data engineers must transform all that unstructured data – emails, service tickets, videoconference recordings, and so on – into something that LMs can use to generate trustworthy content. This is where GenAI data pipelines tools enter the picture.

## Product category

A GenAI data pipeline tool helps data engineers build and manage pipelines that perform the following tasks, using the example of text sources and a vector database target.

**Example of a GenAI Data Pipeline**




- **Extract.** First the pipeline parses and extracts relevant text and metadata from source applications or files, including complex documents that might contain figures and tables. The metadata and structure include elements such as the document title, body, and footnotes.
- **Transform.** Next the pipeline transforms the extracted documents. It divides the text into semantic "chunks" and uses an embedding model to generate vector embeddings that describe the meaning and interrelationships of chunks. It also might filter out sensitive fields or enrich document chunks with data from other systems and data platforms. (Some pipeline tools perform these transformation steps in an intermediate landing zone using an ELTL sequence.)
- **Load.** Finally it delivers the vector embeddings to a target, most often a vector database such as Pinecone and Weaviate or vector-capable platforms such as Databricks and MongoDB.

These platforms index the vectors, also called embeddings, to support similarity searches by a GenAI application that contains the LM.

***GenAI data pipeline tools build and manage pipelines that extract, transform, and load inputs to feed language models***

## Consumption by language models

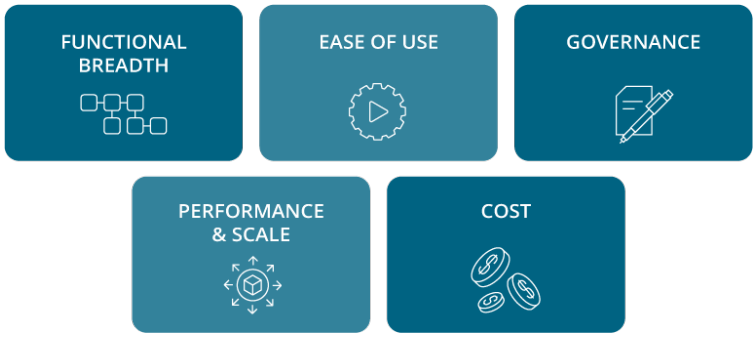
The GenAI application, sometimes operating as an autonomous "agent", often consumes vectors in as part of a process called retrieval-augmented generation (RAG). With RAG, the GenAI application finds and retrieves embeddings based on their similarity to a user's prompt. Then it augments the prompt with that content, thereby increasing the odds that the LM generates an accurate response rather than hallucinating. In addition, data scientists use vector embeddings as part of the process for fine-tuning LMs. They retrieve similar embeddings from the vector database, feed them to the LM, and compare results. This helps them iteratively fine-tune the LM and make outputs more accurate.

Both RAG and fine-tuning help LMs understand and work with companies' domain-specific data. And both need GenAI data pipelines.

## Ask the hard questions

Data engineering leaders should evaluate GenAI data pipeline tools by functional breadth, ease of use, governance capabilities, performance & scalability, and cost. Let's explore each of these evaluation criteria, along with key questions to ask vendors.

**Evaluation Criteria for GenAI Data Pipeline Tools**




***Evaluate GenAI data pipeline tools by functional breadth, ease of use, governance capabilities, performance & scalability, and cost***

# Criterion 1. Functional breadth

Data engineering leaders should evaluate tools' functional breadth by posing the following questions.

### Does this tool enable users to manage the full lifecycle of a GenAI data pipeline?

Your tool should enable your team to design, test, deploy, observe, and orchestrate GenAI data pipelines.

- **Design.** Select and configure pipeline elements such as sources, targets, and transformation tasks.
- **Test.** Test the pipelines, perhaps by comparing the performance of different versions in a development environment, or by checking whether vectors align with the embedding model.
- **Deploy.** Release the pipeline to production. Iterate by branching a pipeline version out of production, revising it, then deploying the new version to production.
- **Observe.** Monitor performance and tune pipelines to optimize pipeline performance and output quality.
- **Orchestrate.** Schedule, execute, and synchronize tasks across pipelines and the applications that consume their outputs.

Your team needs a tool that handles these steps in a single interface to reduce complexity, training requirements, and risk. This also helps data teams streamline pipeline iterations based on what they observe. They can re-design pipelines or conduct new tests, then deploy and orchestrate the new pipelines.

## What sources and targets does it support?

GenAI thrives on a wide variety of data. Your pipeline tool should support all your sources and targets, including both current and planned or potential elements. The sources of text, imagery, audio, and video might include applications such as Microsoft 365, Salesforce, GitHub, Slack, and Zoom.

Targets, meanwhile, might include vector databases such as Pinecone, Weaviate, and Vespa, or vector-enabled data platforms such as Databricks, MongoDB, and SingleStore. The major cloud object stores – Amazon S3, Azure Blob Storage, and Google Cloud Storage – are popular as both sources and targets. Whatever your end points, on premises or in the cloud, you want open APIs and a simple integration process.

## What types of transformations does it perform?

Transforming diverse, unstructured datasets into usable embeddings can require sophisticated transformation methods. Ask your vendor about their support for capabilities such as the following.

- **Normalize** source content to create consistent formats and data ranges.
- **Adjust chunk size** to determine the optimal size for supporting accurate similarity searches. You might want to test one size for a subset of data, review results, then adjust.
- **Support multiple chunking techniques**, such as fixed-size chunking or content-aware methods that split sentences, create hierarchical structures, and so on.
- **Enrich chunks** by appending metadata that describes objects' origin, usage, or content. This helps create and update embeddings in the vector database.
- **Create vector embeddings** based on the logic of various popular embedding models.
- **Filter data**, for example by identifying and redacting sensitive or personally identifiable information (PII).

## What types of multimodal data does your pipeline tool support?

Your pipeline tool should support common file formats such as the following:

- **Text:** PDF, DOC and DOCX, HTML and HTM, and XLS and XLSX
- **Image:** Bitmap, JPEG, GIF, PNG, and EPS.
- **Audio:** MP3, WAV, ALAC, and WMA.
- **Video:** MP4, MOV, WMV, MPEG-4, and AVI

## How does it integrate or interoperate with other elements of the ecosystem?

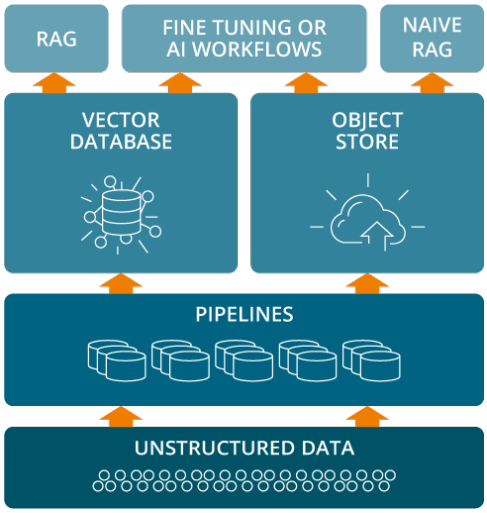
To support diverse applications, workflows, and environments, your pipeline tool should integrate or interoperate with elements such as the following.

- **Development frameworks.** Developers use LangChain, LlamaIndex, and other frameworks to build GenAI applications and RAG processes.
- **Embedding models.** Communities such as HuggingFace provide the embedding models that create vectors to capture the meaning and inter-relationships of chunks.
- **Other tools.** Pipeline tools such as Qlik CDC Streaming, Fivetran, and the open-source Debezium specialize in the ingestion of structured data. dbt, meanwhile, specializes in structured data transformation, and Unstructured.io transforms unstructured data.
- **Containers.** Platforms such as Kubernetes orchestrate the deployment and scaling of applications within containers.
- **Catalogs.** Popular catalogs from Informatica, Alation, Collibra, and other vendors organize metadata that describes both structured and unstructured datasets.
- **Open-source projects.** Modern pipelines rely on open-source software such as Apache Spark, Flink, Kafka, NiFi, and Airflow to move and transform data.
- **Programming languages.** Data scientists and developers build pipelines using languages such as SQL, Python, and Java.
- **Cloud platforms.** Many pipelines and datasets rely on cloud infrastructure hosted by AWS, Azure, and Google Cloud.

## What types of AI implementations does it support?

Companies consume the outputs of GenAI data pipelines in several ways. Your pipeline tool should prepare and deliver data to support the following scenarios.

**AI Implementations**



- **RAG.** As described earlier, GenAI application uses a similarity search to find and retrieve embeddings from a vector database, then adds them to user prompts.
- **Fine-tuning.** Also as described earlier, data science teams retrieve and compare embeddings from the vector database as they fine tune the LM. They also might use content from an object store.
- **Multi-faceted AI workflows.** The GenAI application and other functions (natural language processing, predictive ML, etc.) retrieve content from the vector database or object store.
- **Naive RAG.** The GenAI application searches through key words to find relevant files within an object store. It retrieves them and augments the user prompt with them.

## Criterion 2. Ease of use

Data engineering leaders should evaluate tools' ease of use by asking vendors the following questions.

### What skills and how much training does this product require?

GenAI forces data engineers to learn how to handle unfamiliar elements such as unstructured data, chunking techniques, and embedding models. Evaluate pipeline tools based on their ability to make data engineers proficient in these areas with brief online instruction, delivered live or through recorded sessions. Your tool should not require a data engineer to get significant assistance from a data scientist, ML engineer, or NLP engineer.

### Does the tool take a declarative or imperative approach to coding?

A declarative approach simplifies coding by letting users state what they want – in this case, the type of pipeline they want, including its sources, targets, and transformation techniques – and let the tool decide how to build it. Look for a tool that takes a declarative approach rather than imperative, which instead requires users to specify the exact commands required to build a pipeline.

### What level of automation does it offer?

As always, the more you automate routine tasks, the better you improve productivity. Look for pipeline tools that offer a graphical interface to minimize scripting, and better yet a GenAI chatbot to minimize typing. This interface or chatbot should guide users through the basic steps of pipeline design, offering tips and posing questions about their requirements along the way. Remember, of course, that expert users still must inspect and often revise tool outputs for quality.

### What level of productivity does it support?

User productivity boils down to output per hour. Devise some simple, measurable output metrics – perhaps the number of sources connected, volume of data ingested, or pipelines created – and use them to compare output per hour across tools. You should test how output varies by use case to ensure your team can stay productive when requirements change. Also be sure to measure user productivity for updating pipelines, for example when source or target versions change.

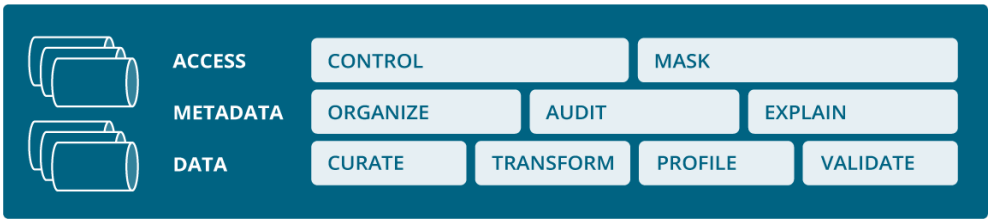
### Is there an active user community to engage?

Emerging product categories such as GenAI data pipelines have a particular need for peer advice. Look for an active community of users that provide validation, support, and continuous knowledge sharing. You might find these communities in Slack forums, at Meetups, or at user conferences hosted by large partners such as Databricks.

## Criterion 3. Governance capabilities

The risks of GenAI range from hallucinations to privacy breaches, biased outputs, and compromised intellectual property. Look for a tool that reduces these risks by governing data, metadata, and user access. The following questions probe the ability of vendors to do this, working from the bottom up in our diagram.

**Governing GenAI Data Pipelines**




### How does the tool help users govern data?

Data engineers and their colleagues govern GenAI inputs by curating, transforming, profiling, and validating data. Your pipeline tool should assist each of these steps.

- **Curate.** The tool should curate source data by enabling users to track and edit their labels, annotations, and other metadata.
- **Transform.** It should give users visibility into transformation tasks such as chunking and enrichment.
- **Profile.** By profiling GenAI inputs at each stage in the pipeline, the tool can help users track lineage end to end.
- **Validate.** The tool should validate GenAI inputs, for example by checking data against schema registries and enabling users to insert custom logic.

Capabilities like these make GenAI inputs more accurate and trustworthy.

### How does this tool govern metadata?

Compare the ability of different tools to organize metadata, enable auditing, and meet explainability requirements.

- **Organize.** Your tool should organize metadata that describes elements such as files, pipelines, and workflows, make it searchable for users, and integrate it with catalogs. It also should map file hierarchies and other relationships between datasets.
- **Audit.** It should maintain an audit log of all pipeline activities. Data engineers, data stewards, and compliance officers need to search, review, and export these records to support both internal and external reporting requirements.
- **Explain.** Data engineers need the option to inspect pipeline inputs, transformation techniques, and outputs. If they ask the chatbot why it recommends a certain chunking method or embedding model, they need an intelligible explanation to share with stakeholders such as data scientists, business experts, and auditors.

### How does it govern access to pipelines and data?

Compare the ability of different tools to organize metadata, enable auditing, and meet explainability requirements.

- **Access controls**. Administrators should be able to enforce role-based access controls that ensure only authenticated users perform only authorized actions on permissible pipelines or datasets. For example, you might not want the data engineer for one business unit to revise pipelines belonging to another business unit.
- **Masking.** Your tool should selectively mask files, tables, or columns from designated parties. For example, data engineers should not be able to view personally identifiable information (PII) such as social security numbers.

These capabilities reduce the likelihood of errors or tampering.

## Criterion 4. Performance & Scalability

Ask the following questions to assess pipeline tools' performance and scalability.

### Can this tool meet service level requirements (SLAs) for the business?

Devise proofs of concept that test each tool's ability to support your most rigorous use cases with low enough latency and high enough throughput. Vendors' documentation and support team should give you ample guidance about what resources their tool needs to perform – and how you can scale to support new sources, targets, and streams in a timely fashion. You also should have granular visibility into CPU/memory utilization metrics and performance KPIs so you can anticipate and prevent bottlenecks. Ask each vendor for proof points about the SLAs they meet for other customers with similar use cases and environments.

### Does it support periodic batch, incremental batch, and streaming delivery options?

To integrate with heterogeneous environments, your pipeline tool should give you multiple options for delivering new or changed data to a target. It should process full batches of source data, incremental batches of source updates, and events within a Kafka topic stream. In addition, when delivering

updates, your tool should have the choice of appending or merging any insertions, changes, or deletions of data objects. Probe vendors about the latency and throughput you can expect with each delivery option.

### How does the tool scale to support rising workloads?

Your pipeline tool should scale quickly, ideally by enabling users to configure new cloud compute nodes in support of new sources, targets, and pipelines. If it adds nodes automatically when workloads spike, you will want the ability to monitor those changes, receive alerts, and even set hard limits to avoid cost overruns. Also look for tools that can process documents or other objects in parallel. You might want guidance within product documentation or the interface itself about how to anticipate and prepare for workload bursts. If your tool can forecast workload behavior based on historical trends, all the better.

### What footprint and workload impact does your tool have on your environment?

Assess the footprint of each tool in your environment. It might be a managed service that consumes few local resources, or it might require installation on a dedicated server on premises. It also might require the installation and maintenance of software agents on pipeline sources or targets. Whatever the implementation scenario, you should estimate its impact on the performance of the other workloads and applications with which it shares resources.

## Criterion 5. Cost

Ask vendors the following questions to understand the cost of investing in their tool.

### How is the product priced?

Software pricing structures vary widely. Many products now favor a subscription model with monthly or annual charges that cover both licensing and support. Those charges depend on a usage metric such as data delivered, nodes allocated or utilized, number of users, or number of pipelines. Whatever the pricing structure, you should probe vendors about the level of transparency, simplicity, and control they provide to users.

- **Transparency**. Your team should understand up front what the software costs for their environment, and exactly how that will vary based on changes to your usage. With per-node pricing, for example, you will want visibility into the dollar implications of workload changes – especially in the case of auto-scaling.
- **Simplicity**. Your vendor should make it simple to understand costs. You should not have to track multiple metrics and re-calculate charges each time you upgrade software, add a target, or reconfigure a pipeline. Keep it simple to minimize your effort and the risk of a surprise.
- **Control**. Your team needs to maintain control over costs by keeping their hand on a few levers. If the tool offers auto-scaling of cloud compute clusters, your administrator should be able to throttle workloads by configuring alerts and setting hard limits. If the vendor charges by number of users or pipelines, you want control over those additions as well.

## How do upfront and ongoing software costs vary based on expected workload ranges?

Work with your business leaders to forecast the likely tool and resource requirements of supporting your target use cases. Use these targets to model your likely tool usage, resource consumption, and software costs over the next 12 months.

## What are the expected costs of learning, implementing, and maintaining this tool?

During the vendor product demonstration and proof of concept, build rough estimates of the worker hours required for your team to get proficient with each tool. Then forecast the time required for them to teach their peers and implement the tool to support your target use cases. Maintenance costs will be harder to forecast, so be sure to script your proof of concept to include as many of the likely future environment changes as possible.

## Next Steps

GenAI creates justified excitement about the opportunity to achieve topline and bottom-line benefits. But it also exposes the Achilles heel of many organizations: inaccurate, ungoverned data. Your data team needs the help of a commercial tool to convert this data into viable GenAI inputs. Data leaders can select the right GenAI data pipeline tool by using the evaluation criteria of functional breadth, ease of use, governance capabilities, performance & scalability, and cost. They can get started by taking three steps.

- **Get peer advice**. Make back-channel checks to get past the slides and demos. Your user community and even personal network might have experts with informed opinions about what you really need – and whether a given tool will help. Find these experts and pick their brains.
- **Evaluate the vendor too**. Behind each tool sits a company and its resources. Assess the track record of its leaders in meeting its commitments and delivering value to customers. Also consider its funding, engineering talent, and notable hires or turnover.
- **Play the long game**. Given the early stage of this market, you might not find a GenAI pipeline tool that meets and exceeds all these requirements today. Compare the ability of tools to meet your most important near-term requirements and assess their roadmap for delivering on the rest.

# About BARC

BARC (Business Application Research Center) is one of Europe's leading analyst firms for business software, focusing on the areas of data, business intelligence (BI) and analytics, enterprise content management (ECM), customer relationship management (CRM) and enterprise resource planning (ERP). Our passion is to help organizations become digital companies of tomorrow. We do this by using technology to rethink the world, trusting databased decisions and optimizing and digitalizing processes. It's about finding the right tools and using them in a way that gives your company the best possible advantage. This unique blend of knowledge, exchange of information and independence distinguishes our services in the areas of research, events and consulting.

**Research**

Our BARC studies are based on internal market research, software tests and analyst comments, giving you the security to make the right decisions. Our independent research brings market developments into clear focus, puts software and vendors through their paces and gives users a place to express their opinions.

**Events**

Decision-makers and IT industry leaders come together at BARC events. BARC seminars in small groups, online webinars and conferences with more than 1,000 participants annually all offer inspiration and interactivity. Through exchange with peers and an overview of current trends and market developments, you will receive new impetus to drive your business forward.

**Consulting**

In confidential expert workshops, coaching and in-house consultations, we transform the needs of your company into future-proof decisions. We provide you with successful, holistic concepts that enable you to use the right information correctly. Our project support covers all stages of the successful use of software.

BARC

BARC

Data Decisions. Built on BARC.

www.barc.com

**Germany**
BARC GmbH
Berliner Platz 7
D-97080 Würzburg
+49 931 880651-0

**Austria**
BARC GmbH
Hirschstettner Straße 19 / I / IS314
A-1220 Wien
+43 660 6366870

**Switzerland**
BARC Schweiz GmbH
Täfernstr. 22a
CH-5405 Baden-Dättwil
+41 56 470 94 34

**United States**
BARC US
13463 Falls Drive
Broomfield, CO 80020
USA