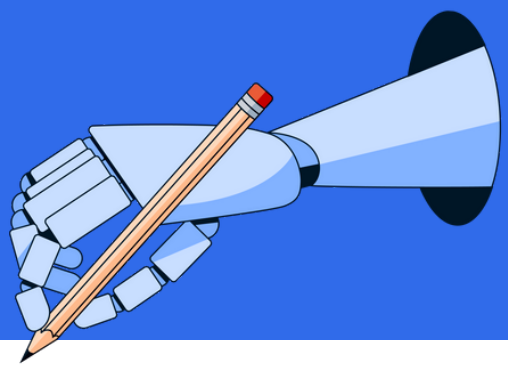
Fivetran Capgemini

Data readiness checklist for generative AI



AI strategy doesn't start with an app or a platform — it starts with a strong data foundation. Building a strong data foundation for AI with data integration best practices is paramount for accurate and timely data to deliver relevant results.

Research shows that supplying subpar data to AI models can lead to severe financial loss. In fact, underperforming AI models built using low-quality or inaccurate data cost companies up to 6% of annual revenue on average.

Here's a step-by-step guide to building a strong data foundation.

☐ Step 1: Consider what problem you're trying to solve

Determine if GenAI is best for your use case. In some cases, simpler predictive modeling and even hard-coded empirical rules could also work. Aligning projects with strategic business drivers — like domain use cases and key metrics — will help you make better decisions.

- Use cases where GenAI excels include:
  - Rapidly creating media, including content, code and digital assets of all kinds
  - Summarizing and extracting insights from large volumes of unstructured data
  - Automation that requires understanding and producing natural language or other unstructured data, like chatbots

☐ Step 2: Identify and inventory data sources

Understanding your data is key for generative AI readiness. This enables organizations to optimize data utilization, break down silos and ensure AI models draw from the most diverse data.

- ☐ **Identify data sources.** Sources should include databases, file systems, cloud storage, external data sources, APIs and even unstructured data like emails or documents. Ask each team in your organization about their unique domain-specific data sources.
- ☐ **Catalog and classify data.** For each data source, document the type of data it contains (e.g., customer information, transactional data or sensor data). Classify the data based on sensitivity, regulatory requirements and business priorities.
- ☐ **Assess data quality**. Evaluate the quality of the data in terms of accuracy, completeness, consistency, timeliness and reliability. This step is crucial for determining the usability of the data and will help determine the relative priority for each data stream.
- ☐ **Document data access and usage.** Record how the data is accessed as well as who has access and for what purpose. You must understand dependencies and potential bottlenecks.
- ☐ **Cadence of data.** Understand how frequently data is updated and how often this data is accessed and used in analysis. This will help you determine how frequently data needs to be refreshed in your model.

Step 3: Identify your destination

Data volume grows by the second. To make sense of it all, organizations need a unified data repository to bring all the data under one roof, ensuring easy access, management and security.

- **Ensure your destination can handle the types of data required.** This will dictate whether you opt for a data warehouse or (more likely) a data lake. If you plan to use unstructured data, a data lake is a better choice.
- **Consider scalability, cost, integration with other elements of the data ecosystem**, control over residency and more. Data warehouses can be more cost-effective at smaller scales while data lakes shine at larger scales.
- **Contemplate the tradeoffs involved in using the destination for other analytics.** Decide whether you may want to use the same platform to support business intelligence or other forms of AI. If possible, you should keep your architecture as streamlined and simple as possible.

Step 4: Integrate data sources into a central repository

Organizations can only model data for AI if it's in one central location. By automatically moving data from various sources into a unified platform, organizations can provide AI models with the rich, diverse data they need to thrive.

- **Integrate the data.** Using a tool like Fivetran, set up connectors to each data source that you will use. Look for capabilities such as change data capture to ensure efficient, real-time data syncs.
- **Ensure the security of your data while in motion.** Make sure your data integration solution features end-to-end encryption, can block and hash sensitive data, offers the deployment options you need and is compliant with relevant regulations.
- **Transform your data.** Data may require a great deal of preprocessing before it’s ready for generative AI. You may, for instance, need to concatenate text fields together.

Step 5: Design and implement retrieval models

A method called retrieval-augmented generation (RAG) adds context to prompts using a vector database and then consults the foundation model. RAG boosts the accuracy and specificity of the AI's outputs, ensuring better results while using resources more effectively.

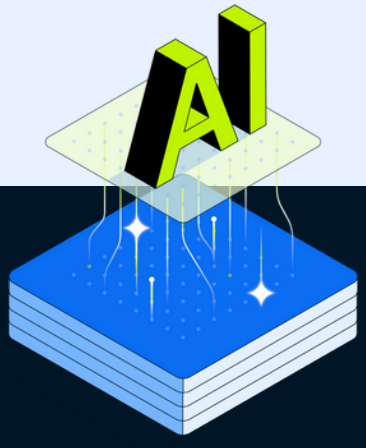
- **Build a pipeline to a vector database**. Your pipeline must convert raw text or other media stored in your central repository into numerical representations that are legible to the foundation model. Choose a vector database that is compatible with your foundation model and easy to use.
- **Build a retrieval model.** The retrieval model must provide a user interface that can accept prompts from users, augment them with context from the vector database, send the augmented prompt to the foundation model and receive outputs.
- **Consider using a knowledge graph to hard-code semantic information.** Knowledge graphs improve the relevance and accuracy of GenAI outputs.

Step 6: Iterate and refine your model

It's crucial to rigorously test models for glitches or errors. By iterating and refining based on real-world feedback, organizations can ensure models are reliable and effective — while also safeguarding sensitive information.

- **Check the factuality of outputs.** Make sure your retrieval model produces accurate and factual results. It helps if your prompts contain helpful context, i.e. prompt engineering.
- **Ensure sensitive information is not exposed.** You should avoid exposing details such as trade secrets, personally identifiable information and other information. You can achieve this through upstream governance and security practices, particularly blocking or hashing sensitive data.
- **Keep iterating.** As with all pieces of software, you are likely to encounter unforeseen issues over the course of normal usage. Building, deploying and maintaining a retrieval model deserves sustained expert attention.

> “By aligning data initiatives with business strategy, integrating diverse data sources and meticulously testing models, your organization can pave the way for transformative innovation. A successful GenAI project can help organizations explore exciting new areas of efficiency, effectiveness and business value.”
>
> Scott Barnes, VP of Insights & Data, Capgemini



Start building a data foundation for AI today.

Start a trial with Fivetran