A QUICK START GUIDE TO DATA READINESS FOR GENERATIVE AI ON AWS

clearscale

Contents

Introduction

GenerativeAI (GenAI) has proliferated in the last few years. GenAI technology gives users the ability to quickly create high-quality content – images, audio, videos, text, etc. – with simple commands delivered through user-friendly interfaces. With platforms like ChatGPT and DALL-E, hobbyists and enterprise users alike can now harness the power of AI/ML for productive and casual use cases. What's more, these tools offer low-cost pricing, which is a big reason why their popularity has skyrocketed.

Back in October 2022, DALL-E was generating over 2 million images every day. And in record-breaking fashion, ChatGPT garnered more than 100 million active users in two months after launching in November 2022. GenAI technology has reached a tipping point, and there's no turning back. The question now is what use cases will emerge and survive going forward.

For obvious reasons, organizations in every industry are interested in leveraging GenAI. The technology opens up exciting opportunities to drive growth, improve efficiency, and reduce costs. The reality is, however, that few are ready to capitalize. GenAI – and AI/ML technology more broadly – may be more accessible than ever, but that doesn't guarantee sustainable, positive outcomes. Tools like ChatGPT and DALL-E have given business leaders a false sense of confidence and distracted their teams from what's more important.

Before investing heavily in GenAI, enterprises first need three things:

1. **A comprehensive data strategy**
2. **A robust data management foundation**
3. **A complete understanding of exactly how and where GenAI can provide value**

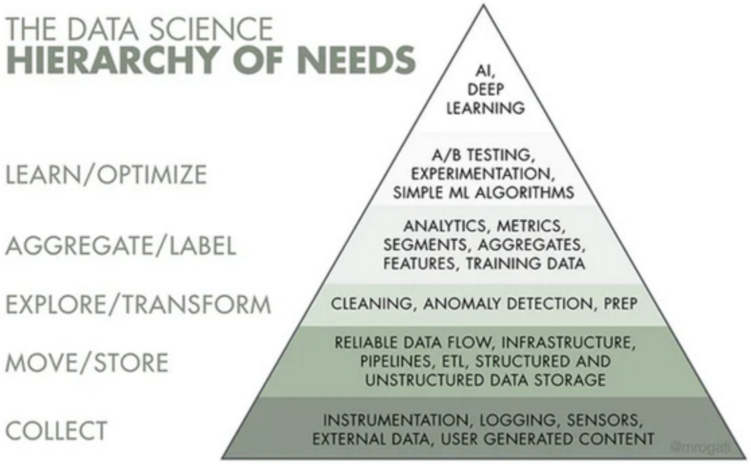
Data Science Hierarchy of Needs

All three are crucial for long-term success with GenAI. Too many companies are jumping ahead and experimenting with GenAI before establishing sound data management practices. As a result, data scientists, machine learning engineers, AI developers, and others at the top of The Data Science Hierarchy of Needs don't have what they need to be effective. Or, in some cases, they have to take on more responsibility on the data management side, preventing them from pursuing what they were hired to do in the first place.

GenAI falls in the topmost layer where AI and deep learning applications live. So, organizations should be able to collect, move, store, explore, transform, aggregate, label, learn, and optimize data before diving into more complex AI and deep learning applications.

Yet, what we're seeing today is organizations hiring data scientists for the top three layers of the pyramid without having the bottom three figured out. Consequently, data scientists have to set up the basic data ingesting, moving, transforming, and preparing architecture before they can apply their skills in a specific domain. This is not the recipe for GenAI success.

In The Data Science Hierarchy of Needs, creator Monica Rogati divides data science into six distinct layers:


THE DATA SCIENCE
HIERARCHY OF NEEDS
AI, DEEP LEARNING
LEARN/OPTIMIZE
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
AGGREGATE/LABEL
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
EXPLORE/TRANSFORM
CLEANING, ANOMALY DETECTION, PREP
MOVE/STORE
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
COLLECT
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

At ClearScale, we believe that having a comprehensive data strategy and data foundation comes from maximizing performance across the following areas in a cloud environment like Amazon Web Services (AWS):

- Data architecture
- Data engineering
- Data science

We dive deeper into each of these areas below and provide questions to help you reflect on your own organization's data readiness for GenAI on the cloud.

Data Architecture

Data architecture encompasses everything needed to get data into and around the cloud. Therefore, data architecture includes components like data ingestion pipelines and database migration tools that can transfer data from a multitude of sources, such as IoT devices and third-party applications, to secure and scalable storage.

Data architecture also includes things like data lakes and data warehouses for unifying data, as well as environments for initial data exploration. Before any data engineering takes place, the right data architecture must exist to handle big data volumes and velocities without compromising performance on any front.

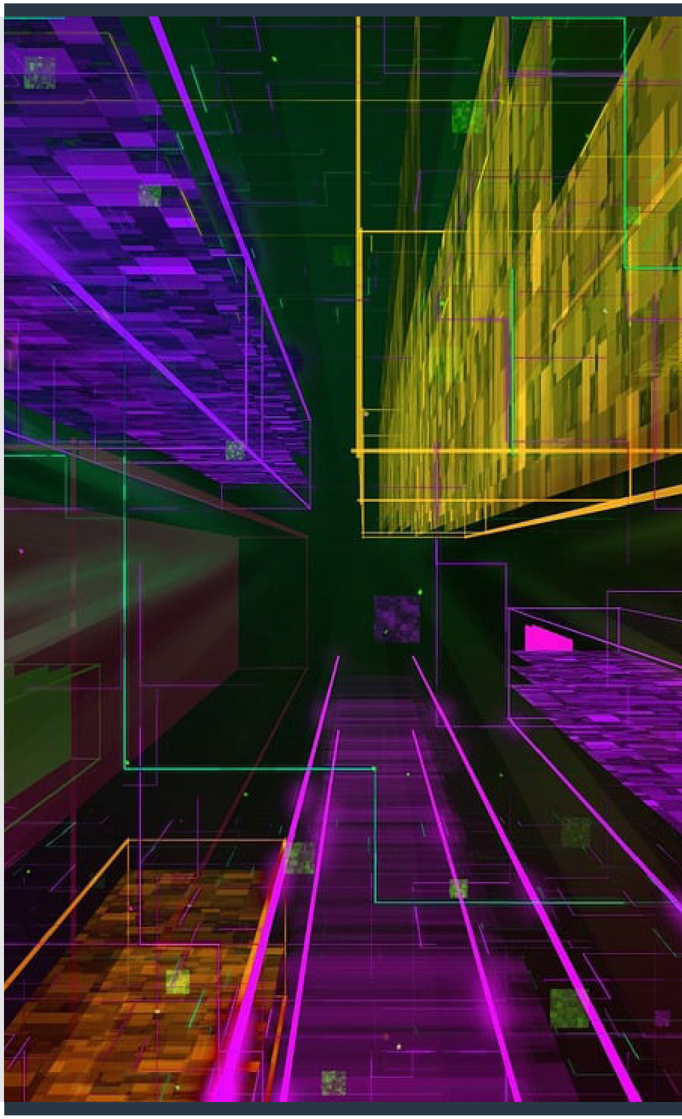
What Data Architecture Solutions Does AWS Provide?

On AWS, solutions architects have access to many services that make it easy to set up and manage data architecture. Many of these services are managed or serverless offerings, which means AWS takes care of all the behind-the-scenes infrastructure work that would otherwise require significant time and effort.

Some of AWS' most popular data architecture solutions include:

- **Amazon Aurora** for storing relational data in a cloud-native, managed database
- **Amazon DynamoDB** for storing key-value data in a cloud-native, managed database
- **Amazon S3** for storing objects in highly available and scalable storage
- **AWS Lake Formation** for setting up new data lakes on AWS quickly
- **Amazon Database Migration Service** for migrating data to and around the cloud

With respect to The Data Science Hierarchy of Needs framework, data architecture represents the first two, and sometimes three, levels of the pyramid. Having performant and reliable data architecture is what sets the stage for efficient data engineering.

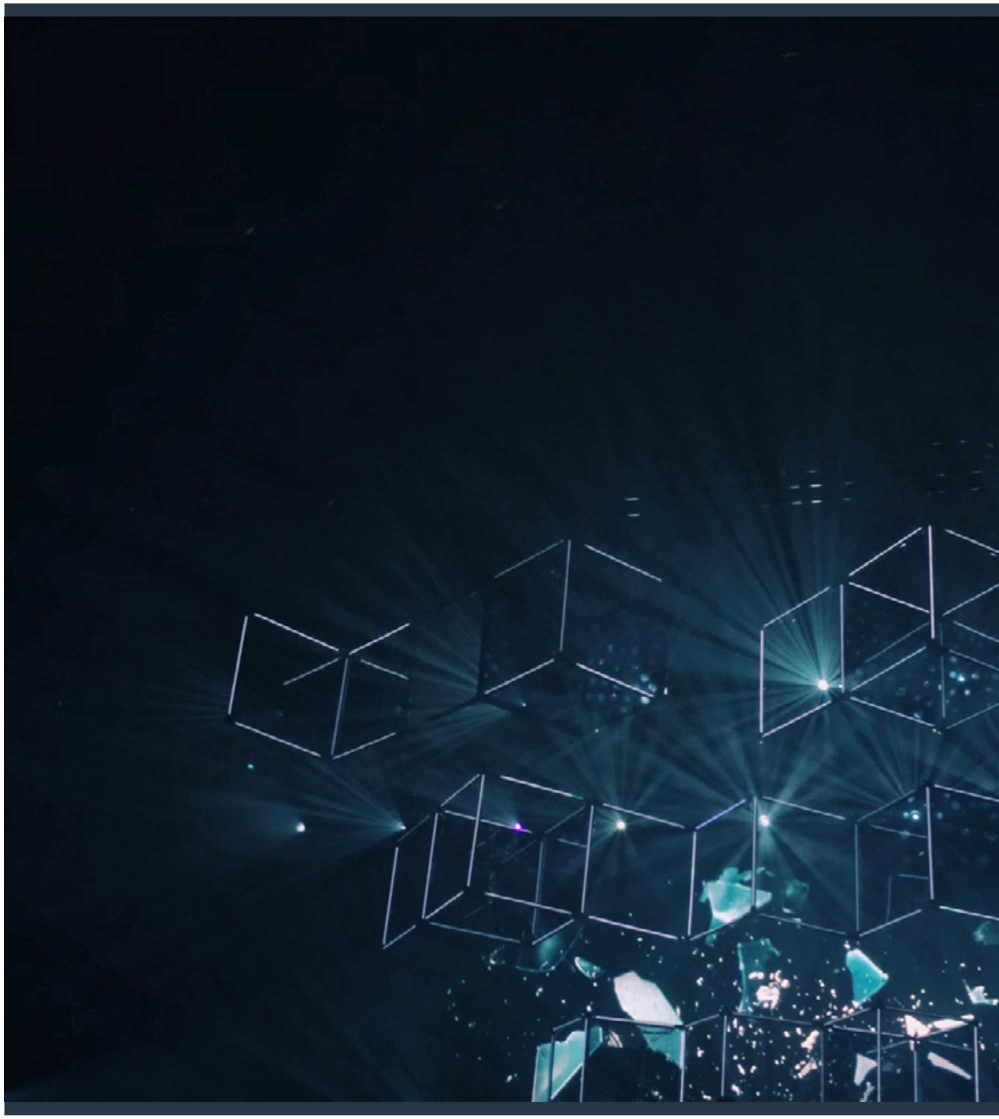


Data Architecture

Data Architecture Self-assessment

Use the following questions to assess your organization’s data architecture:

- Can you ingest data from many different sources into cloud-based architecture?
- Can you ingest any volume of data coming in at any velocity?
- Can you ingest data in real-time?
- Can you store unstructured, semi-structured, and structured data effectively?
- Is your data storage scalable, available, and secure?
- Do your applications and data engineering workflows integrate seamlessly with your data architecture?
- Can you implement fine-grained access controls for your data architecture?
- Does your data architecture enable seamless regulatory compliance and governance?



Data Engineering

Data engineering describes the processing work that happens to data in the cloud so that it's ready for complex use cases. Data engineering may involve data cleaning, categorizing, labeling, and segmenting. The goal of these activities is to improve data quality and subsequent analytical results.

Furthermore, data engineering is where data visualizations and advanced extract, transform, and load (ETL) workflows come into play so that data is prepped for data scientists and machine learning engineers. It's at this point that companies can start to think about how to extract deeper value from their data.

What Data Engineering Solutions Does AWS Provide?

AWS offers a wide range of data engineering services that empower data teams to perform highly customized and automated processing on massive volumes of information. Some of AWS' most popular data engineering solutions include:

- **Amazon Athena** for analyzing data stored in S3 buckets
- **Amazon Kinesis** for ingesting and processing streaming data
- **Amazon QuickSight** for visualizing data in easy-to-understand dashboards
- **AWS Glue** for extracting, transforming, and loading data for downstream analytics
- **Amazon Redshift** for running complex queries on large volumes of data

Again, these services are available as managed offerings with pay-as-you-go pricing. IT teams don't have to worry about things like provisioning or patching servers, so individuals can focus more on the truly differentiated work.

Data Engineering

Data Engineering Self-assessment

Use the following questions to assess your organization’s data engineering capabilities:

- Can you clean, label, segment, and categorize your data efficiently?
- Can you track the entire history of your data, from collection through processing?
- Can you visualize your data in easy-to-understand dashboards?
- Do your data visualizations update in real-time?
- Can you create custom ETL workflows?
- Does your ETL process integrate seamlessly with your data architecture?
- Can you process streaming data at any volume?
- Do you have an organized process for transforming data in stages?
- Is your data secure in transit and at rest while getting processed?
- Can you enhance your data efficiently with third-party information?

Data Science

Data science refers to the practice of studying information within a specific context using software development and statistics together. Data science is a fast-growing field and will continue to expand as organizations bolster big data capabilities.

Work in the data science arena includes activities such as A/B testing, data experimentation, and basic machine learning. Data science effectiveness depends largely on having clean, high-quality, and trustworthy data. On AWS, data science incorporates the collaboration of Sagemaker, or Jupyter, notebooks, data modeling, fine-tuning, and regression analysis.

What Data Science Solutions Does AWS Provide?

AWS has invested significantly in its data science and AI/ML solutions. Some of the most widely utilized today include:

- **Amazon SageMaker** for creating, training, deploying, and fine-tuning ML models
- **Amazon Comprehend** for applying natural language processing technology to unstructured data and text
- **Amazon EMR** for processing massive amounts of data using open-source frameworks
- **Amazon Personalize** for deploying AI-powered personalized recommendations
- **Amazon Forecast** for predicting future outcomes based on ML algorithm outputs

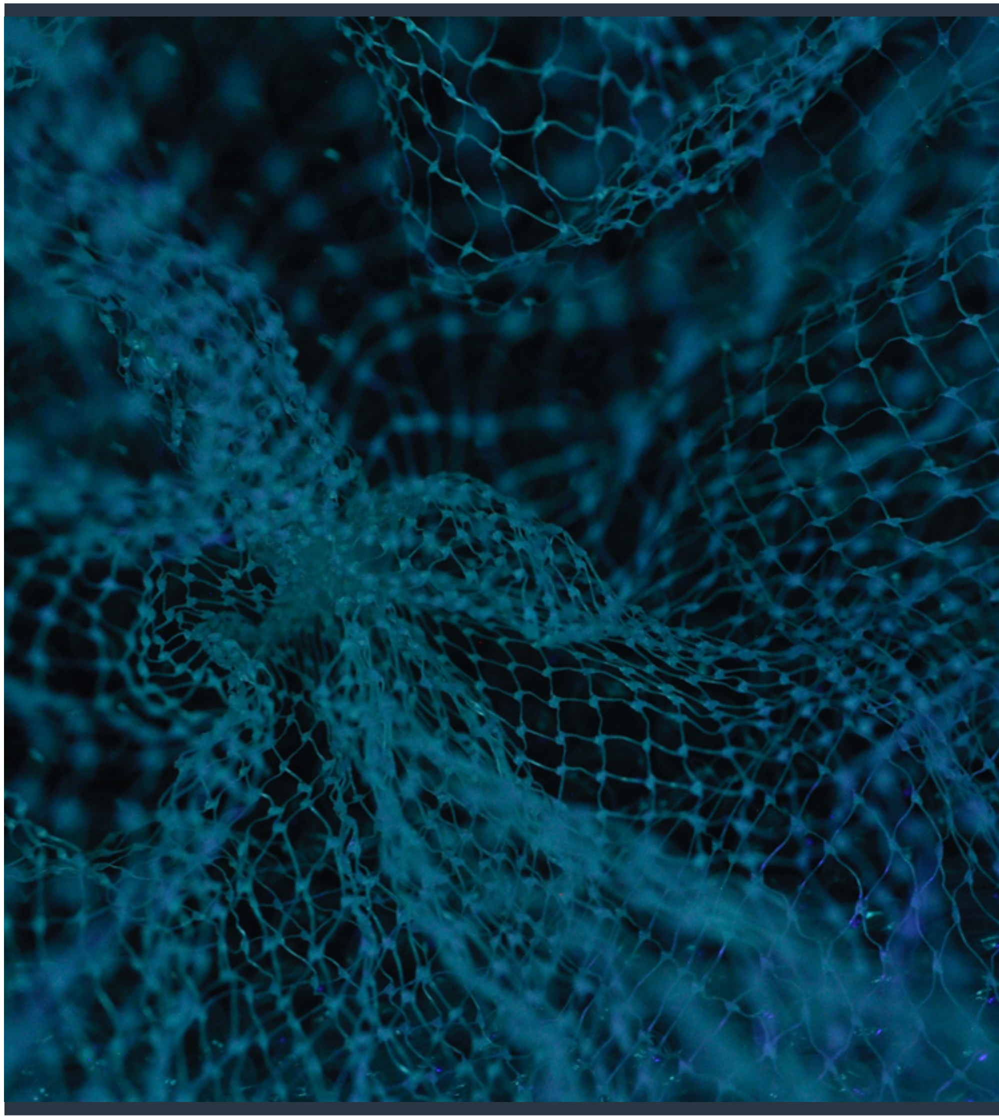
Solutions like these are valuable in many contexts and industries, which is why many organizations are starting to build their data science capabilities on the AWS platform.

Data Science

Data Science Self-assessment

Use the following questions to assess your organization's data science practice:

- Do your data scientists or ML engineers have to do any data architecture or engineering work?
- Do your data scientists and ML engineers have access to all the data they need?
- Is your data complete, high-quality, and trustworthy?
- Is your data science function skilled enough for complex AI and deep learning work?
- Does your organization have a sophisticated MLOps practice?
- Do your machine learning algorithms struggle with problems like data drift and bias?
- Is your organization able to apply ML insights quickly to products and services?
- Are you able to measure the impact of your ML insights?
- What is the ROI of your data science practice?
- Does your data science team have a clear vision of its purpose in the broader organization?



Achieving Data Readiness for GenAI With ClearScale

ClearScale is an AWS Premier Tier Services Partner with extensive experience on the AWS platform. ClearScale has earned 11 AWS competencies, including the Machine Learning and Data and Analytics competencies. These certifications demonstrate not only our technical expertise but also our ability to create solutions that deliver tangible results for our clients.

For instance, we helped the American College of Radiology (ACR) revamp its data architecture to make it easier to work with vast amounts of sensitive patient data. We created a data lake on top of Amazon S3 and introduced ACR to solutions like AWS Config and AWS Control Tower. As a result of these enhancements, ACR's data architecture has become more reliable, secure, and scalable. This foundation not only improved ACR's operations but also paved the way for more advanced data science work.

We are the ideal partner for organizations that need help setting up the foundational data management layers represented in The Data Science Hierarchy of Needs. We can position you for GenAI success on the AWS cloud and enable your team to overcome the biggest challenges associated with leveraging AI.

GenAI AppLink™

ClearScale knows that leveraging GenAI successfully is difficult, especially when trying to incorporate the technology into existing cloud environments. Many organizations struggle to create value with LLMs given the technical complexities involved.

That's why we offer GenAI AppLink – a service designed to effortlessly weave GenAI workflows into existing AWS environments. GenAI AppLink empowers companies to create a bridge that seamlessly interlinks GenAI into any application, setting the stage for long-term value creation with LLMs. Learn more at https://www.clearscale.com/services/aws-generative-ai-services

Ready to get started?

Call us at 1-800-591-0442

Send us an email at sales@clearscale.com

Fill out a Contact Form

aws PARTNER Premier Tier Services

(c) 2023 ClearScale. All rights reserved. The ClearScale trademarks are the exclusive property of ClearScale, are registered with the U.S. Patent and Trademark Office, and may be registered or pending registration in other countries. All other trademarks, service marks, images, products and brands remain the sole property of their respective holders and do not imply endorsement or sponsorship.