quantiphi
Solving What Matters
aws
PARTNER
Premier Tier Services
Unlocking
Generative AI on AWS:
How to Achieve Data Readiness

Share

AWS • June 10, 2024

Unlocking Generative AI on AWS: How to Achieve Data Readiness

What is Data Readiness?
Importance of Data Readiness: Why it Matters for Gen AI
Data Readiness Assessment: Is Your Data Gen AI-ready?
What do Data Readiness Levels (DRLs) Look Like?
Data Readiness Checklist for Gen AI
Data Readiness Journey with AWS Toolkit
Ready to Unleash the Power of Gen AI With Quantiphi
Data Readiness FAQs

Generative AI (Gen AI) is transforming industries rapidly from developing novel drug candidates to creating realistic marketing materials and generating original content or data. However, the bedrock of Gen AI capabilities heavily relies on the quality and quantity of data. Ensuring high-quality, data readiness is paramount before you unleash the power of Gen AI.

This blog post dives into the world of data readiness for Gen AI using the AWS-powered robust suite of services.

What is Data Readiness?

Data readiness signifies how well-prepared and high-quality is your organization's data to support data-rich, informed decision-making, and operating model performance.

Encompassing key elements such as data accuracy, completeness, timeliness, consistency, and accessibility, data readiness is significant to deriving invaluable insights, empowering organizations to respond to market changes swiftly, optimize operational efficacy, and grow innovatively.

Importance of Data Readiness: Why it Matters for Gen AI

Data readiness is essential for organizational success, providing them a competitive edge by leveraging data effectively and adapting to market dynamics. It plays a pivotal role in:

> **Strategic alignment of data and business**
> Aligns organizational objectives with data-driven initiatives to harmonize with business goals and drive growth.

> **Breaking down data silos**
> Eliminates data silos to develop a unified data ecosystem, fostering collaboration, and providing a holistic view of operations.

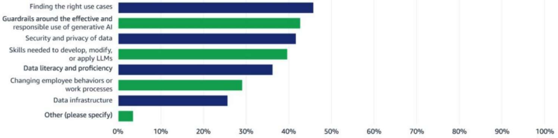
> **Validate data sources**
> Ensures accurate and reliable data by validating the sources, preventing costly errors in decision-making.

> **Optimal selection of data tools**
> Enables you to choose the right tools to manage platforms, streamline operations, and maximize data value.

> **Empowering data-driven decision-making**
> Prepares organizations for data-driven decision-making, fosters a culture of informed choices, and provides proactive insights to navigate complexities and seize opportunities.

Data readiness has become business-critical in today's business landscape, especially for Gen AI models that massively rely on high-quality data. A recent survey highlighted that many organizations struggle with ***accessing quality data and*** identifying the right use cases.

What is the biggest challenge for your organization in realizing the potential of generative AI?
Data quality



Source

The infographic above portrays multiple challenges organizations face in leveraging Gen AI effectively. Considering data as the fuel for your Gen AI engine—dirty or insufficient fuel will lead to a sputtered performance or even engine failure.

Data readiness addresses these challenges by ensuring data quality and quantity, directly impacting the Gen AI model's outputs.

> **Improved Model Performance:** Clean, high-quality data empowers Gen AI models to learn more effectively, leading to more accurate and reliable outputs.

> **Reduced Training Time:** Well-prepared data reduces the time needed to train models, saving valuable resources.

> **Enhanced Generalizability:** Gen AI-ready data ensures models can handle unseen data points, leading to broader applicability.

Data Readiness Assessment: Is Your Data Gen AI-ready?

An organization's data management strategy involves a structured process to assess and enhance data quality. Here's how:

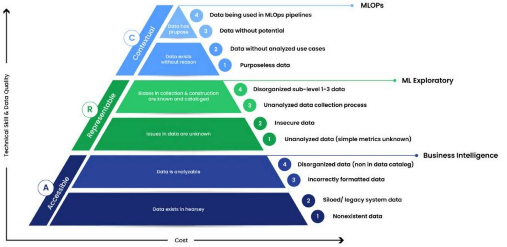
1. **Assessment of Data Accuracy and Reliability:**
   Identify and correct errors, inconsistencies, and discrepancies to ensure data reflects real-world situations.
2. **Evaluation of Data Completeness**:
   Check for gaps or missing data points to avoid flawed analyses and ineffective decision-making.
3. **Review of Data Timeliness:**
   Ensure data is regularly updated and stored to support informed decision-making and operational needs.
4. **Ensuring Data Consistency**:
   Maintain uniformity in data definitions, formats, and standards across various sources and systems.
5. **Assessing Data Accessibility**:
   Make data easily accessible to stakeholders while implementing access controls to safeguard sensitive data and comply with regulations.
6. **Implementing Data Quality Controls**:
   Establish robust quality controls and validation procedures, including automated checks to uphold ongoing data accuracy and reliability.
7. **Strengthening Data Infrastructure**:
   Invest in scalable infrastructure to support seamless data collection, storage, retrieval, and analysis.
8. **Promoting Training and Education**:
   Educate employees and stakeholders on data readiness and introduce best practices to foster a data-driven decision-making culture.

What do Data Readiness Levels (DRLs) Look Like?

Data Readiness Levels (DRLs) are a structural approach to address poor data preparation and representation. The primary goals of DRLs are to enhance the understandability and generality of data at each stage.

To determine the DRL, **data readiness assessment** is performed at different stages of any workflow, or project, for instance, from the onset of a project to subsequent stages to its completion.

With the focus on improving generable data at each stage, DRLs are best visualized as a pyramid, with each letter signifying a new level of data readiness: **Accessibility, Representability, and Context.**


MLOPs
ML Exploratory
Business Intelligence
Data has prupose
Data exists without reason
Biases in collection & construction are known and cataloged
Issues in data are unknown
Data is analyzable
Data exists in hearsey
C
Contextual
R
Representable
A
Accessible
4 Data being used in MLOps pipelines
3 Data without potential
2 Data without analyzed use cases
1 Purposeless data
4 Disorganized sub-level 1-3 data
3 Unanalyzed data collection process
2 Insecure data
1 Unanalyzed data (simple metrics unknown)
4 Disorganized data (non in data catalog)
3 Incorrectly formatted data
2 Siloed/ legacy system data
1 Nonexistent data
Technical Skill & Data Quality
Cost

Source

Each level can be subdivided into more specific targets, and sub-levels as pictured in the below diagram. The objective here is to progress from base level A1 (Accessibility) to top C4 (context) level, enriching data as necessary at each level.

Data Readiness Checklist for Gen AI

The data readiness checklist serves as a framework offering a structured approach to evaluate your data preparedness and unlock its full value with the following steps:

S.no	Factor	Checklist
1	Data Governance	Are data policies communicated and enforced? Is ownership clearly defined, with retention policies in place?
2	Data Architecture	Is there a scalable data architecture meeting present and future needs? Are processing and analytics capabilities aligned with requirements?
3	Data Quality	Are validation processes robust and timely? Is data quality promptly addressed?
4	Data Security and Privacy	Are policies safeguarding data integrity and privacy? Are backup processes minimizing data loss and downtime?
5	Data Analytics and Visualization	Are tools and techniques available for deriving insights? Are dashboards used to monitor KPIs effectively?
6	Data Literacy and Skill	Are employees equipped with data analysis skills? Do data professionals have the necessary expertise?

Data Readiness Journey with AWS Toolkit

Data preparation for Gen AI encompasses five key steps, however, all of which can be addressed using AWS services:

1. **Data Acquisition:**
   How you gather your data is the first step using AWS Glue and Amazon S3.
   - **AWS Glue:** This serverless data integration service crawls and extracts data from various sources, both on-premises and in the cloud. It supports many data sources including relational databases, data warehouses, and even social media APIs. AWS Glue can also schedule crawls to automate data retrieval at regular intervals.
   - **Amazon S3:** It securely stores large datasets of any format in a scalable and cost-effective manner. Amazon S3 offers various storage classes to optimize costs based on your data access frequency.
2. **Data Cleaning and Transformation:**
   Real-world data could be better, how? Services like Amazon SageMaker Data Wrangler and AWS Glue DataBrew provide visual interfaces to clean and transform data efficiently.
   - **Amazon SageMaker Data Wrangler:** This provides a visual interface to data scientists to clean, transform, and explore data efficiently. Additionally, Data Wrangler provides a no-code environment with pre-built workflows for common data manipulation tasks like handling missing values, normalization, and feature engineering.

> **AWS Glue DataBrew:** It is a serverless tool for interactive data preparation with built-in machine learning capabilities. In addition, It offers a visual interface similar to Data Wrangler but with additional features like anomaly detection and data quality checks powered by machine learning algorithms. This allows for the more efficient identification and address of potential issues in your data.

3. **Data Labeling (if required):**
   Supervised learning Gen AI models often require labeled data that can be performed using Amazon SageMaker Ground Truth.
   > **Amazon SageMaker Ground Truth:** It enables building high-quality labeled datasets for your supervised learning Gen AI models using a human-in-the-loop labeling workforce. Ground Truth simplifies the labeling process by providing tools for managing labeling tasks, assigning labels to data points, and evaluating label quality.
4. **Data Augmentation:**
   Machine learning, including Gen AI, often benefits from data augmentation techniques. These techniques involve creating new synthetic data points from existing data to improve model generalizability. While various libraries exist for data augmentation, a well-prepared data foundation on AWS allows for easier implementation of these techniques.
5. **Data Curation and Governance:**
   A secure data lake environment is essential for Gen AI model training using AWS Lake Formation and AWS Glue Catalog
   > **AWS Lake Formation:** It creates a secure data lake to manage access and govern your data effectively by automatically setting up the necessary infrastructure and enforcing security policies. Lake Formation also provides a central catalog for your data assets, making them discoverable and usable across your organization.

   > **AWS Glue Catalog:** It provides a central registry for your data assets, ensuring discoverability and consistency. Glue Catalog allows you to define metadata for your data, such as schema information and ownership, making it easier for data scientists and analysts to find and understand the data they need.

Ready to Unleash the Power of Gen AI With Quantiphi

By leveraging Quantiphi's AWS services, you can build a robust data pipeline that ensures your data is clean, consistent, and ready to fuel your Gen AI projects.

Remember, data readiness is an ongoing process. You must conduct regular **data readiness assessments** using tools to identify data quality issues and utilize **data readiness checklists** to ensure you've addressed all critical steps. As your Gen AI initiatives evolve, you must continuously monitor and improve your data for optimal results.

Data Readiness FAQs

1. **What is AI-ready data?**
   AI-ready data refers to the process of making your data ready for generative AI. This should be well-governed, secured, unbiased, accurate, and of high quality. For generative AI models, your data must have the following characteristics
   > **Understandable with the right context:** Data should be clearly defined and easily interpretable.

   > **Of high quality**: Data must be precise, complete, consistent, and unique.

   > **Well-governed:** Data should be appropriate to ethical and compliant use.

   > **Ease to accessible:** Data should be easy to locate and retrieve when needed.
2. **What is a data readiness checklist?**
   A data readiness checklist outlines best practices to check the accessibility, quality, governance, timelines, and consistency of data. This checklist empowers organizations to better understand their data and pinpoints areas for enhancement. This enables customers to leverage data more effectively, advance decision-making, and gain a competitive edge in the ever-emerging data-driven landscape.
3. **What are the benefits of AI data readiness?**
   Here are the benefits of data readiness:
   > **Enhanced Accuracy:** Data readiness ensures data is accurate, clean, and relevant helping in precise AI predictions.

   > **Improved Performance:** Well-prepared data optimizes AI algorithms, boosting overall performance.

   > **Increased Efficiency:** Data preparedness streamlines model training processes, and reduces time and resources required for AI model deployment.

   > **Better Insights:** Organized data uncovers hidden patterns, deriving actionable insights for informed decision-making.

   > **Enhanced Personalization:** Cleaned data enables tailored experiences, improving customer satisfaction.

   > **Reduced Bias and Fairness:** High-quality, data preparation mitigates biases, promoting fairness in all AI-driven decisions.

   > **Regulatory Compliance:** Prepared data is complaint-ready, adhering to data privacy regulations, and fostering trust with stakeholders.

Written by

Sanchit Jain