EBOOK

The Big Book of AI-Ready Data

data.world

Table of contents

INTRODUCTION:

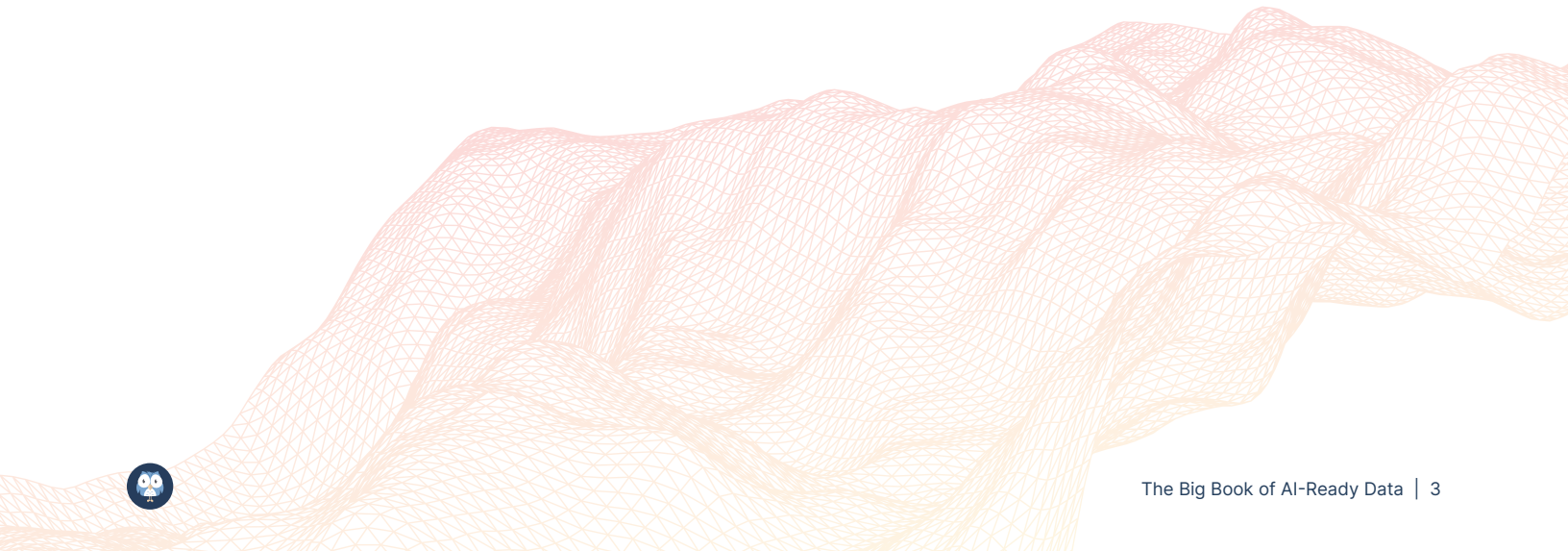
Understanding the role of AI-ready data

In the world of data, the emergence of Artificial Intelligence has signaled a paradigm shift. How do we harness and interpret our vast oceans of data?

AI's potential to transform business and create new opportunities is immense, but it hinges on one critical factor: the readiness of the data we feed into these intelligent systems. This is where the concept of AI-ready data becomes pivotal.

Your goal with data is to meticulously fuel AI systems that can make or break the future of your business. The consequences of feeding poor-quality data into AI models are not just inconvenient. They are potentially catastrophic, steering the entire ship of your organization off-course into dangerous waters where regulatory penalties, incorrect customer insights, and flawed strategies lurk.

In this book, we delve deep into the realm of AI-ready data. We'll explore the challenges that AI-ready data can tackle, the benefits of AI-ready data, and how you can get started enabling it. The demands of AI systems are unique and unprecedented. But you're ready for it – as long as your data is, too.



CHALLENGE:

Poor data quality

Poor data quality is like having a sports car but no fuel. You need quality data to drive your AI solutions to their maximum potential.

Studies have highlighted that up to 96% of enterprises face data challenges, leading to misinformed decisions that could potentially cost businesses dearly.

If the data is not representative of the diverse conditions and populations it's meant to serve, the resulting AI models might perpetuate or even amplify these biases. This is particularly critical in fields like healthcare, finance, and law enforcement, where biased AI has serious real-world consequences, from misdiagnosed illness to predictive policing to unfair job screening processes and beyond.

AI-ready data is meticulously cleaned, processed, and structured to ensure the highest quality. It undergoes rigorous validation processes, like:

- Format checks
- Range checks
- Referential integrity
- Name validation
- Uniqueness checks
- Consistency testing
- Issue tracking systems
- Security testing

Without high-quality, accurate data, businesses can't depend on the performance of their AI models. This is a red flag waving at us to pay attention to the quality of data we feed into our AI systems.

CASE STUDY:

Penguin Random House

Penguin Random House needed a card-catalog-like approach to pulling in and discovering data with data.world. They wanted a single reference for the location of any data set, and a single source of truth for what that data means.

"Imagine managing books without title information, author data, cover images, royalties, or number of chapters. That's what it's like managing data without a catalog. Now, information that once took our data scientists a week to find is discoverable in seconds on data.world."

Rupal Sumaria
Head of data governance
Penguin Random House UK

Learn more →

CHALLENGE:

Scalability with growing data sets

Today, the digital universe is expanding, with an estimated 2.5 quintillion bytes of data being created every day.

The IDC's Global DataSphere Forecast predicts a staggering compound annual growth rate (CAGR) of 23% in global data creation and replication, expecting it to leap to 181 zettabytes by 2025, a significant increase from 64.2 zettabytes in 2020. But the explosive growth in data volume presents significant scalability challenges.

Confronted with such immense volumes of data, traditional IT infrastructures often struggle to cope. The scalability challenge isn't just about handling larger data volumes; it's about efficiently processing them and extracting value. AI systems require a seamless flow of data – from collection and storage to processing and analysis.

CASE STUDY:
WPP

WPP is the largest advertising firm on the planet, a tough title to hold in a world that spends more than $800 billion annually on ads.

They wanted to develop themselves into the world's most advanced and respected creative tech company. The aim was to harness all of the company's assets in a connected way, build stronger collaboration across the enterprise, and help clients win with data.

Through this approach, they hoped to strongly differentiate itself from its competitors in the ad space. To accomplish this goal, WPP customers required a solution that could overcome data silos and foster a genuine data-led culture.

Learn more →

CHALLENGE:

Integrating diverse data sources

Integrating diverse data sources for AI applications is like assembling a complex jigsaw puzzle where each piece comes from a different set. More often than not, data derived from various sources – whether it's structured data from databases or unstructured data from social media – lacks uniformity.

Deloitte's State of AI in the Enterprise survey found that at least 40% of AI-adopting organizations reported low-to-medium sophistication in critical data practices, including data integration.

Mismatched data distorts the AI model's learning process. Inconsistencies in range, format, or timeframe introduce errors directly into the AI system. And that inaccuracy quite easily snowballs into incorrect decision-making.

Scale only compounds the problem. When the volume of data sources swells, integration becomes next-to impossible. For example, Stuart Wagner, Chief Digital Transformation Officer at the US Department of the Air Force, highlighted a phenomenon known as "up classing" or "deanonymization," where combining data inadvertently creates personally identifiable information (PII) that wasn't visible before. There's a delicate balancing act in taking meaningful data from all available datastreams, without overwhelming your AI model.

CASE STUDY:

Power Digital

In late 2022, Michael Murray became Chief Product Officer at Power Digital. His team developed a blueprint for "data activation" — contextualizing data for clear, actionable business use. They leveraged a Large Language Model that could answer complex questions in real-time – bringing data fluidity to the end user.

For Murray and Power Digital, data.world's Knowledge Graph architecture provided the context needed to maximize accuracy from the LLM.

Learn more →

CHALLENGE:

Data privacy and compliance

As AI systems increasingly incorporate sensitive data, privacy and compliance is not just a regulatory requirement. It's a fundamental ethical responsibility.

There are varying and ever-evolving regulations across different regions, like the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA). These regulations impose strict guidelines on data usage, consent, and an individual's rights over their data.

AI systems can accidentally reveal insights about individuals that were not explicit in the original data. That can put an organization into regulatory non-compliance and erode public trust. And cleaning data is a rigorous process. Deloitte found that cleaning inaccurate data to prepare it for feeding into an AI model routinely takes between six and 12 months.

Compliance in AI is a continuous and evolving process, moving just as quickly as AI innovation itself. Models learn, and they evolve. Monitoring and auditing must follow in step. Organizations need a robust data governance framework that includes clarity around policies, audits, and AI operations.

CASE STUDY:

OneWeb

OneWeb is a communications company building a 700-satellite constellation to provide global satellite Internet broadband services to people everywhere.

They wanted to give their satellite engineers access to data from domains distributed around the world, and empower them to discover the data they needed in a self-service infrastructure, while remaining governed and in compliance with regulatory standards.

Learn more →

CHALLENGE:

Time to insight

"Time-to-insight" represents a critical metric on how efficiently an AI system can process data and deliver actionable insights.

The quicker and more accurately it's done, the more valuable the information becomes. In today's fast-paced business environment, the ability to rapidly derive insights from data sets an organization apart from competitors. But the journey from raw data to meaningful insight involves many steps.

Data cleaning and prep alone can consume a majority of the total time spent on data projects; a bottleneck that significantly lengthens time-to-insight. The more complex an AI model, the more processing time is required, especially when incorporating unstructured data like images, videos, and natural language.

Organizational and infrastructural factors also impede time-to-insight. Many companies lack the necessary infrastructure to efficiently process large datasets. They might rely on inefficient legacy systems or face other challenges by data being compartmentalized across different departments. Essentially, wherever the flow of information is hampered, time-to-insight suffers.

CASE STUDY:

Vopak

Vopak, an independent tank storage company, wanted a clearer understanding of their data assets and associated metadata. They wanted to achieve this state without unproductive manual steps that wasted time and resources.

"Data.world is part of the glue that gets data to work for Vopak."

Roel Pot
Data architect, Vopak

Learn more →

For AI systems, those steps include some or all of the following:

Analysis → Interpretation → Insight application and usage → Data collection → Cleaning → Integration

The benefits of AI-ready data

AI-ready data is the secret sauce behind your predictive analytics and machine learning models. AI-ready data gets continually better at making spot-on predictions.

And it doesn't stop at making smarter guesses. It catalyzes organizations to stay ahead of the game and know which moves to make before everyone else does. Top-notch data is a goldmine for innovation, helping you develop groundbreaking ideas that set you apart from the crowd.

Teams with AI-ready data spend less time trying to make sense of messy, disorganized information. Instead, teams have a toolbox where everything they need is readily at their fingertips, including the ability to stay compliant. This efficiency is not just about cutting costs; it's about getting work done faster and smarter.

With AI-ready data, you enable more personalized customer insights. AI-ready data is the next-best thing to having a one-on-one chat with each customer, understanding what they want, and personalizing their experience.

The bottom line: AI-ready data makes everything run smoother.

The three pillars of AI-ready data:

Accurate

Is data correct, reliable, and free from errors?

Explainable

Can the processes and decisions made by AI systems based on this data be explained in human terms?

Governed

Is data managed and regulated according to clear policies and standards (including ensuring privacy, security, and compliance with legal and ethical standards)?

How AI-ready data works

AI-ready data means that data across your entire data stack meets the three pillars: it is accurate, explainable, and governed. Data catalogs can also provide end-to-end AI-readiness.

A **data catalog** serves as a comprehensive inventory of your organization's data assets, providing detailed metadata and context for each dataset. It's like having a meticulously organized library where each book (data asset) is easily discoverable, well-described, and readily accessible.

For AI-ready data, this means not just storing the data but enriching it with contextual information, like where it came from, its quality, and how it can be used. The catalog helps in identifying and organizing AI-relevant data, ensuring that data scientists and analysts can quickly find the right datasets for their AI models.

Data catalogs also further extend into improved data governance and accessibility, so that data is compliant with relevant regulations and policies.

Transforming raw data into AI-ready data:

01 Initially, raw data is collected from various sources, which might be heterogeneous and in disparate formats.

02 This data is then processed – cleaned, normalized, and transformed – to ensure consistency and reliability.

03 A data catalog aids in this process by providing tools for data discovery and lineage, allowing users to track the transformation journey of each data piece.

LLMs and the data catalog platform

Learn more →

Data catalog: Your launchpad for AI-ready data

Adopting a modern data catalog is the first step to creating an AI-ready foundation for activities like self-service analytics, digital transformation and governance, and reducing time to insight by a factor of multitudes.

A data catalog isn't simply an inventory, glossary, or dictionary of your data. It is an active platform where your team can work with data and understand where it's coming from, how to manage it, and who is working with it.

Here's how a data catalog serves as your launchpad for AI-ready data:

Data discovery: A more intuitive search experience (with recommendations, social proof, and filtering) means better information is fed into your ML models.

Data governance: Increased control and the capacity for access-based roles, classifications, and asset types means your data stays compliant with evolving regulations around AI.

Data lineage: Like GPS provides a visual overview of an entire route from start to finish, data lineage provides a start-to-finish overview of data's journey through the lifecycle, so that the data feeding models and algorithms is properly understood.

Integrations: Across the data stack, integrations can bring all key data tools and sources into the ML and AI ecosystem, to make metadata more useful and visible.

Collaboration: A data catalog serves as a foundation for the AI roadmap, as a workspace for the team's daily workflows and access requests.

Cross-functional alignment: A data catalog ensures that data initiatives and AI initiatives line up to support each other, and that there's a shared understanding of the definitions, usages, owner details, and other context to data assets.

How to talk to your boss about the business value of data

Learn more →

AI-ready data and the Knowledge Graph difference

LLMs will surface false information backed by fabricated citations as fact – also known as "hallucinations."

This phenomenon led McKinsey to cite "inaccuracy" as the top risk associated with generative AI.

But new research shows that a Knowledge Graph can have a positive effect on LLM response accuracy in the enterprise - specifically, a triple boost! The research compares LLM-generated answers to answers backed by a Knowledge Graph, via data stored in a SQL database.

3x

improved LLM response accuracy with a Knowledge Graph

0%

returned accurate responses without a Knowledge Graph

Specifically, top-line findings include:

01 A Knowledge Graph improves LLM response accuracy by 3x across 43 business questions.

02 LLMs – without the support of a Knowledge Graph – fail to accurately answer "schema-intensive" questions (questions often focused on metrics & KPIs and strategic planning). LLMs returned accurate responses 0% of the time.

03 A Knowledge Graph significantly improves the accuracy of LLM responses – even schema-intensive questions.

Knowledge Graph benchmark

Learn more →

Assessing your AI-ready data infrastructure

Assessing your current data infrastructure is a critical step towards AI-ready data. Focus on the capacity, scalability, security, and overall efficiency of your existing data systems. Here's a specific questionnaire to guide you:

Storage

- ☐ Can your storage systems handle large volumes of structured and unstructured data?
- ☐ Can your storage scale as data volume increases?
- ☐ Map the data management process with: data warehousing, databases, and data lakes

Processing

- ☐ Can your infrastructure handle complex data processing tasks?
- ☐ List the computational resources available for data processing (include CPU and GPU capabilities)

Integration

- ☐ What external data integration tools are available?
- ☐ What internal data integration tools are available?
- ☐ On a scale of 1-10, how easy is it to consolidate data formats?
- ☐ At the end of the integration process, what are the odds that data integrity remains intact?

Security and compliance

- ☐ Which security measures are in place?
- ☐ Map how you protect sensitive data
- ☐ Check for: encryption, access controls, and audit trails
- ☐ Which specific data protection regulations are you aware of (i.e., GDPR)?

Quality

- ☐ How do you clean data?
- ☐ Validate it?
- ☐ Normalize it?
- ☐ Which tools are in place to support data quality?
- ☐ Which formal processes?

Scalability

- ☐ List measures in place for how infrastructure can adapt to changing data requirements
- ☐ List measures for scaling up resources during high demand
- ☐ List the processes for incorporating new data sources and technologies

Backup

- ☐ How reliable and capable are your backup and recovery systems?
- ☐ What is your average amount of data downtime when an incident occurs?
- ☐ How many systems failures and data loss incidents have you experienced over the past 12 months?

Performance monitoring

- ☐ Which tools and processes are in place for performance monitoring?
- ☐ What are your most common bottlenecks?
- ☐ Which metrics do you measure regularly?
- ☐ Which dashboards track your performance?
- ☐ Which tools?

The six things you can do today to move toward AI-ready data:

01 **Clean your datasets:** Remove or correct erroneous data, handle missing values, and correct inconsistencies.

02 **Automate your data processes:** Automate something you do manually today so you have more time to work on creating business value.

03 **Consolidate your data:** As your data is scattered across different systems and formats, try unifying those formats where you can.

04 **Improve data quality:** Spot check your data for accuracy, completeness, and risk for bias.

05 **Invest in data infrastructure:** Upgrade data storage and processing where necessary, so you can analyze large volumes of data.

06 **Create a culture around data:** Encourage data fluency among your team members so that they can contribute to AI-ready data.

Use cases: The impact of AI-ready data helps real teams solve real problems:

- Productivity increases
- Revenue generation
- De-duplicating information
- Generating reports
- Risk analysis
- Managing third-party information

data.world

AI-ready data: Build your AI future

Your organization will do better when everyone – not just the "data people" – can unlock organizational knowledge. With the data.world data catalog platform, your team can start the march toward AI-ready data today.

Learn more about how data.world can help you build the foundation for AI-ready data and better decisions.

Schedule a demo →