DataHubCloud

Context

The Missing Link Between Your Data Stack and AI Success

A PRACTICAL GUIDE TO DATAHUB CLOUD

DataHub

Introduction

The AI Gap No One Talks About

The AI era has arrived, and companies are investing billions to stay competitive. Yet more than 80% of AI projects never make it past the pilot phase.[1] The question isn't whether AI is powerful—it's why so many initiatives fall short.

The answer? Context.

Most AI systems are built on data foundations that lack clarity, structure, and relevance. Without the right context (metadata, lineage, relationships) AI can't make sense of the data it's fed.

In this guide, we' ll explore:

Why so many AI projects are failing
What context really means (and why it's critical for AI)
How to build an intelligent metadata foundation that powers successful AI
Real-world solutions with DataHub Cloud
How to choose between DataHub Core and DataHub Cloud

Sources:
1. Rand, 2024.

DataHub

Chapter One

The Cost of Contextless AI

AI spending is soaring—but success isn't

Businesses are pouring resources into AI, but the returns have been underwhelming. **Why?**

- Nearly 50% of organizations cite data issues as the #1 barrier to AI success.[2]
- Only 12% say their data is ready for AI. [3]
- Over 80% of AI projects fail, twice the failure rate of traditional IT initiatives.[4]

What's going wrong?

The issue isn't just about having lots of data. It's about the lack of data readiness—having the right data, in the right format, with the right context, at the right time.

Today's data ecosystems are complex. AI needs structure, meaning, and traceability to function effectively . Without this, AI becomes expensive guesswork—or worse, a liability .

Sources:
2. Harvard Business Review, 2024.
3. Precisely and Drexel University's LeBow College of Business, 2024.
4. Rand, 2024.

DataHub

Chapter Two

What "Context" Really Means in AI

Context gives meaning to data

In AI, context refers to the surrounding information that helps models understand and interpret data correctly. This includes:

- **Where** the data came from (lineage)
- **Who** owns it and who trusts it
- **How** it's been used and transformed
- **What** it represents in the real world

Asking AI to operate without context is like giving a tourist an outdated paper map and asking them to drive cross-country.

The roads have changed, but the map hasn't. There's also no way to know which roads are closed, which are safe, which lead to dead ends, or if a faster route is now available.

They're left guessing, not navigating.

DataHub

What are the kinds of context we find around us?

AI systems operate more effectively when they're grounded in different layers of context. There are three main categories:

1. Technical context

Technical context answers the "how" of data. It helps AI systems interpret information accurately by providing structural and provenance details.

It includes:

- **Data lineage:** Provenance information about where data originated, how it was collected, and all transformations it underwent
- **Schema definitions:** Structure, relationships, and constraints of the data
- **Version control:** Tracked iteration of data or models in use
- **Technical dependencies:** Required systems, services, or resources for data processing

2. Operational context

Operational context captures how data behaves within systems, tracking how it's processed, accessed, and maintained. This context helps AI assess data reliability and respond to changes over time.

It includes:

- **Runtime metrics:** Pipeline execution status (success/failure), duration, resource consumption, and error logs
- **Access patterns:** Authentication events, authorization decisions (accept/deny), query patterns, and usage frequency
- **Data SLAs:** Freshness guarantees, update frequencies, and time-to-serve requirements
- **System dependencies:** Inter-service relationships, API dependencies, and infrastructure requirements
- **Operational policies:** Retention rules, archival schedules, and data lifecycle management constraints

3. Business and social context

Business and social context provides the human layer made up of the organizational knowledge and governance that shape how data should be understood and used. It ensures AI aligns with business intent and compliance requirements.

It includes:

- **Collaborative discussions:** Critical insights in Slack channels, Teams threads, and emails where experts explain anomalies and resolve edge cases
- **Unstructured documentation:** Tribal knowledge in wikis and docs explaining why metrics exist and how to interpret them correctly
- **Business glossary:** Standard definitions linking technical assets to business concepts, metrics, KPIs, and domain terminology
- **Ownership structure:** Data stewards, team responsibilities, and who to contact for questions
- **Access control policies:** Rules defining who can view, edit, or use specific data based on roles, sensitivity, and business function
- **Compliance requirements:** Regulatory frameworks and industry standards governing how data must be handled, protected, and retained

DataHub

Chapter Three

Enterprise Context Management With DataHub Cloud

Context gives meaning to data

DataHub Cloud is an enterprise-grade AI & Data Context Platform. It's designed to bridge the gap between data chaos and AI readiness.

DataHub Cloud builds and maintains a **rich semantic model** of your data ecosystem. This contextual layer captures how data is created, transformed, and consumed across teams—mapping ownership, trust, compliance, and business meaning in a unified, queryable graph.

The result? **A reliable source of truth for humans and AI systems.** It powers role-aware discovery, consistent policy enforcement, and enables intelligent automation that adapts to real-time changes across your data landscape.

Built on proven open-source foundations and engineered for enterprise scale, DataHub Cloud offers:

- The only **event-driven architecture** for real-time context visibility and change propagation
- **Unified discovery, observability, and governance** in one integrated platform
- **100+ pre-built connectors** to your most-used data tools
- **Flexible customization and extensibility** without vendor lock-in
- **Automated policy enforcement** to drive scalable, secure governance

Unlike traditional static catalogs, DataHub Cloud is event-driven and dynamic, delivering up-to-the-minute visibility across your entire data ecosystem.

DataHub

Chapter Four

How DataHub Cloud Powers AI-Ready Data Teams

DataHub Cloud is the fully managed, enterprise-ready version of DataHub enhanced with AI-powered capabilities that simplify discovery, observability, and governance at scale. It's built to accelerate the readiness and reliability of your data and AI assets.

With DataHub Cloud, your teams can spend less time wrestling with metadata and more time delivering impact.

For data analysts, developers, and data scientists

Find, understand, and use the right data—faster.

- Get immediate answers to natural language questions about your data using the Ask DataHub chat agent
- Access data where you work using a Chrome Extension for BI tools
- Discover data your way with personalization for multiple business and technical user profiles
- Support AI models and automations with a metadata graph that keeps up with today's data volume and velocity
- Understand data provenance with table, column, and job level lineage graphs
- Use AI-generated documentation and propagation to better understand context
- Stay in the loop with subscriptions to assets, activity, and notifications

For data engineers

Deliver trustworthy, observable data pipelines at scale.

- Provide end-to-end observability with user-created data quality checks and reports
- Surface data quality results and impact analysis across all points in lineage
- Use AI Anomaly Detection for freshness, volume, and column stats
- Easily keep an eye on data quality with assertions and AI-based smart assertions
- Evaluate data contracts and quality checks on demand with API
- Get notified where you work (Slack, e-mail, and more)
- Easily manage data quality with a data health dashboard

With DataHub Cloud, every capability is built to help you scale trust, automate complexity, and enable smarter, faster decisions.

For data governance teams

Move from episodic checks to continuous, intelligent governance.

- Ensure every AI & data asset is accounted for by defining and enforcing documentation standards
- Integrate governance practices early with automated shift-left governance
- Automatically classify your data as it moves and transforms with lineage-driven compliance
- Keep tags harmonized with seamless metadata flow between DataHub and source systems
- Deliver continuous compliance monitoring with forms, impact analysis, and reporting
- Create and implement bespoke compliance approval workflows

For AI and automations

Make data actionable for AI agents and ML systems.

- Enable AI agents to query and reason over metadata with DataHub's MCP Server
- Refresh model predictions as new data becomes available via real-time updates
- Expose metadata through APIs to power autonomous workflows and automations
- Retain and recall training data versions to support reproducibility and responsible AI
- Detect data drift and anomalies before they degrade model performance
- Trace model behavior to upstream changes with end-to-end lineage

DataHub

Chapter Five

DataHub Core v. DataHub Cloud: Which Is Right for You?

DataHub began as an open source project developed at LinkedIn, designed to help modern data teams manage metadata at scale. That project, **DataHub Core**, has since grown into a thriving community of over 15,000 practitioners and developers. But as organizations move from experimentation to production-grade AI, many reach a natural inflection point: the need for more support, scalability, and enterprise-grade capabilities.

That's where **DataHub Cloud** comes in. Built on the same proven foundation as DataHub Core, DataHub Cloud adds enterprise power, security, and scale—without the overhead of managing infrastructure or building custom features in-house.

Why upgrade to DataHubCloud?

Fully managed, enterprise-ready service

DataHub Cloud is fully managed by the same engineers who maintain the open source project.

- SLA-backed 99.5% availability
- Performance - optimized implementation
- Onboarding and rollout support across multiple teams
- Dedicated customer success team to ensure adoption and impact

Why it matters: You focus on value and adoption, not infrastructure and uptime.

Advanced security and compliance

Enterprise environments have higher stakes for data privacy and compliance. DataHub Cloud includes:

- SOC II-compliant infrastructure
- Role-based and attribute-based access controls
- In- VPC Remote Execution Agent for secure metadata ingestion within your firewall

Why it matters: DataHub Cloud meets your enterprise security standards and industry regulatory requirements.

Smarter, more personalized discovery

DataHub Cloud delivers an enhanced metadata search experience with:

- AI chat agent, Ask DataHub makes discovery 10x faster
- Personalized discovery based on user role, behavior, and preferences
- Browser extensions for in-context discovery across BI tools
- No-code automations for asset classification and lineage propagation

Why it matters: Everyone, from engineers to analysts, finds what they need, faster.

Superior observability for data health

In production, observability is every thing. DataHub Cloud goes beyond core metadata with:

- Continuous data quality monitoring
- AI-based anomaly detection
- Data incident tracking and root-cause analysis
- Executive dashboards for tracking data health KPIs
- Contract management to align producers and consumers

Why it matters: Stay ahead of risk, detect issues early, and tie metrics to data reliability.

Comprehensive governance tools

DataHub Cloud empowers governance teams with a full suite of automation and collaboration capabilities, including:

- Human-assisted certification workflows
- Automated policy enforcement based on usage or sensitivity
- End-to-end glossary and ownership workflows
- Shift-left governance that embeds policy in developer workflows
- Sophisticated business glossaries for enterprise taxonomy management

Why it matters: Governance becomes integrated and invisible, not a bottleneck.

DataHub

When should you upgrade?

While DataHub Core is an excellent starting point for metadata management, DataHub Cloud is the clear choice when you need:

- Enterprise-grade uptime and performance
- Dedicated support and managed infrastructure
- Stronger compliance and security
- Cross-team onboarding and adoption at scale
- AI-enhanced discovery and automation
- Proactive data observability and advanced governance

Choosing between Core and Cloud doesn't mean choosing between open and closed. It's about deciding when your organization is ready to go beyond the foundation and start building the future of AI and data with confidence.

DataHub

Chapter Six

Real Results From Real Teams

Organizations using DataHub Cloud see measurable impact in the areas that matter most to enterprise leaders:

10X **faster data discovery** Business users find trusted data in 5 minutes, down from 50	119%+ **more AI/ML models in production** Because every model starts with governed, trusted data	58% **faster incident resolution** Teams pinpoint root cause immediately (no more manual tracing)
17% ↑ **data engineering productivity** AI chat agents make data discovery, metrics debugging, and impact assessment faster	153% **more data assets with complete metadata** Comprehensive context with 75% more data sets with mapped lineage reduces risk of compliance failure by a third or more.	20-25% **reduction in storage costs** Lorem ipsum dolor sit amet, consectetur adipiscing elit.

DPG Media reduced their Snow flake storage costs by 25% after implementing DataHub Cloud.

“We used DataHub’s MetaData Tests to identify unused or duplicate Snowflake tables across business units. Impact Analysis allowed us to safely manage the clean up process. Our cost savings are just the beginning; we still have a long way to go.”

Mathias Lavaert
Principal Data Engineer, DPG Media

DataHub

Better customer and User experiences

- More accurate and reliable AI-driven features in customer-facing apps
- Improved revenue per customer and reduced customer churn
- Personalized, role-specific user interfaces and experiences
- Self-serve data access with contextual metadata and documentation

Stronger collaboration and transparency

- Shared metadata and lineage promote a common language across teams
- Formalized data contracts improve communication between producers and consumers
- Dependency visualization helps teams understand the downstream impact of changes
- Centralized knowledge reduces reliance on tribal knowledge

Chime unifies metadata from Snowflake, Looker, Terraform, and more with DataHub Cloud, gaining end-to-end visibility across their data stack.

"DataHub serves as our guide for navigating the complexities of data cataloging and discovery."

Sherin Thomas
Staff Software Engineer, Chime

Rapid time-to-value

- Accelerated time-to-market for AI models
- Streamlined AI pilot-to-production migration
- Shorter onboarding times for new data engineers and analysts
- Faster data discovery and access through AI-powered search

Long-term strategic value

- Business-adaptive architecture that scales with your needs
- Easier integration following mergers or acquisitions
- Reduced regulatory and compliance risk through built-in controls
- Future-proofed AI and data infrastructure with open extensibility

Notion relies on DataHub Cloud to cut through data noise and empower every team with trustworthy, documented data.

"We rely on DataHub to gain insights and ensure our critical data is reliable. DataHub's managed product takes DataHub to the next level through automation and emphasis on time-to-value."

Ada Draginda
Senior Data Engineer, Notion

DataHub

Conclusion

Future-Proofing Your AI Stack

The path from investment to AI success is clearer than ever, but only if you build the right foundation. DataHub Cloud provides the intelligent context your data and AI need to perform, scale, and deliver results. Whether you're piloting your first model or scaling enterprise-wide, context is your competitive edge.

Ready To Explore DataHub Cloud?

Share a few details about your goals and current challenges, and we' ll schedule a personalized walkthrough tailored to your specific data environment, use cases, and roadmap.

Contact us: **sales@datahub.com**

Let's talk about how we can help you:

- Discover and govern your data with ease
- Get AI models into production faster
- Ensure compliance and collaboration at scale

DataHub

About DataHub

DataHub, by Acryl Data, is an AI & Data Context Platform. Innovated jointly with a thriving open source community of 13,000+ members, DataHub's active metadata platform provides real-time context of AI and data assets with best-in-class scalability and extensibility. The company's enterprise SaaS offering, DataHub Cloud, delivers a fully-managed solution with AI-powered discovery, observability, and governance capabilities. Organizations rely on DataHub to accelerate time-to-value from their data investments, ensure AI system reliability, and implement unified governance—enabling AI & data to work together and bring order to data chaos.

Learn more at **datahub.com**
Follow DataHub on **LinkedIn** and **X**