WHITE PAPER

Improving the Success Rate of AI Projects

An overview of how a MarkLogic Data Hub with embedded machine learning provides organizations with the data platform needed to improve the success rate of AI projects.

MarkLogic®

Introduction

Artificial intelligence (AI) and machine learning (ML) offer so many tantalizing promises. As those fields mature they have the potential to make a huge impact on data science as a whole. They are the key to realizing the future of big data and how it helps businesses. We all know this, which is also why it's so frustrating that the vast majority of AI and ML projects fail.

Although there are numerous examples of where AI has been used successfully, the reality is most AI applications are falling short of business expectations or taking far too long to develop. Market research from IDG indicates that only one in three AI projects is currently succeeding – a nearly 67% failure rate.

> **"Only one in three AI projects is currently succeeding."**

The research also found that even when AI projects are considered successful, they are typically taking more than six months to develop. Moreover, these results may also help partially explain why 84% of organizations digital transformation initiatives fail, according to Forbes.

This doesn't give us a lot of hope for the future of AI and ML in any organization, but we know it can do better. Through many studies and the observation of both successful and failed projects, we are starting to learn the root causes of failure. The next step is learning how to address them.

> **"Data science, broadly defined, has been around for a long time. But the failure rates of big data projects in general and AI projects in particular remain disturbingly high."**
>
> *—Thomas C. Redman, Harvard Business Review*

The Root Cause: Data Integration and Quality Problems

The universe is literally made of data, just waiting to be integrated, transformed, and accessed. If only our universe of data had everything arranged in an ideal format, but unfortunately, that is not the case. Raw data is messy and it takes considerable time and cost to collect and get it clean enough for human consumption, let alone readable by machines.

Existing systems and personnel can process much available raw data, but even human-entered records are rife with inconsistencies and errors. The typical data preparation processes cannot catch many of these easily, and if they make it into a dataset bound for an AI project, they can count for many working hours of additional manual cleaning and reformatting.

Surveys consistently show that most business intelligence (BI) professionals end up playing the role of "data janitor." An Xplenty survey showed that 30% of all BI professionals spend between 50 and 90 percent of their time just on extracting, transforming and loading data (a.k.a, ETL). Another survey from Figure Eight showed that 74 percent of respondents spend at least 25 percent of their time cleaning data.

When your data scientists spend more time cleaning and preparing data than they do actually using it, effort is being wasted and projects are being delayed. You

want your data scientists to be where they'd rather be: building and testing new models and algorithms to help drive improved business performance. Instead, they are spending too much time dealing with dirty data.

Under pressure to deliver better outcomes and results, many organizations overlook foundational issues with data management. When organizations fail to optimally integrate all data across the enterprise, data is not ready for use in AI projects and can contribute to project failure. Unsurprisingly, Datanami reports that 79% of enterprise data is not ready for AI.

Data scientists struggle to adequately prepare and test new models. They spend most of their time just wrangling big data. With so much time spent dealing with foundational data challenges, data scientists have less time to iterate with solution architects and data engineers during a project, undermining the level of collaboration needed for a successful AI deployment.

Data Warehouses and Lakes: Drowning your AI projects

Most big enterprises are caught in a game of data catch-up, with piecemeal legacy data architectures that have evolved over decades. The traditional approach to cleaning up this big data mess involves dumping all of it into massive relational data warehouses or data lakes backed by technologies such as Hadoop.

Unfortunately, many organizations are finding that these traditional approaches are not addressing the problem of data silos, or helping solve data governance issues. While these technologies may serve as repositories for storing massive amounts of data, both have issues that limit their effectiveness for developing AI programs.

Data warehouses, with highly structured data stored in a relational model, limit the agility of your data operations by requiring a strict schema to be defined in advance and optimized for fast analytical queries using SQL. In a relational environment, data integration involves creating a common data model for all the data and writing ETL code to pull the primary data into this format before development can begin. For big data projects, the data modeling and ETL can take months (and in some cases years) before development can begin, severely

> **“The machine learning [and AI] race is really a data race.”**
>
> – Megan Beck, Larry Libert, MIT Sloan Review

hampering a data engineer's ability to operationalize AI programs.

Data lakes often lack capabilities for curating (enriching, mastering, harmonizing) and searching your data, and additional tools are usually required to analyze or operationalize the data. This is a growing problem for many data scientists and engineers since the ability to easily access optimally governed data across the entire organization is foundational for building successful AI programs.

Compounding these difficulties, less-governed data lakes have been shown to open the doors to cyber attacks and information security issues. False data injection and malware obfuscation target these information-rich sources, and with little ability to discern the good data from the bad, data lakes leave themselves vulnerable to compromise. Compromised data sources are amplified when used for AI programs and can have huge implications when mission critical decisions are effected.

AI Continues to Grow in Importance

In spite of the barriers to AI adoption, if organizations do not figure out how to successfully complete and leverage AI projects, they will fall behind. More and more companies are adopting AI, usually to increase efficiency and productivity.

AI is growing into a critical element in the success of enterprises. In some industries, it can optimize sales while decreasing fraud. In others, it can be crucial for compliance and regulation. Wherever AI helps manage and process complex datasets, an organization will work faster and more efficiently.

Take, for example, the financial services sector, where AI and ML adoption grows in importance every year. Regulatory requirements turn compliance into an increasingly heavy burden. A recent report shows that governance and compliance account for up to 20 percent of the operational costs of most major banks. Many of these banks are digitizing the compliance process. With its capacity for financial irregularity detection and replacing manual processing, AI is the perfect tool for this. This makes it even more critical that new AI projects succeed. As we have seen, that success is not a certainty.

As AI and ML become essential elements in core business processes, ensuring the success of new projects makes a big difference in an organization's viability and long-term success. Novel technical approaches to known challenges can better the odds.

Actions to Increase Successful Outcomes for AI Projects

It all starts with clear understanding of goals and objectives for how AI will be used to drive business performance. Once that is understood, there are three actions you can take to improve the success rate of your AI programs:

Deploy an integrated data platform

Use tools that enable better data governance, including core machine learning functions embedded into the underlying database to automate repetitive, non-differentiating tasks. Most organizations are working with legacy systems and disconnected data sets, resulting in data that is fragmented and lower quality. By bringing all data sources onto a single platform, organizations stand to minimize complexity, improve governance and accelerate delivery of AI applications. According to research from IDG, 80 percent of the 200 U.S. and European IT executives surveyed agreed that an integrated data platform would help support their AI initiatives.

Improve data integration, preparation, and governance

AI programs require data of the highest integrity. Without the highest quality data, AI programs are prone to biases. This also places a premium on data provenance and lineage as data scientists must be in a position to explain their models. The growth in data variety, volume

and velocity is creating a “perfect storm”, requiring enterprise architects to modernize data management capabilities or risk capsizing AI initiatives. Gartner predicts an 800% growth in data volume over next five years, with most of the growth in unstructured data. Without a flexible and agile approach to data integration, organizations will continue to produce sub-optimal results with data governance and AI programs.

Better integrate data science and engineering efforts

An integrated data platform combined with codified data governance practices will enable scientists and engineers to spend less time on data wrangling and more time collaborating on AI projects. This, in turn, will lead to greater agility and an increased ability for scientists and engineers to iterate on development of AI applications, leading to better outcomes and returns on innovation investments.

> “Larger and more complex models make it hard to explain, in human terms, why a certain decision was reached (and even harder when it was reached in real time).”
>
> —McKinsey & Company

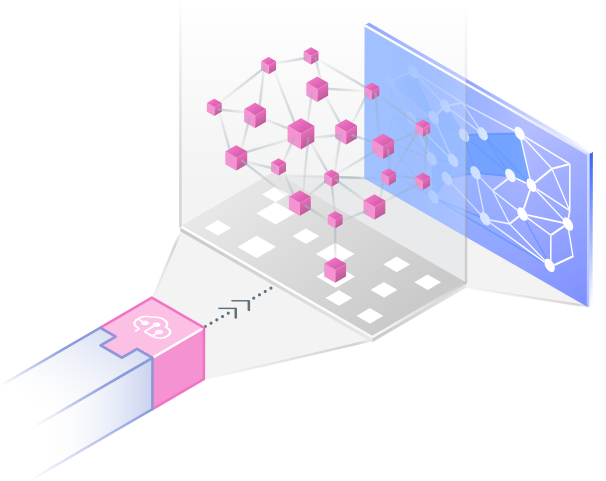
Engaging Embedded Machine Learning with MarkLogic

At MarkLogic, we examined the barriers and obstacles that most often get in the way of an AI/ML project's success and we engineered the best solutions for tackling them. We know the best place to operationalize machine learning is in a data hub where all available data can be governed and curated. The MarkLogic Data Hub enables smart mastering that aids in the discoverability of relationships between records, and gives data the context it needs to make your AI programs more accurate and flexible.

With the release of Data Hub 5.0 and MarkLogic 10, we have embedded machine learning into the core of our solution. Machine learning routines can run close to the data, in parallel across a MarkLogic cluster. This makes it simpler to train and execute models right inside the data hub, where we can handle almost every part of the architecture and process.

The operational data hub is designed to be a valuable data source for downstream applications, including AI. With connectors and APIs to a vast array of technologies, using MarkLogic expands the tools available to you rather than narrowing them.

By embedding machine learning into the Data Hub platform, MarkLogic is helping organizations build the data foundation required to improve the success rate of their critical AI and ML projects.



Benefits of Embedded Machine Learning

MarkLogic enables machine learning in your enterprise. MarkLogic's flexible, multi-model approach is perfect for integrating and storing the highly connected entities that machine learning and artificial intelligence systems need from various data silos and fluidly interacts with other systems to leverage this governed data. The MarkLogic Data Hub with embedded machine learning improves:

Database Operations – With embedded machine learning, MarkLogic runs queries more efficiently and scales autonomously based on workload patterns. With autonomous elasticity, for example, MarkLogic can use models of infrastructure workload patterns to automatically adjust the rules that govern data and index rebalancing.

CASE STUDY

Transforming Workforce Management with MarkLogic

Even when organizations have all the right technology tools, digital transformation initiatives can still come up short. Successful transformations usually combine the right mix of technology, culture, and people. Getting the right people into the right positions to execute is essential, which is why a large financial institution (FI) with over 200 thousand employees decided to invest in advanced technology to enable radical improvement of its processes for evaluating and placing job candidates.

At this FI, recruiters vet thousands of applications and conduct hundreds of interviews each day. This process generates an enormous amount of data that can be a huge challenge for the average person to handle under current processes. Analytics and AI have the potential to radically improve applicant vetting and intelligently match candidates with jobs.

The biggest challenges for this FI were the growing requirements and concerns around data integration, security and transparency associated with collecting more personal and business data about their job candidates and employees.

To deal with these challenges, the FI turned to MarkLogic to enable the ability to upload and create a personal data-driven profile for external candidates and employees. To build these 360° views, important details would need to be acquired from different sources, such as core HR, learning, performance management, skills inventory and compensation systems.

Additionally, MarkLogic also enabled the FI to meet requirements for:

- **Process improvements** from simplified data integration and easier content curation

that lead to accelerated delivery of better machine-learning outputs.

- **Security improvements** gained from data features such as element-level security and redaction, as well as access controls to ensure proper authorization of access to sensitive personal data.

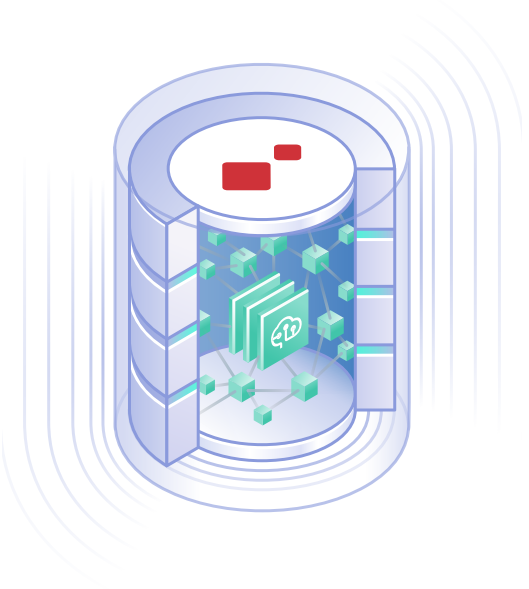
- **Audit improvements** from implementation of data lineage and use of metadata.

Prior to employing MarkLogic, data scientists supporting HR at the the FI were expending considerable effort doing time consuming data preparation activities, such as running multiple queries, cleaning and labeling data elements, and mapping data to create relationships.

With MarkLogic, the FI was able to simplify data integration, reduce time to pull and query new data, and automate the data cleansing process using embedded machine learning. Now the FI's data scientists spend substantially less time wrangling data, and more time developing advanced analytics and AI models to optimize its human resources for driving digital transformation.

Data Curation – Embedded machine learning reduces complexity and increases automation of various steps in the data curation process. For example, with MarkLogic's Smart Mastering feature, machine learning augments the rules-based mastering process so that records are mastered with more accuracy. Models continue to improve as more data is processed—all with less human involvement.

Data Science Workflows – For data scientists, it is now simpler to just do the work of training and executing models. Almost every part of the architecture and process can be handled right inside MarkLogic. This includes data processing, data curation, and the model engineering to build, train, execute and deploy the model.



Conclusions

We have explored the challenges facing AI and ML projects with an eye to learning how to increase the success rate across the industry. Creating a successful AI project requires buy-in and alignment across the enterprise, critically from executives, data scientists, and IT. Everybody needs to understand that AI cannot succeed without the right data, so early, intelligent processing of all data sources needs to be high on the list. Traditional data warehouse and data lake approaches will not be sufficient, and adding niche software to fill in gaps increases the complexity of your architecture and vulnerability of your data.

AI projects require usable data. The MarkLogic Data Hub with embedded machine learning provides the tools needed to automate data preparation and consistently improve the outcomes of your AI projects. Features like smart mastering, semantics and advanced search give data scientists the ability to more effectively govern and explore the big data sources, and confidently deliver projects faster.

> **“We are in the early innings of the game and we are going to see a lot of people strike out that get too aggressive too quickly—and in particular, those that think it is a cure for all data issues. The AI engine is only as smart as the data you put into it.”**
>
> —Gary Bloom, CEO MarkLogic

About MarkLogic

By simplifying data integration, MarkLogic helps organizations gain agility, lower IT costs, and safely share their data. Headquartered in Silicon Valley, MarkLogic has offices throughout the U.S., Europe, Asia, and Australia.

999 Skyway Road, Suite 200
San Carlos, CA 94070

+1 650 655 2300
+1 877 992 8885
www.marklogic.com
sales@marklogic.com


MarkLogic®