MarkLogic®

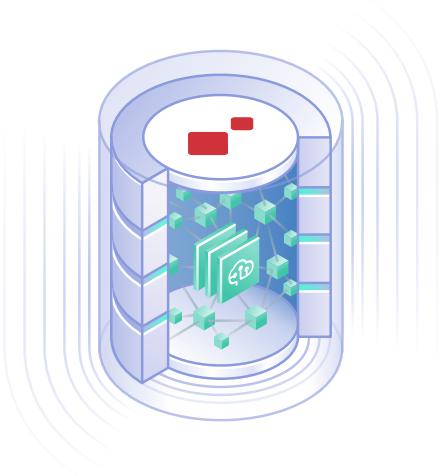
DATASHEET

Embedded Machine Learning

Machine learning has the ability to uncover hidden insights in your data to answer questions you have not thought to ask, but there are many challenges due to complexity and poor data quality.

The MarkLogic® Embedded Machine Learning capability runs at the core of MarkLogic, and enables users to solve some of the big challenges with machine learning by removing complexity and ensuring data governance. As part of the MarkLogic Data Hub Platform, machine learning models have direct access to high quality, curated, governed data. And, it is not just for data scientists.

Embedded Machine Learning is being applied to improve how MarkLogic operates and how data is curated, processes that are completely transparent to users.



Challenges with Machine Learning

- **Lack of quality and governance** – Good data is critical because machine learning can be even more sensitive to data quality since you are using the same data to both train and then execute the model. As a result, any problems with data quality get amplified. Effective machine learning and trust in the outputs requires proper governance.
- **Wild west ecosystem** – The machine learning and AI tools ecosystem is incredibly complex and as security and governance become a priority, it is tough to find people with the right skillsets to build and maintain the systems. According to an article in The New York Times, data scientists spend 80 percent of their time just wrangling data.
- **Low business ROI** – Often times businesses do not understand or trust the outputs of machine learning models to make decisions using them. And, data scientists and the hardware infrastructure they need are not cheap. High costs and poor outputs equate to an overall low ROI.

Key Benefits

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

- **Improving Database Operations** – With Embedded Machine Learning, MarkLogic will run queries more efficiently and scale 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.
- **Improving 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 will augment the rules-

based mastering process so that records are mastered with more accuracy, and models continue to improve as more data is processed—all with less human involvement.

- **Improving Data Science Workflows** – For data scientists, it is now simpler to just do the work of training and executing models right inside MarkLogic, where we can handle almost every part of the architecture and process. This includes data processing/curation, and the model engineering to build, train, execute and deploy the model.

How It Works

MarkLogic's Embedded Machine Learning is a full deep learning toolkit that operates as a run-time library installed right at the core of MarkLogic, in the database kernel. It exposes its functions as built-ins from JavaScript and XQuery, which means these functions run close to the data and are completely integrated.

Embedded machine learning was designed for peak performance not only for CPUs but also for GPUs, and it scales to multi-machine-multi-GPU systems. Additionally, it is designed using a compression technique that dramatically reduces communication costs, reducing inter-node communications and enabling highly scalable parallel training across multiple machines.

Embedded machine learning also supports the Open Neural Network Exchange ONNX format, an open-source shared model representation allowing for framework interoperability and shared optimization. ONNX allows developers to move models between popular frameworks such as CNTK, MXNet, PyTorch, and others.

The toolkit leveraged to build MarkLogic Embedded Machine Learning was originally developed by Microsoft in conjunction with Facebook and AWS and released under the name Cognitive Toolkit, or CNTK. Microsoft used CNTK to develop keystone products like Skype, HoloLens, Cortana, and Bing.

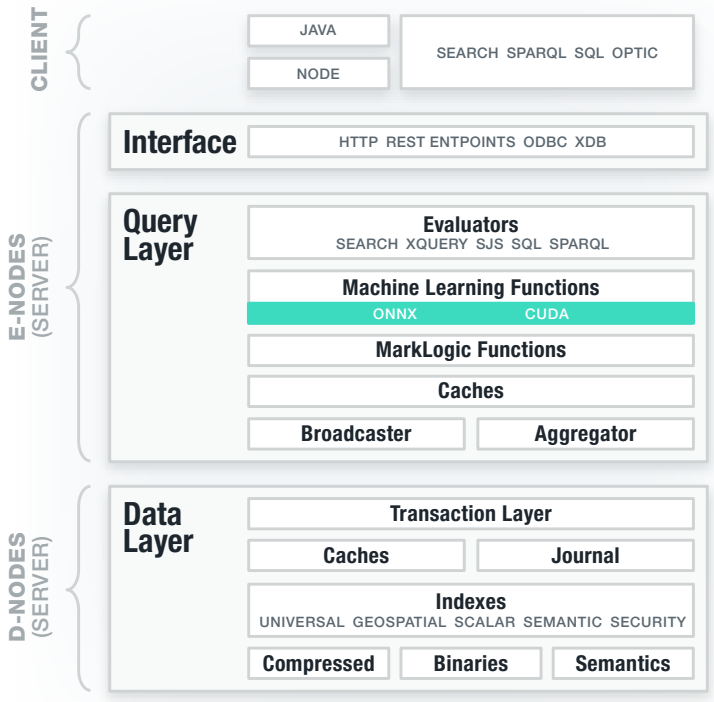


Figure 1: Architecture showing Embedded Machine Learning functions built in to the query layer in the E-nodes

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.

MARKLOGIC CORPORATION
999 Skyway Road, Suite 200 San Carlos, CA 94070
+1 650 655 2300 | +1 877 992 8885 | www.marklogic.com | sales@marklogic.com