

IRI Voracity

PAGE 14

MACHINE LEARNING
IN ANALYTICS AND
ANONYMIZATION

BDOQ
BIG DATA QUARTERLY

SUCCESSING **WITH** DATA SCIENCE **AND** MACHINE LEARNING

Best Practices Series

EXPANSIVE DATA FUELS DATA SCIENCE AND MACHINE LEARNING

Best Practices Series

WITH GROWING ATTENTION devoted to AI, machine learning, and IoT, what we've come to know as big data has become an even broader version of itself. In recent years, big data was seen as an unstoppable force of nature that would either overwhelm enterprises or propel them to new heights. This next generation of big data—we'll call it *expansive data*, pulsing through systems in real time, powering processes unseen to human eyes, and adapting and learning as it goes along—is going to reshape enterprises in ways not even anticipated. This requires attention to new types of tools, platforms, and approaches to deliver value to today's data-hungry businesses.

Expansive data will represent ever-growing volumes of information, potentially increasing within enterprises at a rate of up to 36% a year, according to Dresner Advisory Services. Platforms

supporting this growth include Amazon Web Services S3, Spark SQL, Hive, and Hadoop. Additional tools popular in enterprises are Apache Spark and TensorFlow.

This next generation of big data—we'll call it *expansive data*, pulsing through systems in real time, powering processes unseen to human eyes, and adapting and learning as it goes along—is going to reshape enterprises in ways not even anticipated.

Expansive data places even greater demands on enterprise infrastructures, processes, and the managers and administrators responsible for making it all work. That's because organizations are leaning more heavily than ever before on their data assets and analytics capabilities, and initiatives such as AI and machine learning, to help them compete.

Edge computing is also a defining factor in expansive data. There is likely to be greater activity at the edges—expansive data means more

processing may be distributed across IoT networks. Data can be ingested, processed, and even stored within edge devices and systems, and, if it is deemed critical on an enterprise scale,



moved to centralized data centers or clouds. Edge computing continues to extend its capabilities and encompasses a broad assortment of devices and systems that may require real-time interactions and responsiveness, including kiosks, autonomous cars and trucks, and sensors embedded across IoT.

With expansive data surging across all points of the enterprise, infrastructures could be quickly overwhelmed with ingestion, processing, and storage demands. Expansive data could also be expensive data without proper preparation. Fortunately, none of this is happening in a vacuum, and other developments may be helping organizations manage the challenge. Thanks to the ubiquity of cloud-based services, from infrastructure to platform to applications, the power and capacity to support even bigger data environments are readily available.

A new generation of database tools and platforms—led and enabled by machine-learning initiatives—is supporting the continuous, relentless data growth. Hadoop, the big data framework that made massive-scale data analytics a reality for every company that needs it, is beginning to show its age. While Hadoop was once seen as the single cure-all for big data challenges 10 years ago, today's expansive data calls for a variety

Thanks to the ubiquity of cloud-based services, from infrastructure to platform to applications, the power and capacity to support even bigger data environments are readily available.

of tools, platforms, and frameworks to help enterprises better manage their data. Nonetheless, the Hadoop Distributed File System can either support or be a part of data lake architectures, opening up a new mission for these environments.

According to a 2018 survey conducted by Unisphere Research, a division of Information Today, Inc., 44% of enterprises had Hadoop in production, which represents a downward shift from 2016, in which 55% reported using the framework (“2018 Next-Generation Data Deployment Strategies Report”). In addition, the survey found general satisfaction levels with Hadoop are mixed: Only 14% consider themselves to be “extremely satisfied” with Hadoop, while 64% are either dissatisfied or lukewarm toward the framework. While Hadoop provided one-of-a-kind functionality in its early days—such as parallel processing and management of a variety of data types—other technologies and solutions also now share these capabilities without the skill levels that Hadoop demands.

Predictably, the growth of expansive data is likely to track closely to that of IoT itself.

Accordingly, next-generation data technology initiatives represent new approaches to data management. The Unisphere Research survey found notable growth in the adoption of data lakes—places to store diverse datasets without having to build a model first. Their adoption continues to rise as data managers seek to develop ways to rapidly capture and store data from a multitude of sources in various formats. Overall, 38% of organizations are employing data lakes as part of their data architecture, up from 20% in a survey conducted 2 years prior. Another 15% said they were considering adoption. Data lakes are growing to impressive levels as well—close to one-third, 32%, support more than 100TB of data, the survey found.

With the relentless rise of IoT, AI, machine learning, and cloud-based services, enterprises are now challenged with accommodating and delivering value from the expansive data that surges through their systems. Data warehouses and Hadoop represented solutions for the pre-IoT, pre-AI enterprises. Today's opportunities and challenges call for the next generation of platforms and tools to bring it all together.

—Joe McKendrick



IRI Voracity
An Insatiable Appetite for Data

Machine Learning in Analytics and Anonymization

THE INCREASING NUMBER OF applications for machine learning testify to its ability to improve the speed and accuracy of informational assessment from ever larger sources of data. Users of the IRI Voracity data management platform can leverage two machine learning modules: one for predictive analytics, and another for protecting sensitive data. Many more are possible.

PREDICTING MALIGNANCIES

A common use of machine learning involves training a computer to evaluate

data sets and create prediction models from trends in that data. Machine learning builds off traditional statistics and rapidly creates larger and more advanced models.

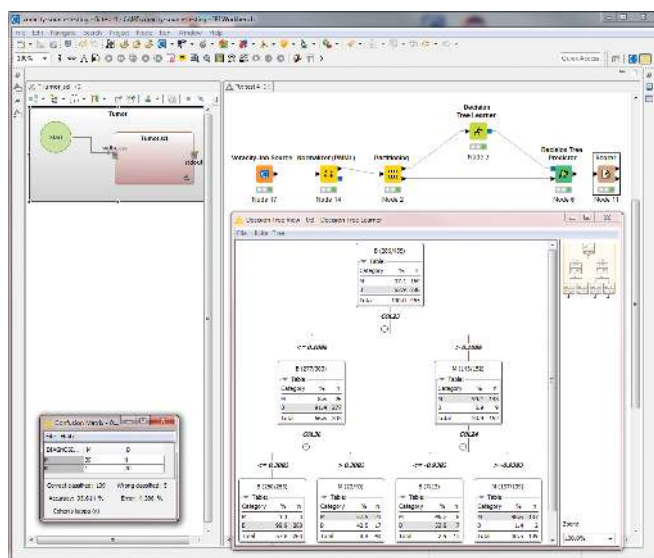
Many machine learning-related modules are included in KNIME, a popular open source data science platform that runs with Voracity in Eclipse. In this KNIME workflow, a Voracity data wrangling node feeds tumor measurement data into a KNIME decision tree node to improve breast cancer prediction accuracy:

Here, Voracity prepared raw data containing 20 different measurements of

breast tumors, including their overall size, shape, and features of the cells' nuclei.

Within seconds, the wrangled results flow into a decision tree to help determine if a tumor is likely to be malignant or benign.

The “Decision Tree Learner” node goes through different variables and creates multiple binary trees. Each tree determines if a given factor is likely to be a cause for a malignant tumor before it tries the next variable. Once the tree is built, the predictive model using those variables is tested for accuracy. In this case, it was about 95%, so the model should continue to be a reliable predictor future for data sets, too.



KNIME decision tree analysis of tumor data wrangled in IRI Voracity node

FINDING A SKING PII

Personally identifiable Information (PII) in documents like Word and PDF can be difficult to discover, delete, or de-identify en masse. This is particularly true when items like names or addresses — which do not match patterns or lookup values — can only be found in their Natural Language (Processing) context.

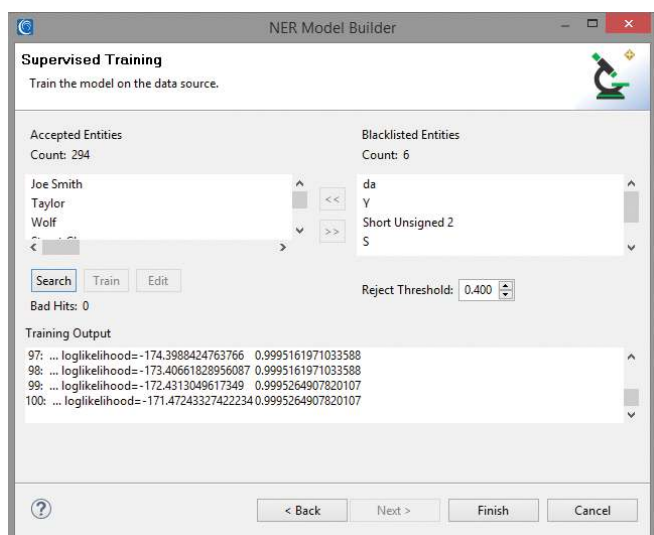
To address this challenge, IRI DarkShield technology in Voracity supports the use, and training, of Named Entity Recognition (NER) models to find those items. NER models are built from custom training data to improve search results.

The graphical front-end for DarkShield (and the rest of Voracity), called IRI Workbench, includes a wizard for either creating a NER model from existing annotated training data, or for training a model with actual documents that DarkShield parses. The latter employs a semi-supervised, iterative machine learning and annotation process.

Model training in this way improves the accuracy of PII search results when performed on a representative subset of documents. 15,000 sentences are considered a good minimum for teaching the machine to find named entities.

For more information email voracity@iri.com.

www.iri.com/voracity



Using Machine Learning in IRI Workbench to train IRI DarkShield NER models