opentext™

WHITE PAPER

OpenText AI: Guide to the technical architecture of OpenText Magellan

This white paper explains the technical architecture and basic components of OpenText™ Magellan™, which provides artificial intelligence (AI), machine learning, advanced analytics, text mining and big data processing combined with enterprise-class business intelligence, data discovery and reporting. OpenText AI adds value to Enterprise Information Management (EIM) through innovation and insight, delivering the Intelligent and Connected Enterprise.

opentext™

Contents

opentext™

Introduction

Many organizations are using AI, or cognitive computing, to automate repetitive tasks, spot patterns, predict trends and discover ways to streamline their business or derive profit. However, building an effective cognitive analytics system requires high-powered data science talent, which is often scarce. Moreover, IT teams are quickly realizing that assembling the components of a cognitive analytics system from disparate proprietary technologies is costly, time-consuming and complicated and may be difficult to customize.

Instead of bogging enterprises down with installation and integration complexity, Magellan delivers a ready-to-go AI-powered analytics platform, which includes machine learning, data discovery, text analytics and sophisticated visualization and dashboarding, drawing from the proven, widely-used business intelligence and analytics components of the OpenText™ Analytics Suite and the powerful text mining capabilities of OpenText™ Magellan™ Text Mining.

Pre-integration of all the components serves as a “force multiplier” that lets businesses make the most of precious data science talent and extend the benefits of AI-enriched analytic insights to a wide range of user roles across the organization. This gives businesses a cost-effective and timely method of leveraging machine learning to drive their critical decisions. Magellan uncovers insights from big data and big content and empowers IT teams, operational users and business analysts to share findings, make more informed decisions and act with more impact.

What is Magellan made of?

Magellan bypasses many of the hurdles that enterprises face when implementing AI and analytics because it is built on four major integrated components that serve a variety of user types with an environment most appropriate to their technical skill level. Rather than requiring the organization to acquire, install and integrate different technologies to achieve a comprehensive set of AI capabilities, Magellan provides everything out of the box. Magellan's four main components are

- Magellan data lake
- Magellan Text Mining
- Magellan™ Data Discovery
- Magellan™ BI & Reporting

The Magellan data lake is built on top of the Apache Hadoop® platform and Apache Spark™, the powerful open-source platform built for processing big data and machine learning. Magellan leverages Spark components that are pre-integrated and purpose-built to deliver a full AI and data science platform. Data scientists can use the Magellan Data Science Notebook with the data lake to create, save and process custom machine learning algorithms using programming languages, such as Scala, Python, SQL and R. These languages are familiar environments that help make it easier for developers and data scientists to get to work. They can train these models using machine learning and following advanced data science methodologies.

Magellan Text Mining provides the functionality to acquire data from a large variety of sources, including Twitter and other social media streams, news articles and documents stored across content silos within the enterprise. With its advanced capacities, it can even enrich and offer an understanding of that data via natural language processing and AI. Magellan Text Mining can identify key concepts, noun phrases and named entities within text, as well as extract topics, subjectivity, emotion, and tone through text classification, and sentiment and emotion analysis, providing powerful, pervasive text analytics across an enterprise's entire population of unstructured content.

Magellan Data Discovery enables business analysts to go from raw data to answers in seconds by applying advanced and predictive analytics with a few clicks. Without any coding, business users can apply advanced analytical techniques, such as crosstabs, Venn diagrams, correlation, profiling, bubble charts or maps, to predictive analytics and machine learning techniques, such as anomaly detection, association rules, clustering, decision trees, Naive Bayes classification, linear and logistic regression, and pattern mining—all done via a visual, drag-and-drop interface. This functionality is pre-integrated with the Magellan data lake, allowing data scientists to publish models they create, which has the potential to scale their work and make AI more accessible to business users. Business users can have visualizations recommended to them based upon the data they have selected, through Smart Data Discovery, helping them to quickly create accurate and representative visual insights.

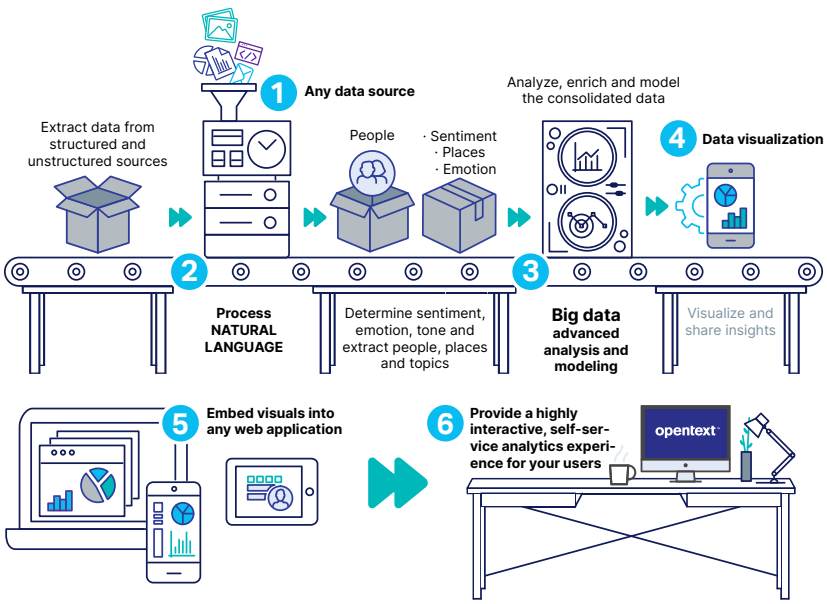
Magellan BI & Reporting can embed the insights it generates into any application or any device as intuitive and highly-interactive data visualizations. Magellan BI & Reporting is engineered for high availability, reliability, deep security integration, API-driven pliability and performance to support everything from a modular, service-based deployment to a full-scale enterprise rollout to millions of users. These advantages allow Magellan data scientists and business users to promote rich analysis and content to a large audience in virtually any manner, regardless of platform.

Magellan BI & Reporting, Magellan Data Discovery, Magellan Text Mining, and the Magellan Data Science Notebook each provide multiple deployment methods; on-premise, hybrid, or in the cloud. Cloud-native editions for these products are built on a modern, containerized architecture supported by Docker and Kubernetes, and provide high levels of agility, performance, and scalability. These editions can be easily deployed to the OpenText Cloud or to another cloud platform of choice.

How it all works together

At a high level, the overall process can be described in six steps:

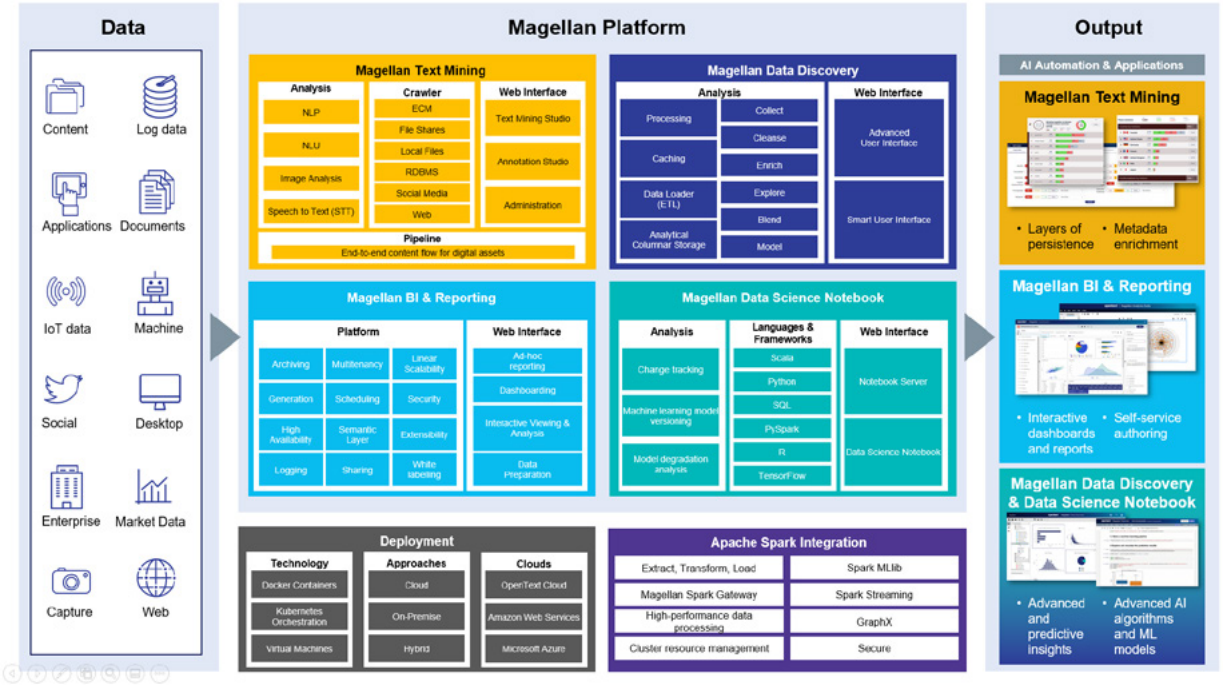
1. **OpenText™ Magellan™ Text Mining Crawling Services acquire** data from both structured and unstructured sources.
2. When textual data is involved, **Magellan Text Mining processes** content ingested from the various sources, applying natural language understanding to extract semantic metadata, such as key terms and named entities (people, places, dates, and concepts), performing text classification (or categorization) against native or custom taxonomies and analyzing subjectivity, emotion, and tone. Editorial and semantic metadata from all documents are then federated and ready to be used and blended for advanced text analytics.
3. **Data scientists use the Magellan Data Science Notebook with the data lake to apply advanced algorithms** and create models to understand the data at a previously inaccessible level.
4. **Business users can now use Magellan Data Discovery and its Smart Data Discovery functionality to create visualizations and analyze** the prepared data models as well as data in the Hadoop layer.
5. All of this can be easily shared and socialized **as a wide range of appealing visualizations via Magellan BI & Reporting. These can be embedded into web pages or portals with minimal effort via modern (HTML5-based) web page elements.**
6. Users can explore and interact with the insights delivered by Magellan on a self-service basis. There is no need to summon a data scientist to build a new query each time someone just wants to adjust a variable slightly.



The power of Magellan makes true integration of a wide variety of data sources into value-driven and user-adopted applications possible—often a challenge so vast that some organizations don't even attempt it. Magellan also enables a collaborative and complete decision-making process across an entire organization by ensuring that each role within the audience, regardless of skill set or desired depth of access to the data, has an approachable route to navigate this vast amount of knowledge and insight.

Magellan integration architecture

Magellan was thoughtfully designed to be rich in "bridges" or integration points that facilitate data flows, transformations and enrichment throughout the platform. The diagram below shows that each of Magellan's core components is built from vetted open-source technology enriched with OpenText innovations to produce a robust, enterprise-grade solution.



With the Magellan data lake, data scientists can use MLlib, Python, R and a full platform of data science tools to create advanced algorithms and models that truly capture the value in the data using AI. This includes the Magellan Data Science Notebook, which is based on Jupyter™ Notebook, an industry-leading notebook solution that enables progressive development and formulation of models. The Magellan Data Science Notebook also includes model versioning capabilities that lets data scientists easily track changes to notebooks and maintain multiple Machine Learning model versions. They can then train, fit and publish these models using the Magellan Spark Gateway.

The Magellan Spark Gateway enables Magellan Data Discovery to expose published and fit machine learning models, created and curated by data scientists directly from within the notebook, to business users who can then execute those models against massive data sets.

This gateway connects Magellan Data Discovery directly to Spark, eliminating additional steps to transfer the models' data elsewhere for analysis. Real fit models can be executed directly by business users, giving them unprecedented access to AI-empowered data.

The Magellan Text Mining Crawling Services and OpenText™ Magellan™ Text Mining Pipeline enable connectivity directly to unstructured information sources, such as Enterprise Content Management, document repositories (including file shares) or financial systems. Additionally, they can connect to less traditional unstructured sources of information, such as Twitter, websites, forums or chat streams. Each crawler within Magellan can be configured individually and processed separately by the pipeline based on its unique requirements.

OpenText™ Magellan™ Text Mining Studio is an application to manage controlled vocabularies, such as taxonomies and authority files, and allows users to create and manage controlled vocabularies that classify content and terms. In addition to taxonomies provided with Magellan, custom taxonomies can be created using knowledge engineering or machine learning. Also, new classification models can be attached to a taxonomy to drive text classification, named entity recognition or sentiment and emotion analysis. This enables Magellan to load data directly into the data lake or send it to Magellan Text Mining for natural language processing, letting users extract vast amounts of information from text and spoken information. The Magellan Text Mining data is then merged with other sources and can be used for analysis via Magellan BI & Reporting components, data models or Magellan Data Discovery or to go directly into the data lake for use by data scientists.

Magellan BI & Reporting can access all the Magellan platform components individually or all at once, depending on the need. This means that any needed view of the data can be exposed at any point along the data pipelining process and used for advanced and interactive analysis by end users at any skill level.

Magellan BI & Reporting's data models use a technology called the Open Data Access (ODA) framework, which allows direct access to specialized tools, such as Apache's Hadoop ecosystem storage solutions, HBase® and Apache Hive™. Magellan BI & Reporting also includes pre-built proprietary ODA connectors that enable access to the Magellan Text Mining data and Magellan Data Discovery. Magellan BI & Reporting can embed any content directly into any application component, enabling Magellan to expose any level of the data to any level of the audience, seamlessly from any application.

With an end-to-end approach to integration across the entire ecosystem, Magellan enables organizations to take advantage of their valuable information and finally realize the real promise of AI by making it ubiquitously accessible for today's larger audiences of tech-savvy users.

Summary

By making the integrated capabilities of an entire AI and data science platform easily accessible, Magellan solves most of the challenges of moving data and integrating access and connectivity to make AI truly usable. Adding MLlib, Python, R and a toolbox of Magellan Analytics Suite products helps organizations get the full value of their most important asset—information.

About OpenText

OpenText, The Information Company, enables organizations to gain insight through market leading information management solutions, on-premises or in the cloud. For more information about OpenText (NASDAQ: OTEX, TSX: OTEX) visit: opentext.com.

Connect with us:

- OpenText CEO Mark Barrenechea's blog
- Twitter | LinkedIn

opentext.com/contact