KNOWLEDGENT WHITE PAPER

How to Design a Successful Data Lake

Information through Innovation

www.knowledgent.com

Executive Summary

Business users are continuously envisioning new and innovative ways to use data for operational reporting and advanced analytics. The Data Lake, a next-generation data storage and management solution, was developed to meet the ever-evolving needs of increasingly savvy users.

This white paper explores existing challenges with the enterprise data warehouse and other existing data management and analytic solutions. It describes the necessary features of the Data Lake architecture and the capabilities required to leverage a Data and Analytics as a Service (DAaaS) model. It also covers the characteristics of a successful Data Lake implementation and critical considerations for designing a Data Lake.

Current Enterprise Data Warehouse Challenges

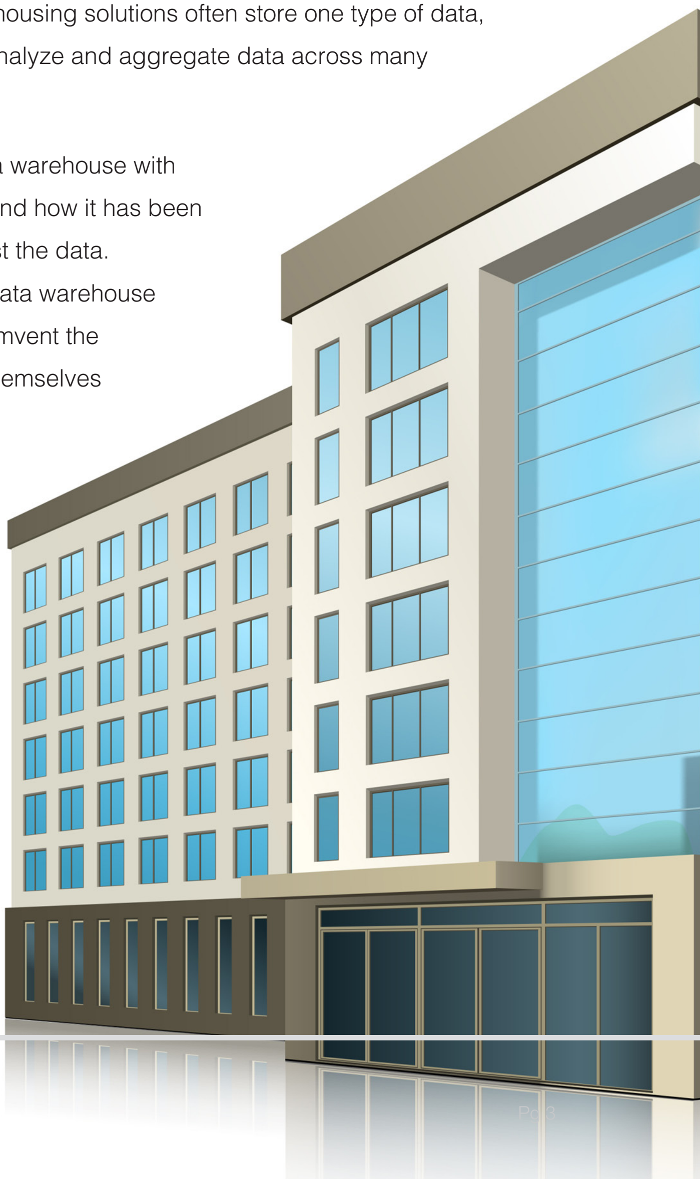
Business users are continuously envisioning new and innovative ways to use data for operational reporting and advanced analytics. With the evolution of users' needs coupled with advances in data storage technologies, the inadequacies of current enterprise data warehousing solutions have become more apparent. The following challenges with today's data warehouses can impede usage and prevent users from maximizing their analytic capabilities:

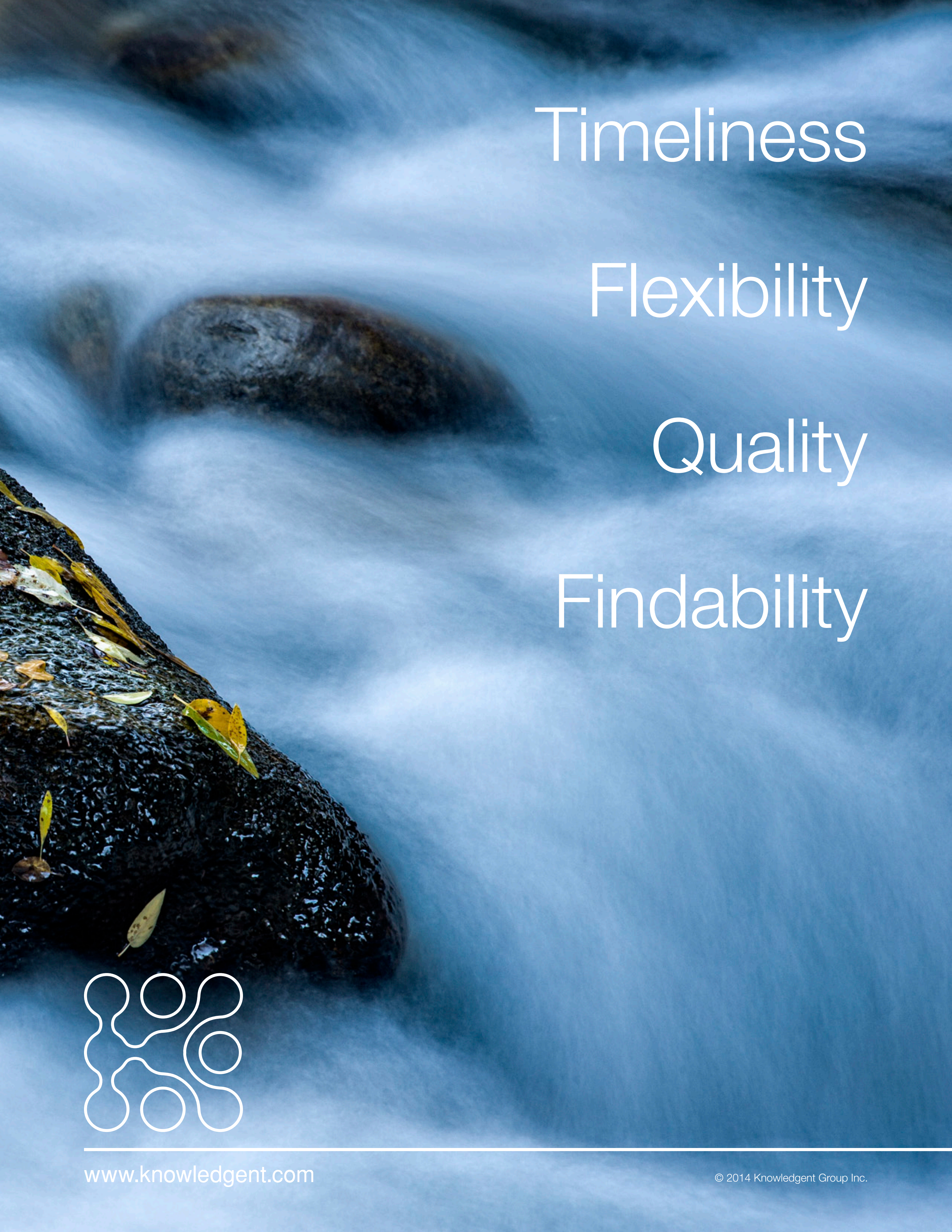
- Timeliness. Introducing new content to the enterprise data warehouse can be a time-consuming and cumbersome process. When users need immediate access to data, even short processing delays can be frustrating and cause users to bypass the proper processes in favor of getting the data quickly themselves. Users also may waste valuable time and resources to pull the data from operational systems, store and manage it themselves, and then analyze it.

- Flexibility. Users not only lack on-demand access to any data they may need at any time, but also the ability to use the tools of their choice to analyze the data and derive critical insights. Additionally, current data warehousing solutions often store one type of data, while today's users need to be able to analyze and aggregate data across many different formats.

- Quality. Users may view the current data warehouse with suspicion. If where the data originated and how it has been acted on are unclear, users may not trust the data. Also, if users worry that the data in the data warehouse is missing or inaccurate, they may circumvent the warehouse in favor of getting the data themselves directly from other internal or external sources, potentially leading to multiple, conflicting instances of the same data.

- Findability. With many current data warehousing solutions, users do not have a function to rapidly and easily search for and find the data they need when they need it. Inability to find data also limits the users' ability to leverage and build on existing data analyses.





Timeliness

Flexibility

Quality

Findability

www.knowledgent.com

Advanced analytics users require a data storage solution based on an IT “push” model (not driven by specific analytics projects). Unlike existing solutions, which are specific to one or a small family of use cases, what is needed is a storage solution that enables multiple, varied use cases across the enterprise.

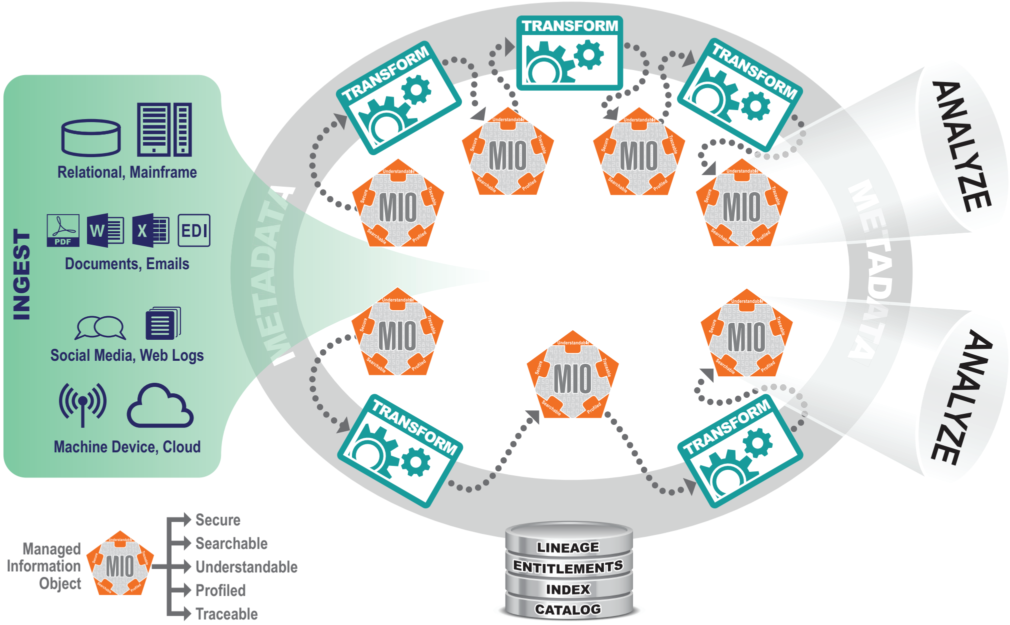


Figure 1: The Data Lake Architecture

This new solution needs to support multiple reporting tools in a self-serve capacity, to allow rapid ingestion of new datasets without extensive modeling, and to scale large datasets while delivering performance. It should support advanced analytics, like machine learning and text analytics, and allow users to cleanse and process the data iteratively and to track lineage of data for compliance. Users should be able to easily search and explore structured, unstructured, internal, and external data from multiple sources in one secure place.

The solution that fits all of these criteria is the data lake.

The solution that fits all of these criteria is the data lake.

www.knowledgent.com

The Data Lake Blueprint

The Data Lake is a data-centered architecture featuring a repository capable of storing vast quantities of data in various formats. Data from webserver logs, data bases, social media, and third-party data is ingested into the Data Lake. Curation takes place through capturing metadata and lineage and making it available in the data catalog (Datapedia). Security policies, including entitlements, also are applied.

Data can flow into the Data Lake by either batch processing or real-time processing of streaming data. Additionally, data itself is no longer restrained by initial schema decisions, and can be exploited more freely by the enterprise. Rising above this repository is a set of capabilities that allow IT to provide Data and Analytics as a Service (DAaaS), in a supply-demand model. IT takes the role of the data provider (supplier), while business users (data scientists, business analysts) are consumers.

The DAaaS model enables users to self-serve their data and analytic needs. Users browse the lake's data catalog (a Datapedia) to find and select the available data and fill a metaphorical "shopping cart" (effectively an analytics sandbox) with data to work with. Once access is provisioned, users can use the analytics tools of their choice to develop models and gain insights. Subsequently, users can publish analytical models or push refined or transformed data back into the Data Lake to share with the larger community.

Although provisioning an analytic sandbox is a primary use, the Data Lake also has other applications. For example, the Data Lake can also be used to ingest raw data, curate the data, and apply ETL. This data can then be loaded to an Enterprise Data Warehouse. To take advantage of the flexibility provided by the Data Lake, organizations need to customize and configure the Data Lake to their specific requirements and domains.

Potential Pitfalls of the Data Lake Solution

Even a flexible, next-generation solution like the Data Lake is subject to its own set of challenges. Although a large volume of data is available to users in the Data Lake, problems can arise when this data is not carefully managed, including:

- Lack of data governance. Without the structure and controls to manage and maintain the quality, consistency, and compliance of data, a Data Lake can rapidly devolve into a Data Swamp.
- Poor accessibility. Although the data might be available, its value is limited if users are unable to find or understand the data.
- Poor data quality and lineage. Users need to have the context of the data and know where it comes from to trust the data completely.
- Lack of data security. Data loaded to a Data Lake without any oversight can lead to compliance risks.

To maximize the value of the Data Lake and avoid these pitfalls, organizations need to ensure that their Data Lake implementations address specific critical success factors.

Ingest

Curate

Apply

www.knowledgent.com

Characteristics of a Successful Data Lake Implementation

A Data Lake enables users to analyze the full variety and volume of data stored in the lake. This necessitates features and functionalities to secure and curate the data, and then to run analytics, visualization, and reporting on it. The characteristics of a successful Data Lake include:

- Use of multiple tools and products. Extracting maximum value out of the Data Lake requires customized management and integration that are currently unavailable from any single open-source platform or commercial product vendor. The cross-engine integration necessary for a successful Data Lake requires multiple technology stacks that natively support structured, semi-structured, and unstructured data types.

- Domain specification. The Data Lake must be tailored to the specific industry. A Data Lake customized for biomedical research would be significantly different from one tailored to financial services. The Data Lake requires a business-aware data-locating capability that enables business users to find, explore, understand, and trust the data. This search capability needs to provide an intuitive means for navigation, including key word, faceted, and graphical search. Under the covers, such a capability requires sophisticated business ontologies, within which business terminology can be mapped to the physical data. The tools used should enable independence from IT so that business users can obtain the data they need when they need it and can analyze it as necessary, without IT intervention.

- Automated metadata management. The Data Lake concept relies on capturing a robust set of attributes for every piece of content within the lake. Attributes like data lineage, data quality, and usage history are vital to usability. Maintaining this metadata requires a highly-automated metadata extraction, capture, and tracking facility. Without a high-degree of automated and mandatory metadata management, a Data Lake will rapidly become a Data Swamp.

- Configurable ingestion workflows. In a thriving Data Lake, new sources of external information will be continually discovered by business users. These new sources need to be rapidly on-boarded to avoid frustration and to realize immediate opportunities. A configuration-driven, ingestion workflow mechanism can provide a high level of reuse, enabling easy, secure, and trackable content ingestion from new sources.

- Integrate with the existing environment. The Data Lake needs to meld into and support the existing enterprise data management paradigms, tools, and methods. It needs a supervisor that integrates and manages, when required, existing data management tools, such as data profiling, data mastering and cleansing, and data masking technologies.

Keeping all of these elements in mind is critical for the design of a successful Data Lake.

Analytics
Visualization
& Reporting

Designing the Data Lake

Designing a successful Data Lake is an intensive endeavor, requiring a comprehensive understanding of the technical requirements and the business acumen to fully customize and integrate the architecture for the organization's specific needs.

Knowledgent's Big Data Scientists and Engineers provide the expertise necessary to evolve the Data Lake to a successful Data and Analytics as a Service solution, including:

- DAaaS Strategy Service Definition. Our Informationists leverage define the catalog of services to be provided by the DAaaS platform, including data onboarding, data cleansing, data transformation, datapedias, analytic tool libraries, and others.

- DAaaS Architecture. We help our clients achieve a target-state DAaaS architecture, including architecting the environment, selecting components, defining engineering processes, and designing user interfaces.

- DAaaS PoC. We design and execute Proofs-of-Concept (PoC) to demonstrate the viability of the DAaaS approach. Key capabilities of the DAaaS platform are built/demonstrated using leading-edge bases and other selected tools.

- DAaaS Operating Model Design and Rollout. We customize our DAaaS operating models to meet the individual client's processes, organizational structure, rules, and governance. This includes establishing DAaaS chargeback models, consumption tracking, and reporting mechanisms.

- DAaaS Platform Capability Build-Out. We provide the expertise to conduct an iterative build-out of all platform capabilities, including design, development and integration, testing, data loading, metadata and catalog population, and rollout.

Conclusion

The Data Lake can be an effective data management solution for advanced analytics experts and business users alike. A Data Lake allows users to analyze a large variety and volume when and how they want. Following a Data and Analytics as a Service (DAaaS) model provides users with on-demand, self-serve data.

However, to be successful, a Data Lake needs to leverage a multitude of products while being tailored to the industry and providing users with extensive, scalable customization. Knowledgent's Informationists provide the blend of technical expertise and business acumen to help organizations design and implement their perfect Data Lake.

For more information on how Knowledgent can customize and configure a Data Lake solution, visit www.knowledgent.com.

About Us

Knowledgent is a data and analytics firm that helps organizations transform their information into business results through data and analytics innovation. We help clients maximize the value of information by improving their data foundations and advancing their analytics capabilities. We combine data science, computer science, and domain expertise to enable our clients to implement innovative, data-driven business solutions.

Knowledgent operates in the emerging world of big data as well as in the established disciplines of enterprise data warehousing, master data management, and business analysis. We not only have the technical knowledge to deliver game-changing solutions at all phases of development, but also the business acumen to evolve data initiatives from ideation to operationalization, ensuring that organizations realize the full value of their information.

For more information, visit www.knowledgent.com.

Knowledgent™
Innovation Through Information

New York, New York • Warren, New Jersey • Boston, Massachusetts • Toronto, Canada
www.knowledgent.com

For more information contact marketing@knowledgent.com.