COGNIZANT

Cognizant

Semantic 'Radar' Steers Users to Insights in the Data Lake

By infusing information with intelligence, users can discover meaning in the digital data that envelops people, organizations, processes, products and things.

KEEP CHALLENGING™

Executive Summary

In the past several years, we've seen Amazon grow to nearly $75 billion in sales without operating a single storefront, in part by using information about customers' past purchases to recommend new offers to them. More recently, Airbnb managed 10 million stays in 2013 without owning a single hotel; instead, it encourages members to offer space in their own homes to rent out to others and publishes online ratings of guests and hosts.

These are only two examples of how businesses are transforming customer experiences, establishing new business models and redefining their markets. At the heart of their success is an innate ability to continuously understand and manage the digital data generated by people, organizations, processes, products and things[1] – what we call a Code Halo™.

In an increasingly "data-rich, meaning-poor" business world, companies everywhere are attempting to learn the value of signal and the cost of noise when coming to grips with their big data and analytics initiatives. Extracting meaning from this data empowers companies, brands and employers to better understand and meet customer needs, and to create more personalized experiences for them. Doing so requires new tools to manage and understand the torrents of data that will come from the projected tens of billions of devices as the "Internet of Things" links everything from cars to clothing.[2] It also requires algorithms that filter data "noise," generate new insights and create personalized and enriching experiences that help Code Halo pioneers like Pandora and Netflix treat us as if we were individual markets of one.

Combining streams of data from different sources into a "data lake"[3] can help steer business decision-makers to the insights they need more quickly and efficiently. At its core is high-performance, low-cost data storage technology to store and manage frequently used, large volumes of data. Multiple analytics engines speed the delivery of a personalized experience to each customer at the point of interaction. Semantic models standardize the definition and mapping of internal and external data. Domain expertise,

such as knowledge of which patterns in the data represent risk or opportunity, is embedded in the semantic models[4] to guide business users to insights. Finally, self-service features help business users rapidly find and analyze data without delay.

This white paper highlights the challenges that organizations face when using today's data architectures to find and act on business-changing big data insights. It describes how a "data lake" can help, and the central role that semantic technology plays by slashing the time required to find, integrate, share and analyze the data. It finally describes how enterprises can begin creating their own data lakes.

Today's Information Architectures: Too Slow, Expensive and Hard to Understand

A data lake eliminates many of the roadblocks that now stand between business users and business insights. Among these are:

- **Onboarding and integrating data is slow and expensive.** Creating personalized customer experiences and tracking complex markets requires collecting, integrating, analyzing and responding to data from many different sources. However, the ETL (extract, transform and load) function requires custom processes and technologies for each source system, causing delays when skilled staff is busy on other projects.
- **Optimizing data for different users is difficult and costly.** Departments such as finance, sales and manufacturing may organize and define data differently. Finance, for example, may organize customers based on their annual purchases or credit ratings, while sales organizes them by geography, and manufacturing by the products they purchase. The cost and delay of building data marts tailored to the needs of each unit can make it impossible to quickly uncover vital insights.
- **Moving and reorganizing data introduces delays and cost.** Data storage is becoming steadily less expensive, but it still represents a capital and operational expense. Hampered by insufficient funding and the high costs of onboarding data, organizations often wait until there is a defined need for a set of data before loading it into an analytics database. By then, a competitor may have already captured the data and acted on the analysis.
- **Data provenance and meaning is often lost in translation.** Before knowing what data to query, a user must understand what that data represents and how it is used. One example: "This number (X) represents defect rate per thousand, which we use to justify our premium pricing for a customer." As data knowledge is transferred from expert to designer to developer, meaning is often replaced with technical syntax and rules that the business user cannot understand, effectively removing all-important context from "corporate memory." Maintaining the provenance and meaning allows the data to tell a story that matters to a real-life decision-maker.
- **Who transformed what data, and why?** In order to ensure that users aren't making decisions based on faulty data, it is essential to understand who changed the format of the data, cleansed it to eliminate inconsistencies or changed values to comply with standards. Such management artifacts are often tracked only in spreadsheets and don't capture the latest changes made to live data. This problem extends to tracking changes to data after it is onboarded, increasing the costs and complexity of performing data analysis.
- **Models are too rigid or incomplete.** Excessive standardization of the vocabulary used to define and organize data may make it harder to conduct specific types of analysis. On the other hand, the lack of a model makes it difficult or impossible to identify what data is available, how to access it and how to integrate it with other data. That means delays when users must search for the required data, determine who owns it, request permission to access it, and integrate it with other data.
- **The speed of change.** With every new competitor, product offering, business model or distribution channel, an organization must rapidly assimilate and query new types of data, organized and optimized in new and innovative ways. Lengthy data design and standardization processes are not effective in today's business environment.

All of these issues are solvable with a "smart" data lake strategy. Figure 1 depicts best practices that overcome key points of friction that organizations encounter when using traditional approaches.

Data Lake Essentials

Filling the data lake to generate unique customer experiences and unlock market insights requires new processes and technologies across the data and query management lifecycle (see Figure 2, next page). At each step, semantic technology plays a central role (see sidebar, page 7).

- **Data ingestion:** This involves finding and importing the data needed for Code Halo analysis. Structured and unstructured data can be collected through traditional scheduled batch loading, data streaming or real-time, on-demand queries for internally- and externally-hosted data.

The Data Lake Defined

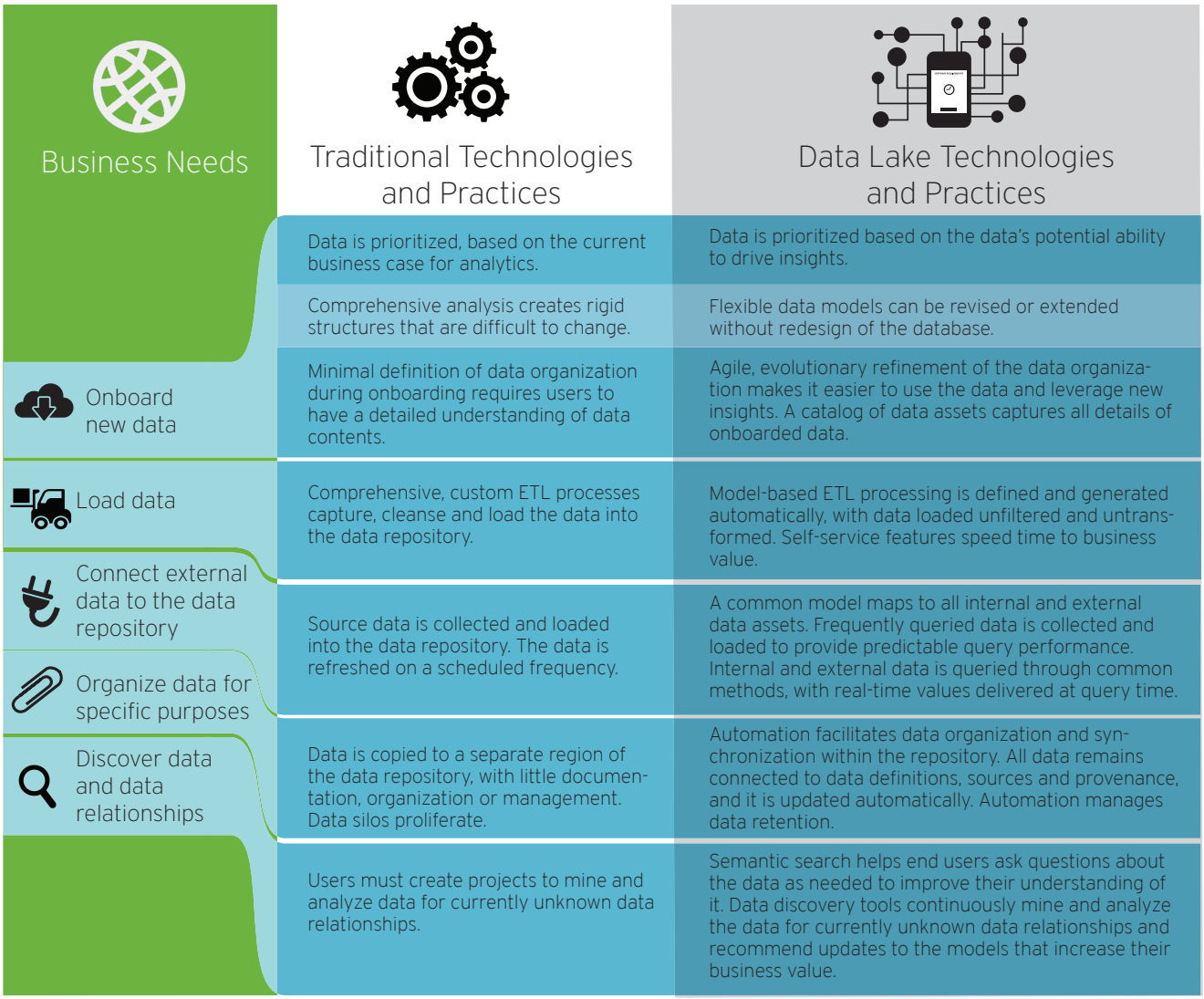

Figure 1

The intelligent data management layer directs the model-driven movement of data, while provenance capabilities capture and track details about its ingestion, transformation, movement and the queries run against it.

Organizations can quickly analyze the data needed without loading every conceivably useful bit of data beforehand. The combination of semantic models and automated data movement greatly reduces the need for manual coding of ETL processes by defining all the available data, how it is organized, where it is located and what data movement/transformations must be performed. During ingestion, the data is semantically tagged and categorized so it can be easily identified for future needs.

Pre-modeling also lets business users perform self-service data mapping and queries of new data that will only be needed for a short time or that are awaiting integration with formal models. They can then use this data for rapid, on-demand Code Halo analysis to meet urgent needs without waiting for help from IT.

Data Lake Components

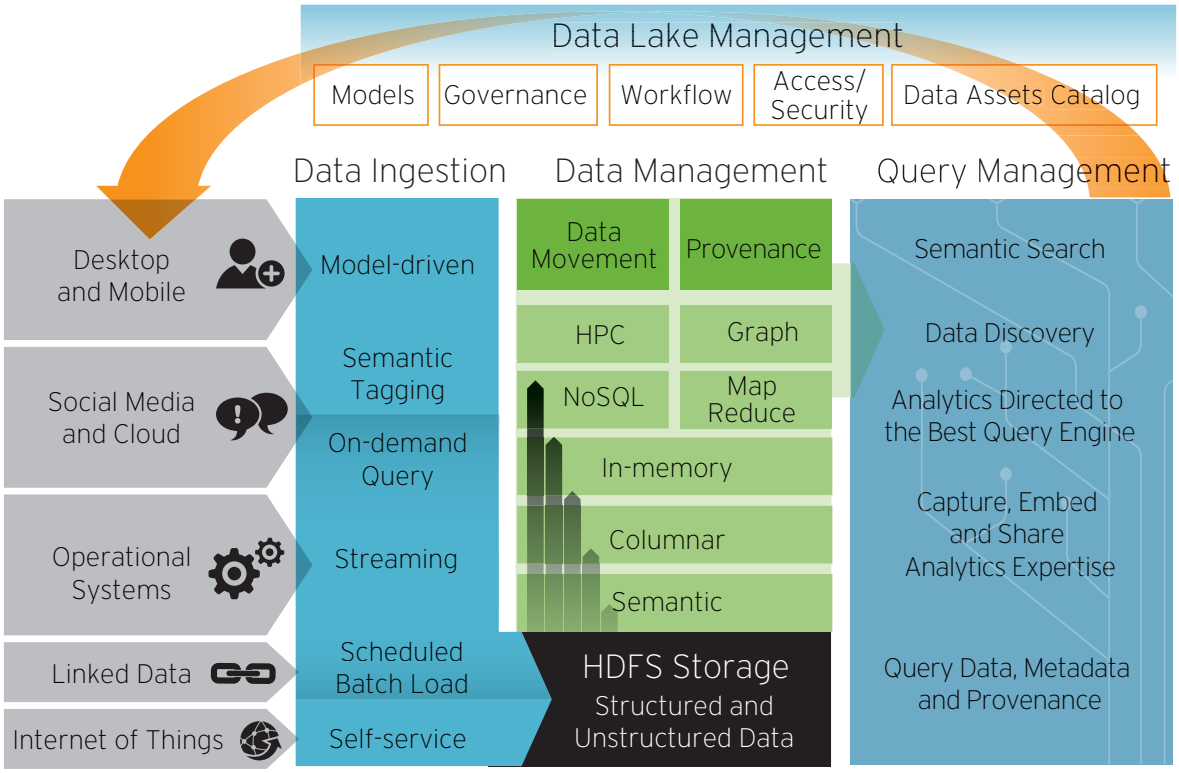

Figure 2

By performing a semantic analysis of the data during data loading, organizations can capture an understanding of the content, which makes it easier to embed intelligence in the data. This intelligence might include which characteristics or patterns indicate the likelihood that a customer will "churn" from one cable provider to another, or that a homeowner is at increased risk of falling behind in loan payments. A trigger can generate an offer to customers that meets their needs and enhances the relationship.

- **Data management:** Data lakes benefit from modern, low-cost, high-performance storage technologies, such as the Hadoop Distributed File System (HDFS), that provide high availability and automated failover. Specialized database and analytic engines, such as high-performance computing, NoSQL and semantic, can be configured to speed Code Halo analytics at the lowest possible cost.

  The intelligent data management layer directs the model-driven movement of data, while provenance capabilities capture and track details about its ingestion, transformation, movement and the queries run against it.

  The data management layer supports the various engines required for different types of analytics, including future technologies, as an organization's use of Code Halos matures. It also allows the engines to access and analyze data directly from high-performance storage, eliminating the need for custom ETL processes.

- **Data lake management:** Like any other corporate resource, the data lake requires management and governance to ensure that it operates as efficiently, consistently and securely as possible. The management layer directs data ingestion and movement workflows while managing access and balancing performance, cost and security to deliver the best results at the lowest cost.

  The model manager allows analysts to embed expert knowledge, such as a complex calculation or set of logic, into the ontology model for use by non-experts. For example, a researcher can easily share a new risk-assessment model for home loans with underwriters in the field. A genomics specialist can share a model that maps new biomarker indications to patients' genetic backgrounds to help researchers at clinical trial sites test a new cancer treatment.

Quick Take

Semantics Maps the Data Lake

A smart semantics model can help organizations make use of their Code Halos by capturing meaning, as well as the related domain expertise, from structured, unstructured or semi-structured data. The building blocks for such a model are standards and technologies, such as:

- **The Resource Definition Framework (RDF),** which organizes data in a graph structure, reducing development time and cost while delivering business value sooner.
- **The Web Ontology Language (OWL),** which provides a comprehensive model of data definitions and relationships that is human- and machine-readable.
- **The SPARQL Query Language,** a SQL-like query language for semantic data that can leverage ontological relationships to execute smarter questions across multiple databases in a single query.
- **Inferencing,** which makes it easier for users to construct queries by capturing and embedding expertise in the ontology model.

Proper governance is required to ensure the data is accurate and consistent and that it complies with organizational and industry standards. This should cover both master and reference data, and take advantage of industry standards, such as the Financial Industry Business Ontology (FIBO)[5] in banking/financial services and CDISC/RDF[6] in life sciences. The model manager captures governance-based standards to ensure consistency across models that reference shared data. Business units will organize views of required data that include enterprise- and business unit-specific data. The business unit-level data models will inherit the definitions of enterprise standard models while adding business unit-specific data relationships and terminology (see Figure 3).

The model manager enables organizations to construct data views to support a specific analytic or class of analytics, without the cost of building data marts or custom data movement processing.

The model manager enables organizations to construct data views to support a specific analytic or class of analytics, without the cost of building data marts or custom data movement processing. By capturing the details of data relationships and data movement that directly support the analytics process, organizations can enable automation that reduces delays and costs.

The data catalog manager captures the details of both internal and external data assets, describing the content, definitions and organization of the data and supporting multiple views of shared data. End-users can define workflows for scheduled data movement.

Different Data Models for Different Users

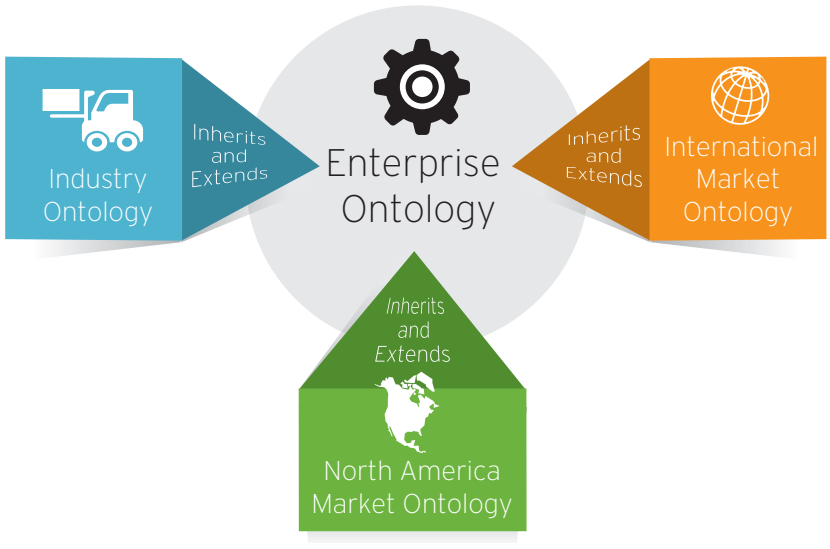

Figure 3

Access management features help users find data assets available in the data lake, automate registration and access to new data assets, and provide secure access to the data. Monitoring functions ensure that everything from the storage technology to the analytics and reporting functions is working properly, with tools to support intervention when necessary.

- **The query management layer:** This helps users understand what data is available and is most likely to deliver the richest Code Halo insights. Semantic search uses natural language queries to analyze the data and metadata to best identify the data that will support an analytical process. To reduce the time and effort required to find and model business-critical data, data discovery functions analyze the data to find and link related data concepts. Optimized data structures speed analytics, with queries directed to the analytics engine that delivers the best performance.

  The query management layer thus enables "smart" analytics in which users can define not only which data to use, how to transform it, which analytics engine to use and which analysis to run, but also what type of chart to plot the results to, what file format to save the results in and who to mail the results to. Workflow tools allow users to automate the analytics process.

- **Rules-based data sourcing:** This helps users fine-tune their queries using a context such as metadata and provenance. For example, a user could specify a query that says, "Read the data from the data lake, unless that data is more than X hours old, in which case it should be read directly from the source system."

  While the data lake is most often used by analysts or business decision-makers, true Code Halo value is achieved when it responds to a customer's click with an immediate, personalized response. This is enabled by features that expose analytics processes as a Web service, or through an application program interface (API) that is callable from a Web page or mobile app.

Moving Forward

Building an intelligent semantic model on top of your current information architecture may sound daunting. But it can and must be done, if you hope to discover and act on the next game-changing insight before your competitors do. You can get started with these important but gradual changes:

- **Prioritize the onboarding of data** by its ability to create truly individualized customer experiences or business-changing insights into market needs or operations.
- **Onboard data assets as they become available, without waiting for a specific use case.** Those uses will emerge when you least expect them, and when they do, you'll need the data immediately. At a minimum, map the data to the semantic model for easier access.
- **Load the data without filtering or transforming it, since new data rules may override old rules.** Filtering and transformation rules can be applied as the data is moved to an analytics engine or during query execution.
- **Model the data using familiar terminology.** This makes it easier to change the model as needed without physically moving the data. Customize models for specific business groups, encouraging them to ensure its accuracy and completeness.
- **Enable search mechanisms** that make it easier for business users to see what data is available and accessible.
- **Balance legal and compliance needs** for security with the imperative to improve the customer experience through analytics.

Code Halos are powerful sources of industry-changing insights. Too often, this knowledge is difficult to find within the volume and complexity of the data. A data lake provides the radar that allows end users to quickly see past the chaos of too much data and focus on the actionable insights they need.

Footnotes

1 For more on Code Halo thinking, see Malcolm Frank, Paul Roehrig and Ben Pring, "Code Rules: A Playbook for Managing at the Crossroads," Cognizant Technology Solutions, June 2013, http://www.cognizant.com/Futureofwork/Documents/code-rules.pdf, or read our book *Code Halos: How the Digital Lives of People, Things, and Organizations are Changing the Rules of Business*, by Malcolm Frank, Paul Roehrig and Ben Pring, John Wiley & Sons, April 2014, http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118862074.html.

2 "Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units by 2020," Gartner, Inc., Dec. 13, 2013, http://www.gartner.com/newsroom/id/2636073.

3 James Dixon, "Pentaho, Big Data, and Data Lakes," Pentaho, October 2010, http://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.

4 Thomas Kelly, "Semantic Technology Drives Agile Business," Cognizant Technology Solutions, August 2013, http://www.cognizant.com/InsightsWhitepapers/How-Semantic-Technology-Drives-Agile-Business.pdf.

5 "Financial Industry Business Ontology," Object Management Group, http://www.omg.org/hot-topics/fibo.htm.

6 "Representing Regulations and Guidance in RDF," Phuse Wiki, http://bit.ly/18WAxxf.

About the Authors

Thomas Kelly is a Director in Cognizant's Enterprise Information Management (EIM) Practice and heads its Semantic Technology Center of Excellence, a technology specialty of Cognizant Business Consulting. He has 20-plus years of technology consulting experience in leading data warehousing, business intelligence and big data projects, focused primarily on the life sciences and healthcare industries. He can be reached at Thomas.Kelly@cognizant.com.

Robert Hoyle Brown is an Associate Vice President in Cognizant's Center for the Future of Work, and drives strategy and market outreach for the Business Process Services Practice. Robert is responsible for looking at how Code Halos will change enterprise business process services in the coming years and is a regular contributor to the blog www.unevenlydistributed.com, "Signals from the Future of Work." Prior to joining Cognizant, he was Managing Vice President of the Business and Applications Services team at Gartner, and as a research analyst, he was a recognized subject matter expert in BPO, cloud services/BPaaS and HR services. He also held roles at Hewlett-Packard and G2 Research, a boutique outsourcing research firm in Silicon Valley. He holds a Bachelor of Arts degree from the University of California at Berkeley and, prior to his graduation, attended the London School of Economics as a Hansard Scholar. He can be reached at Robert.H.Brown@cognizant.com.

The authors would like to thank Editorial Director Alan Alper for his insights and contributions to this white paper.

COGNIZANT

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world's leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 75 development and delivery centers worldwide and approximately 187,400 employees as of June 30, 2014, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.

World Headquarters
500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277
inquiry@cognizant.com

European Headquarters
1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 (0) 207 297 7600
Fax: +44 (0) 207 121 0102
infouk@cognizant.com

India Operations Headquarters
#5/535, Old Mahabalipuram Road
Okkiyam Pettai, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060
inquiryindia@cognizant.com

Cognizant