



Surviving the Petabyte Age: A Practitioner's Guide

Executive Summary

The concept of “big data” is gaining attention across industries and the globe. Among the drivers are the growth in social media (Twitter, Facebook, blogs, etc.) and the explosion of rich content from other information sources (activity logs from the Web, proximity and wireless sources, etc.). The desire to create actionable insights from ever-increasing volumes of unstructured and structured data sets is forcing enterprises to rethink their approach to big data, particularly as traditional approaches have proved difficult, if even possible, to apply to structured data sets.

One challenge that many, if not most, enterprises are attempting to address is the increasing number of data sources made available for analysis and reporting. Those who have taken an early adopter stance and integrated non-tabular information (a.k.a. unstructured data) into their pool of analysis data have exacerbated their data management problems.

A second challenge is the shrinking timeframe in which a business stays focused on a particular topic. Thanks to the highly integrated and communicative global economy, and the great strides made in expanding communications bandwidth, both good and bad news circumnavigate the globe at a much faster pace than ever before.

The amount of time it takes for news to become common knowledge has shrunk, thanks to:

- An emerging network of social media and blogs that potentially makes everyone a publisher of good and bad news.
- A rapid increase in the number of people who are untethered from traditional information receptacles and now have a highly mobile means of collecting and ingesting information.
- The meteoric rise of desktop tools housing a significant portion of information. Organizations need to understand the information and processes involved in the dispensation of desktop-managed information (mostly Microsoft Access and Excel). This information is most likely to be found in the form of:
 - Copies of operational data (including both sources and targets).
 - Copies of operational data that is enriched (including the processes and sources used for enrichment, as well as the targets that receive the enriched information).
 - Processes bypassing the systematized processes (including the bypassed processes, the sources used for these processes, the actors in these processes and the results of these processes).

This whitepaper lays out the concept of a business information model as a vehicle to organize information into separate categories, which directly influences the creation, capture or extraction of business value and elevates it to a heightened focus. We will cover four main topics:

1. Why companies dealing with big data in today's Petabyte Age¹ need to stratify information so that trustworthy, relevant, actionable and timely data can be found at a moment's notice.
2. A business model that can be used to stratify information.
3. A new definition of partitioning and a business process for formulating the partitions. Partitions should deal with stratifying information based on its contribution to organizational data, as well as the more traditional technical partitioning that is conducted for performance and maintenance reasons.
4. Methods of rolling out an information infrastructure aligned with this new partitioning definition. The realities of this new environment are that the maintenance of a traditional enterprise information model happens at the speed of business and is in direct opposition to maintaining the focus of information that directly contributes to enterprise value.

Three Issues to Solve

The Petabyte Age² is creating a multitude of challenges for IT organizations, as they find that their well-honed, carefully constructed informa-

tion models cannot be maintained fast enough to appease their business constituents. Moreover, once constructed and populated with information, these models require new technologies to interface with the data. Adding insult to injury, all this data is largely introspective and serves merely to support the status quo. When disruptions occur, insights can only be gleaned from this data over a sufficient passage of time; in the meantime, insights are derived from what is largely called unstructured and semi-structured data, as well as data obtained from outside the organization via social media, blogs, Web sites and a host of other sources that don't fit into the neatly organized tools devised for insight generation.

A major shift is transforming the basic tenets of data-driven insight generation. This shift requires a new way of combining and synthesizing data used for navigating the highly integrated and communicative global economy.

Overcoming this challenge requires organizations to solve three important issues (see Figure 1):

- **Data depth:** How to derive insight from structures that contain billions or more instances of data. These can include sessions in a Web log, entries obtained from social media, entries from RFID activities or mobile-sourced activities. One thing is sure: The sheer size of these pools of data will continue to grow, resulting in technical hurdles that challenge traditional methods for efficiently and effectively using such large pools of like data. Most solutions that deal with big data attempt to meet this challenge.

Data Challenges of the Petabyte Age

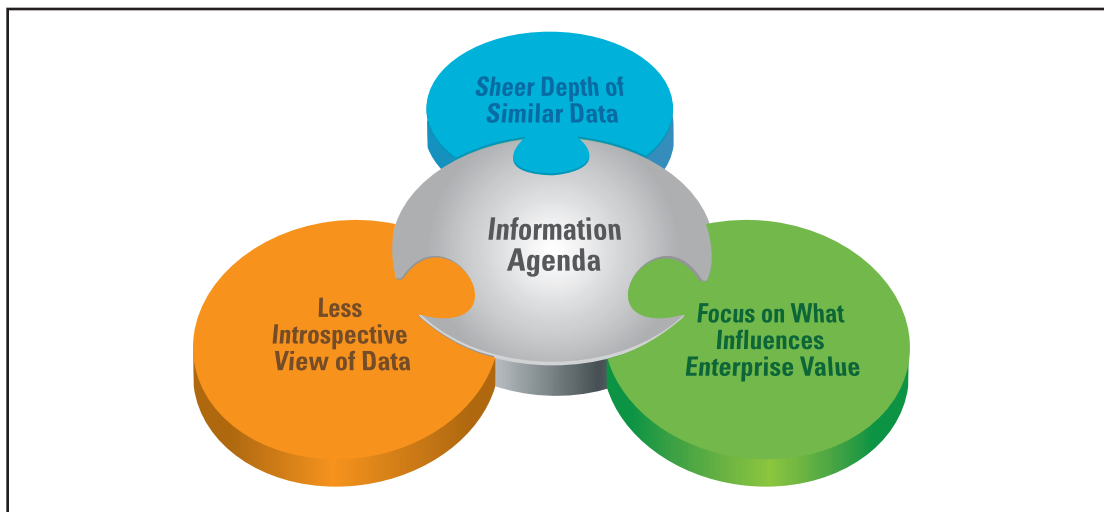


Figure 1

- Focus on enterprise value:** How to quickly determine which data requires the most focus at any point in time. Thanks to our tightly connected global economy, news travels around the world more quickly than ever, which requires rapid rethinking of enterprise strategies and tactics. This requires the ability to quickly change which data is focused upon. Traditional information models that are constructed to synthesize business knowledge from the deluge of available data impede the nimbleness required to meet the needs of the modern-day enterprise.
- Less introspective view:** How to make the whole information fabric less introspective. Using information derived from inside the organization can predict future trajectories only if the status quo is assumed. However, when there is a high degree of turbulence, knowledge obtained from internally-generated information is woefully inadequate in the short term; insights are obtainable only after sufficient time has passed and several cycles have been interpreted. The resulting organizational missteps are covered regularly in the news media. What is required is an ability to wield information as an early-warning system for understanding changes in enterprise trajectories. Such data sources are external to the enterprise until enough time has passed for a history of data points to be inferred from internal data.

Sheer Depth of Similar Data

Specialized tools have emerged to address this issue of enormous pools of similar data. These tools originate from the realization that the time-honored structured query language tools, as well as other tools built around database technologies, are ill-equipped to efficiently deal with billions, if not trillions, of rows of data. Spawned from Google's attempt to deal with the data accumulated from all the interactions that occur with the Google software suite, a whole new framework built around the MapReduce technology has been borne, and an emerging suite of tools has begun to appear on this new stack of technologies.

There will no doubt be a refinement of the techniques that are maturing to deal with this concept of big data. The only thing we can be sure of is that the big-data business issues addressed by MapReduce and the related suite of technologies are not going away.

Just as the technologies available for launching the initial collection of Web sites were immature, so are the tools for developing solutions for big data. Much has been said about how technology has taken a major step back from what is commonly available for business intelligence and data warehousing solutions – but this is much less a statement about the problem of big data than it is about the immaturity of the technologies available for solving the big-data problem set.

Converting Big Data Into Value

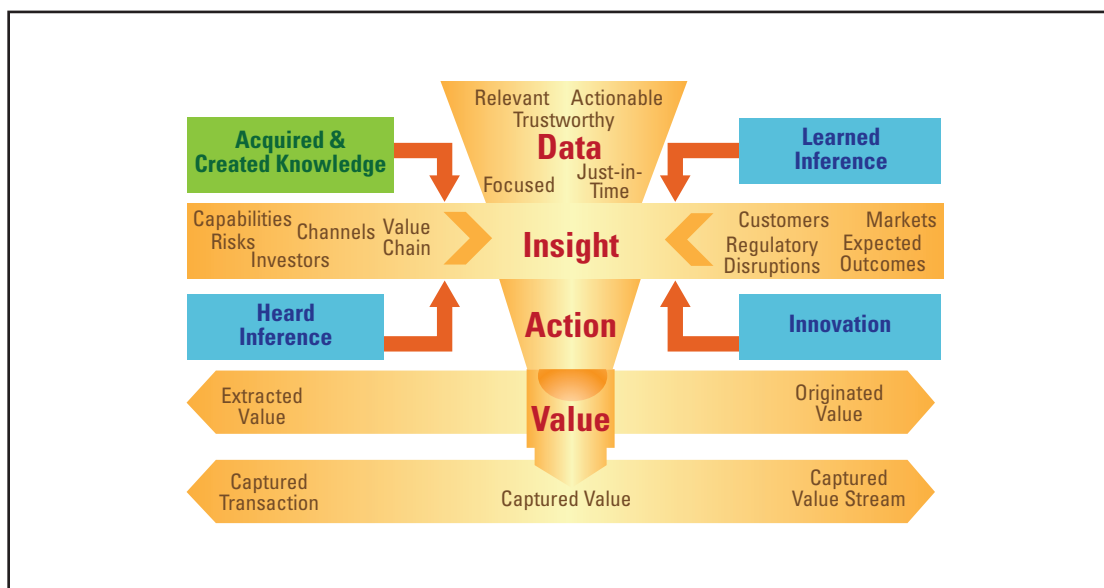


Figure 2

Managing Opportunity and Risk

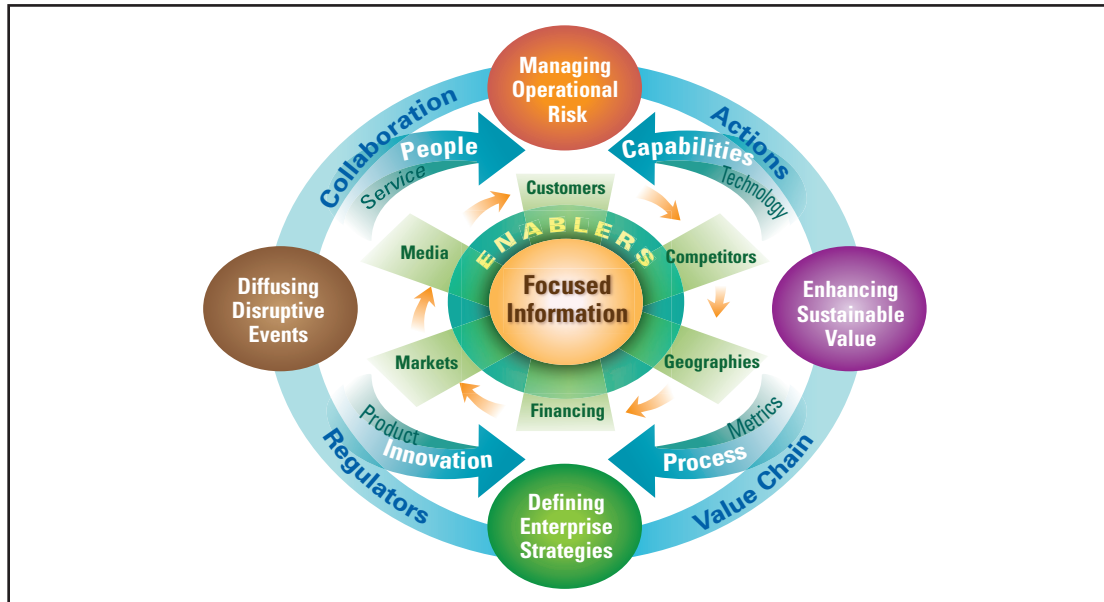


Figure 3

Interestingly, the problem of large pools of data is the primary issue, which today is tackled by introducing technologies to tackle each of the challenges outlined above independently. Companies that thrive in the Petabyte Age will be able to consolidate the technologies so their business constituency is faced with a single interface that addresses their full complement of informational needs.

Focus on Influencers of Enterprise Value

The intent of business intelligence is to take actionable, relevant, trustworthy and timely data; put it through a model that aligns with key business challenges (customers, geographies, channels, investors, markets, etc.) as the means to gain insight; and derive an action plan to extract, originate or capture organizational value (see Figure 2, previous page). Furthermore, captured value can be a one-time event (i.e., a temporary supply shortfall of a competitor) or a permanent value stream. While captured transactions are acceptable, captured value streams are more desirable.

Data is converted into insight by using acquired and created knowledge (obtained from both inter-

nal and external sources), learned inferences, heard inferences and innovations, some of which will serve as disruptions to others in the participating marketplaces.

It is the business model itself that must provide the focus into what is pertinent to the business at a particular point in time and that serves as the point of contention. The enterprise business models used as the basis for synthesizing information as the means of gaining insight are devised to map all data rather than “tiering” data into focus areas. Examples of focus areas include the following:

- Directly relates to creating or protecting extracted, originated or captured enterprise value.
- Does not directly contribute to value but is mandatory for business operations.
- May not be mandatory for business operations but is mandatory for regulatory purposes.
- May not be mandatory for the above categories but is mandatory for archiving.
- Was once important but is now relegated to historical trivia.

To create or protect extracted, originated or captured enterprise value, the information deemed worthy of focus must be sufficiently broad in scope so that both the opportunities and risks are exposed in all dimensions of the business model.

To create or protect enterprise value, the information deemed worthy of focus must be sufficiently broad in scope so that both the opportunities and risks are exposed in all dimensions of the business model.

For example, in the illustrated business model in Figure 3 (see previous page), operational risks, disruptive events, enterprise strategies and sustainable value sources will be managed by managing:

- People, as well as the services they provide.
- Processes and the metrics used to manage the processes.
- Innovations – specifically, the products released into the marketplace.
- Capabilities aligned with technologies.

Information will be managed in this model, along the following dimensions (i.e., the enablers):

- Customers, or the customers, prospects and visitors who can be tapped for enterprise value.
- Media, both traditional and emerging (social media like Facebook and Google+) that can influence enterprise value.
- Markets participated in for originating, extracting or capturing enterprise value.
- Financing, or the source of funds used for investments and cash flow used to originate, extract or capture enterprise value.
- Geographies and sovereign nations from which enterprise value will be originated, extracted or captured.
- Rivals in markets and geographies that compete for customers, market coverage and funding sources.

A Less Introspective View of Information

Only expected trends can be tracked using internal information. Disruptions will eventually appear in internal data, but their trajectory will only be evident after two or more cycles of information make their way into the internal data stream. This means:

- It will take a minimum of three days for new sales trajectories to make themselves known to a daily sales system. By that time, any progress that competitors have made in capturing value from your largest customers is removed for immediate transactions (i.e., captured transactional value) and, in many cases, is gone forever (i.e., captured value streams).
- In cases where data is reported less frequently, such as financial results, it will take weeks or months for such situations to be exposed,

at which point it is much more difficult to remediate.

Disruptions make themselves known through external data much more readily than internal data. However, there are also problems with external data, including the fact that this data is much more loosely defined and that the sheer number of information sources are more extensive and change more frequently in scope and content.

An example of an external data source that can be captured is Twitter. All Twitter content is capable of being captured, and a competitor's promotion that is broadcast on Twitter can be immediately exposed. In order to listen for a Twitter message, however, a handful of literally billions of 140-byte messages will be the potential source of this information. And Twitter is only one of many information sources that can expose such calls to action.

Early warning systems are not a new phenomenon. Just as those that are deployed for catastrophic weather and natural disasters, early warning systems for businesses should be launched to warn of disruptions to the orderly management of the strategies and tactics of enterprises that ultimately extract, originate or capture value.

Integrating this information into a meaningful early warning system requires a new way of examining information. In the Petabyte Age of ubiquitous and proliferating data, the integration of information must be done immediately, or else the value of such integration is worth significantly less than when it was initially exposed.

Several years ago, computer scientists discovered that code was more nimble if it was decoupled from its underlying model, which gave rise to the SOA and REST architectures; similarly, a process can decouple the modeling of data from the ability to publish alerts, dashboards and access to consumers. This post-discovery means of utilizing data has been written about by Forrester and others and is the basis of many advanced tools in the marketplace today. The reason for such an approach is to discover anomalies prior to the normal publication cycle.

A number of technical solutions are emerging to deal with publishing data at a moment's notice. Most of these solutions are covered under the topic of "virtualized data warehouses," which will be covered in a separate whitepaper. Momentum for virtualized warehouse technology has picked up, as all vendors in the space have positioned themselves to offer "perfect solutions."

Stages of Information Management

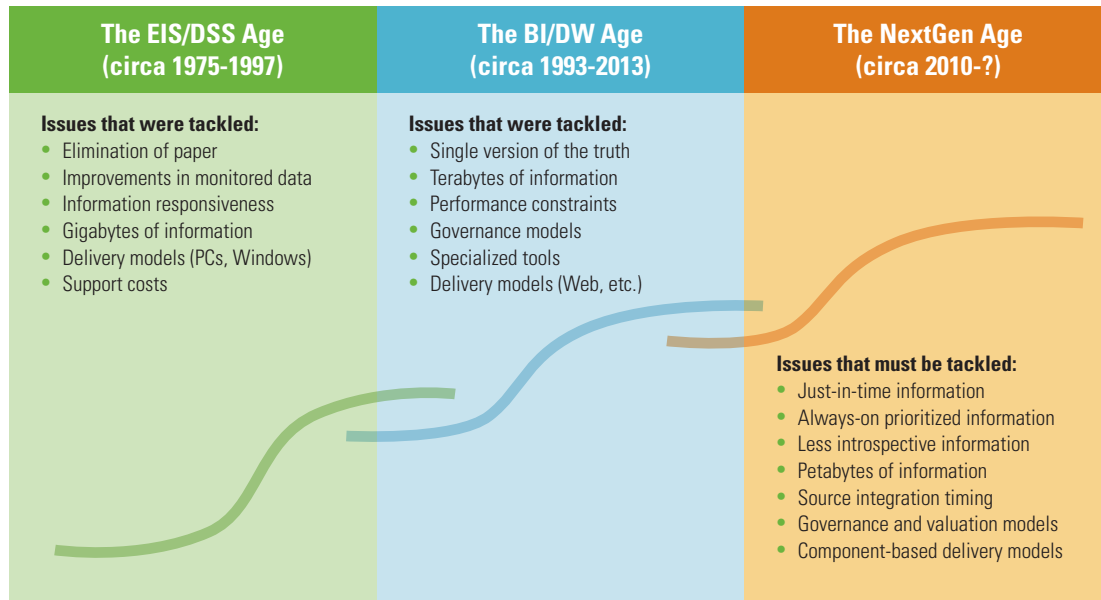


Figure 4

A Framework for the Petabyte Age

Roughly every 15 to 20 years, the disciplines of delivering enterprise information for creating business-critical insight and improving the overall decision-making process undergo radical change (see Figure 4). We are in the midst of such a major shift. These cycles tend to share the following characteristics:

- They are ushered in with the availability of tools that are greatly reduced in price or are open source and displace much of the functionality of the products being replaced (e.g., in the late '90's, such products like Pilot and Comshare were displaced by market upstarts like Javelin and Excel).
- There are referenceable cases of enterprises that have successfully utilized next-generation solutions for translating raw data into insight.

Challenges that must be tackled as part of this next-generation age are:

- The ability to deliver prioritized, just-in-time information through an always-on interface (i.e., mobile).
- The ability to combine information generated inside the organization (introspective) with information made available elsewhere. It is important to note that information made

available elsewhere rarely comes in neat bundles of tables that are easily integrated using readily available scripts.

- The ability to integrate new sources of information at a moment's notice. This requirement challenges the basic tenets of the enterprise information model and ETL processes that have matured over the past 20 years.
- The ability to embrace changes (i.e., additions and deletions to the information fabric used to steer, organize and ultimately produce enterprise value by proving that the technology arm can responsively deliver trustworthy information). Disciplines such as process governance, data governance, information centers of excellence that manage a catalog of components and information lifecycle management³ are enjoying renewed popularity because they are cornerstones of this renewed responsiveness to the knowledge worker community.

What is important in the new disciplines associated with insight generation is that they are centered on focusing on information, whether or not it is traditional, internally sourced information. Many of the information sources will require techniques associated with big data (billion-plus row tables), but all of it will require assistance in

focusing on the information dilemma for the foreseeable future (i.e., finding which information is critical for a specific business need is much akin to finding the proverbial needle in a haystack).

Much work has been done to create an information lifecycle for managing performance of analytical and operational systems. However, partitioning strategies have rarely been relegated to partition information into the following schemes:

- Information that is directly attributable to generating or protecting revenue for an enterprise.
- Information that may not be strategically or tactically significant to generating revenue but is mandatory for business operations. Much financial data (not financing, which is often a cash position) falls into this category.
- Information that may not fall into the above two categories but is required for regulatory purposes.
- Information required for archival purposes.
- Information that may have once fallen into the above categories but has been relegated to historical trivia.

The process of partitioning information into areas deserving focus (called “focus partitioning”) is completed by determining the following:

- **Step 1:** Taking inventory of information used in the organization. Information will come from one of five categories:
 - Downloaded and enriched through processes managed entirely from desktop systems.

- Available in official operational systems.
- Available from unofficial operational systems (normally Microsoft Access and Excel).
- Introspective but document-centric information (contracts, e-mail, etc.).
- Information that is sourced outside the organization (social media, blogs, newswires, etc.).

- **Step 2:** Create an information component inventory, assigning each information component to a segment of the business information model and determining its priority in generating value to the organization. Also, identify information that is required but not available as part of this exercise.
- **Step 3:** Assign the information inventory to the partitions of the business information model (i.e., directly contributing to enterprise value, required for operations, etc.).
- **Step 4:** Align potential initiatives with the partitioned information inventory and determine the impact to improving enterprise value by tackling these initiatives, thereby creating a roadmap to this prioritized information fabric critical to capturing, extracting or originating enterprise value.

It is important to note that as much as we think that the business stakeholders don’t have the data they need to perform their job, in reality there is always a means to obtain and utilize information required for determining and executing on the strategic, tactical and operational needs of the

Template for Capturing, Aligning Information Components

Information Component Data Entry						
Information Component (from Data Inventory)	The reason this information made the data inventory (this list)	Priority	Primary Issue	Secondary Issue	Availability	Focus Area
						Maintain Customers Obtain Customers Maintain Capabilities Obtain Capabilities Maintain Financing Obtain Financing Financial Excellence Technology Excellence

When capturing the focused information that is used in a big data initiative, it is important to align the data back to the business information model. The template above is a vehicle that can be used to capture the focused information exposed through a big data initiative and ensure alignment and proper placement in the business information model.

Figure 5

References

Mark Albala, "Enhancing Agility: Enabling Information Intelligence for a Turbulent World," 2010.

Mark Albala, "Post Discovery Intelligent Applications: The Next Big Thing," 2009.

Mark Albala, "Information and Execution Agility: The New Imperative," 2009.

Boris Evelson, "Information Post Discovery - Latest BI Trend," blog post, Forrester Research, May 18, 2009.

About the Author

Mark Albala is Practice Director of Cognizant's North American Enterprise Information Management Consulting and Solution Architecture Practice. This practice provides solution architecture, information governance, information strategy and program governance services to companies across industries and supports Cognizant's business intelligence and data warehouse delivery capabilities. A graduate of Syracuse University, Mark has held senior thought leadership, advanced technical and trusted advisory roles for organizations focused on the disciplines of information management for over 20 years. He can be reached at Mark.Albala@cognizant.com.

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 50 delivery centers worldwide and approximately 130,000 employees as of September 30, 2011, 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
Email: inquiry@cognizant.com

European Headquarters
1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102
Email: 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
Email: inquiryindia@cognizant.com