Understanding the Information Architecture, Data Management, and Analysis Challenges and Opportunities of the Internet of Things

As the hodge-podge of IoT's connected and instrumented devices reaches maturity, organizations need a robust enterprise information architecture to collect, manage and analyze its rich, real-time data. Here's how to get started, with a framework, implementation strategy and use cases.

Executive Summary

We live in an age of information explosion, driven by technology that has progressed from monolithic mainframes, to distributed computing, to on-premises and hybrid distributing computing, and towards multi-tenant cloud environments. Further, a new paradigm has emerged where connectivity has a multitude of channels – mobile, tablets, sensors and monitors – that yield an abundance of data and intelligence about those devices and their users.

This interconnected world of disparate devices, communication and transmission of large volumes of data across various formats is collectively referred to as the Internet of Things (IoT). The great promise of IoT is that it will allow these devices ("things") to provide data about themselves that can be communicated and controlled remotely – even automatically. This allows for more direct integration between computer-based systems and the physical world.

Many have referred to IoT as the third wave of the Internet's evolution, moving beyond today's widespread mobile access that connects several billion people, and on to a massive new world of tens of billions of connected sensors and devices,[1] according to Gartner Inc. research. These can range from smart refrigerators, thermostats, personal fitness equipment and cars, and on to heart monitoring implants, railroad safety monitors, field operation devices in a variety of industries and even "smart cities." It is a future that will require the collection, storage and real-time analysis of vast amounts of machine data across various formats.

The question is, how can this powerful new technology be practically and profitably set up and deployed to benefit companies and customers? The

IoT Framework

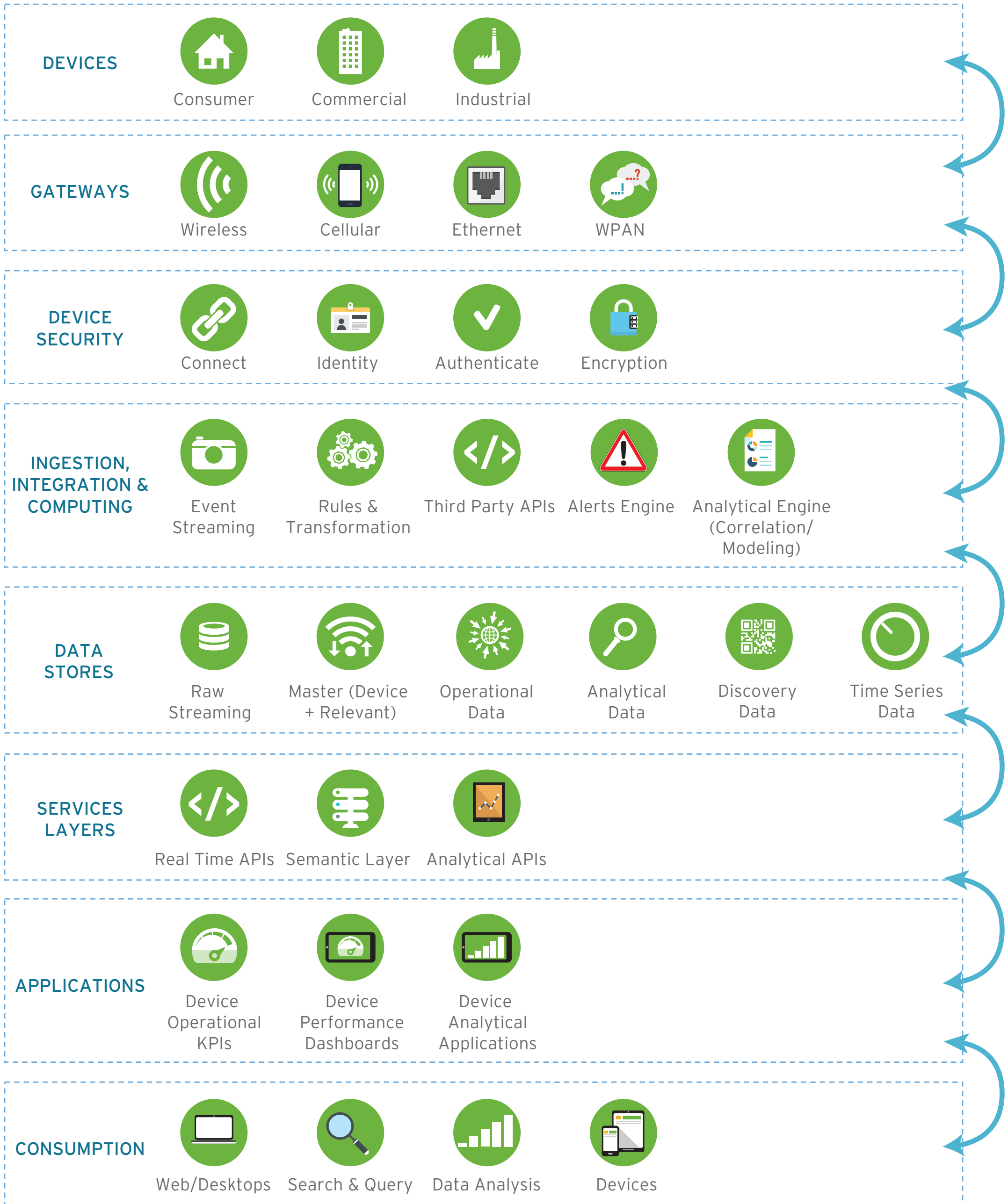

Figure 1

purpose of this white paper is to provide insights on IoT, and to provide a blueprint for practical courses of action. We'll look at the following:

- An IoT framework, as it's key to building blocks and capabilities.
- Representative use cases by industry.
- An IoT maturity model, which is critical in gauging adoption approaches.

The IoT Framework

Most of IoT is machine-generated data, so it's useful to think of it as a large-scale information architecture with complex spatial data, extremely fast speeds of data movement and numerous data sources.

Device Types

Three types of device data can be comingled with enterprise data sets to achieve IoT insights. They include data from the following:

- **Consumer home devices:** These include items used on a day-to-day basis by consumers, such as appliances, meters and sensors that monitor and frequently communicate such things as light and temperature.

- **Commercial-grade devices:** This covers such industries as automotive, healthcare, electronics, high-tech and med-tech, where devices transmit data based on consumer interaction and usage.
- **Industrial-grade devices:** These include devices that assist in critical business operations for security, operations, logistics and control. Examples here are healthcare diagnostic machines, manufacturing equipment, transportation logistics, cameras and sensors.

Connectivity

IoT devices can connect to the network using Bluetooth, cellular, Wi-Fi or a hardware connection, sending messages using a defined protocol. One of the most popular and widely supported protocols for IoT applications is message queue telemetry transport (MQTT), but plenty of alternatives exist, including constrained application protocol, XMPP and others.

Security

With the exponential increase in connected devices interacting and exchanging data with each other, security solutions are likely to multiply. There is a need to ensure that communication flows are authentic and authorized, enabling system and device manufacturers, as well as service providers, to integrate the right level of security without compromising the user experience. Here, it is critical to create layers of security implementations, integrity checks, authentication and secure key management at the device level. Extremely important is the right level of encryption and tokenization[2] to securely transmit hack-proof sensitive data. This will become ever more important as IoT matures.

Standardization

In a true IoT system, diverse devices and systems share information and interact across devices and business applications. However, industrial control today is dominated by proprietary interfaces and equipment designs. Bridging these devices will require some form of standardization of messages, data and delivery formats without disrupting the key functioning of the devices.

Ingestion and Integration

Machines and devices are not traditional IT systems. In order to realize the full potential of IoT, they will need to be configured to produce data themselves, and not merely the other way around. Integration technology needs to adapt as well, making sure that it can deal with streaming and unstructured data, including many instances where data needs to be processed "in flight" as it moves from a particular device to data repositories. And contrary to classical enterprise integration, IoT integration is based on time-series processing and data correlation logic, along with timely data synchronization. This requires a type of integration where correlation of device data with other device data leads to immediate notifications. Only with this kind of integration can users take tactical and strategic actions informed by IoT intelligence.

Contrary to classical enterprise integration, IoT integration is based on time-series processing and data correlation logic, along with timely data synchronization.

Typically, IoT data sources feature velocity and volume thousands of times greater than social media sources. To substantiate this hypothesis, if we take an order of magnitude of even 10 billion devices, each generating millions of events per second in click streams, logs, sensory data and other forms of device data, compared to millions of responses to posts/tweets on social media per day (that also depends on the number of posts, which usually don't go over five), we can gauge the disruption IoT brings to the table. It is often too big to fit in memory – and most types of IoT data analysis are not summarizations that allow records to be discarded – which precludes the use of NoSQL[3] database platforms or in-memory databases. Thus, a distributed platform is needed, one that can reliably process and store many gigabytes per second. Many valuable IoT data sources individually generate tens of gigabytes of complex records per second without interruption, and many applications of that data combine multiple data sources. These records must be parsed, processed, indexed and stored at massively fast transmission rates if they are going to be analyzed in real time or near real time.

Data Layer: Complex Spatial Data Models and Analysis

A characteristic of most IoT data is that it captures measurements of the real world. Most of this data is sourced automatically from smart objects instrumented with sensing, computing and communication capabilities. Events in data streams can be correlated and contextualized across diverse data sources, based on when and where they happen. The data typically involves complex geospatial geometry, such as the paths people take or the interactions of different types of sensors. Many of these spatial data types are complex, and the analytics are frequently spatial-join operations across these data sources. Spatial joins involve ways to link disparate data via context, semantics or other probabilistic discovery mechanisms compared to a deterministic approach in relational database management systems.

It is important to identify the characteristics of a database that make it suitable for typical IoT applications. Requirements here fall into these general categories:

- **Device master data repository housing different types of devices,** as well as necessary relevant information that can be integrated with other data repositories to gain insights.
- **Continuous machine-scale ingestion, indexing and storage**: Even a modest data source may generate millions of complex records per second on a continuous basis, which usually can continuously stream into a data landing zone for storage and processing.
- **Operational (real-time) queries and analytics, which extract value from IoT data.** This is all about minimizing the latency (time lag) from data ingestion to online queries and actionable analytics. For many applications, the value of the data is highly perishable, with an exponential decay measured in seconds. IoT queries and analytics are rarely summarizations, stream processing rarely works and there is the need to support ad hoc queries in something like a SQL interface. Depending on the use case, these queries can be merged into

Potential IoT Uses Across Industries

Industry	Representative Use Case
Manufacturing	A manufacturing company can use all the data generated, processed and gathered from IoT devices not only to implement manufacturing lean principles but also to fine-tune methodologies, concepts (including Six Sigma), processes and strategies to finally achieve maximum output with minimum input. Instances where enablement can be provided include: • Real-time operational KPI monitoring of machine diagnostics for performance, breakdown and timely maintenance extending its usability and throughput. • Provides 360-degree visibility into shop floors, supply chains, warehouses and distribution, delivering real-time data streams that can be used to identify new patterns, optimize processes, gain and maintain complete operational control and drive new levels of efficiency across the manufacturing industry and adjacent sectors.
Insurance	Many automobile insurers can gain added insight into the driving habits of their customers. Through the use of smart devices within customer vehicles, insurers now have access to a breadth of data that will allow them to provide more personalized service while simplifying their processes. By combining diverse spatial data on vehicle speed, road conditions, accidents, driving distance, time of day, weather conditions and vehicle make, insurers are able to build new offerings, improve services and provide usage-based plans for better risk coverage and smoother claims processes. Risks can be reduced with more timely and accurate data.
Healthcare	Health risks can be averted, and costs contained, via remote patient monitoring of wearable devices for vital conditions, with data streamed quickly to provide timely insights into patient progress. If there is a need for immediate medical attention, real-time notifications can be sent to the nearest hospitals or pharmacies.

Figure 2

IoT's Evolution, from Inception to a Well-Optimized State

Level	Description	Causes
0: Use-Case-Based Pilot Integration	Based on a use case, device data will be integrated with the enterprise ecosystem with the right tenancy (on-premises/ cloud).	Synergies of device data are not yet established and opportunities for seamless integration are still in discovery mode.
1: Stabilization	The ecosystem of device data and enterprise data has been harmonized with some level of repeatable synergy successes. Some multi-tenancy options are in the exploration stage.	Device data has established some level of integration with enterprise data, with optimization in latency, storage and analysis.
2: Standardization	Device data has been standardized in acquisition, integration and consumption patterns. Coexistence with enterprise data repositories is in place, with tenancy guidelines.	Repeatable synergies and learnings ensure that there are defined standards and norms for onboarding device data from acquisition to analysis. This ensures timeliness of the onboarding-to-analysis process. There is some interoperability among various resources and data providers and consumers.
3: Optimization	Onboarding, availability and consumption of all device data across NoSQL, master data management (MDM) stores, enterprise data platforms and other types of data platforms are available without much delay, to address operational and strategic insights on devices.	Organizations are continuously measuring effectiveness of nonfunctional service level agreements for device latency and operational metrics, and are discovering newer insights to enhance operability. Optimizations in effective data access, integration and analytics is accomplished.
4: Governance	Agile governance to manage technology, data and analytical paradigms for newer devices, their operational metrics and analytical metrics are well orchestrated with well-defined guidelines for tenancy.	Device data is a governable asset and has a defined set of processes and procedures, especially when it comes to managing device data assets with the right architectural patterns and tenancy decisions. More increased focus is on resource discovery, reasoning and knowledge extraction on existing and new devices.

Figure 3

one type of database, or be kept separately in their respective work areas.

- **IoT data is all about spatio-temporal relationships and join operations.** To support speed and scale, there is the need for a true spatial database for normal complex operations, or a true time-series database for very simple uses.
- **Supporting data platforms for discovery.**

Consuming Patterns for IoT: Real-Time Operational Queries and Analytics

IoT implementations typically require timely queries of live ingested and historical data. The resulting analytics are not summarizations of data sets or simple event graphs, nor are they stream processing. This is real time in the sense of an online transaction processing (OLTP) database, without the complex transactions, and requiring much greater scale. The challenge of typical IoT architectures is not that different from other technologies. The issue revolves around finding components for the architecture that weave together the above capabilities simultaneously.

Potential Use Case Variants

Many excellent use cases and technology patterns abound across industries, several of which are good candidates for an IoT implementation. Figure 2 (on page 4) details some of the possibilities. In each case, potential technology considerations include advanced analytics, in-memory (high-speed computing), real-time ingestion, data layer and semantic standardization of devices via APIs and ontologies.

Assessing IoT Maturity

The various steps in maturity of an entire IoT ecosystem involve realizing the optimal synergy among such elements as devices, sensors, networks, data repositories and standardization APIs, all with seamless integration, and offering the ability to serve different types of analytics. This synergy is illustrated by the maturity progression shown in Figure 3 (on preceding page).

In dealing with large volumes of distributed and heterogeneous IoT data, issues related to interoperability, automation and data analytics will require common description and data representation frameworks and machine-readable and machine-interpretable data descriptions.

Making Sense of Critical Technology Intersections

There are numerous data management technologies that must be coordinated to enable proper IoT functionality. Figure 4 details what these technologies can accomplish and future potential considerations for optimizing IoT deployments.

IoT Technology Challenges/Solutions

Technologies	IoT Consideration
Real-Time Ingestion	• Continuous streaming of IoT device data in raw form. Potential technology candidates for this category include open source: Apache Kafka. • Near-real-time event processing with some transformation, filtering and decision-driving rules. Apache Kafka or Flume with Yarn and HBase are potential candidates for this category. • Complex event processing with correlations and aggregations at ultra-low latency. SPARK with HBase and HDFs usually work better in this space.
Data Repository	There are several options for storing IoT data. Depending on the needs for throughput, latency and volume to add new event data types, NoSQL for the lower latency and higher throughput and HDFS for batch mode analysis can be considered. Time-series databases are also gaining popularity due to latency-based analysis having high performance and clustering demands.
In-Memory (High-Speed Computing)	Operational intelligence for IoT requires a computing platform that can store, update and continuously analyze data sets representing dynamic real-world entities or business assets. In-memory computing, which can perform these functions at scale and with extremely low latency, provides the computing power required.
Advanced Analytics	Here, there is the need for a scalable machine-learning library consisting of algorithms and utilities - including classification, regression and clustering – to perform predictive analytics on large sets of device data. Depending on the scalability and provisioning needs, the same could be done in a cloud environment. In-memory analytics technologies such as SAP HANA and SPARK[6] are potential technology candidates here.
Semantic Integration	To achieve IoT standardization, organizations will need a more intelligent way to enable new devices to be recognized and profiled and to be able to transmit data that can be consistently interpreted. A semantic model enabling rapid onboarding via right device ontology and evolving rapidly without much overhead will provide value here. In dealing with large volumes of distributed and heterogeneous IoT data, issues related to interoperability, automation and data analytics will require common description and data representation frameworks and machine-readable and machine-interpretable data descriptions. Data annotations and semantic descriptions can be used at different levels, and semantic annotations can be applied to various resources in the IoT.

Figure 4

Looking Forward: Recommended Approach

To leverage the IoT's virtues, organizations need an implementation framework informed by best-of-breed use cases, influenced by a strategy guided by continuous maturing, from technology selection through implementation and testing and ongoing refinement and governance. To achieve these goals, we advise organizations to:

- **Consider best-of-breed use cases.** Each industry is unique; they rarely rely on the same types of platforms. Moreover, most big data management platforms are unable to accommodate the scale and required real-time speed of IoT. Custom implementations, each built with specific technologies, are typically required to bring IoT to its most effective maturity.
- **Determine agreed-upon standards of connectivity and security.** This is necessary to ensure a viable IoT future, one that can communicate and collaborate rather than exist in ecosystem silos. Assess appropriate technologies for particular uses. Integration and coexistence of technology, platforms and locations is critical since there is no one technology or platform that can solve all IoT challenges and requirements. Further, each of the particular functions of an IoT implementation – data gathering, storage, high-speed computing and analytics – requires unique sets of technologies best suited to the task. Start small. There is a natural progression in an IoT implementation, which should start with a pilot use case. (For more, read "Transcending the Hype: A Transformative IoT Emerges.") Maturity and success will develop as device data is added to the enterprise ecosystem, and will progress through optimization and agile governance.

Within the discipline known as the Internet of Things, the opportunities inherent in real-time data gathering, analysis and action are abundant. Those companies that stake out an IoT position in their respective industry sectors will find themselves ahead of the competition in product development, customer service, risk avoidance and predictive analytics.

Footnotes

[1] http://www.gartner.com/newsroom/id/3165317.

[2] A process where a sensitive piece of data is substituted by its nonsensitive equivalent to prevent misuse of confidential information.

[3] A NoSQL database environment is, simply put, a non-relational and largely distributed database system that enables rapid, ad-hoc organization and analysis of extremely high-volume, disparate data types.

[4] A high-volume, low-latency throughput open source message brokering engine for real-time data feeds.

[5] A column-oriented database more suited for sparse data sets simplifying storage and performance needs for data querying and analysis.

[6] An open source engine that combines SQL, streaming and complex analysis at high processing speeds.

About the Author

Ajay Raina is a Principal Architect (Director) within Cognizant's Analytics and Information Management Practice. As a key leader, he advises banking, financial services, healthcare and life sciences clients, among others, on enterprise information management strategies. His forte is establishing stability, optimization and modernization of enterprise information architecture for data management and analytics initiatives. In pursuit of enterprise information management excellence, Ajay provides strategic oversight, thought leadership, delivery guidance, technology enablement and solution definitions, blending in leading and proven practices in information management initiatives. He has 20-plus years of information management experience in leading data warehousing, MDM, big data and analytics engagements. Ajay can be reached at Ajay.Raina@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 100 development and delivery centers worldwide and approximately 221,700 employees as of December 31, 2015, 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.

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

© Copyright 2016, Cognizant. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the express written permission from Cognizant. The information contained herein is subject to change without notice. All other trademarks mentioned herein are the property of their respective owners.

Codex 1824