Untangling the Internet of Things

Driving Value with Treasure Data

February 2014

A White Paper by

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

The interest over the last few years in big data is rapidly moving to a subset of such data—that generated by the Internet of Things. The aim is to place ever smaller and cheaper sensors in almost every object conceivable and connect them all through the Internet in order to monitor and predict physical reality. This apparently simple idea will revolutionize business and IT support needed in the next few years. While the current hype is undoubtedly huge, the potential disruption if even 20% of the hype happens is enormous.

Although much of the popular market attention has focused on the devices, the real monetary value emerges from the business' ability to effectively use the data they generate and on IT's skill at capturing, storing and managing it. The first section of this paper broadly explores the uses and challenges of the Internet of Things, drawing on examples of the business uses that are emerging, while the second section positions the data processing needs in the context of traditional data warehousing, big data and cloud computing.

The third section of the paper positions a new managed service cloud offering, Treasure Data, which is structured rather like a traditional three-stage data warehouse, but is optimized for many of the aspects of data from the Internet of Things. Finally, the paper provides some initial considerations and next steps for the journey to gaining real benefit from this emerging phenomenon.

Contents

3 *The value is in the data—business view*

4 *The devil is in the (data) detail—technical view*

6 *Treasure Data as a hybrid solution to IoT data*

7 *Next steps and conclusions*

Sponsored by:
Treasure Data Inc.

www.treasuredata.com

The hype machine for the Internet of Things was running at full throttle at the 2014 Consumer Electronics Show in Las Vegas. From wearables that monitor your (and your dog's) every vital sign to in-vehicle computers that report on your driving skills, from iBeacons that track your precise in-store location to home appliances that save you going to the store at all, manufacturers have found ways to add intelligence and communications to almost every inanimate and animate thing that exists.

Welcome to the Internet of Things (IoT), a simple, decade-old idea that technology has only now made feasible. It is a wirelessly interconnected network of physical Things / devices that send data about one or more environmental variables either automatically or in response to requests. Some sensors are so small that Australian scientists are attaching them to bees to track their movements and investigate colony collapse disorder. One is ingestible and powered by your stomach acids, sending nothing more than an ID to confirm it has been swallowed, along with your medicine. Others are sophisticated miniature computers, broadcasting extensive data about complex operating environments like jet engines. In overview, the IoT is simply physical objects containing sensing machines connecting to control and collection machines, exchanging data about their environments. While some observers equate the IoT and M2M (machine-to-machine) concepts, I believe it makes more sense to see M2M as referring to the underlying communication mode, whereas IoT is more a vision of fully instrumenting the physical world, recording its states and events, in order to better understand its behavior and predict or manage its future states.

A number of studies in 2013 offer indications of the extent of this phenomenon. Gartner predicted[1] that by 2020 there will be 30 billion devices, mostly product-based sensors, with unique IP addresses connected to the Internet, up from the 2.5 billion mostly cellphones and PCs in 2009. Cisco Systems' consulting arm, together with Beecham Research, put the 2020 number at 50 billion[2]; 10 billion—about 1% of all the objects in the world—are already connected. IDC suggested[3] that there "will be approximately 212 billion 'things' globally by the end of 2020…includ[ing] 30.1 billion installed 'connected (autonomous) things'". From a financial perspective, McKinsey predicts[4] the economic impact of the Internet of Things will be $2.7 to $6.2 trillion per year by 2025. Gartner opts[1] for a total economic value-add of $1.9 trillion dollars in 2020.

The Internet of Things will drive significantly larger volumes of data into businesses, requiring extensive analytics to extract real value.

These ranges of figures indicate it is too early to predict the endpoint for IoT. Their sizes, however, indicate the importance of the concept. And Google's recent acquisition of Nest, a maker of smart smoke alarms and thermostats capable of re-programming themselves based on people's behavior, for $3.2 billion speaks volumes. Larry Page's anodyne comment that Google *"are excited to bring great experiences to more homes"*, seems designed to avoid the thought that extensive ambient data collection by such devices might be among Google's key interests.

And focusing on data, such devices are going to drive an even bigger explosion of data than we've seen so far from social media sources, with the ability of these devices to continuously take measurements at ever shorter intervals. Already, some utility companies are measuring consumption at 15-minute intervals—a 3,000-fold increase over manual meter readings. This explosion of data, in turn, leads to the need for extensive analytics to drive the expected innovative opportunities for business, both traditional and emerging.

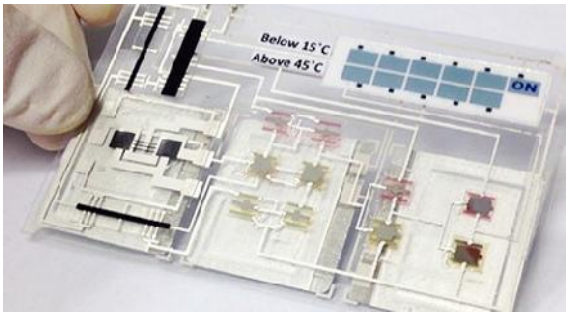
The value is in the data—business view

While much of the above hype is around the Things (on the Internet) themselves, the real commercial value is to be found in the data that the Things produce and the innovative applications that use that data to decrease costs, increase revenues or attract/retain customers. More broadly, societal value emerges from the potential to better understand, predict and control real-world physical phenomena from personal health to climate change. A few examples will make clear both the envisaged benefits and the longer-term unanticipated and highly disruptive impacts.

The commercial value of the Internet of Things lies in the data produced and the innovative business applications it enables.

A well-established IoT ecosystem has already sprung up around the sensors that manufacturers of trucks and automobiles have been building into their vehicles for some years now. Originally designed to ease and predict maintenance, these on-board diagnostics (OBD) sensors include tachometers, thermometers, fuel gauges, accelerometers, pressure gauges, and vibration sensors, to name but a few. While maintenance is, of course, interesting to some, the real value emerged when it was recognized that combining such data with GPS data and other public or business data offered new business opportunities. Logistics can begin to re-route deliveries in real time to minimize fuel costs and optimize for weather, traffic conditions, high value customers and more. Auto insurance companies can reduce claims exposure by offering new products to reward safe driving as measured by speed, sharp maneuvers, sudden braking, etc. At the extreme, fully automated, driverless automobiles are made possible (or inevitable?) by the array of sensors they embody and the data they collect—overturning the operating models of auto repair shops, insurance companies and even hospitals.

In retailing, sensors in products, in stores and on the customers themselves (smartphones are the pioneering species of the Internet of Things) first drive deep integration of clicks and mortar. Packaging can detect when food is spoiling and alert staff or customers to avoid health issues. Shelves can monitor stock levels and trigger automatic reordering via the supply chain system. Shopping patterns can be tracked individually, interest levels determined via facial analysis, and personalized offers made to specific customers. Items placed in the shopping cart can be noted and payment taken on exit without physical scanning at the checkout. At home, sensors in refrigerators and cupboards continue the monitoring and alerting process. The end-state could fundamentally disrupt shopping patterns as regular physical trips to the grocery store become unnecessary. However, significant security and privacy concerns must be addressed by device manufacturers and providers of connectivity. A recent news report[5] highlights that *"these days, even a refrigerator can fall prey to a cyber attack"*.



Health care is emerging as a focus area for reaping the benefits of the IoT, with the efficiency of treatment for chronic conditions a prime target. Sensors monitor the vital signs of at-risk patients at home and alert doctors and nurses to emerging problems, enabling proactive and less expensive interventions. McKinsey estimates[4] the global cost of chronic illness could be over $15 trillion annually by 2025, a figure that could be reduced by 10-20 percent with remote monitoring. Individual responses to medications, as well as compliance to prescriptions, can be tracked, allowing personalization of treatment, eventually at a genomic level. While reducing health care costs and improving quality of life are key drivers, there are potential downsides. Privacy also becomes a key concern here; the collection of personal, health and behavioral data in the home opens up the possibility of a wide range of abuses of individual freedom. The business of health insurance is open to disruption.

The breadth of impact of the IoT should be evident from the above examples. In each case, the value emerges from the data collected and the novel uses to which it is put. Every industry stands to be affected. In addition to the areas mentioned, Cisco[6] identifies the following industries or functions as major beneficiaries of the IoT:

The Internet of Things has the potential to transform business processes in every industry and affect people in every walk of life.

- *Manufacturing:* the creation of smart factories through the addition of connectivity to machines and applications increases productivity, reduces inventory and cuts supply-chain costs. (This shows IoT applied within the enterprise.)
- *Marketing:* location-based services built around smartphone data in public and within stores allows individually targeted offers and messages at the right time and place to optimize impact
- *Utilities and communities:* the implementation of smart grids and smart buildings allows ongoing monitoring and optimization of resource delivery, both by the producers and consumers, improving reliability, sustainability and economics of use

Other areas include transport logistics and healthcare, as previously mentioned. Also according to the same report, the five factors driving value are (i) increased asset utilization and cost reduction, (ii) employee productivity, (iii) supply chain and logistics efficiency, (iv) improved customer lifetime value and (v) new revenue streams and reduction in time to market.

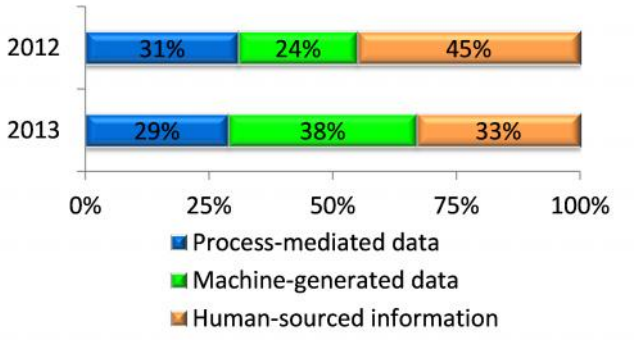


Figure 1: The current rise in importance of machine-generated data

Of course, in the face of such change, individual companies stand to gain royally or lose dramatically. Understanding the role and extracting the value of the data generated by the IoT will determine in which category you eventually emerge. And if you think that this is all too futuristic, the 2013 big data survey conducted by EMA and 9sight[7] showed a significant switch in percentage volumes from human-sourced information to machine-generated data in a single year. Both terms are defined more fully in the next section, but we note that IoT data is, by definition, the externally generated subset of machine-generated data. While the survey did not probe the difference between these two classes, the 12-14% switch shown in Figure 1 is likely to derive from IoT devices, given the ongoing market focus.

The devil is in the (data) detail—technical view

The machine-generated data produced by the Internet of Things poses a number of challenges to any organization that wants to extract value from it. The now traditional 3 Vs—volume, velocity and variety—are, of course, factors that must be considered. But they do not really help us to understand the challenges. A simple approach is to consider, first, how machine-generated data differs from the traditional process-mediated data used to run and manage business and, second, how this IoT data differs from the social media and related data (which I call human-sourced information) that has been the focus until now of the so-called big data movement.

Process-mediated data has traditionally been generated and collected by traditional business processes such as buying an item or cashing a check. Such data is characterized by its well-defined and long-lasting meaning and structure, and its known, internal sourcing. It is the basis for the operational applications and business intelligence / data warehousing architectures and technologies we have used for decades. Deeper thought about such data shows that it comes from two sources: (i) people entering it on keyboards and (ii) machines, such as ATMs, telephone exchanges, and more, generating

it as a byproduct of human actions. As we've moved into the era of big data, both of these once-unseen, ultimate sources of data have become recognized as direct and important. Human-sourced information, the current focus for much of predictive analytics, comes directly from external (for the most part) social media sites and is very loosely defined in both structure and content.

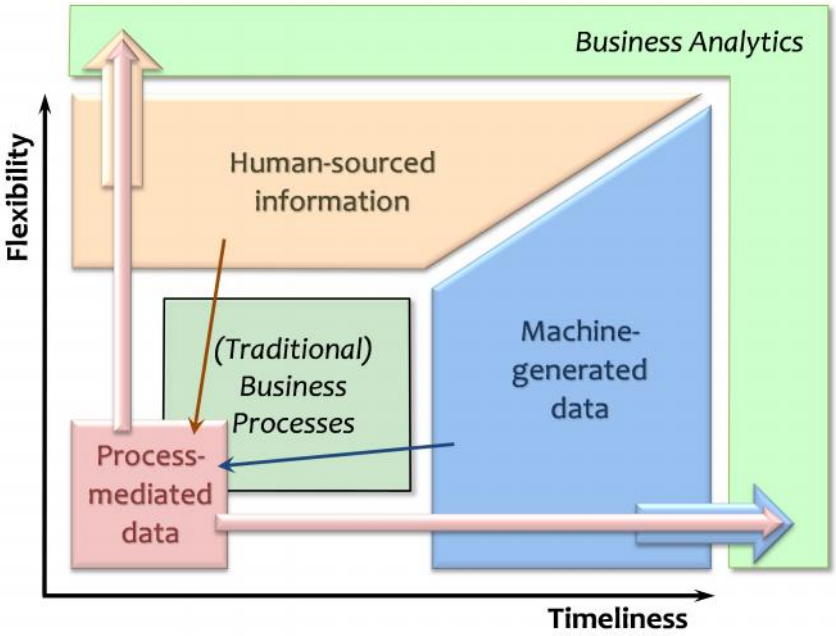


Figure 2: The tri-domain information model

The relationships between these three types of data are shown in Figure 2. This plots these data types against two key characteristics of data, timeliness and flexibility, which determine how we can process and analyze data. The two solid-line arrows show how business processes and applications traditionally create our managed representations of what is happening in an around the business. Human-sourced information and machine-generated data, which we seldom noticed or stored in the past are stored in transactional and business intelligence databases with reduced flexibility and timeliness. However, as business accelerates and increases focus on external data, the more flexible and timely human-sourced information and machine-generated data become ever more important and are thus physically stored, often in non-relational stores such as Hadoop or NoSQL stores. The thicker arrows emphasize that modern business analytics must combine these three data types to ensure that the quality and consistency of process-mediated data is applied to the flexibility and timeliness of the new data sources. Further discussion of these important issues can be found in my new book, *"Business unIntelligence—Insight and Intuition Beyond Analytics and Big Data"*[8].

Machine-generated data is now arriving in increasing quantities and speeds from the IoT, in addition to internal machine sources. Its characteristics—which determine how we must store and process it—are:

1. Loosely or only partially defined in meaning when it arrives in the enterprise

   Because Things are often created and deployed by third parties, the data they provide needs to be understood (or modeled) in the context of the enterprise wishing to use it, including what is being measured, the units of measure used, and any relevant limitations.

2. Well-structured, typically in a name-value format, but changeable in both names and value ranges

   As new Things emerge or existing Things are upgraded, new or changed data becomes available. Systems receiving such data must be highly flexible and agile in adapting to such changes, often by maintaining the data in name : value pairs as long as possible.

3. Sourced in high volumes and high speed externally to the enterprise[9]

   The number of Things and their ability to send messages at high rates, even second by second, demands a powerful and expandable receiving environment.

4. Analytical processing is the primary means of extracting value

   In general, individual IoT data instances are of limited interest; analytics, from operational to predictive are needed to make sense of the data and drive valuable decisions and actions on its basis.

These characteristics offer a challenge to the implementer of an environment to gather, manage and process IoT data. On one hand, this relatively well-structured data resembles process-mediated data, especially that in a data warehouse environment, suggesting traditional relational and BI technology. On the other hand, the volumes and variability drive thinking towards big data tools like Hadoop and NoSQL. Furthermore, the experimental nature of the uses and external sourcing of IoT data demands consideration of cloud solutions for both agility and positioning. As a consequence, both traditional relational vendors as well as newer big data tool vendors are offering solutions in this space. However, one can envisage that a hybrid approach might offer the best of both worlds.

The name-value nature of IoT data makes it amenable to more traditional analytics, although its variability in structure can be a challenge in a strict relational model.

Treasure Data as a hybrid solution to IoT data

In response to the analytical nature of processing and the relatively well-structured data involved, Treasure Data is a managed service cloud offering that is internally organized very much along the lines of a traditional three-stage data warehouse. First, data is acquired from the relevant sources. Second, data is stored in a robust and flexible environment. And third, data is made available to a variety of users through their preferred BI tools. However, a closer look at the technology shows how volume and variability needs are taken into account.

Data acquisition focused on feeds

In a traditional data warehouse, data acquisition (usually as ETL—extract, transform and load—or some variation) is historically structured around batch loading of files from the operational environment and their integration and reconciliation. Here, the emphasis is on streaming, filtering and transforming data in relatively small batches or as individual records as fast as possible from applications that collect IoT messages into the store, through a component called Treasure Agent or through an open-source version, Fluentd. This agent technology focuses on real-time ingestion of data, an approach that is particularly well suited to the IoT. Because of the stand-alone and independent nature of such messages, filtering and transformation is relatively simple, lacking the type of integration and reconciliation seen in traditional ETL. Bulk loading from a variety of more traditional sources is, of course, also supported.

Data storage focused on agility

Treasure Data has developed a proprietary data store, called Plazma, which combines characteristics of the Hadoop and relational worlds to offer a high level of agility with access to the data through standard SQL. Built in an enhanced Hadoop environment, but eschewing HDFS which is poorly structured for analytic use, Plazma is a columnar database, which stores the data in a compressed format that optimizes storage use (10-20x compression) and accelerates I/O speed by 3-4 times. It is also schema free, meaning that no prior data model need be defined, which provides the agility required for handling rapidly changing data.

The ingestion and storage functionality offered by Treasure Data are well suited to the processing needs of data originating from the Internet of Things.

To further support agility, Treasure Data is implemented entirely as a managed service in the cloud (on Amazon S3), enabling rapid deployment with implementation and maintenance support and no upfront capital expenditure. This cloud implementation is also attractive because IoT data originates on the Web, and is particularly suited to companies whose operational and other analytical activities are also cloud-based. However, more traditional businesses will have to transport internally-sourced data to the cloud environment, with attendant transport speed and cost impacts.

Data usage focused on user need and preference

SQL access is either in batch via Hadoop Hive or in real-time through the Treasure Query Accelerator, based on an enhanced cloud query engine. In addition, basic data exploration is provided through Treasure Viewer. But the main drive is to enable users to stick with the tools they already use, such as Excel or Tableau. By providing both ODBC and JDBC drivers, access is available through a wide variety of BI and analytic tools. In this area, Treasure Data differs little from traditional BI.

Next steps and conclusions

With the physical hardware and connectivity needed for the Internet of Things rapidly falling into place in 2014 and succeeding years, the stage is set for companies with viable business needs and opportunities to at least experiment with the technology and, where the returns are most obvious, to move rapidly to production. A cloud-based approach provides an ideal environment for either move, with its rapid deployment and minimal up-front cost.

Agility in every sense of the word is vital in this market. New and upgraded Things are appearing with unprecedented regularity. In many instances, the potential business cases will be unclear, although we can be certain that analytics will play a central role. Therefore, aspiring market leaders need to experiment with many possibilities in parallel, at as low as cost as possible, to succeed at speed or fail fast.

The Internet of Things is simply physical objects containing sensing machines connecting to control and collection machines, exchanging data about their environments.

The first step is to notice that many of the most successful uses of IoT data are unrelated to the initial reasons for making the measurements. Often, they bring together two or more sets of unrelated data, often collected for completely different purposes, to allow another behavior or situation to be inferred. Such correlations can be uncovered through extensive data ingestion from multiple sources followed by deep predictive analytics. The type of solution offered by Treasure Data enables this first step with a minimum of investment.

If a valuable opportunity emerges, a cloud solution based on a proven three-tier data warehousing architecture provides an easy and obvious route to expansion and a move to real production. If no opportunity appears, backing off is equally easy. A new experiment can be immediately substituted or the entire approach dropped at minimum sunk cost.

Some of the most valuable outcomes from IoT data usage emerge only when this data is combined with traditional process-mediated data from financial and other internal systems. The challenge here—technically, financially and legally—is that cloud-based and internal data must be combined. Early consideration of such potential issues is vital. Will results from the cloud-based warehouse be brought inside the firewall or vice versa? Another issue which deserves early and close attention is the impact on privacy of combining different IoT feeds and of combining IoT data with internal data. These data governance issues will be ignored at your peril. Indeed, recent exposures of excessive governmental data collection and the theft of enormous volumes of personal and financial data in the commercial field are driving significant public concern. The clear danger is that the real opportunities offered by the Internet of Things to address real-world problems from climate change to social inequality will be swept aside by such insidious and undisciplined behaviors.

On the IT side of the equation, one of the most insidious myths circulating at present is that big data and IoT data projects will be destroyed by the bureaucracy of existing BI teams and environments. While this danger does, of course, exist in some organizations, it pales into insignificance in compari-

son to the chaos that can ensue when such big data projects are built and managed as stand-alone skunk works. Although IoT data may not warrant all of the data quality processes required for financial data, for example, the fact that both types of data will end up together on users' desks demands more control and management than might at first appear necessary. The three-tier approach adopted by Treasure Data is clear evidence of the parallel between traditional BI and emerging analytics. The recommended approach is to ensure a shared responsibility and chain of command for both initiatives.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His new book, "Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data" (http://bit.ly/BunI-Technics) is published in October 2013.

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and IT implications of deep business insight solutions that combine operational, informational and collaborative environments. A regular contributor to BeyeNETWORK, TDWI and other publications, Barry is based in Cape Town, South Africa and operates worldwide.

---

[1] *"Gartner Says It's the Beginning of a New Era: The Digital Industrial Economy"*, Press Release, October 7, 2013, https://www.gartner.com/newsroom/id/2602817

[2] Infographic: *"The Internet of Things"*, http://share.cisco.com/internet-of-things.html

[3] *"The Internet of Things Is Poised to Change Everything, Says IDC"*, Press Release, October 3, 2013, http://www.businesswire.com/news/home/20131003005687/en/Internet-Poised-Change-IDC

[4] *"Disruptive technologies: Advances that will transform life, business, and the global economy"*, McKinsey Global Institute, May 2013, http://www.mckinsey.com/insights/business_technology/disruptive_technologies

[5] *"Refrigerator among devices hacked in Internet of things cyber attack"*, Salvador Rodriguez, Los Angeles Times, January 16, 2014, http://www.latimes.com/business/technology/la-fi-tn-refrigerator-hacked-internet-of-things-cyber-attack-20140116,0,5757808.story

[6] *"Embracing the Internet of Everything to Capture Your Share of $14.4 Trillion"*, Cisco, 2013, http://www.cisco.com/web/about/ac79/docs/innov/IoE_Economy.pdf

[7] *"Operationalizing the Buzz: Big Data in 2013"*, EMA and 9sight Consulting, November 2013, http://bit.ly/BD-survey13

[8] *"Business unIntelligence—Insight and Intuition Beyond Analytics and Big Data"*, Barry Devlin, October 2013, Technics Publications, New Jersey. http://bit.ly/BunI-Technics

[9] In some industries, such as Manufacturing, for example, a larger percentage of such machine-generated data originates internally.