Machina Research

Research Note

Why NoSQL databases are needed for the Internet of Things

Emil Berthelsen, Principal Analyst

April 2014

The issue	Big Data in the Internet of Things is snowballing. The amount of data is ever-increasing and becoming more and more varied. The impact on traditional relational database management systems (RDBMS) will be significant. Databases need to adopt and meet these new IoT requirements with greater data processing agility, multiple analytical tools including real-time analytics, and aligned and consistent views of the data. This Research Note examines what database capabilities will be required to address data managed in the Internet of Things, and how NoSQL systems like MongoDB, Cassandra and HBase are meeting this challenge.
Our view	***The traditional relational database management systems will continue to have a role in the Internet of Things when processing structured, highly uniform data sets, generated from a vast number of enterprise IT systems and where this data is managed in a relatively isolated manner. When it comes to managing more heterogeneous data generated by millions and millions of sensors, devices and gateways, each with their own data structures and potentially becoming connected and integrated over the course of many years, databases will require new levels of flexibility, agility and scalability. In this environment, NoSQL databases are proving their value.***

The significance of the Internet of Things is not that more and more devices, people and systems are 'connected' with one another. It is that the data generated from these 'things' is shared, processed, analysed and acted upon through new and innovative applications, applying completely new analysis methods and within significantly altered timeframes. The Internet of Things will drive Big Data, providing more information, from many different sources, in real-time, and allow us to gain completely new perspectives on the environments around us.

The difference between machine-to-machine (M2M) and the Internet of Things (IoT) is driven by significant changes in the handling of data. In M2M, machine-generated data generally reflects well-defined data sets, communicated within established protocols and formats, and delivers well-defined alerts and notifications when values exceed their parameters. Applications in M2M make efficient use of this data as these applications have been developed hand-in-hand with what the characteristics of

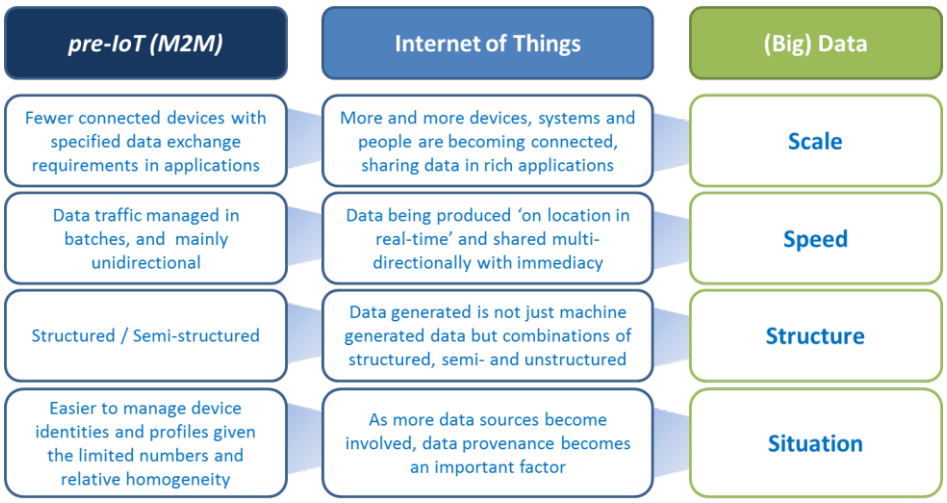
the data. In effect, application and data are intrinsically designed as one to meet the specific purposes of the application in fairly robust yet static model.

In the Internet of Things, the world of data and applications requires significantly more flexibility, agility and scalability. These requirements are driven by a limited number of underlying needs. This Research Note explores these requirements and underlying needs in more detail, including:

- **Diversity of devices and data –** Data generated from an exponentially growing number of diverse sensors, devices, applications, and things will be accompanied by a growing diversity in the structure and scale of that data – and more and more sources of additional data ranging from data sourced from corporate systems to crowdsourced data will need to be combined with this data.
- **Flexible and agile systems** – As the Internet of Things becomes a more open system where new sensors, devices, applications and things are added and connected, and 'evolve' what we term 'Subnets of Things,'[1] the supporting applications and databases that manage the data generated from these 'things' will need to remain flexible and agile, allowing businesses to meet their requirements without having to reinvent the application every time.
- **Sophisticated analytics** – Where 'simple' alerts and notifications were the backbone of M2M systems, analytics becomes the cornerstone of the Internet of Things, requiring enhanced and multiple analytical approaches to address the requirements of the applications.

It should come as no surprise that developments in the Internet of Things and databases have run parallel with discussions around Big Data, a topic explored in other publications by Machina Research.[2] The relationship between IoT and Big Data, and the changes from M2M to IoT are captured in Figure 1.

Figure 1: M2M, IoT and Big Data connected [Source: Machina Research, 2014]



[1] Machina Research created the term 'Subnets of Things' in context of our 'Big Data in M2M: Tipping points and Subnets of Things' White Paper released in February 2013, in referring to an island of interconnected devices, driven either by a single point of control, single point of data aggregation, or potentially a common cause or technology standard

[2] Machina Research has published a series of Research Notes, a Strategy Report and a White Paper on the topic of Big Data as this remains a significant opportunity area in M2M and IoT for operators, system integrators, and solution and platform providers.

What is important to take notice of is that improvements in technologies as illustrated by M2M/IoT Application Platforms[3] and NoSQL databases are enabling these developments to take place.

1 What is NoSQL?

SQL stands for Structured Query Language, designed for managing data in relational database management systems (RDBMS). NoSQL, referred to as 'Not Only SQL', was designed for managing data which did not necessarily have the structure of RDBMS. Some of the leading players in RDBMS are well-established names such as Oracle, IBM, Microsoft and SAP.

Instead of data being structured in 'fixed' relational columns, data could be stored in objects of, for example, graph format (Neo4J and Virtuoso), key value format (Riak and DynamoDB) or wide column formats (Accumulo, Cassandra and HBase). One additional and widely adopted format of NoSQL database is the document-based solution as presented by MongoDB and CouchDB.

To illustrate the fundamental difference between RDBMS and a document-based NoSQL solution, the former stores data in highly structured relational databases where data schemes are 'fixed,' and curated data has to conform to specific characteristics (field length, characters, etc.) defined. NoSQL databases, in a nutshell, allow for data to be stored in what essentially may be defined as form-free formats, i.e. without a rigidity or structure of RDBMS, and allowing for all types of structured, semi-structured and completely unstructured data to be stored in an object. The significant benefit of this approach becomes clear when considering the vast amounts of data generated by 'things' which may exhibit an equally extensive range of structures and qualities and from which diverse data needs to be pulled together and analysed in order to create the IoT. This diversity may range from smaller differences in data structures (for example different field lengths or which data fields are captured) to more substantial differences such as data without any immediate structure as exhibited by images, audio files, random content in text messages, Facebook likes, and so on. NoSQL provides significant flexibility with regards to data management.

2 Extending data model schema and analytical processing

Another challenge presented by the Internet of Things involves the extensibility in the software. As businesses recognize and realize new opportunities from the expanding estate of sensors and devices implemented and the data generated, additional requirements from applications and supporting

[3] For more detailed information about M2M/IoT Application Platforms, see Machina Research's White Paper on "The Emergence of M2M/IoT Application Platforms" published in September 2013

systems emerge. As seen with the development in M2M/IoT Application Platforms,[4] managing and developing applications and addressing scalability are significant and new capabilities for platforms in the age of the Internet of Things. These developments will drive innovations in databases, data analytics, and the structure of data storage.

Responding to the Internet of Things with relational database management systems (RDBMS) is an option but presents a limiting factor which in time will become a significant obstacle to realization of the full opportunities available from all types of data. Databases in the Internet of Things require the flexibility of the NoSQL approach, allowing as explored earlier, different types of data to be stored but more importantly, the agility and flexibility to adapt the underlying data models to new and changing business requirements and applications.

Applications and data models follow the requirements of businesses, and another key development area needed will be in data analytics. No longer limited to historical data analysis, data analytics moves to as close to real-time analytics as possible, introducing and requiring tools such as data streaming analysis, and with the growing numbers of multiple data sources, Complex Event Processing. In the Internet of Things, data analysis will require multiple analytical approaches, and in some cases, significant value is achieved from real-time data analytics (for example, in mission critical or life-saving scenarios where information may guide rescue services within dangerous environments) or from historical analysis for predictive maintenance services.

The Internet of Things generates significantly more data to be stored, and while cloud based services have certainly provided an efficient and highly scalable solution, developing tools to manage distributed databases (rather than located on a single server) emerge as one more capability required from the platforms. In this corner, Hadoop with its HDFS architecture allows for hundreds and even thousands of nodes to become part of the data cluster, a new and distributed way of managing data.

Finally, the location of data storage and storage optimization will become important factors. The speed and frequency with which data will need to be accessed by different applications may change, from instance to instance, the optimal configuration of data storage. Essentially, the ideal database structure required to support (for example) off line historical analyses is different to that required to support real time analyses. However, specific datasets may be required to support either off line or real time applications (or, more realistically, an ever evolving mix of real time and non-real time applications that changes over time as new IoT applications are developed). The implications that an evolving mix of real time and off line applications accessing related datasets may have on optimal database structure is just one illustration of a wider dynamic: ever changing application requirements in the Internet of Things will result in a need to dynamically manage the optimal structure of associated databases on an ongoing basis.

[4] For more detailed information about M2M/IoT Application Platforms, see Machina Research's White Paper on "The Emergence of M2M/IoT Application Platforms" published in September 2013

3 Data integration becomes the challenge

How to handle data interrogation is not a new challenge, but it is a significantly greater one given the volume of data involved, its diversity, multiple storage locations and different analytical processes. The unified view of data becomes the challenge that data analytics platforms will need to address.

This challenge involves, at a fundamental level, the development and application of semantics technologies, recognising, for example, common vocabularies or linked data. At more advanced levels, data integration involves mapping and overcoming the heterogeneity of the data, and presenting this with a unified view.

The additional challenge comes from unifying the data at two levels, the first of which is clearly in terms of the actual data produced by sensors and devices such as temperature readings, on/off status, vibration, etc.

At a second level, and perhaps significant when looking to group sensors and devices according to any number of criteria that may be chosen, the task then becomes one of augmenting data as it arrives in supporting platforms or the databases, and ensuring that these data identifiers (or metadata) are correctly aggregated and processed alongside associated data, enhancing the capability of platforms in verifying and contextualising data.

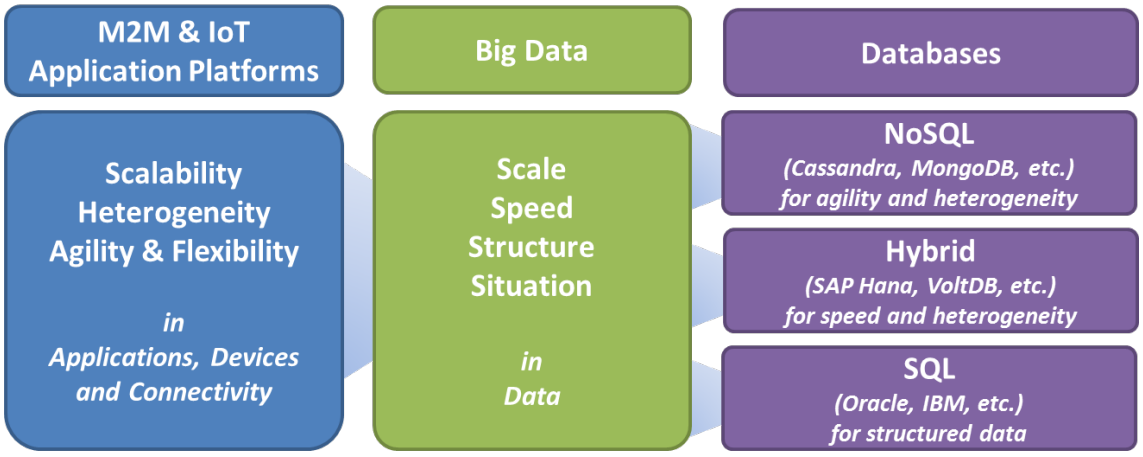
4 Conclusions and recommendations

Platforms managing data and applications require significantly more flexibility, agility and scalability to meet the requirements of businesses in the Internet of Things. These requirements will be dynamic, constantly changing as the opportunities from implemented sensors and devices, and their data are identified. Businesses will look to gain and maintain competitive advantage from innovative applications, requiring quicker application development and agile data models.

In this environment of the Internet of Things and Big Data, Machina Research makes the following recommendations:

- **Enterprises should carefully consider different types of databases** as illustrated in Figure 2 to address the data generated from the exponentially growing number of diverse sensors, devices, and things as well as the growing diversity in structure and scale of that data. In some instances, SQL and hybrid databases will meet the requirements of businesses, and in others, as addressed in this Research Note, a NoSQL database will be the way forward.

Figure 2: New Capabilities in the Internet of Things [Source: Machina Research, 2014]



- **Providers of databases will need to ensure that their solutions are flexible and agile**, meeting the dynamic requirements of business without having to reinvent the application or data model every time, and having robust integrations with M2M/IoT Application Platforms
- **A diverse approach to data analytics will pay dividends.** For providers of data analytics, the ability to support multiple analytical approaches to address the varied requirements of business applications and data will become an increasingly important feature in the future.