Harnessing Data in the Internet of Things

Strategies for managing data in a connected world

Summary

Catalyst

The number of internet connected devices will explode over the next few years. By 2020, well over twenty billion devices ranging from complex interactive systems to tiny sensors, will be collecting and transmitting data over the internet. The volume of data generated by the internet of things (IoT) will be immense and organizations that are planning to implement IoT projects will have to radically rethink the way they transmit, store, manage and exploit the data that is produced.

This report examines the potential that the IoT offers in enabling organizations to develop deeper, more fine-grained and timely insight from the massive volume of data that it will generate and the steps that organizations need to take in order to drive new insight from big data.

Ovum view

The growth of the internet of things will transform the way we think about and use data. Traditional notions of “volume” and “velocity” will be redefined as the internet of things grows in scale and ambition. A few years ago, big data could be defined as a million records and a few hundred megabytes; this volume of data already seems trivial. Before too long, “big” in the context of data volume will be measured in terms of billions of records and petabytes of data. Rates of 1000 messages per second have traditionally been regarded as significant, but today we’re already looking at systems that can produce hundreds of thousands of data points per second.

While this explosion in the volume and velocity of data creates challenges, it also presents us with opportunities to garner insights about the behavior of systems and people that we’ve never previously had. Organizations in industries like electricity supply, manufacturing, logistics and retail that can effectively harness and exploit the data created by the IoT will enjoy a significant competitive advantage.

There is no single “best” approach to data management in the context of IoT; in most cases organizations that implement large-scale IoT networks will have to adopt multiple architectural models in order to effectively manage their networks and the data they produce. When it comes to building the right foundation to store and manage IoT data, organizations need to focus on technologies that offer flexibility and scalability on one hand while providing sophisticated storage and query services on the other.

Key messages

- We are at the beginning of a huge wave of growth in IoT
- IoT will challenge traditional approaches to data storage and analysis
- Data management and analysis will increasingly take place at the edge of the network
- IoT calls for a different type of database technology

Recommendations

Recommendations for enterprises

Get ready for IoT

The growth in the creation of IoT networks will be dramatic over the next decade. No industry vertical will be able to ignore the potential benefits that organizations will be able to drive out of the internet of things, or immune from the threat posed by competitors who begin profiting from IoT first. While it is easy to see how manufacturing, logistics and utilities organizations are already being transformed by IoT, its effects will be felt in every sector from retail to healthcare.

Not every database is designed to harness IoT data

When looking at how best to manage your IoT data, you need to select a database platform that is optimized for the volume, velocity and variety of data that your IoT network will generate. Databases that are optimized for handling time-series and geospatial data will provide a significant price-performance advantage of general purpose database engines.

Consider the potential value of deploying data management technology within and on the edge of the network

As the volume of IoT data grows, it will be increasingly necessary to push data processing, management and analysis both into and out to the edge of the network. You should be looking for solutions that support multiple possible deployment scenarios. The database technology that you choose should be capable of scaling up to support the volume of data you plan to centralize on one hand, and it should be able to scale down so that it can be embedded within a small chip-based device on the other.

Reevaluate your current database architecture and data management strategy

The data you garner from IoT will not replace the transactional data that currently resides in your core database systems, but its value will be limited unless you can integrate all of the information you have in a coherent way. This doesn't mean that you have to throw away your existing database architecture, but that you will have to extend it in order to get the maximum benefit from IoT data.

Evaluate your vendor choices

A host of technologies will emerge over the next few years from both commercial vendors and the open source community. When evaluating technologies it's crucial that you think beyond the application of technology to a single application and consider how it will fit in with your existing environment and how well it will serve your future needs.

Recommendations for vendors

Offer your clients choice, don't push them down an architectural dead-end

There are many different deployment scenarios for IoT, and no single model will dominate. On the contrary, we expect that most IoT networks will make use of a number of different approaches, so it is important that your offering is flexible enough to support a range of different scenarios.

DON'T overlook the hackers and hobbyists

While the hardware hacking community may not drive a lot of revenue in the short to medium term, a small number of the hundreds of small "hobbyist" projects will turn into giant deployments. To be a successful vendor in this field you will have to invest some effort in working with, and at times supporting, small innovative projects.

Remember that not every “thing” in the internet of things will have a high-performance processor

While the cost of processors continues to fall across the board, many IoT projects will use the least expensive processors. Cost is one factor, but there are other factors, such as power consumption, that will have a major bearing on processor choice.

You should certainly be offering an “embeddable” version of your data management technology, but you should also offer simple, secure, API’s that can be invoked by the simplest of processors.

We are at the beginning of a huge wave of growth in IoT

A Connected Data World

Up until now, "big data" has largely been made up of human-generated transactional and interaction data that is held in relational databases. As the deployment of IoT networks continues to grow, the balance will shift firmly in the direction of very large volumes of data in the form of the sensor readings, alerts, and events that are generated by connected devices.

Traditional transactional data will continue to play a crucial role of course, but the availability of fine-grained data from connected devices will create new opportunities to understand systems, human behavior and our environment. When combined with transactional data, this new source of insight will enable organizations to gain a deeper understanding of customer behaviors, optimize their existing systems and better identify and manage risk.

New applications and new markets

Remote sensors and monitoring devices have been with us for decades, they have been used within manufacturing plants to monitor processes, within pipelines to monitor the flow of commodities like oil and gas, and in safety critical systems. Traditionally though, the cost of deploying and managing sensors has limited their use to systems where there is a very strong incentive either in terms of safety or cost. The dramatic change in the economics of sensing that we've seen over the past few years has now opened up a myriad of new potential applications.

In logistics, packages can now be tracked in real-time through the use of hand-held GPS and GPRS enabled devices. In urban environments the flow of pedestrians through city streets can be monitored using the blue-tooth signatures of people's mobile phones, buses and trams can broadcast their location, providing travelers with up to the second scheduling information.

An explosion of end-points

The number of devices capable of gathering data and connecting to the internet is growing dramatically. In 2009, 300 million smartphones were sold. In 2013, that figure rose to over 1 billion. Across the globe utility companies are rolling out smart metering projects that will see tens of millions of smart meters being deployed in people's homes and business premises. By 2020, nine out of ten new cars will have connectivity built into them to enable tracking, fault monitoring and to provide a channel for value-added services like mapping, real-time information, and entertainment.

The growth in IoT is being boosted by the lowering in cost of devices and communications

The growth in the deployment of sensor networks is being fuelled by a sharp reduction in the cost of the technology that makes up devices, and a similar decline in the cost of transmitting data. Whereas in the year 2000 the cost of building a wirelessly connected device would have easily exceeded $1000, it is now possible to build one for under $100.

While the "mainstream" will account for the vast majority of devices, hackers and hobbyists will help to drive innovation

The vast majority of new connected devices will come from mainstream sources. Utility companies will represent the largest group deploying new devices as they roll out their smart metering programs, manufacturers and retailers will also drive a significant volume of new deployments; however it is important not to overlook the work that is being done at a much smaller scale by hackers and hobbyists. Over the past five years a number of "open hardware" projects have been launched, which provide low cost development platforms, as well as tools, tutorials and collaboration spaces. Projects like Arduino (www.arduino.cc) and mbed (www.mbed.org) are spurring a wave of innovation that will not have a significant impact in the short term, but will certainly grow in importance over time.

An Explosion of Data

New projects are springing up all over the world that will be generating significant volumes of new data about systems, behaviors and the environment.

Gathering air quality data

In London, United Kingdom, a small non-profit organization, Change London, is setting out to deploy 10,000 low-cost air quality monitoring stations across the city. Each monitoring station will measure a number of different phenomena, typically carbon monoxide, nitrogen dioxide, ozone, particulate matter, temperature, humidity, and noise levels. Typically, each device will take many measurements of these phenomena each minute, transmitting averaged data once a minute. When the network is fully deployed, it will be recording a little over 100,000,000,000 (100 billion) readings a year, which will correlate to more than 400 terabytes of storage per year.

The collected data will be used to provide a real-time air-quality map for the citizens of London and will also be made available to researchers that are developing environmental models. One of the key research areas lies in understanding how pollution propagates within an urban environment. The modelling needs to incorporate data from multiple sensor locations and combine external data, such as topography and real-time traffic data, from other sources.

While general purpose databases may well be able to store this volume of data, the Change London team is adopting a specialized time-series database platform in order to ensure that the platform is capable of supporting future growth.

As the project evolves, the Change London team is keen to look for other ways in which the data might be exploited in real time. The team is looking at developing an alerting mechanism that is triggered when air quality falls below a certain level and a real-time reporting system that would help the operators of hybrid vehicles decide when to run on electricity vs internal combustion engines while travelling through areas of high pollution.

It is important to note that by comparison with some IoT projects, the Change London project is relatively modest in terms of the number of nodes, and the volume of data.

Smart Metering: Millions of nodes, billions of data points

Across the world, electricity companies are rolling out smart metering initiatives to better manage their networks and as a way to offer clients more transparency and greater choice in how to use electricity, tariff selection and payment options.

Over the next decade, millions of smart meters will be deployed, each of which will collect data on electricity consumption down to a very fine level of detail. Of course, electricity companies don't need to know on a second by second basis how much electricity every individual property or business is using, although some manufacturing businesses' consumption needs to be monitored at very regular intervals. The fine grained data is important for a couple of reasons. First, it provides the consumer with a real-time trace of their electricity consumption, helping them manage their use and secondly, it enables electricity providers to offer more flexible tariffs, based on demand.

The electricity companies also need to address the privacy concerns of their customers. The minute to minute electricity usage of a home can identify a number of things about user lifestyles. For example, potential burglars would have a good idea when the property is vacant. To address these privacy concerns, smart meters need to store usage data locally in a secure database and should only send appropriately aggregated data back to the electricity supplier. If a consumer is on a basic tariff that is based on a fixed cost per kilowatt hour, the meter only needs to transmit the total number of kilowatts consumed during a given billing cycle. If they are on a time based tariff where they pay a lower price for electricity consumed overnight, the meter needs to transmit details of the consumption during the standard time window and the discounted time window for each billing cycle.

The Chinese company, Kaifa, entered into a partnership with AMT-SYBEX and IBM to design a smart metering solution that uses technologies developed by each company, including IBM Informix TimeSeries database. A benchmark demonstrated that the system is capable of scaling to support 100 million meters. Just as importantly, the system is capable of processing a day of data from 10 million meters, each sending data every half an hour, in under 100 minutes. Key to this performance is the use of a database that is optimized for time-series data.

Dealing with Exabytes of data

The number of devices gathering and transmitting data will explode over the next five years, presenting challenges and opportunities never seen before. A UK based company, Hildebrand, is developing a platform that will enable organizations from a range of vertical sectors to capture and exploit huge volumes of data from IoT networks. The volume and velocity of the sensor data that will be generated by many of these networks will overwhelm general purpose database management systems (dbms). Hildebrand adopted a dbms that can cope with the volume and rate of data while allowing that data to be simultaneously analyzed and queried. Hildebrand partnered with IBM to incorporate IBM's Informix database, which ships with specialized features that enable it to store, process and query massive volumes of sensor data.

Hildebrand has applied this technology to a retail environment, using it to track the movement and behavior of shoppers across 25 UK-based shopping malls. The system tracks 300,000 shoppers and collects 250 million data points each day. This information is used by the shopping malls to identify the number of visitors and insights into the type of shopper their mall attracts. In turn, retailers can use the information to determine which malls to invest in, and which locations are likely to deliver the highest revenue.

The internet of things will challenge conventional approaches to data storage and analysis

The three examples above highlight the wide spectrum of data volumes that different IoT projects are likely to produce. In all of them, however, it is clear that we will need a new approach to the way we store and analyze the data generated by the internet of things.

Challenges for Enterprise Data Management & Analysis

IoT marks a shift in the domain of Big Data management and analysis. IoT will create challenges in all aspects of data management. While traditionally centralized database (data warehousing), data management and analytics architectures will continue to play a valuable role, they are not well suited to handle the large volumes of raw and intermediate IoT data flows spun-off from sensors and mobile devices. The problem is no longer the absence of enough data – it's making sure only important data is moved or analyzed.

Raising the bar in terms of volume and analysis

Traditional database management technology has evolved over time to be increasingly optimized for the management of transactional data. This data is typically well structured, predictable in terms of volume and well suited to the traditional relational database engines that lie at the heart of all leading enterprise databases.

The IoT will produce data in far greater volumes than ever before and the structure of that data will be much less standardized and predictable than the transactional data most organizations are used to handling. The nature and source of the data will create additional challenges and organizations will have to consider privacy and data protection alongside the challenge posed by its volume.

Moving petabytes of data can become costly

The scale of the data means that moving it from point to point will be costly in terms of money and time. The ability to push processing and analytics into the network becomes an important priority.

The key, of course, lies in ensuring that unimportant data is discarded and important data is retained. This in itself will create challenges for many organizations, as one of the leading scientists from the Square Kilometer Array project noted, "We are always worried that we will decide on Monday that we can discard a certain part of the data, only to be told by another researcher on Tuesday that the information we decided to throw away is now vital for their research". The ability to adapt the algorithms that are used to filter and aggregate data at the edge of the network will be essential for some sensor networks.

IoT Data will have more value when tied with other sources of data

We often tell our clients that data is "sociable" because it is always far more interesting when connected with other pieces of data from different sources. This is particularly true of data generated by the internet of things. For example, data about pollution levels can be combined with traffic management systems to adapt the system to reduce the adverse effects of urban traffic on air quality. Retailers can use fine grained data about footfall to offer customized offers and discounts to shoppers.

Existing data management and analytic architectures will have to be adapted to embrace data capture from network devices and rather than creating distinct islands of data, organizations can draw insights from all of their data sources as if they were a single whole.

Data management and analysis will increasingly take place at the edge and in networks

The traditional approach to managing data by collecting it at the edge and centrally storing and processing it will be inappropriate for many IoT networks. This is due to the volume of data or because of privacy concerns and data ownership. As the size and complexity of IoT networks grows, data management will be pushed out from the center into the network itself and onto the devices that sit at the edge.

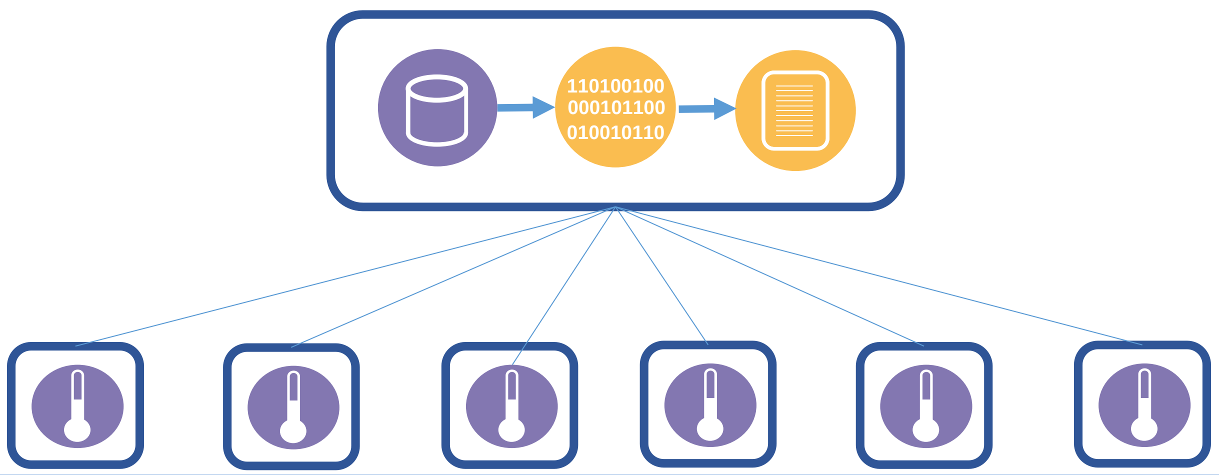
New architecture challenges

There are a number of different architectures and system topologies that will be applied to IoT networks. New architecture challenges warrant a combination of approaches that are based on various requirements, such as environmental drivers, privacy concerns, local data visibility requirements, and communication infrastructure optimization.

Simple centralized topology

Figure 1 displays the simplest topology for IoT networks, in which remote sensors transmit data to a central service for storage, processing and reporting.

Figure 1: Centralized topology



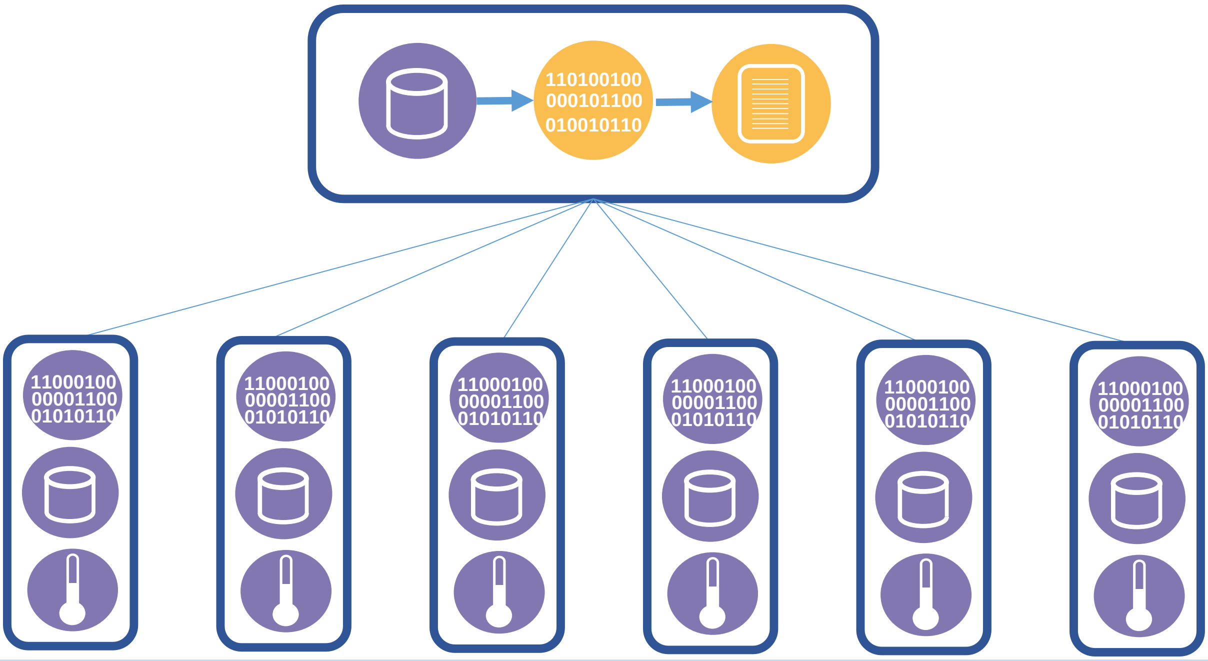
This topology is only well suited to relatively small networks, where a modest amount of data is being transmitted by each end-point. The remote devices may be extremely simple, consisting of one or more simple sensors, a low cost processor and a communications module. The communication between the sensors and the center may be via radio, fixed lines, or using cellular technology.

In many cases, the remote devices will sense multiple phenomenon and take multiple readings which are filtered and aggregated before transmission.

Local filtering and aggregation

In this model, data gathered at the edge of the network is pre-processed prior to transmission, largely related to filtering and aggregation. For example, 100 sensor readings a second may be filtered and averaged to transmit a single average for a given time period. It might also involve the application of more complex algorithms that operate on large amounts of data.

Figure 2: Local filtering and aggregation



The volume of data being processed may be trivial or significant. In significant cases, the local device will store and manage the data before aggregation and transmission.

Federation of data storage and processing

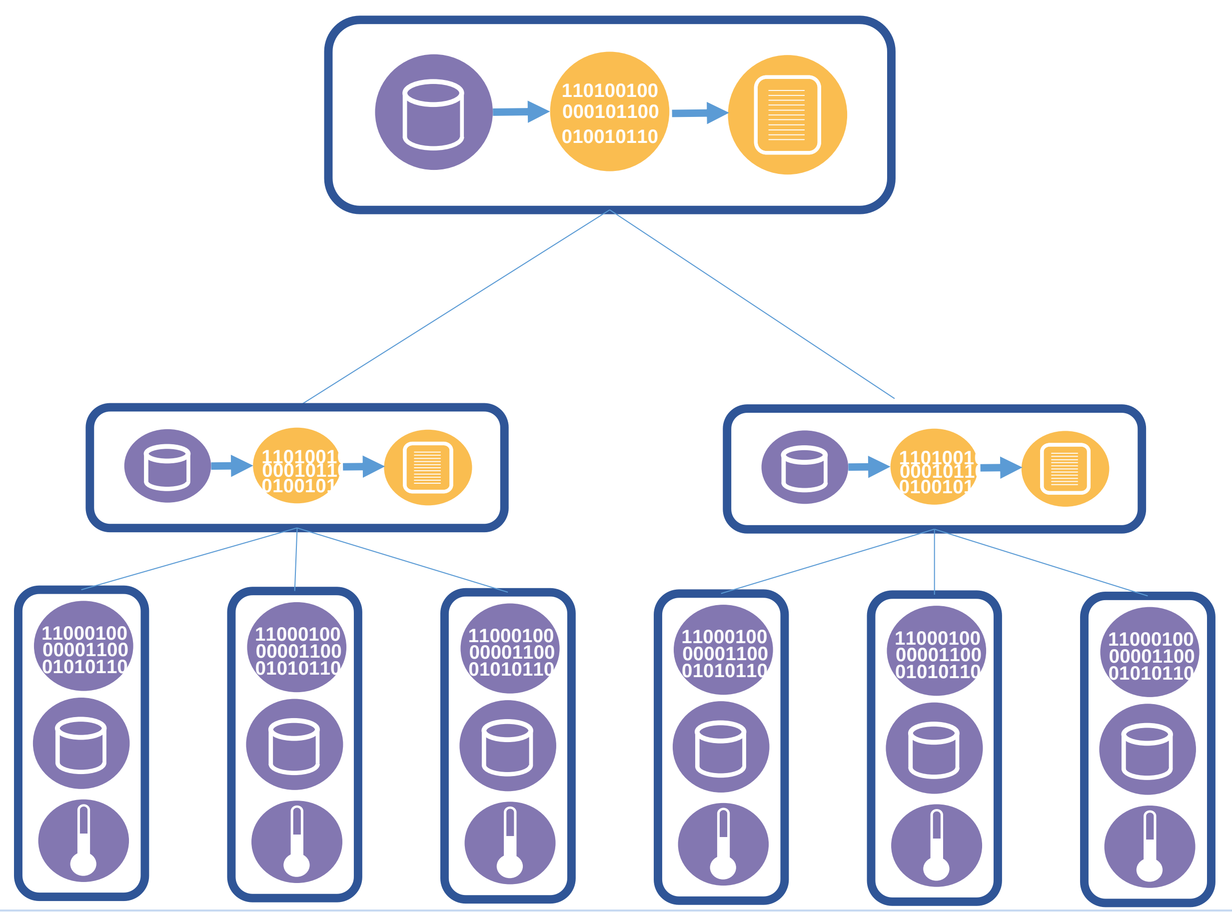
In many cases, a federated approach will be mandated. The cost of long distance transmission might restrict the viability of sending data back to the center from every node. Gateway devices can be deployed within the network to collate and aggregate data from multiple end points for retransmission to the center. There may be a localized requirement to monitor local conditions, combined with a need to view and monitor the data centrally.

The ability to query and process data locally will be essential in some scenarios

In some cases, gateway devices will only transmit a small proportion of their data to the center and will provide a mechanism for the central service to query remote data that span multiple gateway devices.

In the case of smart metering, the meters will typically store a significant volume of consumption data locally, often over 12 month of usage data. In order to address the privacy concerns of consumers, most smart metering initiatives allow customers to mandate the minimum information necessary for bill calculation that is sent back to the utility company.

Figure 3: Federation of data storage and processing



Pushing data management into the network has benefits, but also presents challenges

By federating the data storage, management and processing, organizations that are building IoT networks can account for the cost and complexity of data transmission while ensuring the right data is centralized. Not only does this take pressure off the communications network, it also reduces the amount of storage and data processing power required at the center.

The other key benefit of managing the data locally is that it provides an opportunity to provide users with insights and alerts that are of interest to them. A smart washing machine may need to transmit information about the overall health of the appliance to a central system, but it also needs to provide the local user with information that enables them to use the appliance properly. Events like "washing cycle completed" or "detergent level low" are relevant locally, but are unlikely to be significant to merit transmission to the central application. Critical data about failures or impending failures can be sent to the center to trigger the deployment of an engineer.

This approach can create a number of challenges that aren't effectively addressed by mainstream database technologies.

The transition from the "data warehouse" to the "data landscape"

The dispersal of IoT data to the edge of the network means that the management of that data also needs to be dispersed. Database operations such as tuning and configuration are impacted differently in a centralized environment versus distributed. The ability to manage multiple data stores across a complex data landscape will depend on the flexibility and sophistication of the data storage technology.

The internet of things calls for a different type of database

The volume, complexity and growing geographical dispersal of IoT data calls for a database that is optimized to handle the type of data that IoT devices will generate. The ideal database should be capable of handling IoT data, accessible via a number of interfaces and embeddable within devices that are deployed at the edge of the network.

Key considerations in selecting a database technology for the IoT

Some important things to consider when selecting a database technology to support your IoT network are outlined below. This is not intended to be an exhaustive list, but presents a number of capabilities important in a potential database technology.

Hybrid unstructured and unstructured database support

The ideal IoT database is flexible enough to handle unstructured, time-series, geospatial as well as traditional relational data. It should be possible for these types of data to co-exist within a single database management system.

Optimized for embedded deployment and low resource utilization

The devices that sit at the edge of the network make use of a wide range of technologies and many will not be based on "standard" X86 compatible processors. IoT devices will be built around various processor architectures, some using less powerful processors in order to optimize power consumption and keep costs down. Some IoT projects are being built around processors that have as little as 512 bytes of RAM.

The ideal database should be embeddable within software developed for various processors and accessible via simple API calls from less powerful processors that lack the ability to run an embedded database.

One of the scarcest resources in many IoT environments is likely to be network bandwidth, either because it is simply not available or because it is expensive. The ideal IoT database supports the transparent compression of data before it is transmitted over the network.

Good embedded software should require little or no configuration at run-time. Features like self-tuning and automatic recovery from failure are essential.

Integration with the cloud

Given the flexible, scalable, nature of cloud-based infrastructure, we expect a lot of IoT data to be stored and processed in the cloud. The ideal IoT data platform can be deployed either on premise or in a public, hybrid or private cloud environment. The platform should be capable of being administered via both a web-based interface and API calls.

It should also be possible to manipulate the data itself either via traditional client/server style API's and driers or restful interfaces.

Support for real-time analytics and events

The ability to analyze sensor data in real time is key to putting the information to work. We would also like to be able to specify event-triggers to save the need to manually code threshold checks or event monitoring.

Appendix

Author

Gary Barnett, Chief Analyst, Software

Gary.barnett@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

CONTACT US

www.ovum.com

askananalyst@ovum.com

INTERNATIONAL OFFICES

Beijing
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
San Francisco
Sao Paulo
Tokyo

IML14446-USEN-01