Internet of Things & Creation of the Fifth V of Big Data

Shaurya Shekhar

SCOPE, VIT University, Vellore, India

Abstract: ***The Internet of Things (IoT) is a term widely used nowadays to refer to the development of an intrinsic network of objects and appliances with Internet connectivity, which allows them to send and receive data. All of this data can be collectively referred to as Big Data, which basically comprises of large data sets and can be analyzed computationally to reveal patterns and trends relating to human behavior and interactions. In addition to the existing 4 Vs of Big Data (namely Volume, Variety, Velocity and Veracity), we can then derive a fifth V, called Value, which will help in engineering products better suited for a wider group of customers. Thus, we create a loop by connecting IoT (for sourcing of data) to Big Data (collection of data) to creating value from it and then back to IoT (for designing products with better functionality and greater acceptance).***

Keywords: Big Data, Database, Data Security, Information, Value, Veracity, Velocity

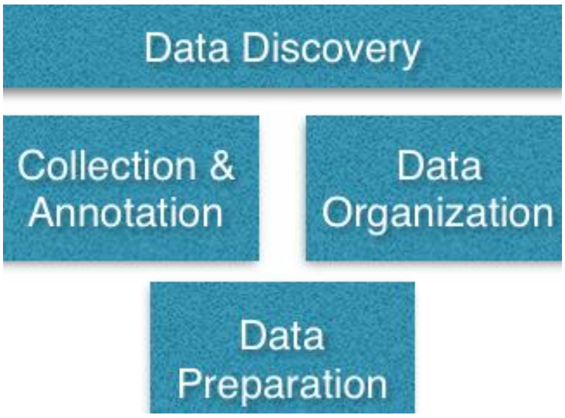
1. Proposed Cycle For Creation Of Value From Big Data

It is proposed that we create value from Big Data using 3 concrete steps, which have been explained in detail further on:

(1) Data Discovery [1]
(2) Data Integration [2]
(3) Data Exploitation [3][6]

The first step in the process which we are proposing is Data Discovery, which is itself composed of three steps:

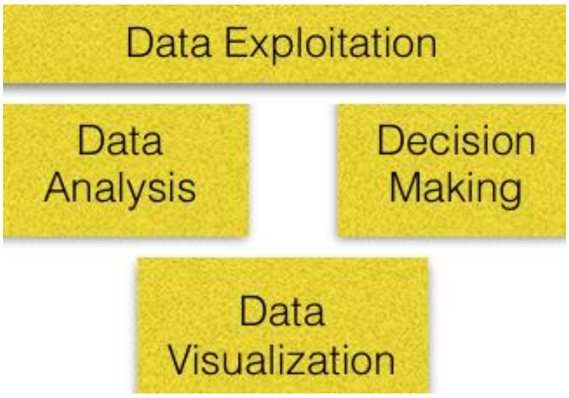
1) Data Collection & Annotation: An inventory of data sources and all the metadata that describes them is created.
2) Data Preparation: Access-control rules are set up. The syntax structure as well as the semantics of all the data collected from various sources is identified.
3) Data Organization: This step deals with the organization of data into a pre-defined structure which makes the next step less cumbersome and considerable less complicated.



The second step in the process refers to the Integration of Data from various sources into a common data representation scheme. A data type is chosen beforehand to ensure that all the received data types can be suitably condensed into a different type of data with minimal data loss. This makes further analysis of the different types of the data received easier.

The third step in the process refers to the Exploitation of Data, which is the essence of creating Value from Big Data, which then allows us scope for improvement of devices & services. This is achieved in three steps:

1. Data Analysis: Analysis of integrated data takes place by various methods. This is made easier by the fact that all the different types of data have been integrated and incorporated into a single form and type in the previous step.
2. Data Visualization: Presentation of analytical results to the decision maker takes place in the form of interactive mathematical models which support scope for further exploration, refinement and predictions on the basis of cosmetic as well as major changes which might be proposed.
3. Decision Making: The last and final step of the process results in the decision maker determining what actions (if any) need to be carried out on the basis of the interpreted results.



2. Data Storage Mechanisms

When we talk about IoT, one of the first things that comes to mind is a huge, continuous stream of data hitting company data storages[4] . Data centers must be equipped to handle this additional load of heterogeneous data.

In response to this direct impact on big data storage infrastructure, many organizations are moving toward the Platform as a Service (PaaS) model instead of keeping their own storage infrastructure, which would require continuous expansion to handle the load of big data. PaaS is a cloud-based, managed solution that provides scalability, flexibility, compliance, and a sophisticated architecture to store valuable IoT data.

Cloud storage options include private, public, and hybrid models. If companies have sensitive data or data that is subject to regulatory compliance requirements that require heightened security, a private cloud model might be the best fit. Otherwise, a public or hybrid model can be chosen as storage for IoT data.

3. Data Security

The types of devices that make up the IoT and the data they generate will vary in nature – raw devices, varied types of data, and communication protocol – and this carries inherent data security risks. The heterogeneous IoT world is new to security professionals, and that lack of experience increases security risks[5]. Any attack could threaten more than just the data – it also could damage the connected devices themselves.

IoT data will require organizations to make some fundamental changes to their security landscape. As the IoT evolves, an unmanaged number of IoT devices will be connected to the network. These devices will be of different shapes and sizes and located outside the network, capable of communicating with corporate applications. Therefore, each device should have a non-repudiable identification for authentication purposes. Enterprises should be able to get all the details about these connected devices and store them for audit purposes. All internal and external core routers/switches should be instrumented with X.509 certificates for creating trusted connectivity between public and private networks.

A multi-layered security system and proper network segmentation will help prevent attacks and keep them from spreading to other parts of the network. A properly configured IoT system should follow fine-grained network access control policies to check which IoT devices are allowed to connect.

4. Data Integration Mistakes To Avoid

Since, the entire foundation of this research paper lies in the second part of the process, that is to say the Data Integration stage, where-in data from different sources is integrated for further interpretation.

We shall now look at 5 mistakes which need to be avoided while integrating data:-

(4) Invest In Enterprise Grade Hadoop And Data Integration Technology[7] - When selecting Hadoop for the enterprise, it has to stand up to rigorous stresses of security, administration and monitoring. In addition there should also be continued dedicated development and support resources that enhance the chosen technology. Similarly, the data integration technology that is chosen should lend itself to scale with the Hadoop technology. The data integration technology should deliver tools to streamline development, enforce quality and shorten time to implement. This reduces custom coding from proliferating, a common precursor to tough maintenance and data transparency. As with any enterprise technology standard big data integration tools should be able to work with a variety of heterogeneous big data languages and sources while abstracting the user base from the complexity of implementation. The data integration technology should allow enterprises to work with newer Hadoop standards as and when they mature and switch between multiple underlying big data standards without business risks and disruption.

(5) Focus On Modern Big Data Architecture Patterns- Many enterprises approach big data architecture as an extension of their existing data warehouses. Big data architectures (including reservoirs, etc..) frequently co exist with traditional data warehouses, but to build one along similar principles of economic data storage will restrict the value of data within the big data store. Specialist storage and engineered platforms built for performance complement the big data reservoirs which are mainly used for data exploration. Properly engineered, big data reservoirs can hand off subsets of often used data to engineered platforms to improve speed and performance. Modern big data architectures emphasize data streaming for real time data ingestion into the big data platform, data enrichment and transformation using native big data query languages (examples are Pig Latin, HiveQL, MapReduce etc...) and are fully orchestrated and governed to minimize risk.

(6) Prioritize Efficient Data Ingression & Data Transformation- It is the ability to ingest real time data into the data reservoir. This ensures that the data used for decision making is up-to-date and business analytics reflects the latest data. Sub second latency differentiates average user experience from excellent customer experience providing timely insights. These data ingestion tools should be non-invasive when capturing data and not impact source technology performance. Once within the data reservoir, the transformation technology should be transparent and not inject proprietary code onto the Hadoop nodes. It should provide facilities for modular, team based development. It should be portable across platforms, or in other words, abide by the “design once, run anywhere” mantra. Some of these criteria are satisfied by traditional data management technologies. However to see success in big data projects all these criteria are necessary specifications for the selected tool.

(7) Incorporate Pervasive Data Governance- Big data

reservoirs are generally considered the blackbox playpen of data scientists. While this was true in the first wave of Hadoop projects this is no longer the case. In fact, proper emphasis should be laid in ensuring transparency in Hadoop clusters. If the upsides of storing full raw data in the data reservoirs are profits and customer experience, the downside of a data leak to large amounts of data is costly litigation and irreparable reputation damage. Managing metadata across every technology in the data management landscape is key to govern data. It allows complete data provenance, allowing business and IT accountability for the data that passes through the systems and business decisions. Good governance depends as much on technology as on the organization's culture and business processes. However selecting the right governance technology is critical in enabling the business govern its data. A good governance technology brings data transparency, accountability and helps identify areas of process and performance improvements. In the integrated big data platform, it is important that the governance tool cuts across multiple technologies (data bases, data warehouses, data quality and enrichment technologies, data integration technologies, business intelligence and analytics technologies) to efficiently fulfill governance requirements. The governance technology should service both the business user and the technology users.

(8) Maximize The Potential Of Your Hadoop Cluster[8] - If you think of Hadoop and NoSQL as just commodity data stores you miss the big advantage they provide through their compute capabilities. The gains that you achieve through data storage are lost if you do not efficiently utilize the big data platform for processing. To do this, you should offload compute intense queries into the big data store by generating code that is native to the underlying big data standard. This allows you to use your big data investment both for storage and processing, and as your data volume and storage scales up, you do not have to invest in additional processing hardware. To do this, the data integration technology should not use middleware, or a processing platform that is proprietary and exists outside of target/source data bases. Traditional Extract Transform and Load (ETL) technologies that are designed for relational databases normally have middleware based architecture which counteracts and nulls any big data advantages. Modern tools that have an ELT (Extract, Load and then Transform on target/source) based technology are best suited for big data integration.

5. Summary

The industrial internet of things (IIoT) is an exciting outcome of the digital revolution that is changing the way we live and work. Many organisations already focus on how to benefit from it, but extracting maximum value requires a big data approach[9].

The IIoT is defined as "a universe of intelligent industrial products, processes and services that communicate with each other and with people over a global network". This connected web is becoming increasingly ubiquitous across a wide range of industries, from oil and gas, utilities and transportation through to the medical field.

While these devices can be very beneficial for organisations and consumers every day, even greater value may be derived by using analytics to generate insights from the vast datasets generated by IIoT.

Previous research has found that industrial organisations see a considerable upside in IIoT as a complement to big data analytics. The more conservative forecasts in this research estimate that this activity could be worth $500 billion by 2020.

The research found that nearly three out of four (73%) companies are already investing more than 20% of their technology budgets in big data analytics and almost as many expect to increase spending in that area within the next year. Exploiting the data from IIoT is a key part of those budgets.

As the IIoT takes off, it has been estimated that the number of internet-connected devices is likely to multiply to tens of billions. One reason for this growth is that the combination of IIoT and big data analytics promises to drive further operational efficiencies along with more innovation and, ultimately, new sources of revenue.

Operational savings and revenue opportunities

Through the IIoT, operations technology and information technology will blend together and become more intelligent through the use of sensors, analytics and machine applications – a development that will share and create even more data. The insights gleaned from this big data and new types of data can bring many benefits for businesses, including operational savings and revenue opportunities.

Research has indicated that predictive maintenance can generate savings of up to 12% over scheduled repairs, leading to a 30% reduction in maintenance costs and a 70% cut in downtime from equipment breakdowns. For a manufacturing plant or a transport company, achieving these results from data-driven decisions can add up to significant operational improvements and savings opportunities.

And there are many more big data applications. For the healthcare industry, the ability to collect health data in real time and then apply advanced analytics to it can inform decisions that lead to improved care management and address risk factors early – ultimately providing more positive experiences and outcomes for patients.

Some hospitals put a tag on patients admitted to emergency wards, which collects and connects their data to a dashboard. The tag monitors a patient's location, treatment cycle and journey through the hospital, enabling hospitals to provide better healthcare and faster discharge times.

Improving customer service

Water companies can combine data from embedded connected sensors on its pipes and treatment facilities with data from a wide range of business systems to provide deep insights in real time. The utility company can track in near real time and, in many cases, predict the cost risk and performance of its assets, enabling speedier response times and saving money while improving customer service.

As an example of revenue generation, an energy company can optimise its wind turbines by analysing thousands of data points as they collect wind speeds every second. Using this information, the company can constantly make adjustments to optimise its turbines – for instance, adjusting the pitch of the blades to maximise electricity generation. In practice, we have observed that such a data-driven optimisation process can lead to a 5% improvement in power output and thus boost revenue without significant upfront investment.

In a different twist on this theme, an operator of mobile phone networks in Europe sells data relating to the movement of people to retailers, giving them hour-by-hour insights into consumer behaviour. That data not only helps retailers target consumers with relevant communications, but it is also another source of revenue for the mobile network operator.

Bringing it all together
The key challenge for organisations is to drive actionable insights and, ultimately, incremental value from data generated by IIoT. To do so, organisations must think about how they collect, store and analyse that data.

Today, businesses are examining their current data systems and looking for ways to complement them with emerging tools and technologies, creating hybrid technology environments that can meet their business needs. Such environments help companies analyse their data in real time to make decisions that can benefit customers, improve their business performance and disrupt their industries.

The good news is that setting up a big data platform to leverage IIoT can be done incrementally without needing a big up-front investment. I recommend that organisations begin implementing their big data strategy in an agile way with a small proof of concept. This approach will show them which data combinations generate the most value before they scale into a broader enterprise solution. Cloud-based approaches are also becoming more common because they provide greater flexibility to scale up and down as the business need evolves.

Whatever approach businesses take, they must consider how the data can derive actionable insights that can really drive value, such as increasing productivity, generating savings, and uncovering new income streams. In many industries, IIoT will become a key source of feedstock for big data platforms, and the exploitation of this data is an area that is set to grow quickly.
Data-driven opportunities are available and growing by the moment. Businesses should explore how to capitalise on them effectively, and try to do so before their competitors do.

References

[1] Lin, K., & Levis, P. (2008, April). Data discovery and dissemination with dip. In *Proceedings of the 7th international conference on Information processing in sensor networks* (pp. 433-444). IEEE Computer Society.
[2] Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246). ACM.
[3] Chang, C. I. (Ed.). (2007). Hyperspectral data exploitation: theory and applications. John Wiley & Sons.
[4] Korfhage, R. R. (2008). Information storage and retrieval.
[5] Pfleeger, C. P., & Pfleeger, S. L. (2002). *Security in computing*. Prentice Hall Professional Technical Reference.
[6] Ives, Z. G., Florescu, D., Friedman, M., Levy, A., & Weld, D. S. (1999, June). An adaptive query execution system for data integration. In *ACM SIGMOD Record* (Vol. 28, No. 2, pp. 299-310). ACM.
[7] Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.
[8] Ren, Z., Xu, X., Wan, J., Shi, W., & Zhou, M. (2012, November). Workload characterization on a production hadoop cluster: A case study on taobao. In *Workload Characterization (IISWC), 2012 IEEE International Symposium on* (pp. 3-13). IEEE.
[9] LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT sloan management review*, *52*(2), 21.