Where is the real value in Big Data?

Dated 28 January 2016
By Pete Ianace

Taking a look at all the activity related to Big Data one should ask the question, how much of Big Data is actually useful. By applying just a little common sense we discover only a small amount.

I have been working with data for over 40 years and if we go back to pre-Internet days we experienced what we called data overload and we discovered then that data itself wasn't valuable but only a small slice of that data proved to have a direct impact on actual business decisions. With history in mind what has really changed in solving the most critical issue is related to finding the data that is actually useful. Well volume has certainly increased, but what is important to deal with is that much of the growth in volume comes in the form of unstructured data.

So let me start with what is unstructured data using the definition from Webopedia. The term unstructured data refers to any data that has no identifiable structure. For example, images, videos, email, documents and text are all considered to be unstructured data within a dataset.

While each individual document may contain its own specific structure or formatting that based on the software program used to create the data, unstructured data may also be considered "loosely structured data" because the data sources do have a structure but all data within a dataset will not contain the same structure. This is in contrast to a database, for example, which is a common example of "structured" data.

So looking back in history we are talking about data overload with an added new twist called unstructured data, which represents much of the new volume being generated. I would suggest that companies that bring a combination of strong data analytical expertise along with a good grasp of both industry standards and compliance rules can offer precise filtering solutions that can identify the most valuable data for the user.

Peeling back the onion a bit more

While there are numerous solutions emerging that address the filtering and analytics of structured data such as Splunk Enterprise that collects, indexes and harnesses all of the fast-moving machine data generated by applications, servers and devices - physical, virtual and in the cloud. In the case of what Hadoop brings to the table there are many others that have debated its pluses and minuses and I will leave that topic to them. My view is that the real challenge is to provide cost effective solutions that address the much more complex world of filtering and real-time analytics of unstructured data.

While the volume of all data types is expected to grow 800% in the next five years, 80% of that growth will be unstructured data. I would suggest that companies that possess skills and capabilities that include data modelling, analytics, OCL, and ontology have a leg up when it comes to delivering solutions that leverage both structured and unstructured data. As of today the jury is still out on who will be the players that will offer compelling solutions that address the holy grail of finding the needle in the haystack in the growing world of Big Data.

Big Data, What role does ontology play?

An ontology formally represents knowledge as a hierarchy of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts.

Ontologies are the structural frameworks for organizing information and are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it. The creation of domain ontologies is also fundamental to the definition and use of an enterprise architecture framework.

It is important because it eliminates the need to integrate systems and applications when looking for critical data or trends.

How is it applied and what are the important elements that make it all work?

Ontology uses a unique combination of an inherently agile, graph-based semantic model and semantic search to reduce the timescale and cost of complex data integration challenges. Ontology is rethinking data acquisition, data correlation and data migration projects in a post-Google world.

Sharing common understanding of the structure of information among people or software agents is one of the more common goals in developing ontologies. For example, suppose several different Web sites contain medical information or provide medical e-commerce services. If these Web sites share and publish the same underlying ontology of the terms they all use, then computer agents can extract and aggregate information from these different sites. The agents can use this aggregated information to answer user queries or as input data to other applications.

Making explicit domain assumptions underlying an implementation makes it possible to change these assumptions easily if our knowledge about the domain changes. Hard-coding assumptions about the world in programming-language code makes these assumptions not only hard to find and understand but also hard to change, in particular for someone without programming expertise. In addition, explicit specifications of domain knowledge are useful for new users who must learn what terms in the domain mean.

Often an ontology of the domain is not a goal in itself. Developing an ontology is akin to defining a set of data and their structure for other programs to use. Problem-solving methods, domain-independent applications, and software agents use ontologies and knowledge bases built from ontologies as data.

What is the difference between a Taxonomy and an Ontology?

In the world of information management, two common terms that people use are "taxonomy" and "ontology" but people often wonder what the difference between the two terms are.

On the technical side, ontologies imply a broader scope of information. People often refer to a taxonomy as a “tree”, and extending that analogy I’d say that an Ontology is often more of a “forest”. An ontology might encompass a number of taxonomies, with each taxonomy organizing a subject in a particular way.
A taxonomy generally is limited to a specific subject area, for example Products or Medical Conditions. Taxonomies are valuable when you want to add structure/context to unstructured information to make that information more easily searchable, For example, if a taxonomy is used to

tag documents in a search index, then when a user does a keyword search of this content, the Taxonomy can be presented on the left hand side of the search results as filter options for the end user. Multiple taxonomies can be combined together as filters to make for a powerful drill down search experience. This is what you see on many leading ecommerce sites like Amazon.

Ontologies can be thought of more like a web, with many different types of relationships between all concepts. Ontologies can have infinite number of relationships between concepts and it is easier to create relationships between concepts across different subject domains .For example, you could create a relationship between the topic of "Wood" in a materials taxonomy and "Chair" in a products taxonomy. Relationship types could be "example of", "Purpose of" or "Part of". Ontologies would be used when wanting to create a more sophisticated information model that might be deployed to do advanced natural language processing or text analytics. Ontologies would allow you to better understand things like cause and effect between two concepts within a corpus of information. Ontologies can also power question answering engines: for example, if I search for "Who was the 16th president of the USA?" an engine leveraging ontologies could return a specific result of "Abraham Lincoln"

Ontology in its simplest terms:
- What is the data
- What does it mean
- Where is it from
- Why do we need it – Once we know that, the real data we need is at hand.

Pete Ianace is Chief Operating Officer/Executive Vice President at No Magic, Inc. a provider of modelling solutions for enterprises

Published by IDM http://idm.net.au/article/0010860-where-real-value-big-data

Submitted by: Alan Kong - PROV