Data Analytics for Linked Data in Real Time

Shankarayya Shastri, M. Pooja and R.M. Soumyashree

Abstract--- *Big data is a term that has risen to use describing data that exceeds the processing capacity of conventional database systems. With many solutions entering the marketplace it's all too easy to focus on the technology that allows big data processing for real-time business analytics and yet to lost sight of the long-term goal of integrating many data sources to provide even potential.*

Keywords--- *Big data, IOT, Unstructured Data, Social Web, Linked Data.*

I. INTRODUCTION

NOW a day's Big data is one of the hottest topic in IT and education sectors. Unfortunately, just as cloud computing was previously hyped big data is in danger of being over-hyped to the point that it becomes an nuisance for some and confusing for many. But there is, undoubtedly, some value to be derived from analyzing ever-increasing volumes of data that were previously inaccessible, or just too difficult to process – looking for patterns and information that were only previously available to huge corporations because of the costs involved and then presenting that to the business in a way that can be used to drive new or developing business areas.

Now, through a combination of commodity hardware, cloud computing scale and open source software, big data processing and effective presentation is becoming accessible for small startups [1].

A. What's so Big About Big Data?

Right now we are generating more data than at any point in our history. In just the four years (up to 2016) the volume of data increased by a factor of six to an estimated 988 exabytes [2]. The "digital universe" was expected to pass 3.8 zettabytes in 2018 [3], is expected to reach 4.7 zettabytes in 2020 and may approach 8 zettabytes by 2025 [4]. To get some example of just how significant this explosion in data is, think of a single megabyte of information being equivalent to a square metre of land mass. In 2020 we would have covered the world with data. By contrast, in 1920 we would have covered an area the size of Madagascar, and in 2020 we will need 1700 globes to represent the volume of data we will have generated.[5]. Or, to visualise this volume in another way, the 988 exabytes of data in 2010 is roughly equivalent to a stack of novels from the Sun to Pluto and back [2]. This data explosion shows no sign of decline and is likely to accelerate with new data types (sensor data: record structure; and social networks and transactional) with greater access to networked devices (such as smart meters and smart phones with geopositioning data). That data is accumulated from a variety of sources but there are two in particular that are driving this explosion, combined with decreasing storage costs:

1. The "Internet of things" with a variety of sensors collating information on our activities and our environment (the number of connected devices globally is expected to rise 11-fold from 4.5 billion in 2010 to 50 billion in 2020 [2]).
2. The social web of networks sharing information about our activities, interests, location, likes and dislikes. In addition to those consider the private data stores created on our financial transactions, 'phone calls [6], health records [7], CCTV [8], etc. together with other online activities generating text, audio, video, click-streams, log files and more.

The McKinsey Global Institute describes big data as "the next frontier for innovation, competition and productivity" [9] but, put simply, it's about analysing masses of unstructured (or semi-structured) data which, until recently, was considered too difficult, too time consuming or too expensive to do anything with. Not only is some of that data now required to be kept for regulatory or compliance purposes (e.g. compliance with Safe Harbor in the United States or with data discovery laws in European Union nations) but there is also insight to be gained from looking for patterns in the way that we interact. For example: carbon information may be required in countries where sustainability legislation is being introduced; information about financial transactions and bank liquidity may be required to comply with financial regulations; scientific breakthroughs create new information sources (either new discoveries or new sensors); location information can help to optimise the physical position of moving assets (human or inanimate); and social graph information can be used to help companies better understand the ways in which they work to improve knowledge-worker productivity [10].

B. The Problem with Big Data

In solving one problem, our new-found ability to understand and then exploit big data has created another. Over many decades scientists established models for data processing based on efficient, structured, databases and the new world of unstructured data does not fit. Consequently we run the risk of creating pillers of data, each with their own limitations and benefits. This paper puts forward a view that the key to extracting value from big data lies in a related

Shankarayya Shastri, Department of CSE, SDM-CET, Dharwad. E-mail:Shankarayya Shastri

M. Pooja, Department of CSE, VTU, Belgaum. E-mail:mallasure.pooja@gmail.com

R.M. Soumyashree, Department of CSE, VTU, Belgaum. E-mail:soumya.rm28@gmail.com

DOI:10.9756/BIJSESC.8238

concept; the concept of linked data exploitation. Linked data offers the potential to create massive opportunity from countless data sources, both open and closed. And, with linked data as a broker, we have the ability to extract new data from old, creating insights that were previously unavailable, and facilitating exciting new scenarios for data processing.

II. Approaches to Data Management

1. Transactional data
2. Analytical data
3. Unstructured data
4. External (open) data sources

In the UK, the government was launched an Open Data Institute to support businesses in the exploitation of open data, based in the East London Tech City (Silicon Roundabout) and led by Professors Sir Tim Berners-Lee and Nigel Shadbolt [16]. Typically the Government-owned data sites provide metadata to describe the datasets, information about the datasets and tools for access to the datasets[17]. In the UK, new licensing constructs have been created to allow data to be used freely and flexibly [18].

The UK's plans include [19]:

1. Linked data services to track healthcare impacts and improve medical practice.
2. Enabling citizens to access their personal medical records.
3. Real-time and planned information on trains, buses and road networks for more efficient transportation and logistics.
4. Allowing third parties to develop applications for businesses and consumers using data sets such as weather data and house price information.

Met Office data has already been made available, data from the Land Registry will be released in March 2012, Department for Transport data will be made available in April 2012 and the NHS released its data in September 2012 [15].

External data sources offer massive potential for commercial use but it is worth noting that they are outside the control of the organization consuming them and their quality may be unknown. Feedback mechanisms and standards will be created for improving the quality of open data but, for now, they will be viewed with caution by some organizations. Even so, there is significant interest in open data and national media organisations have already made extensive use of public data to report on the news through a new medium of data-driven journalism, with examples including the Guardian Data Store3 . By locating data, filtering/interrogating, combining/mashing-up, visualising and reporting stories based on data, this new model for journalists is just one example of the massive social implications of big data [20][21].

III. Characteristics of the Various Data Models

Unfortunately, as each data model has been refined, limitations have emerged that result in the creation of a silos, each suited to a particular type of system and with its own benefits and restrictions. Each data type has its own management requirements, and its own strengths and weaknesses. No one data model is likely to consume the others so we need a solution that acts as broker access to all type of data, minimising the overhead of maintaining several disparate data sources.

Table 1: Data Model Characteristics

Data model	*Attributes*	*Flexibility*	*Age of data*	*Quality of data*
Transactional	known	Fixed schema	short	High
Analytical	aggregated	Fixed schema	long	Typically high
Unstructured	unknown	implied	Random	Low
External	variable	variable	variable	variable

IV. A Functional View on Big Data

One, simple, description of big data is that of a "firehose" that requires filtering and plumbing to ensure that the right data is received at the right time [22]. Whilst that provides an (extremely) high level view, there are certainly a number of core functions that any big data solution should provide.

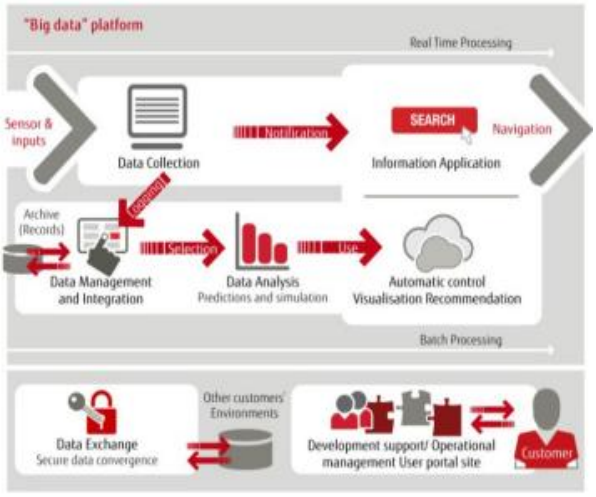

Figure 1: A functional view of a big data platform

V. Linked Data

Whilst there are exceptions, big data tends to be unstructured (i.e. has no formal schema) and metadata (data about data) becomes important (for example location data can help to make some sense of the data in that it provides some structure). Linked data provides some significant advantages in tying together different records to provide a view of the bigger picture. To give an example, humans can relate different types of information on a web page (for example: a profile page about someone, with some status updates and some geo-tagged pictures) but machines can't. If, instead of linking to the container (i.e. a web page), we link to the data within that container, we can create machine-readable relationships [28]. In short, linked data sees the web as an immense database that can be mined to link to data, rather than document-based resources; however it may be helpful to look in more detail at how linked data works.

A. How Linked Data Works

There are three rules to linked data [29]:

1. Linked data uses HTTP URIs – not just for documents as with "traditional" websites but for the subjects of those documents – places, products, events, etc.
2. Fetching data using the URI returns data in a standard format, with useful information such as who is attending an event, where a person was born, etc.
3. When the information is retrieved, relationships are defined, for example a person was born in a particular town, that a town is in a particular country, etc.

Importantly, those relationships are also expressed using URIs, so looking up a person links them to the town where they were born, and that links to a region or a country, etc. A model is required to identify the data within a resource and a file format to encode it in – one such format is Resource Description Framework (RDF) [28], a standard for data interchange on the web4 . RDF extends the linking structure of the Web to use Universal Resource Indicators (URIs) to name the relationship between things as well as the two ends of the link [30]. It provides a framework (structure) for describing resources (assets) [31]. Effectively, using addresses to identify assets RDF is to a web of data what HTML is to a web of documents[5]. Or, to put it another way the Internet is about connecting computers, the world wide web is about connecting documents and now we are focusing on a "giant global graph" to connecting the things that those documents are about [32] – i.e. the data, information and material information (or knowledge) that influence our decisions [10]. RDF has an elegant solution for identifying data – using a grammar of subject, predicate and object [28] as shown in the example in Figure 2:

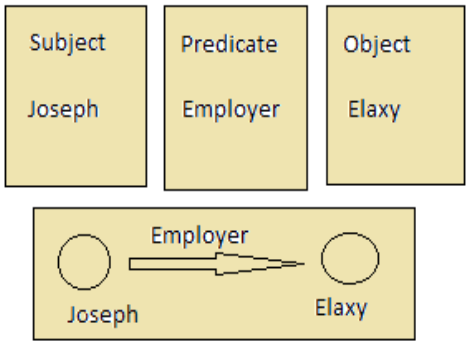

Figure 2: Linked Data Triples

RDF is both flexible (in that data relationships can be explored from many angles) and efficient (it is not linear like a database, nor is it hierarchical like XML) but it's still just a framework – to define and classify entities and the relationships between them ontology is required. The Web Ontology Language (OWL) is based on RDF and provides an extended vocabulary to describe objects and classes5. Importantly, this allows for inference: the creation of new triples based on existing triples to deduce new facts based on stated facts [31]. It is specifically defined to add capabilities to the web that may be distributed across many systems, scale, remain compatible with web standards, and be open and extendable[33].

As with human communication, vocabulary can emerge according to uptake. This can be for specific domains of interest, or to introduce additional generic ways of describing relationships between information. For example, the Simple Knowledge Organisation System (SKOS) is based on RDF but is designed to express hierarchical information6 – broad/narrow terms, preferred terms, and other thesaurus-like relationships [31].

Whilst RDF is not intended for human readership [34] we can still use triples to build a graphical representation of the relationships [28].

In Figure 3, the round nodes are resources. Resources may be typed with a class. Square nodes are classes. A class is also a resource (i.e. metadata is also data)[35].

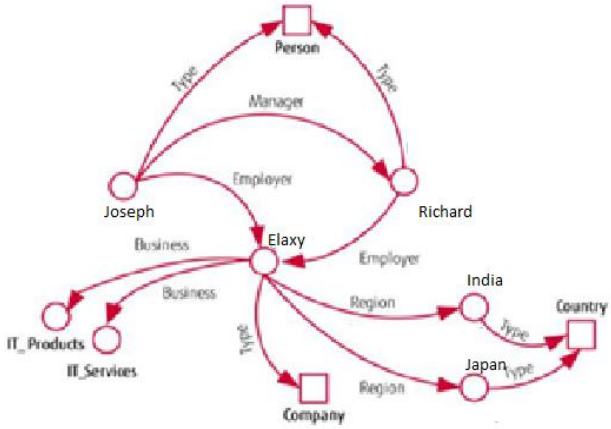

Figure 3: Linked Data Graph

Each resource is represented as a web resource, named with an HTTP URI (textual properties in the content, hyperlinks to related pages) Resources may be distributed and refer to other resources with URIs.

In Figure 3, the triples are:

<Joseph><hasEmployer><Elaxy>

<Richard><hasEmployer><Elaxy>

<Joseph><hasManager><Richard>

<Elaxy><hasBusiness><IT_Products>

<Elaxy><hasBusiness><IT_Services>

<Elaxy><inRegion><India>

<Elaxy><inRegion><Japan>

<Joseph> and <Richard> are of type <Person>, <Elaxy> is of type <Company>, and <India> and <Japan> are of type <Country>.

Using this structure, which is both human and machine readable, we can navigate and query data. For example, we can query based on triple matching and we can build complex queries that are then filtered.

B. The Linked Open Data Cloud

Using linked data based on a combination of RDF, OWL and SKOS, we can see how our data might easily be joined with other data sets. For example, extending the graph shown in Figure 3, there may be a database of companies that contains more information about Elaxy; or a database of countries that has information on currency, GDP, government structure, etc. Quickly, many linked data sets lead to others and create a cloud of linked, open, data – the semantic web [36].

VI. SUMMARY AND CONCLUSIONS

At the start of this paper, we drew comparisons with cloud computing in that big data is becoming massively hyped. Just as with "cloud", which is now approaching acceptance as a business model to the point that it has become the normal way to provide computing and application services, there will be a period of consolidation before we truly understand the application of "big data" solutions and "big data" becomes just "data" (i.e. business as usual). Even so, the volumes of data that we process, their variety in structure and the need to make timely decisions will lead to the creation of a new class of business systems.

These business systems include a number of capabilities that should be considered core to a "big data" solution, including:

1. Data collection and detection
2. Data management and integration
3. Data analysis
4. Information application, automation and visualisation
5. Data exchange
6. Development support and operational management
7. Exploitation of cloud computing architectures

Because there is limited value that can be obtained from any one source, aggregation of multiple sources is key to unlocking the potential in our data; however an unfortunate side effect of the ability to analyse increasing volumes of data using commodity infrastructure and cloud services is the emergence of unstructured and semi-structured data which does not fit well in our traditional structured databases and so creates issues around data management.

With many solutions entering the marketplace it's all too easy to focus on the technology that allows processing of new data streams but simply focusing on Hadoop or other No SQL technologies is not enough. It's important not to lose sight of the long-term goal of integrating many data sources to unlock even more potential in data – and the current technology landscape is a barrier to meeting business expectations which include:

1. Greater accuracy (derived from larger data sets).
2. Immediacy (near-real time data, from new data sources).
3. Flexibility (not constrained by database structure).
4. Better analytics (the ability to change the rules).

Linked data has the potential to provide a new architectural pattern for mapping and interconnecting, indexing and feeding real-time information from a variety of sources. Critically, it can be represented in human or machine-readable form and also allows new relationships to be inferred from existing data, creating new insights for further analysis. Linked data integrates all data – whether that's previously inaccessible big data that's got the IT industry in a buzz; structured data in traditional databases; or the increasing number of external (open) data stores.

For enterprises there are number of considerations (and some challenges) to address around data integrity, integration, data management, data replication, data quality, data storage and, critically, data security but none of these should be insurmountable. They do require careful planning though.

Once we've linked the myriad data stores at our disposal then we can generate new data from old, trade information, and extract new knowledge in support of more services, and even new business models, for example based around mash-ups, analytics, geospatial representation and augmented reality.

Exploiting linked data potentially offers massive opportunity in the integration of myriad data sources, both open and closed. With linked data acting as a broker, we have the ability to extract new data from old, creating insights that were previously unavailable, and facilitating exciting new scenarios for data processing.

REFERENCES

[1] E. Dumbill, "What is big data", 2012.

[2] Townsend, Eddie. UK Future Internet Stategy Group: Future Internet Report, Technology Strategy Board, 2011.

[3] B. Woo, D. Vesset, C.W. Olofson, S. Conway, S. Feldman and J.S. Bozman, "Worldwide Big Data Taxonomy", IDC report, 2011.

[4] Gens, Frank. IDC 2012 Predictions: Competing for, 2011. [Online]http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf.

[5] Sanderson and Mike, "Linked data for executives: building the business case", London:British Computer Society, 2011.

[6] Lawson and Stephen, "Nokia Siemens brings big data analytics to mobile carriers", 2012. [Online] http://www.cio.co.uk/news/3337309/nokia-siemens-brings-big-data-analytics-to-mobile-carriers/.

[7] Everyone 'to be research patient', says David Cameron. BBC News, 2011. [Online] http://www.bbc.co.uk/news/uk-16026827.

[8] Best, Jo. Big data: cheat sheet. Silicon.com. 2011. [Online]http://www.silicon.com/management/ceoessentials/2011/12/19/big-data-cheat-sheet-39748353/.

[9] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition, and productivity", Information as Strategy, 2011.

[10] B. Hopkins, B. Evelson, B. Hopkins, B. Evelson, S. Leaver, C. Moore, and M. Cahill, "Expand your digital horizon with Big Data", 2011.

[11] Woodward and Alys, "IDC European Software Predictions", 2012.

[12] Sand CDBMS 6 at Dunnhumby - Big Data, Big Users, Security and Analytics, 2010. [Online]15 http://www.youtube.com/watch?v=7iN1bfqWxCk.

[13] Swabey and Pete, "Getting relevant. Information Age", 2007. [Online]http://www.information-age.com/channels/information management/it-case-studies/277256/getting-relevant.thtml.

[14] Du Preez and Derek, "European Commission launches open data strategy", 2011. [Online]http://www.computing.co.uk/ctg/news/2131718/european-commission-launches-strategy-europe.

[15] Open Data Institute to help drive innovation and growth. Technology Strategy Board (Innovate UK), 2011.

[Online]http://www.innovateuk.org/_assets/0511/open%20data%20institute%2029nov11%20final%20(2).pdf.
[16] About Data.Gov. Data.Gov.
[Online] http://www.data.gov/about.
[17] UK Government Licencing Framework. The National Archives, 2011. [Online] http://www.nationalarchives.gov.uk/information-management/uk-gov-licensing-framework.htm.
[18] Open data measures in the Autumn Statement. UK Cabinet Office, 2011. [Online]http://www.cabinetoffice.gov.uk/news/open-data-measures-autumn-statement.
[19] Bradshaw and Paul, "How to be a data journalist", The Guardian Data Blog, 2010.
[Online]http://www.guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide.
[20] Lorenz and Mirko, "Data-driven journalism: what is there to learn?", Slideshare.net, 2010.
[Online]http://www.slideshare.net/mirkolorenz/datadriven-journalism-what-is-there-to-learn.
[21] Wang and R. Ray, "Twitter", 2012.
[Online]https://twitter.com/#!/rwang0/statuses/160140204873744385.
[22] Wu and Michael, "Searching and filtering Big Data: the 2 sides of the "relevance" coin", Lithosphere, 2012.
[Online]http://lithosphere.lithium.com/t5/Building-Community-the-Platform/Searching-and-Filtering-Big-Data-The-2-Sides-of-the-Relevance/ba-p/3807 4.
[23] Kim and Ryan, "Run Keeper builds a fitness network with Healh Graph API", GigaOm, 2011.
[Online]http://gigaom.com/2011/06/07/runkeeper-builds-a-fitness-network-with-health-graph-api/.
[24] Taylor and Colleen, "Kaggle gets $11M to crowd source big data jobs", GigaOm, 2011.
[Online]http://gigaom.com/2011/11/03/kaggle-funding-max-levchin/.
[25] Barton and Robin, "Code of conduct: The relentless march of the algorithm", The Independent, 2012.
[Online]http://www.independent.co.uk/news/business/analysis and features/code-of-conduct the-relentless-march-of-the-algorithm-6288080.html.
[26] Treanor and Jill, "Ultra-fast trading blamed for flash crash", The Guardian, 2011.
[27] [Online]http://www.guardian.co.uk/business/2011/jul/08/ultra-fast trading-blamed-for-flash-crash.
[28] Acuna and Antonio, "Linked data for executives: building the business case", London: British Computer Society, 2011.
[29] Berners-Lee and Tim, "Talks: Tim Berners-Lee on the next web", TED. 2009.
[Online]http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html.
[30] Resource description framework. W3C semantic web, 2004.
[Online]http://www.w3.org/RDF/.
[31] Lovinger and Rachel, "RDF and OWL: A simple overview of the building blocks of the semantic web", Slideshare, 2007.
[Online]http://www.slideshare.net/rlovinger/and-owl.
[32] Berners-Lee and Im, "Giant Global Graph. Massachusetts Institute of Technology Decentralised Information Group", 2007.
[Online] 21 November http://dig.csail.mit.edu/breadcrumbs/node/215.
[33] Herman and Ivan, "Web Ontology Language (OWL)", W3C Semantic Web, 2007.
[Online]http://www.w3.org/2004/OWL/.
[34] W3Schools, RDF Tutorial, W3Schools.
[Online] http://www.w3schools.com/rdf/.
[35] Menday and Roger, "A perspective n DaaS", Fujitsu Laboratories of Europe Limited, 2011.
[36] Cyganiak, Richard, Jentzsch and Anja. "Linking Open Data cloud diagram",
[Online] http://lod