Keynote Paper:
Tracking Big Data: Real-Time Qualities for OBOR Businesses

Dr. John Hamilton
*Chair Professor Management and Governance, JCU, Cairns, **AUSTRALIA***
John.Hamilton@jcu.edu.au

ABSTRACT
Big data capture presents additional business intelligences for corporate leaders and businesses when considering their leading-edge 'One Belt One Road' (OBOR) – now termed Belt Road Initiative (BRI) project. Big data can offer disruptive changes when applied to latest technical and software innovation deliveries. Big data integrates social, mobile, analytics and cloud information. Its analysis offers an over-the-horizon, time-shift jump into tomorrow's competitiveness - with linkages often targeted towards profit generation. Big data values are extractable and interpretable through behavioural approaches. Big data qualities can also be extracted and interpreted and can be used to guide improvements in a business's deliverables. Real-time big data tracking remains highly complex and this presentation shows that both values and qualities components can be extracted and interpreted. In BRI applications a Hadoop MapReduce approach, or a similar text-extraction approach, can be applied across any of China's foreign (and/or domestic) businesses as they work to improve their international projects. Big data values and qualities research agendas remain essential if China is to maintain and develop its dominant BRI partnering nation status.

Keywords: Big Data, Text Capture, Audio Capture, Image Capture, Values Extraction, Mining, Quality, Belt Road Initiative, One Belt One Road

1. Introduction

Big data captures databases, unstructured data, and semi-structured data, from a variety of data sources including text, audio, image/video, and selective other digital sources. Big data can be analysed using intelligent, business software intelligence tool extractions – particularly with a view to providing business insights. The volume of big data, the velocity (or speed of change in the data), the veracity (or uncertainties within the data), and the value propositions across the available big data, each deliver additional complicating features to the software analytics applied.

This presentation investigates open source big data. Open source software (OSS) big data manipulation began in the 1950's when computers first developed into rapid, but simple, manipulation tools. By the 1960's lead universities shared, and further developed the software. In the 1970's operating systems still remained limited in their scope and applications – still only allowing certain software modifications. Across the 1980's OSS developments remained small-scale - with free GNU operating system and software made available across OSS communities. Throughout the 1980's and 1990's patenting, fees, and business modelled hardware-software packages, restricted OSS commercial tasking deliverables. During the 1990's the acceptance of Linux, and the internet jointly allowed OSS advancements into web tools. In the 2000's OSS became actively-accepted as large tech companies developed new tools and specific-purpose business-ready programs. During the 2010's OSS continues to grow - as many seek access to the talented OSSS developers that can help quickly solve their software, or tool, or hardware development problem(s).

GitHub
GitHub is an OSS company that offers Git - a distributed version control system, for repository hosting. Git tracks content changes and provides ways to share content with other developers. GitHub offers web-based hosting as its version control. It accepts a wide variety of programming languages - including Java, JavaScript, Python and Ruby. Each GitHub project houses all the activities undertaken in an individual GitHub supported repository. These Github repositories (repos) house all of the individual project's developments as contributes by all participants. A core team and creator manages the project. Some OSS participants are allowed to watch a project's development activities. Some OSS developers (OSSDs) fork and sample the project main branch code - seeking new solutions, or to add ideas. Some OSSDs pull-sample, solve a specific point, and feedback. Some OSSDs review and test the code additions. Some

OSSDs enter as new contributors, and so on. Thus each GitHub project has numerous and cross-purpose participant OSSD contributions.

Seffinga, Lyons and Bachmann (2017) assessed the blockchain activities (including transaction fees and computational services) of GitHub. These Deloitte authors found 86,034 GitHub blockchain projects existed on 12th Oct 2017. Across 2016 26,885 new blockchain projects were added to GitHub - with only 8% of projects being active, and only 5% of forked projects surviving (and blockchain projects existing for an average of only 1.22 years. Here, Python is a common language deployed for blockchain's decentralized virtual machine platform beyond cryptocurrency protocol Ethereum (Buterin, 2013).

The GitHub community in 2017 reached 24 million developers from 200 countries, working across 67 million repositories, using 337 unique programming languages, and sharing code sharing code across 25 million public repositories. JavaScript remained the most popular language followed by Python, Java, Ruby, PHP and C++ (Octoverse.github.com/). Around half the 100 largest companies in the United States (by revenue) use GitHub Enterprise to build software and to power other industries – from finance to retail. GitHub continues to grow as shown in Figure 1.

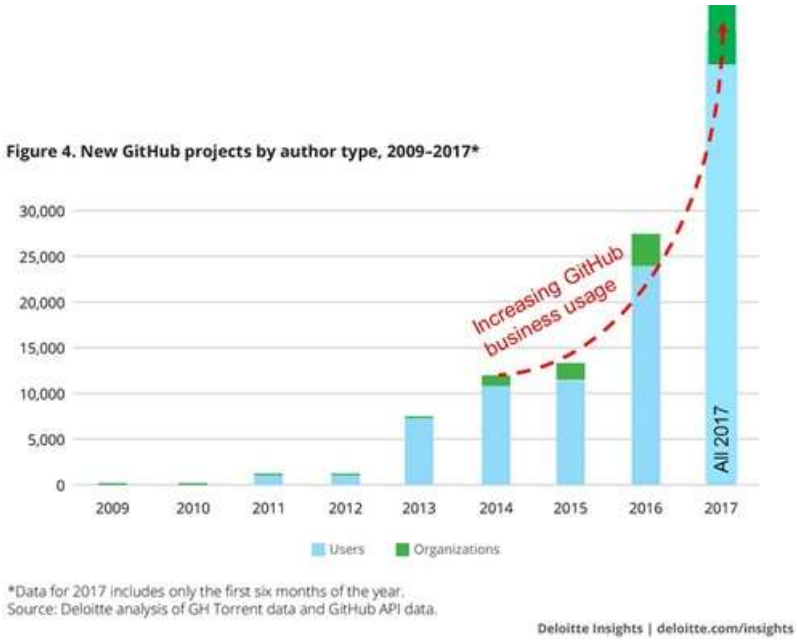

Figure 1: New GitHub projects to end 2017 (modified from Deloitte Insights, 2017)

Blake-Plock (2017) adds that value can be extracted from the Forbes big data landscape of apps, intelligences, analytics, ops, services, and visuals. Approaches capture both structured and unstructured data - from social, image, html, email, databases, image, audio and the like, and place these into storage that is then analysed as required for further intelligence, insight and understanding.

2. GitHub and Big Data

In 2017 at the 21st ICIT conference, we reported on an BRI pathway to in the future deliver business qualities, and to do so in stages. We now advance this work, through an understanding of GitHub and its software construction processes. GitHub has around one OSSD per three repos (Deloitte, 2017). Hence, OSSDs selectively-choose when and how much to contribute towards some aspect or aspects of an OSS project. Hence, by benchmarking the OSSD roles contributing towards a project's activity it is possible to then generate ways to compete a project faster.

With GitHub OSSDs being in short supply, and with multiple contribution pathways adding/deleting/informing into the project repositories, it is also possible to assess this big data for the

real business value components, and to determine the dimensional qualities pertinent to a business product or service collectively embedded within these data sources.

GitHub splits its OSSDs into groups based on their behaviours. Watchers receive notifications of content changes. Stars indicate a liking for the project. Contributors request to contribute into the project. Commits incrementally deliver pull-request or document revisions/changes. Forks offer independent development solution approaches. Version releases meet project milestones. Open issues address identified coding problems. Closed issues fix coding problems. Open pulls house OSSD problems not fixed (or fork commit problems still developing further). Closed pulls are pull requests adopted into master branch.

Using path analysis we can assess these for different programming languages and determine their total effects on the project's activity. Our studies indicate each programming language offers different key pathways that contribute strongly to a project's activity levels. This further suggests cross-promotional strategies through selected sites including: wikis, forums, Facebook, conferences, HackerNews, GitHub Explore, fan pages, and the like may draw additional OSSDs into the project. In particular we seek those OSSDs that show collaborative project value – captured into the project as: performance, quality, service, economic value, and emotional perceptions. We also note the project's activity level achieved is point-in-time dependent, and cyclic, and that creating a hype of interest through SMS boosts, at specific times, may be of further use when attempting to speed a project.

The mining of big data also needs to be filtered. So we need to choose whether we mine against the active repository branch, or the project's activity level, or the pulls, or the issues, or the releases, or the commits, or the forks, or combinations, and the like. This depends on what intelligence is specifically being sought. We also note that in terms of measurement data this depends on each path models': correctness, language, size (number) of repos studied, normality and removal of outliers (such as an interesting repo with variable activity), longevity (duration) of project, recognition of project, repo ongoing activity. These, and other specific aspects, represent other big data filter considerations.

3. Extraction of Useful Language

Useful language can be extracted from big data provided the target purpose is known. As GitHub is an emotional, free-presence place, we apply perception and behavioural approaches. Hence, we trial the most active GitHub projects (based on number of forks) across differing languages. We adopt the Google Hadoop MapReduce approach, and only tackle text extraction, sorting and re-grouping into value dimensions (performance, quality, servicing, eco-value, emotive satisfiers) that are then path modelled, and benchmarked across consumer activities as a motive-to consumption-to-gratification behavioural process. We add real-time feedback loops components (to improve ongoing consumer contributions), and benchmark the system against selected opposition projects. We then use business targets, and pursue ways to then improve the current net-project activity-level. This approach is summarized in Figure 2.

Another text approach is by engaging deep learning algorithms. But as the simpler Hadoop MapReduce approach can bow incorporate image capture through OCR, relative edging, etc. audio via text conversion, and smaller video via audio to text, and image interpretation, etc.

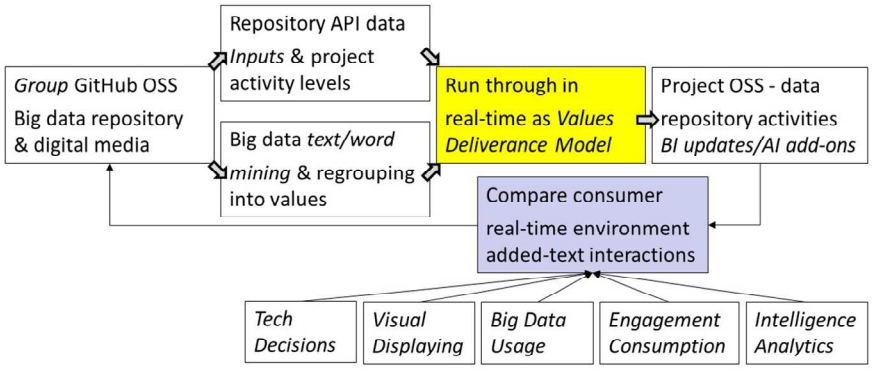

Figure 2: Behavioural Values Extraction

4. Big Data Qualities

Big data qualities can be extracted as intrinsic, contextual, product and service quality domains as discussed in Hamilton and Tee's (2018) ICIT paper. In 2017, the process of big data OSS mining for BRI Digital Qualities was deliverable across 4 stages, and 5 levels. In 2018, it is now deliverable from digital sources in one overarching step that engages all Hamilton and Tee's (2018) Table 1's cells at once. Appropriately filtered data capture across transactional, transformational, and authentic leadership approaches, plus a Hadoop MapReduce approach, with appropriate data storage capacities, suitable collation, strong processing power, plus understanding and application of behavioural values and business qualities can unleash new near real-time business monitoring and subsequent consumer targeted adjustments.

Table 1: Big data OSS mining delivers across 4 stages of BRI Digital Quality (Hamilton & Tee, 2018)

Quality advancement systems	Stage 1	Stage 2	Stage 3	Stage 4
Renewal systems	Digitization to data clouds	IoT digitization	Industrie 4.0 Entrepreneurship	Hype cycles delivery
Engagement systems	Flatten digitally interconnected organization	Specialization to selling organization	Precision, remote, specialized controls	Efficient service value networks (SVNs)
Collaborative systems	Workforce collaboration	Talent management	Consumer value focused	Consumer personalization

5. BRI Fast Rail Example

The BRI transport linkage is the belt corridor connecting land bridge BRI participating nations. Each emanates from China, and threads specifically across chosen Eurasian countries to its final destination. Rach belt corridor build involves, uses, deploys and engages: (1) Chinese finance, systems and knowhow, (2) Chinese products, (3) Chinese technologies and InternetPlus devices, and (4) primarily Chinese (expatriate) labour. Each BRI transport linkage belt also reliably-develops each partnering nation's interconnectiveness, provides ongoing new jobs, increases export/import trade capabilities, and increases the national wealth of all parties.

For the above system to truly work today it should be digitally, conjointly, and swiftly managed. This requires instant leadership-workforce interconnectivity, fail-safe cross-referencing communications systems, big data analysis, and extensive decision support systems. The creation and development of these connectivities systems into world-class consumer values deliverance systems - that can always deliver high quality solutions. This requires systems that both monitor, and deliver decisions, in near-real-time. Hence the purpose of this address.

These data captures first involve structured and unstructured mobile, social, web, dashboard, email, and database, and cloud sources. Next data extraction is delivered by including the desired components from such relevant data sources. Here, extracted text, text conversions, OCRs, and digital measures from around the transport system are interpreted, along with consumer commentaries. These initial steps are shown in Figure 3. The Hadoop MapReduce processes and Figure 2's staged values and qualities solutions are then applied to the text components extracted.

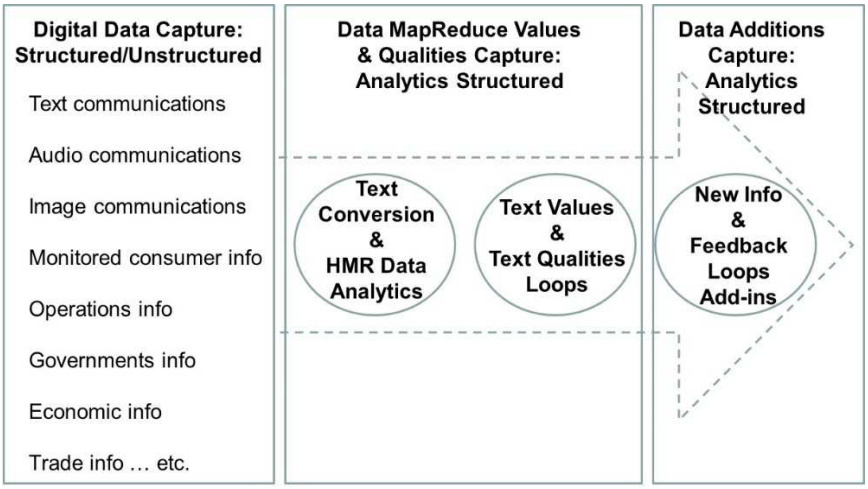

Figure 3: Data capture considerations

Next add-ins, and relevant feedback loop component updates, are near-real-time included, and this cyclic process is comparison-assessed for change(s). China's BRI fast rail operators can then digitally assess the pathways, and modify their overall system as required to better fit a specific BRI partnering nation's perceptions.

Extracting tagged HTML text involves coding to remove tags and scrape text. Depending on specific website functionalities, key major programming languages sometimes offer their own HTML parser. Others target specifics. For example, Octoparse captures all text web data, and can follow rules on what and how to extract text string including depth and breadth, but not image, video, audio or text-filing. Import.io can also slice/dice contents into a table, where further queries can be added. Scrappy can quickly screen scrape and web crawl to extract structured data.

Verma et al. (2016) discuss such large data capture approaches against statistical, text, audio, video, sensor, and bio-metrics data and within big data social media analytics, content analytics, text analytics, audio, and video analytics. They too recognize the need to consider storage and management, but especially data-retrieval functionality.

Google Cloud Platform allows the writing of specialized functions to transform data and chain it into a pipeline. Google Cloud offers some Machine Learning APIs that transcribe text-in-mages, audio, landmarks, identifications (facial recognition), language translation; and even helps re structure previously constructed prose. Specific Github OSSD API's also offer similar text capture outcomes. Hadoop allows the facility to read/write binary files. Hence, unstructured data (audio, image, video) is also binary-stored for further processing from the HDFS. The isolate HDFS values components are next gauged for relative and path importance. Positive-change, negative-change, or no-change are noted, and the system is intelligently and digitally recalibrated as and where required to improve its behavioral deliverables.

Value deliverance across any BRI system is vital from the consumers' (or users') behavioural perspective. Where the consumer (or user) positively recognizes the values provided by the system then degrees of satisfaction, loyalty and trust generally follow.

Similarly, qualities dimensions can be captured and gauged against standards or settings, and even mapped in terms of the consumers' (or users') ensuing satisfaction, loyalty and/or trust of the system. Hence, how does the BRI fast rail system ensure that it is always delivering ongoing positive values, and highest qualities, across each of its partnering nations, and why is this necessary?

If positive values and highest qualities are not perceived to be delivered towards the BRI partnering nation consumers (or users) of the fast train system, then disagreements may arise between partnering nations and China. The instantaneous spotting of a consumer (or user) perception change is a key prevention pathway towards negating such inter-nation disagreements.

Fast train, digital big data sorting with near-real-time values analysis, conducted as described above, and similar to as described in Figure 2, and with language translation being sufficiently precise, can deliver the continuous feedback-modifying update systems necessary to deliver ongoing consumers (and users) values improvements. This feedback loop system can also extend to workers and operators, to managers, and even to leaders.

The fast train system's deliverance, maintenance and improvements to its four qualities dimensions include:

(1) Intrinsic qualities being representative, unique, and/or irreplaceably-distinct characteristics of the system.
(2) Contextual qualities covering economic, social, physical components embedded within the system.
(3) Product qualities encapsulating multifunctional (yet simple) problem solutions, suitable-pricing, credible and appropriate workings and acceptable results within the system.
(4) Service qualities that are tangible, reliable, responsive, assuring and empathetic, and with components that jointly: appeal, perform-to-specification, and meet perceptions about the system (Wang et al. 1995, Wang, 1998).

The extraction, delivery and relative importance of each of these four dimensional qualities requires similar digital values collations, but they are measured and interpreted differently – depending on the capabilities of the system, the qualities targeted and the analysis applied.

This near-real-time tracking of big data is vitally important is maintaining positive consumer or user perceptions regarding the system's values and qualities, and this approach can be applied to China's foreign and domestic BRI partnering nation businesses.

6. Conclusion

Big data capture can be used to for additional business intelligence purposes. It has application for corporate leaders and businesses when working across borders on a leading-edge 'One Belt One Road' (BRI) project. Big data values and qualities research agendas are essential if China is to maintain and further develop its dominant BRI partnering nation status.

Big data intelligence deployment is often a disruptive change applying latest technical and smartest software innovations. It integrates social, mobile, analytics and cloud information. Its relevance interlinks with the hype cycle process, and learning is applied against these new innovation horizons. Big data analysis represents an over-the-horizon, time-shift approach and it is a jump into tomorrow's near-real-time competitiveness adjustments. Both the big data values components and their dimensional qualities can be interpreted, applied to the business' situation, and can generate business improvements that also deliver with linkages towards higher profit generation. Thus, big data, its analytics, and its interpretation conjointly offer new competitiveness avenues for digital leaders' wishing to maintain a leading-edge business positioning, and for those in pursuit of business highest qualities delivery.

Big data is highly complex, but its qualities tracking in near-real-time is approaching the commercial deliverables stage. The near-real-time tracking of big data is achievable using the approach discussed. The understanding of big data remains vitally important is maintaining positive consumer or user perceptions regarding a business system's pursuit of its values and qualities deliverables. A Hadoop MapReduce approach or a similar approach can be applied across any of China's foreign and/or domestic businesses as they work internationally on BRI partnering nation projects.

The near-real-time tracking of big data is achievable using the approach discussed. The understanding of big data remains vitally important is maintaining positive consumer or user perceptions regarding a business system's pursuit of its values and qualities deliverables.

Big data can be mined for business intelligence purposes - such as security, or precision. It has IoT integration applications that are of benefit to corporate leaders and businesses when approaching leading-edge 'One Belt One Road' (BRI) projects. In BRI applications a Hadoop MapReduce approach, or a similar text-extraction approach, can be applied across any of China's foreign (and/or domestic)

businesses as they work to improve their international projects. Ongoing big data values and qualities research agendas remain essential if China is to maintain and develop its dominant BRI partnering nation status. Thus big data BRI research agenda remains ongoing, as does research into big data values and dimensional qualities, and all are important if China is to maintain its internationally-dominant BRI partnering nation status.

References

Blake-Plock, S. (2017). *Where's the value in Big Data?* Forbes Pub. Available from: <https://www.forbes.com> [Accessed 8 Feb 2018].

Buterin, V. (2013). *Ethereum white paper*. GitHub repository. Available from: <https://github.com/ethereum/wiki/wiki/White-Paper> [Accessed 27 Feb 2018].

Deloitte (2017). *Over 26,000 Blockchain Projects Began in 2016*, Deloitte Report. Available from <https://www.coindesk.com/deloitte-report-over-26000-blockchain-projects-began-in-2016/> [Accessed 27 Feb 2018].

Hamilton, J.R., and Tee, S (2018). BRI: Mining business big data qualities. *22nd International Conference on ISO & TQM*, BIT, Zhuhai, China, 2-4 April, Vol. 22, No. 1, pp.1-7.

Octoverse. (2017). *The State of the Octoverse 2017*, Available from: <https://Octoverse.github.com/> [Accessed 28 Feb 2018].

Seffinga, J. Lyndon Lyons, L & Bachmann, A. (2017). The Blockchain (R)evolution - The Swiss Perspective. *Deloitte White Paper*. pp. 1-40.

Wang, R.Y., Storey, V C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE transactions on knowledge and data engineering*, Vol. 7, No 4, pp.623-640.

Wang, R.Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, Vol. 41, N0. 2, pp. 58-65.

Verma, J., & Agrawal, S. Patel, B., & Patel, A. (2016). Big Data Analytics: Challenges and Applications for Text, Audio, Video, and Social Media data. *International Journal on Soft Computing, Artificial Intelligence and Applications*, Vol.5, No.1, pp. 41-51.

Author's Background

Dr. John R. Hamilton *is Chair Professor of Management and Governance. He researches competitiveness, innovation and strategic futures. He has extensive corporate national (and international) leadership and management experience. He consults on online and/or offline engaging interactive environments, and develops capabilities for business-consumer real-time interfacing. Current research interests include: big data business value extraction, digital leadership, value-deliverance, social networks, corporate and virtual intelligences, cloud business scenarios, major-events management, tracking, and interactive learning. John's Acute Futures Group (AcuteFutures.com) engages international R&D task teams (Hong Kong, Indonesia, Singapore, and Australia) in pursuit of latest digital value-deliverance systems for global and futures-focused organizations.*