

## **BRI: Mining business big data qualities**

Prof. John Hamilton

Dr. SingWhat Tee

*College of Business Law and Governance; James Cook University, Cairns, Qld. Australia 4870*

[John.Hamilton@jcu.edu.au](mailto:John.Hamilton@jcu.edu.au)

[SingWhat.Tee@jcu.edu.au](mailto:SingWhat.Tee@jcu.edu.au)

### **ABSTRACT**

*Digital data capture from a business' big data sources remains difficult, but techniques are emerging. This paper reviews, describes, and connects usable big data components as an ingredient towards delivering an understanding of the qualities a Chinese business may astutely utilize as they expand along China's Belt Road Initiative (BRI) pathway. It suggests digital data across the business, web and social domains can provide sorted text extracts through Hadoop MapReduce algorithms, and these outputs can be qualities grouped to deliver additional business understanding concerning the qualities they aim to deliver. Similarly, qualities aspects can also be linked to audio, video and image files - but often for slightly different purposes. Hence, text mining appears the most important big data link to business quality. A digital 4 stage 5 level big data digital qualities system then strategically frames a pathway for qualities outcomes for BRI business ventures.*

**Keywords:** Big data, quality, data mining, Hadoop MapReduce, China, one belt one road

### **1.0 Introduction**

Big data is revolutionizing business operations - with 89% of businesses rely on text information and 68% using some form of predictive analytics as a means to improve their competitiveness. The big data market is likely to exceed \$84B over the next decade. This 17% compounded growth is a strong driver for ongoing business change. China has embraced this vision, and is aiming to be the global leader in artificial intelligence (AI) by 2030.

Today, several areas China is already creating horizon-shifting digital advancements (Hamilton, 2017). For example, at the micro level, Da-Jiang Innovations Science and Technology's compact, lightweight Mavic Pro drone flies four-miles and from hundreds of feet up its 4K camera and cutting-edge software follows subjects, avoids mid-air obstacles, and automatically returns-to-base. Other micro Chinese designs include Momenta's driverless-car software, Face++'s facial-recognition scanner and VIPKID's online-tutoring system.

At the macro level Chinese super-fast trains now target 500km/hr travel and Chinese state-run space contractors are refocusing this technology towards a 'high-speed 4,000km/hr flying train aimed at replacing key destination flight travel. Similarly China's digital technologies is linked through its 'Internet Plus' (encompassing mobile-internet, cloud computing, big data, IoT with traditional industries) Its Internet Plus has expanded across its Long March 2D rocket development. Its recent rocket launch delivered special cameras into space to provide detailed land surveys from orbit.

On a global terrestrial and maritime scale China's one-belt one-road strategy is providing a global framework across 65 belt-connected countries, and across it's the major shipping trade roads, including key ports. Here, China is scoping, and creating a vast array of complex builds that require extensive IoT and sensor connectivities, along with extensive embedded arrays of intelligent big data strategies.

Big data AI then generates a closer understanding of the behaviors, preferences and flaws within and around this system. It also supports Chinese businesses as they seek to expand their enterprises across this preferred belt and road strategy. Here the capture of social media data, browser logs as well as text analytics and sensor data can provide a much more complete picture of the dealings of their foreign entities and the domestic interactions with their home country customers.

This complex strategic AI development process also draws on emergent technological innovations. This encompasses the leading-edge adoption of technological advancements. It incorporates digitally-driven,

disruptive change (Utesheva, Simpson & Cecez-Kecmanovic, 2016), and the intelligent incorporation of emergent IoT digitally connection solutions incorporated draw on highest quality standards.

Chinese businesses can assess when to adopting emergent technological approaches by assessing their specific industry against Gartner's digitally-relevant 'Hype-cycles' (Gartner, 2018). This visual assessment enables Chinese businesses to choose their entry point, commercial viability of maturity, adoption of which technologies and applications, and how this chosen technology (or application) is likely to trend over time.

## **2. Data Variety**

Quality data capture is vital for all substantive businesses. It also encompasses a richness of sources such as text, imagery, videos, audio-files, and the like. However, measuring quality is not simple, and it remains a substantive analytical obstacle in facilitating the effective use of such large volumes of data.

Incompatible data formats, non-aligned data structures, and inconsistent data semantics all create significant analysis challenges. These also generate a diversity of approach opinions within the corporate and the OSS community. Currently, further mechanisms are being pursued to convert latent, unstructured text, imagery, videos and audio information into suitable numerical indicators that also deliver computationally. All these data forms potentially offer corporate analytical significance, and so each solution requires compatibility.

Thus for China, processing big data remains a challenge – from storage to management interpretations. Techniques can capture descriptive, estimative, patterning, predictive, prescriptive, optimization and benchmarking. Approaches can be gamified, economic, simulation, modelling or agile.

In developing BRI strategies each business must provide scaling capabilities through expandable data storage - so it can incorporate the ever-changing and ever increasing array of useful data streams - whilst still delivering the necessary information and knowledge to improve its services, operations, and/or capabilities.

## **3. The Incorporation of Social Networks**

Social media data streams are now prevalent sources from which to gauge trends in mining and analysis. In the West these streams include Facebook, Instagram, Twitter and other micro-blogging sites. In China, Renren is Facebook, Weibo replaces Twitter, and WeChat is a key mobile text/voice-message/image-posting/social communications service).

Outside China, Facebook, Instagram, Twitter can provide instantaneous engagements. Hence business algorithms must be capable of working across these dynamically-changing, time, format, and space constraints. For example Twitter delivers around 360K tweets per minute or around 200B tweets per year (Anon, 2018a). Facebook receives around 510K comments, 293K status updates and 136K photo uploads per minute (Anon, 2018b). Thus streaming capture and analysis data algorithms can offer real directive information sources for business, but these largely remain works-in-progress.

## **4. Opinion Mining**

Sentiment analysis offers a behavioral decision analysis opinion mining approach to assess such data sources. For example natural language tweets, comments and updates can be positively or negatively classified and used in learning algorithms to systematically identify, extract, quantify, and study subjective information. It can be used to sense attitudinal polarity change regarding a topic. These behavioral trends can be further mapped against smiles, likes and emotional cues (Anon, 2018c).

Websites offer another digital extraction addition. Here, structured and unstructured webpage data can be category and relationally extracted to guide and refocus ongoing digital business processes. Content Grabber, and Dexi.ios, like many other website extraction tools, engage web scraping agents (Anon,

2018d). Dexi.ios adds browser support allowing scraping and interaction with data from any website. However to date, the discovery pattern techniques available to transform and combine this extracted data into useful new datasets may still be considered as work-in-process solutions.

As technologies advance and their ensuing life cycle flow-ons first become more digitally productive greater consumer reach and engagement is achieved (Reeves, Zeng & Venjara, 2015). This in turn elicits behavioral feedback loops into the business. This data can be mined too.

## 5. Text Measurement

The Hadoop MapReduce approach assesses vast amounts of multi-terabyte data in parallel using large clusters of nodes that can both compute and store (Bifet, 2013). MapReduce (typically as Java C++, Python executables) and its Java Hadoop Distributed File System run concurrently - effectively scheduling tasks on the active data nodes across the aggregating cluster(s) (Cosentino, Izquierdo & Cabot, 2017). MapReduce first splits the input stream into independent data components. It task-reduces, and then map-sorts. Its master JobTracker and its one slave TaskTracker-per-cluster-node combine to schedule tasks, monitor them, re-execute any failed tasks and deliver these data components into a useful output file system (Anon, 2018e). ([https://hadoop.apache.org/docs/r1.2.1/mapred\\_tutorial.html](https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html)).

Other text approaches include Apache S4 real-time-combines data streams and processing-elements (Neumeyer et al., 2010), and Twitter's Storm (Anon, 2017a) ensemble learning classifiers approach which scale, parallel and process data feeds. Still other approaches combine aspects of these approaches. For example Lambda architecture processes data using batch layering (Hadoop) and stream-processing methods (Storm) to deliver greater fault tolerance, scalability, and dynamically changing inclusions (Anon, 2017b).

Thus, API (functions/procedures to access features/data) text assessment approaches (such as Hadoop MapReduce, S4 or Storm algorithms) rapidly process large amounts of data by parallel splitting/processing data as independent subsets. These are then reduced to deliver final interpretable outputs. Here, pattern techniques are added, but to date their effective-use and update-discovery patterns are still works in process.

## 6. Image Measurement

Mining image information engages algorithms that target-each, input and process the image input code and allow this digital information to be used later. Image information is either:

- (1) content-based information - analyzing large structured and unstructured text/audio information
- (2) structure-based information - showing linkages between image edges, and locating consumer preferred hub-points (nodes).
- (3) community-based information - providing current behavior patterning and/or helping predict emergent community properties.
- (4) social-correlation-based information - showing behavior connection links and relative strengths that can be leveraged to deliver greater brand awareness and sales (Gandomi & Haider, 2015)
- (5) optical-character-recognition (OCR) to explain image purpose.

Overall, image mining offers a more general perspective of a consumer's engagement with a business.

## 7. Audio Measurement

Audio analytics capture communication. They can analyse active calls, deliver cross-sell (or up-sell) consumer targeted marketing and/or deliver near-real-time feedback. These transcript or large vocabulary continuous speech recognition approaches first transcribe the audio speech content, then engage text-based analytics to find each search term in the transcribed file.

Audio analytics can also deploy interactive voice responses to communicate with consumers. These phonetics-based approaches first create a words sequence from the text and then compare these against input search terms. Thus audio analytics is still under-developed, but it is workable in controlled environments - such as within language training software packages.

## 8. Video Measurement

Video data is typically connected with security or the monitoring of buyer behavior. It houses huge amounts of data, but it must be automatically sifted and leveraged using API server-based or edge-based video analytics to locate and extract useful intelligence items.

Thus, video streaming data still requires quality near-real-time digital analysis tools.

## 9. Hype Cycles

Gartner's (2017) 'hype cycle' (Figure 1) captures latest emergent technologies within a relevant industry the context, as a five-phase, technologies-influence life-cycle. Hype cycles hold special relevance across the early stages of commercialization, commencing with an innovation trigger initiation - where a potential technology breakthrough emerges, but as yet has no commercial viability or usable product.

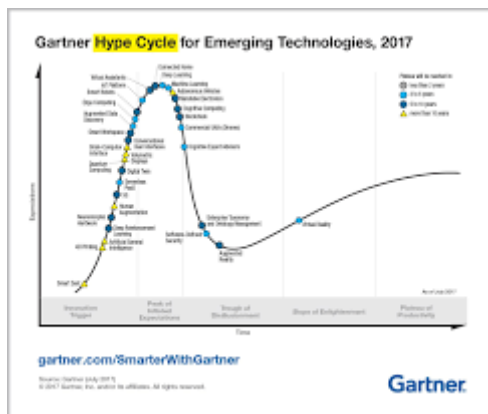


Figure 1: A Gartner 2017 hype cycle (available at: [https://blogs.gartner.com/smarterwithgartner/files/2017/08/2017\\_Infographic\\_R6A.jpg](https://blogs.gartner.com/smarterwithgartner/files/2017/08/2017_Infographic_R6A.jpg)) Emerging-Technology-Hype-Cycle-for-2017\_Infographic\_R6A.jpg)

As others tackle this new innovation and attempt to build their commercial products or solutions an expectations peak emerges. However, as commercialization implementation successes wane and many attempts fail, a disillusionment trough emerges with few persistent commercialization survivors remaining on task and improving their initial product. Next an enlightenment slope emerges as the technology begins to crystalize into an understood package, and finally a productivity plateau develops as more businesses enter and fund pilot product activities, and those achieving successes emerge and develop second and third generation improvements or product versions that commercially extend the initial innovative technology capabilities into successful mainstream businesses with clear payoffs and broad market applicability.

## 10. Big Data Management and Quality

Thus, across the above examples, the measurement of big data involves capture processes including: the acquisition of structured and unstructured data, the management selection of appropriate data measures, the operational assurance of valid data inclusions, the ongoing data research and development (R&D) additions, the production of outcomes data (as a usable format), the meaningful dashboard data distribution mechanisms, and the inherent legalities that accompany such measurement processes.

This suite of data management measures is likely particularly useful when deployed as potential near real-time process controls and where the relevant embedded data qualities can be gauged against Wang and Strong’s 2006 dimensions of: intrinsic (completeness, unambiguousness, meaningfulness, and correctness) and contextual (accessibility, and usefulness) motives (Haug & Stentoft Arlbjørn, 2011) along with product representation and service accessibility data measurement dimensions. The combination of these big data measurements can likely be modelled for total effects contributions. Recent assessments suggest little quantitative research has been done in this area (Chen, Chiang & Storey, 2012, Liebowitz, 2016, Sugumaran, Sangaiah & Thangavelu, 2017).

Business is now recognizing the values inherent in big data. High quality data is a necessary condition for the business analysis. This requires the development of a big data quality framework. The big data quality framework shown in Figure 2 first links data management encompasses the acquisition, the management, the production, and the distribution of the extracted big data components.

In Figure 2, the data quality dimension defines how these big data components meet the product and service qualities (Wang et al. 1995; Wang 1998; Kahn et al. 2002). Further, the analytics involved draws on the modelling, the analysis-approach and the interpretation of this grouped corporate data. Thus, the Figure 2 digital qualities approach is also likely highly-influenced by the business’ digital leadership and management style.

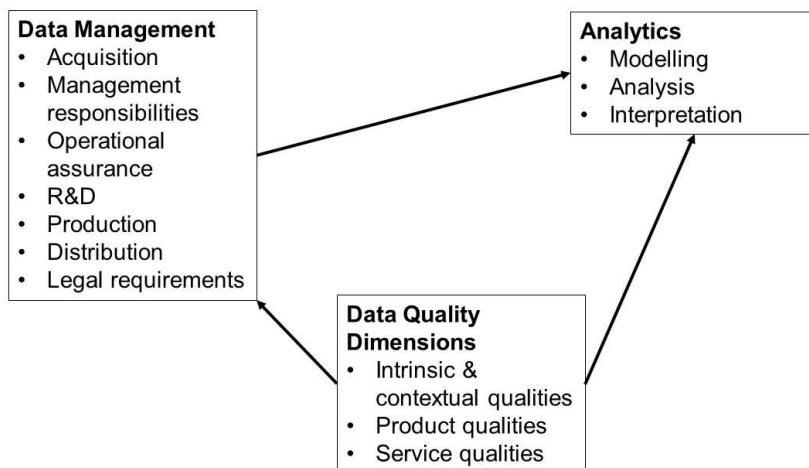


Figure 2: Big data quality framework

## 11. Delivering Digital Qualities

Digital qualities leadership draws on a conglomeration of transformational, transactional and authentic leadership ideals (Prince, 2017). Digital qualities leadership can be instituted at various levels of increasingly complexity. The technologies and strategic advancement systems to support and assess these complexities develop across four strategic stages as shown in Table 1. These qualities enhancements are deployed sequentially across five levels of development beginning at the top left with strategic digitization. Each subsequent qualities level then requires ongoing and increasing reliance on emergent technology systems, and each draws increasingly on associated digital intelligences.

For example, consider the deliverance of data mined digital quality as applied through Table 1. First the digital corporate leader renews and transforms the corporate’s strategic digital capabilities by extending existing capabilities into external cloud information sources. Here, the corporate’s internal intrinsic, contextual, product and service quality measures are engaged. This leadership also incorporates suitable corporate analysis tools that can assess such base-level big data capture. Here, corporate qualities outcomes are digitally-assessed against the industry’s quality standards and/or against the corporate’s quality standard measures.

Table 1 Digital quality: Staged business developments that draw on technologies and strategic advancements

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

The second qualities enhancements level (light grey) delivers more connectivities and information but it requires more corporate digitization expenditure. Further IoT digital information is provided via process and operational IoT and e-business connectivities. This links qualities aspects and speeds decision making at the corporate executive management level. With better information now available to corporate executive management, less lower-management processing levels are required. Hence a hierarchical dissolution of some management levels also ensues – delivering management efficiencies, and moving qualities assessments closer (and quickly) to the CEO.

Similarly, the remaining three qualities enhancements levels deliver a connected, informed workforce, technological innovations, high corporate precision, a service value networks approach with consumer personalized services and loyalty, and greater profit-driven revenue streams delivering high-quality, consumer-relevant outcomes.

Thus, a digital corporate leader (Prince, 2017) can potentially create a leading-edge high-quality corporate by digitizing, capturing, analyzing and executing the quality triggers available within its available big data.

## 12. Discussion

Chinese businesses seeking to lead in the digital age and seeking to extend into BRI ventures need to harness their big data and draw competitive positioning intelligence from these sources. Capturing data and extracting it from diverse business sources remains a challenge, but each data format – be it structured or unstructured, can be digitally accessed. Text remains the most utilized data source. Program combinations like Hadoop MapReduce offer a solid means to sort these acquired text sources. The astute interpretation of this text is the current stage of research. These authors currently believe behavioral assessments offer a strong pathway towards collating and interpreting such extracted and sorted text data into useable business enhancing performance data.

Delivering BRI qualities engages these digitized solutions, and links them into a 4 stage 5 level qualities system. This allows the Chinese BRI business ventures to strategically include their appropriately sourced big data executed quality triggers and then strategically frame their individual business’ qualities outcomes.

GitHub’s OSS Hadoop systems can now capture text, image, audio, video, and even sensory data, and machine data - thereby facilitating a suitable software process to mine big data, to isolate its components, and to form desired information groupings – provided suitable data storage and computing power is



available. Hence, if BRI developments are to be state-of-the-art in their capabilities this Hadoop derived aggregated data now requires the development of subsequent intelligent analytics, and the development of real-time business adjustment systems.

### 13. Conclusion

This paper's objective is to review, describe, and connect usable big data with corporate digital quality leadership potential and consider it against BRI business ventures.

Digital data capture from a business' big data sources remains difficult. Some techniques are now available. This paper reviews, describes, and connects usable big data components as an ingredient towards delivering an understanding the qualities a Chinese business may astutely utilize as they expand along China's BRI pathway. Digital data across the business, its web and its social domains offers text components which can be extracted through Hadoop MapReduce algorithms. These outputs can then be qualities-grouped to deliver insight concerning the qualities a business aims to deliver. Similarly, qualities aspects can also be linked to audio, video and image files - but often for slightly different purposes. Hence, text mining appears the most important big data link to business quality.

To fully digitize a business extensive process and connectivities changes are likely required. This change typically requires a move to an all-encompassing digital leadership approach by the CEO and the business' executive. It requires extensive systems changes and further expenditure on new software. Hence, to introduce digitization across a business, a structured, 4 stage, 5 level qualities-focused approach is suggested.


### References

- Anon. (2017a). Accessed 20<sup>th</sup> Dec 2017 at: <http://www.storm-project.net/>
- Anon. (2017b). Accessed 20<sup>th</sup> Dec 2017 at: ([https://en.wikipedia.org/wiki/Lambda\\_architecture](https://en.wikipedia.org/wiki/Lambda_architecture)).
- Anon. (2018a). Accessed 1<sup>st</sup> Feb, 2018 at: <http://www.internetlivestats.com/twitter-statistics/>.
- Anon. (2018b). Accessed 1<sup>st</sup> Feb 2018 at: <https://zephoria.com/top-15-valuable-facebook-statistics/>.
- Anon. (2018c). Accessed 2<sup>nd</sup> Feb 2018 at: [https://en.wikipedia.org/wiki/Sentiment\\_analysis](https://en.wikipedia.org/wiki/Sentiment_analysis).
- Anon. (2018d). Accessed 2<sup>nd</sup> Feb 2018 at: <https://www.capterra.com/data-extraction-software/>.
- Anon. (2018e). Accessed 3<sup>rd</sup> Feb 2018 at: [https://hadoop.apache.org/docs/r1.2.1/mapred\\_tutorial.html](https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html).
- Bifet, A. (2013). Mining big data in real time. *Informatica*, Vol. 37, No. 1. pp. 15-20
- Chen, H., Chiang, R.H., & Storey, V.C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, pp. 1165-1188.
- Cosentino, V., Izquierdo, J.L.C., & Cabot, J. (2017). A Systematic Mapping Study of Software Development with GitHub. *IEEE Access*, Vol. 5, pp. 7173-7192.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, Vol. 35, No. 2, pp. 137-144.
- Gartner. (2017). Accessed 19<sup>th</sup> Feb 2018 at: <https://www.gartner.com/technology/research/methodologies/hype-cycle.jsp>.
- Hamilton, J.R. (2017). Digital age corporate leader approaches. *21<sup>st</sup> International Conference on ISO 9000 and TQM at BNUZ, Zhuhai, China*, 26-28 Sept, Vol 21, No.4.1, pp. 1-5.
- Haug, A., & Stentoft Arlbjörn, J. (2011). Barriers to master data quality. *Journal of Enterprise Information Management*, Vol. 24, No. 3, pp. 288-303.
- Kahn, B.K., Strong, D.M., & Wang, R.Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, Vol. 45, No. 4, pp. 184-192.
- Liebowitz, J. (2016). *Big Data and Business Analytics*. Auerbach Publications, pp. 1-269.
- Neumeyer, L., Robbins, B., Nair, A., & Kesari, A. (2010). S4: Distributed Stream Computing Platform. *ICDMW '10 Proceedings 2010 IEEE International Conference on Data Mining Workshops*, Washington DC USA, 13 Dec, Vol 10, No.1, pp. 170-177.
- Prince, K. (2017). Industrie 4.0 and Leadership. *17<sup>th</sup> International Conference on Electronic Business*, Dubai, UAE, 4-8 Dec., Vol. 17, No. 1, pp. 132-139.
- Reeves, M., Zeng, M., & Venjara, A. (2015). The self-tuning enterprise. *Harvard Business Review*, Vol. 93, No. 6, pp. 77-83.
- Sugumaran, V., Sangaiah, A.K. & Thangavelu, A. (2017). *Computational Intelligence Applications in Business Intelligence and Big Data Analytics*. Taylor & Francis CRC Press, pp. 1-362

- Utesheva, A., Simpson, J.R., & Cecez-Kecmanovic, D. (2016). Identity metamorphoses in ruption: A relational theory of identity. *European Journal of Information Systems*, Vol. 25, No. 4, pp. 344-363.
- Wang, R.Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, Vol. 41, NO. 2, pp. 58-65.
- Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, Vol. 12, No. 4, pp. 5-33.
- 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.

### Authors' Backgrounds



**Prof. 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.



**Dr. SingWhat Tee** is Senior Lecturer of Information Systems at James Cook University. He researches the impacts of information systems on global citizens and their environments. His research focuses on the criticality of operational data, value analysis, and decision-making information systems. Current research interests include: big data, modelling organizational systems; intelligent data/information modelling, social networks, and the new information technologies systems that influence knowledge transfer and experiential learning.