# THE USE OF BIG DATA IN AI DEVELOPMENT AND APPLICATIONS

Yew Kee Wong

School of Information Engineering, HuangHuai University, Henan, China.

***ABSTRACT***

*In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in volume, velocity, variety and veracity (the four V's of big data), which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyse some of the use of big data for the artificial intelligence development and its applications in various decision making domains.*

***KEYWORDS***

*Big Data, Artificial Intelligence, Data Analytics, Business Intelligence, Decision Making*

## 1. INTRODUCTION

Big data is a new driver of the world economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education [1]. While the data complexities are increasing including data's volume, variety, velocity and veracity (the four V's of big data), the real impact hinges on our ability to uncover the `value' in the data through big data analytics and artificial intelligence (AI) technologies.

Big data analytics and AI poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse, complex, and of a massive scale. Potential breakthroughs include new AI algorithms, methodologies, systems and applications in big data analytics that discover useful and hidden knowledge from the Big Data efficiently and effectively [2]. Big data analytics and AI development must also be team effort cutting across academic institutions, government and society and industry, and by researchers from multiple disciplines including computer science and engineering, health, data science and social and policy areas.

Today, big data has become capital. Think of some of the world's biggest tech companies. A large part of the value they offer comes from their data, which they're constantly analysing to produce more efficiency and develop new products. Recent technological breakthroughs have exponentially reduced the cost of data storage and compute, making it easier and less expensive to store more data than ever before. With an increased volume of big data now cheaper, more accessible, and can make more accurate and precise business decisions [3].

## 2. WHAT IS BIG DATA

The Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer. A buzzword that is used to describe immense volumes of data, unstructured, structured and semi-structured, big data can inundate a business on a day-to-day basis. Big data is used to analyse insights, which can lead to better decisions and strategic business moves [4]. The definition of big data: "Big data is high-volume, and high-velocity or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." The characteristics of Big Data are commonly referred to as the four Vs:

### Volume of Big Data

The volume of data refers to the size of the data sets that need to be analysed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. In other words, this means that the data sets in Big Data are too large to process with a regular laptop or desktop processor. An example of a high-volume data set would be all credit card transactions on a day within Asia.

### Velocity of Big Data

Velocity refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. An example of a data that is generated with high velocity would be Instagram messages or Wechat posts.

### Variety of Big Data

Variety makes Big Data really big. Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. An example of high variety data sets would be the CCTV audio and video files that are generated at various locations in a city.

### Veracity of Big Data

Veracity refers to the quality of the data that is being analysed. High veracity data has many records that are valuable to analyse and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.

Data that is high volume, high velocity and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labelled Big Data [5].

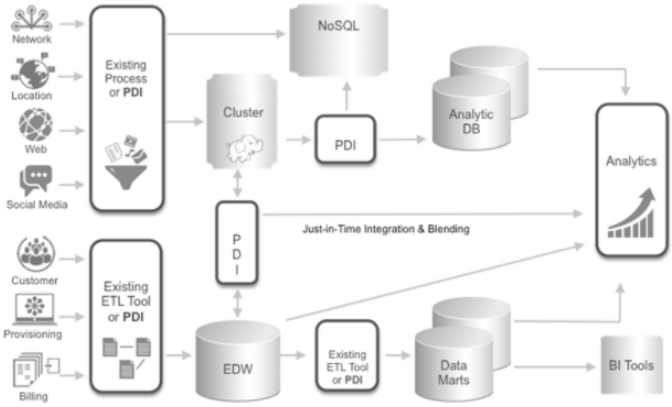


Figure 1. Big Data Architecture. (arccil.com)

## 2.1. Types of Big Data

There are 3 types of big data; unstructured data, structured data and semi-structured data.

**Unstructured data:**
Any data with unknown form or the structure is classified as unstructured data.

**Structured data:**
Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

**Semi-structured data:**
Semi-structured data can contain both the forms of data.

Dealing with unstructured and structured data, data science is a field that comprises everything that is related to data cleansing, preparation, and analysis. Data science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing, and aligning data [6]. This umbrella term includes various techniques that are used when extracting insights and information from data.

**Big data benefits:**

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data, which means a completely different approach to tackling problems.

## 2.2. What is Big Data Analytics

Data analytics involves applying an algorithmic or mechanical process to derive insights and running through several data sets to look for meaningful correlations. It is used in several industries, which enables organizations and data analytics companies to make more informed decisions, as well as verify and disprove existing theories or models [7] [8]. The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

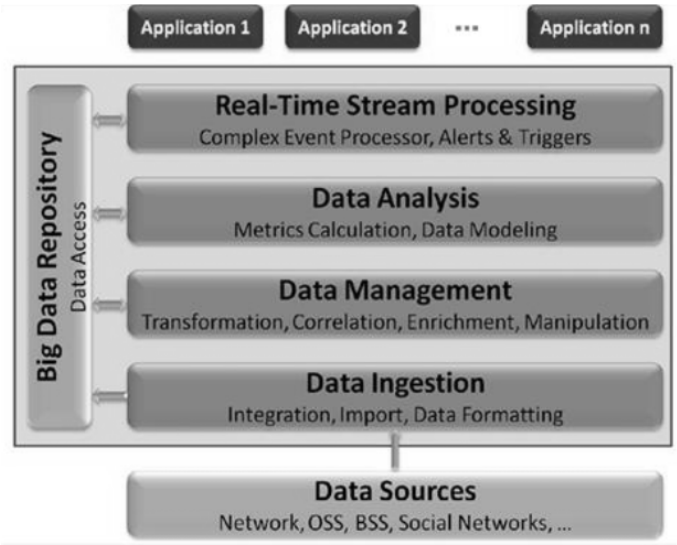


Figure 2. Big Data Analytics Architecture.

## 3. AI DEVELOPMENT USING BIG DATA ANALYTICS

There are many AI researchers from various industry are trying to find solutions in using big data to develop more efficient and effective AI technologies [9]. These are some of the industry which are highly sensitive when applying big data.

**Healthcare industry**
The main challenge for hospitals is to treat as many patients as they efficiently can, while also providing a high. Instrument and machine data are increasingly being used to track and optimize patient flow, treatment, and equipment used in hospitals. It is estimated that there will be a one percent efficiency gain that could yield more than US$63 billion in global healthcare savings by leveraging software from data analytics companies [10].

**Travel industry**
Data analytics can optimize the buying experience through mobile/weblog and social media data analysis. Travel websites can gain insights into the customer's preferences. Products can be upsold by correlating current sales to the subsequent browsing increase in browse-to-buy conversions via customized packages and offers. Data analytics that is based on social media data can also deliver personalized travel recommendations.

**Gaming industry**
Data analytics helps in collecting data to optimize and spend within and across games. Gaming companies are also able to learn more about what their users like and dislike.

**Energy industry**
Most firms are using data analytics for energy management, including smart-grid management, energy optimization, energy distribution, and building automation in utility companies. The application here is centred on the controlling and monitoring of network devices and dispatch crews, as well as managing service outages [11]. Utilities have the ability to integrate millions of data points in the network performance and gives engineers the opportunity to use the analytics to monitor the network.

**Big data best practices**

To help various industries on their big data analysis and AI development, we have put together some key best practices to keep in mind. Here are some guidelines for building a successful big data foundation with the AI technologies integration.

Table 1. Big Data best practices breakdown.

| Development stages | Detail operations and analysis breakdown |
|---|---|
| Align big data with specific business goals | More extensive data sets enable you to make new discoveries. To that end, it is important to base new investments in skills, organization, or infrastructure with a strong business-driven context to guarantee ongoing project investments and funding. To determine if you are on the right track, ask how big data supports and enables your top business and IT priorities [12]. Examples include understanding how to filter web logs to understand e-commerce behaviour, deriving sentiment from social media and customer support interactions, and understanding statistical correlation methods and their relevance for customer, product, manufacturing, and engineering data. |
| Ease skills shortage with standards and governance | One of the biggest obstacles to benefiting from your investment in big data is a skills shortage. You can mitigate this risk by ensuring that big data technologies, considerations, and decisions are added to your IT governance program. Standardizing your approach will allow you to manage costs and leverage resources. Organizations implementing big data solutions and strategies should assess their skill requirements early and often and should proactively identify any potential skill gaps. These can be addressed by training/cross-training existing resources, hiring new resources, and leveraging consulting firms. |
| Optimize knowledge transfer with a centre of excellence | Use a centre of excellence approach to share knowledge, control oversight, and manage project communications. Whether big data is a new or expanding investment, the soft and hard costs can be shared across the enterprise. Leveraging this approach can help increase big data capabilities and overall information architecture maturity in a more structured and systematic way. |
| Top payoff is aligning unstructured with structured data | It is certainly valuable to analyse big data on its own. But you can bring even greater business insights by connecting and integrating low density big data with the structured data you are already using today.<br><br>Whether you are capturing customer, product, equipment, or environmental big data, the goal is to add more relevant data points to your core master and analytical summaries, leading to better conclusions. For example, there is a difference in distinguishing all customer sentiment from that of only your best customers. Which is why many see big data as an integral extension of their existing business intelligence capabilities, data warehousing platform, and information architecture [13].<br><br>Keep in mind that the big data analytical processes and models can be both human- and machine-based. Big data analytical capabilities include statistics, spatial analysis, semantics, interactive discovery, and visualization. Using analytical models, you can correlate different types and sources of data to make associations and meaningful discoveries. |
| Plan your discovery lab for performance | Discovering meaning in your data is not always straightforward. Sometimes we don't even know what we're looking for. That's expected. Management and IT needs to support this "lack of direction" or "lack of clear requirement."<br><br>At the same time, it's important for analysts and data scientists to work closely with the business to understand key business knowledge gaps and requirements. To accommodate the interactive exploration of data and the experimentation of statistical algorithms, you need high-performance work areas. Be sure that sandbox environments have the support they need—and are properly governed [14]. |
| Align with the cloud operating model | Big data processes and users require access to a broad array of resources for both iterative experimentation and running production jobs. A big data solution includes all data realms including transactions, master data, reference data, and summarized data. Analytical sandboxes should be created on demand. Resource management is critical to ensure control of the entire data flow including pre- and post-processing, integration, in-database summarization, and analytical modelling. A well-planned private and public cloud provisioning and security strategy plays an integral role in supporting these changing requirements. |

# 4. INTEGRATING BIG DATA WITH AI APPLICATIONS

In today's world, we are seeing many AI applications integrating with big data analytics in various industries. Here are some of the critical real-time industry which are heavily depend on advanced AI processing technologies and big data analytics operation.

**Big Data for Financial Services**
Credit card companies, retail banks, private wealth management advisories, insurance firms, venture funds, and institutional investment banks all use big data for their financial services. The common problem among them all is the massive amounts of multi-structured data living in multiple disparate systems, which big data can solve [15]. As such, big data is used in several ways, including:

- Customer analytics
- Compliance analytics
- Fraud analytics
- Operational analytics

**Big Data in Communications**
Gaining new subscribers, retaining customers, and expanding within current subscriber bases are top priorities for telecommunication service providers. The solutions to these challenges lie in the ability to combine and analyse the masses of customer-generated data and machine-generated data that is being created every day.

**Big Data for Retail**
Whether it's a brick-and-mortar company an online retailer, the answer to staying in the game and being competitive is understanding the customer better. This requires the ability to analyse all disparate data sources that companies deal with every day, including the weblogs, customer transaction data, social media, store-branded credit card data, and loyalty program data [16].

## 4.1. New Business Models With Big Data Element

Standard big data gives you new insights that open up new opportunities and business models. Getting started involves three key actions:

**Integration**
Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as Extract, Transform and Load (ETL) generally aren't up to the task. It requires new strategies and technologies to analyse big data sets at terabyte, or even petabyte, scale. During integration, the developer needs to bring in the data, process it, and make sure it's formatted and available in a form that your business analysts can get started with.

**Management**
Big data requires storage. The storage solution can be in the cloud, on premises, or both. The user can store your data in any form they want and bring their desired processing requirements and necessary process engines to those data sets on an on-demand basis. Many people choose their storage solution according to where their data is currently residing. The cloud is gradually gaining popularity because it supports your current compute requirements and enables you to spin up resources as needed [17].

**Analysis**
Creative investment in big data pays off when the analysis and act on the data. Get new clarity with a visual analysis of the varied data sets. Explore the data further to make new discoveries. Share the findings with others. Build data models with machine learning and artificial intelligence. Put the big data to work.

## 5. CONCLUSIONS

Big data brings big benefits, but it also brings big challenges such new privacy and security concerns, accessibility for business users, and choosing the right solutions for your business needs. The advanced big data analytics and algorithms with various applications show promising results in artificial intelligence development and further evaluation and research is in progress.

## REFERENCES


[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, (2015). Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8), pp.228-232.

[2] A. Gandomi and M. Haider, (2015). Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2), pp.137-144.

[3] C. Lynch, (2008). Big data: How do your data grow?, Nature, 455, pp.28-29.

[4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, (2015). Significance and challenges of big data research, Big Data Research, 2(2), pp.59-64.

[5] R. Kitchin, (2014). Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1), pp.1-12.

[6] C. L. Philip, Q. Chen and C. Y. Zhang, (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275, pp.314-347.

[7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, (2014). Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7), pp.2561-2573.

[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, (2014). On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285, pp.112-137.

[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, (2014). Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1, pp.114-126.

[10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, (2013). A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, pp.17-22.

[11] Z. Huang, (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.

[12] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, (2015). Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79, pp.3-15.

[13] I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, (2014). The rise of big data on cloud computing: Review and open research issues, Information Systems, 47, pp. 98-115.

[14] L. Wang and J. Shen, (2013). Bioinspired cost-effective access to big data, International Symposium for Next Generation Infrastructure, pp.1-7.

[15] C. Shi, Y. Shi, Q. Qin and R. Bai, (2013). Swarm intelligence in big data analytics, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li and X. Yao (eds.), Intelligent Data Engineering and Automated Learning, pp.417-426.

[16] M. A. Nielsen and I. L.Chuang, (2000). Quantum Computation and Quantum Information, Cambridge University Press, New York, USA.

[17] M. Herland, T. M. Khoshgoftaar and R. Wald, (2014). A review of data mining using big data in health informatics, Journal of Big Data, 1(2), pp. 1-35.

**Author**

**Prof. Yew Kee Wong (Eric)** is a Professor of Artificial Intelligence (AI) & Advanced Learning Technology at the HuangHuai University in Henan, China. He obtained his BSc (Hons) undergraduate degree in Computing Systems and a Ph.D. in AI from The Nottingham Trent University in Nottingham, U.K. He was the Senior Programme Director at The University of Hong Kong (HKU) from 2001 to 2016. Prior to joining the education sector, he has worked in international technology companies, Hewlett-Packard (HP) and Unisys as an AI consultant. His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.