2

Data Protection

When Legal Meets Data Analytics

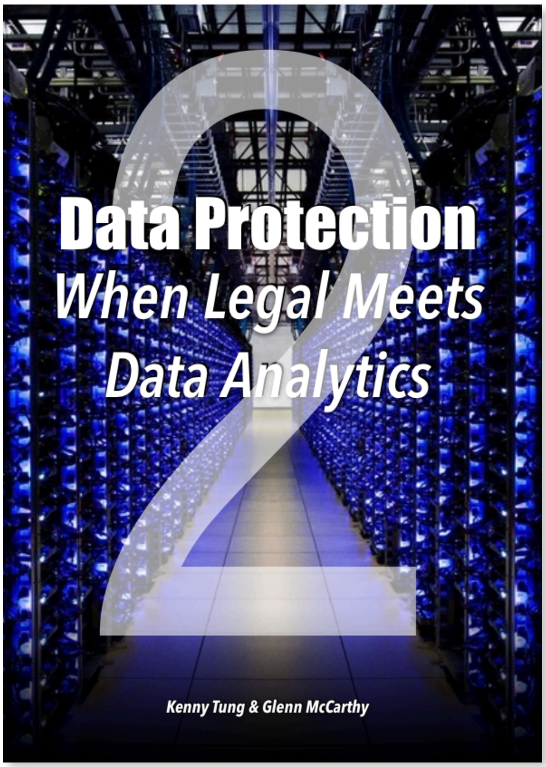
Kenny Tung & Glenn McCarthy

in-Gear LEGALYTICS

Cover photo: Google Data Center
Published by LegalBusinessWorld

Complete Series

Read the complete series Data Protection - When Legal Meets Data Analytics by Kenny Tung & Glenn McCarthy. Go to the Legal Business World [eBook library]() and read the book online or download your copy. *(As of March 2022 available: Part 1 & Part 2).*



Summary

In Part 1, this (e)book investigates the data protection challenge from the perspectives of Legal and business strategy. Part 2 below starts with a survey of the scale that data has grown into and the use cases for data analytics, followed by a deeper look into programmatic advertising and the representative data issues. After enumerating a variety of new technology coming into the data and analytics business, the piece explores the trade-offs for businesses in approaching data and analytics, including drivers such as business models and data industry dynamics, laying the groundwork for the recommendations in Part 3. Part 2 also includes a insert on data taxonomy (page 17-21).

This (e)book series is based on a workshop taking an interdisciplinary perspectives of data protection delivered at a Legal Function Transformation Round Table subgroup on Oct.22, 2021.

******

III. Technology Perspective

Data Proliferation - Quantity Has a Quality of Its Own [1]

At the turn of the millennium, an individual or team of human analysts using commonly available tools could comprehend and manage the realm of existing data to solve most business problems and make important decisions. Around that time, the total data produced annually in the world amounted a few exabytes (10^{18} bytes) [2] and while the growth of internet data and computers was signaling the future of Big Data, the vast majority of corporations did not work with "big" data. After all it was not much more than 25 years ago when the fax machine overtook snail mail. In the 1990s, lawyers would still reference thick volumes on massive book cases; finance people typed data from paper into the revolutionary Lotus 1-2-3 spreadsheet; engineers worked with slide rulers and scientific calculators; and the general public would look up the location of a book in the library card catalog and then search the book's index in hope of finding the information they seek.

One aspect of this quaint history was that it was possible for humans to search the universe of available data, analyze the information obtained, and reach a conclusion. Businesses drew conclusions from what is today a small sample of customer surveys or the hand-written observation of a mechanic, and companies deployed human beings as the main resource, plus some tools, to handle the volume of information. Before businesses sport data warehouses and even data lakes, most activities were managed through spreadsheets and attachments between computers, regularly printed out for hardcopy filing.

Today this is no longer the case. As recently as the 2015, a single commercial aircraft collects data on more than 300,000 parameters, and a Boeing 737 flying from New York to L.A. will create 240 terabytes ($240x10^{12}$, think 240,000 YouTube resolution movies) of data for every engine hour.[3] "Autonomous vehicles will generate as much as 40 terabytes of data an hour from cameras, radar, and other sensors—equivalent to an iPhone's use over 3,000 years—and suck in massive amounts more to navigate roads, according to Morgan Stanley."[4] At the tip of the spear, Tesla announced on AI Day 2021 that it was rolling out its own designed chip that features 362 teraFLOPS of processing power, or a capacity to perform 362 trillion floating-point operations per second (FLOPS), a measure of computer performance. Multiplying 25 chips per training tile and linking 120 tiles through multiple servers will arrive at the astronomical computing power of 1.08 exaFLOPS or 10^{18} FLOPS.[5] To provide some context, albeit between apples and oranges, the estimated range of firing potential per second in a human brain at the neuron level is from 86 billion to 17.2 trillion ($86x10^{9}$-$17x10^{12}$), and at the synaptic level from 100 trillion to 20 quadrillion (10^{14}-$2x10^{16}$).[6] This provides a measure of the scale of the data flow that full self-driving needs to crack the last mile and the long tail of scenarios in driving reality.

As of December 2021, there were nearly 3 million product listing for cosmetics items for sale on the two largest e-commerce marketplaces in China, and Amazon USA listed over 75 million total consumer items for sale. WeChat Pay claims to serve 72 million merchants in China and processes over 1 billion transactions every day.[7] In 2021, global spend on programmatical advertising was estimated to have been US$155 billion dollars,[8] and the average person encountering between 6,000 to 10,000 ads every single day.[9]

Tiktok has over a billion users and Meta (formerly Facebook) over 1.7 billion. It is estimated that globally two trillion searches each year (or 63,000 searches per second) take place on Google.[10] In the year 2021 alone, the world generated an estimated 80 zettabytes ($80x10^{21}$) of data.[11] When compared to the 0.005 zettabytes ($5x10^{18}$) [12] of information created between the dawn of civilization through 2003, it becomes understandable that humans with traditional tools and regular expressions can no longer keep up with the data sphere.

This data proliferation is the outcome of not only increasing activities in the world being captured, measured, digitized, organized and archived, but also the rising commercial complexity in turn driving demand for more diverse data, variables, metadata, types of analytics.[13] As the demand for data and insight outstrips the supply of data scien-

tists, analysts, software engineers, system architects and other IT professionals, the ecosystem is pivoting toward developing products and platforms to serve consumers of data in more self-help modes.
The growth of data fuels analytics, predictive tasks and augments decisions, representing a massive opportunity. However, data proliferation, especially before reaching an ecosystem equilibrium, also presents risks and challenges in terms of governance and compliance, a problem that can balloon very quickly. While some regulatory vectors may point toward individual privacy or competition concerns, at a macro level, regimes across the spectrum have woken up to the threats of the unbridled network effect of information platforms. For enterprises, having access to the sea does not default on attempting to boil the ocean; on the other hand, the power to observe the ecosystem and its elements with more granularity should not be wasted, provided that one takes up the responsibility commensurate with such power.

Along the same line, machine learning (ML) and other algorithmic engines in this data gold rush will increasingly deliver powerful correlations but not necessarily causality. Analysts must still be mindful of the warning, “lies, damned lies, and statistics,” preserve a sense to avoid the folly of garbage-in-garbage-out exercises as well as stay vigilant of human biases. Big Data requires not only statisticians as part of the data-to-decision team, but also other T-shaped professionals relevant to the domain to maintain quality, sanity and make data analytics useful to users and others whom it will touch.

Instead of coding only if-then-else regular expressions to generate outputs from inputs, data analytics also deploys ML to process voluminous inputs and outputs through neural networks to derive algorithms for applications. While ML delivers options and alternatives to solve problems and augment decisions, this approach has yet to address the demands for explainability in many fields. Therefore, tools like the job to be done (JTBD) discussed in Section II. In Part 1 of this (e)book are complementary to the insight journey and safeguard against conflating causality with correlation.

Data Value Propositions – Uses for the Force

The universe of enterprise data value propositions continues to expand but is fairly well-known. Across the spectrum of large corporate business, a number of Big Data enterprise application areas has emerged. Some top-of-mind use cases include:

Intelligent Manufacturing – To optimize further the production process to drive quality and reduce labor and other costs with measures such as predictive maintenance.
Supply Chain Optimization – To boost cash flow and increase profits by reducing inventory and ensuring products are in stock in the right quantity, at the right place and right time, e.g., data analytics optimizing products and components logistics schedules through orders prediction.

Transportation and fleet management – To use telematics data and routing software to optimize vehicle speed to save gasoline, improve safety (e.g., favoring routes that minimize turns across oncoming traffic), track and augment driver behavior in connection with auto insurance rates setting, and optimize valuable airport equipment scheduling to reduce capital investments.

Financial performance applications – To automate budgeting and planning, streamline reporting, and provide critical business information such as profitability at product and customer level.

Financial crime detection – To require anti-money laundering algorithms as a mandated piece of financial service infrastructure.

Sales Force Effectiveness – To highlight customer opportunities and measure sales by agency, geography, product, and numerous other dimensions.

Human resource candidate screening – To automate front-end candidate selection process with ML, saving labor costs while enabling proper analysis of millions of resumes to streamline identification of suitable candidates.

The health and life sciences industry – To use Big Data analytics for everything from diagnosing and predicting diseases to remotely monitoring patients to discovering new drugs.

Customer service chatbots – To save costs by reducing human interactions while creating more predictable communication scenarios, all the while tracking a digitized trail of useful and valuable data.

Customer retention – To deploy churn models to help firms such as those in telecom and content subscriptions industry to predict when

customers may consider leaving and to trigger a promotion or other retention measures.
Smart city – To optimize truck routes with smart garbage cans, reduce electricity costs through smart street lights, improve traffic flows through smart traffic lights, and even elevators that dynamically group people through facial recognition and direct them to specific elevators going to adjacent floors.

But perhaps the most exciting (and concerning) of all the enterprise data use cases is digital marketing. Customer profiling combined with programmatic advertising and the tracking of the customer journey or path to purchase and other activities goes a long way to solving John Wanamaker's quandary mentioned in Part 1 and reducing waste in marketing spend.

Customer Data Platforms and Programmatic Advertising

In top-of-mind surveys, data protection is associated with personally identifiable information (PII).   For consumer-facing businesses, PII presents both the greatest business opportunity and the greatest compliance risk. Many newer categories of data are valuable as descriptions surrounding human behavior and raw materials for analytics engines to generate predictions and businesses to apply prescriptions. On the opposite side of the ledger like the U.S. FCPA top ten chart, there is for the E.U. General Data Protection Regulation (GDPR) since 2018 a top fines chart with the latter record standing at €746 million against Amazon in 2021.[14] Across jurisdictions, examples of causes of sanction include cookie consent, inability to guarantee transparency and fairness in automated decisions based on PII, and even failure to explain properly data processing practices in privacy notices.

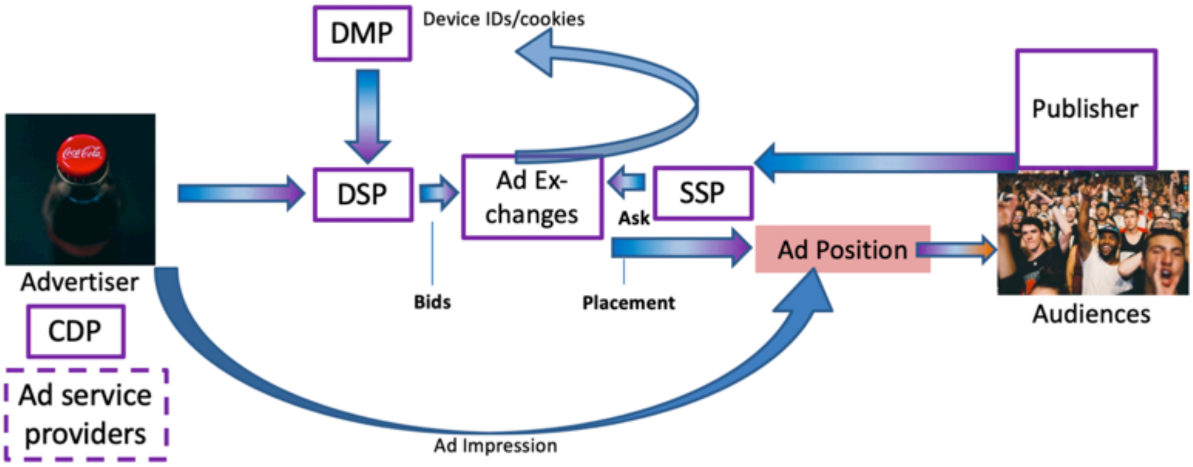
Today many brands have built, are building or planning to build customer data platforms (CDPs). They are undertaking the costs, and increasingly risks, of doing so because they want to know who their customers are and to maintain direct connection to and contact with them. Brands are growing increasingly uncomfortable with third parties, platforms like Amazon, Alibaba, or Facebook, knowing who their customers are and possessing more robust customers demographic and shopping history data than they do. Brands and retailers have come to understand the plight of disintermediation by platforms, e.g., bookstores and other retailers by Amazon, traditional media by Google as well as new and disruptive media. The resulting Hobson's choice looks something like paying exorbitant fees for traffic from the platforms or living with the platform pointing the consumers eyeballs at competitors' or private label offerings. Platforms do not appear to offer exclusive access to the data of consumers. Access to customers is rented. Keywords are auctioned. There are the few exceptions like Apple which has been able to negotiate with Amazon, but these are very few and far between.

As brands open up their own direct-to-consumer (DTC) e-commerce sites, they have the ability to both capture end user data and to earn higher margins by selling directly to end customers. Brands and retailers want to have this first party connection to customers in order to learn as much as possible about their habits, traits, preferences, and behaviors. Through direct communication with consumers they can target promotions, share relevant product information, and build community and loyalty. This channel can also feed into the exploration of a customer's JTBD. Brand owned CDPs are central to this operation.

CDPs receive data from multiple sources and program to resolve this data to a single identity. The function is to form a comprehensive profile of that individual. Information that companies commonly collect or centralize in a CDP includes name, email addresses, phone numbers, age, gender, device type and ID, apps used, geolocation, browser type and potentially even browsing history, as well as affinity to specific sports, movies, and food, etc. Over time a brand may collect thousands of pieces of data about an individual, including data purchased from third parties such as home address, automobile ownership type, and employment income. With increasing varieties of information such as geolocation data from pervasive mobility and the IoT, CDPs can yield a 360-degree view of the end consumer.

The great shift toward CDPs is driven by AdTech and the emergence of digital and programmatic advertising. Customers profile in a CDP enables a brand to market and sell directly to them, and also use this technology to create and send anonymized versions of consumer profiles to a data management platform (DMP). DMPs are platforms that can run software to collect and manage data that enable businesses to identify target segments, and target specific users and contexts in online advertising campaigns. Facebook, Amazon, Alibaba, and JD.com are best known examples of DMPs, but there are many others. Generally, these DMPs then interface with supply side platforms (SSPs) to purchase targeted or micro targeted ad impressions. An SSP is software used by digital publishers to sell ad inventory programmatically to advertisers without human intervention. Yet these DMPs are no ordinary cogs in the wheels of target advertising. Meta (formerly Facebook), Amazon, Apple, and Alphabet (Google), and in China the equivalent monopolies of Alibaba and Tencent comprise trillions of dollars of market capitalization, with a big part of that value stemming from their ownership and control of consumers data and detailed demographics profiles.

This trend creates two main areas of concerns, especially for FMCG businesses. The first presents a form of disintermediation from the pool of potential customers, resulting in excessive advertising costs such as tolls on these platforms and marketplaces, only to reach parity with competitors which also pay up for similar data access from the platforms. "In 2021, Google, Facebook, and Amazon account for 64 percent of U.S. digital ad spending.[15] The second is disruption, where the platforms and marketplaces themselves become direct competitors. Amazon now has nearly 100 private label brands like Amazon Basics and is a video content producer and distributor.



CDP has PII, and DMP generally does not, as the latter masks the PII. When brands use DMPs, they are renting access to end customer anonymized profiles - lower risk in terms of data compliance, but also knowing less about the actual customer.
In the past few years brands have been moving to own directly their end customers data in order to not rely on 3rd party platforms. However, under China's new Personal Information Protection Law (PIPL), they will face higher technology standards and stricter regulation. This creates a **strategic business decision:** being disintermediated from the end customer vs. investing to comply with regulations that apply to consumer internet companies.

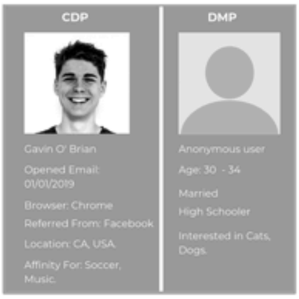


Image Source: https://headerbidding.co/customer-data-platform/

Apple is a distributor of music, games and apps and a payments company.

Doordash is opening kitchens, professional food preparation and cooking facilities for delivery-only meals, not exactly to serve restaurants listed on its platform.

With globalization and trade driving the means of production through many commoditized and modularized supply chains, ownership of customers data and their hearts and mind is the new king. Data analytics not only composes a picture of the world in richer pixels (e.g., by labeling variables for ML), but also facilitates emerging developments (what do people want next week) and predictions in optimal time frame to support product development. Even an erstwhile Marxist regime turned factory of the world is giving this credence. Thus, while the platforms offer a cost-effective, quick way for go-to-market with built-in traffic and services, they are also powerful and sometimes monopoly rent seekers, and in some cases intend to partake of much of someone's lunch.

To Be or Not to Be

With the background music of disruption and innovator's dilemma playing louder than ever, the approach to data represents perhaps the starkest opportunities and threats to businesses, and finding and keeping the optimal balance in generating and managing their own data versus leveraging third party platform data can be existential. This goes beyond the routine financial analysis of owning versus leasing. Getting this right can make the difference in innovating new products ahead of competition, mastering the customers journey from exploration to brand identity, achieving efficiencies that create competitive advantage and highlighting insights that raise barriers to entry. Getting it wrong can lead to diluting customer satisfaction and defection, tired, also-ran brands, reduced profitability and in some cases, the dreaded loss of market position to a disrupter.

As businesses digitize operational aspects to keep up with ecosystems and develop digital business models, software and systems are now ubiquitous across the enterprise and growing IT budgets seem a given in the commercial arms race. However, an under-appreciated aspect of digitization is the growth of the cost of maintaining and analyzing data. While storage of data is now considered cheap, refining the "new oil" (data) into fuel (actionable insight) and ensuring proper data governance, residency, and privacy protection can be costly.

Inherent in this challenge is the need to integrate the costs and risks of owning, managing, and maintaining data and data channels into the enterprise overall strategy. This includes compliance with the laws of today and anticipating the regulatory equilibrium tomorrow. All this is taking place in an uncertain ecosystem where regulations rapidly shift as business models are being created. Governments make rules and view the behaviors of yesterday by the standards of tomorrow, and some countries are now aggressively handing out penalties and often compounded in reputational damage.

A Data Taxonomy

A short twenty years ago, corporations generated internally most business data in structured databases like tables. In the ensuing years, the volume and types of data have greatly expanded the universe of business data. Alongside Internet of Customer or Content (IOC) - data from marketing, social media, and e-commerce - the Internet of Things (IoT) pipes in data from sensors and machines. The proliferation in latter data sphere accelerates as innovative ways to capture senses like smell and motions and convert the signals into data stream. Increasingly, companies capture data generated externally, process and store them inside the firm. The explosion in the volume of external data has created multiple opportunities to gain insights into consumer behavior, trends, and preferences as well as to track competitors, supply chains, regulatory drivers, and to monitor streams of data from weather to market news around the world.

The spectacular growth of external data has been more so in unstructured data than in structured data. The structure through which businesses traditionally handle data typically adheres to a predefined format or data model for ease of analysis and storage, such as in the rows and columns of a relational database or in a traditional data warehouse.

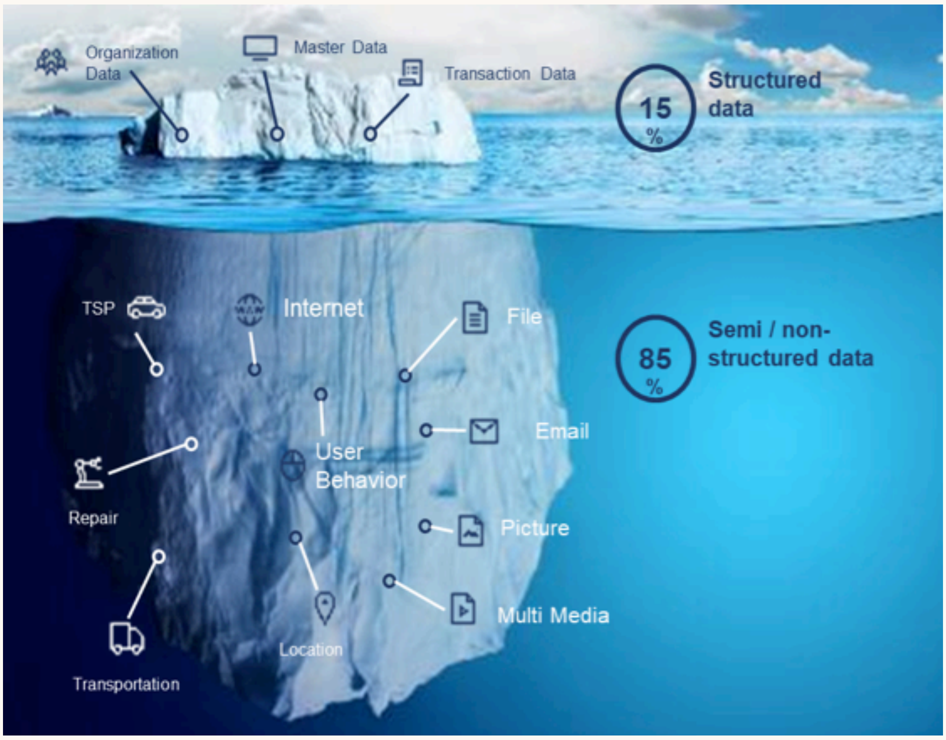
Typical fields of structured data include names, dates, addresses, credit card numbers, inventory information, price, and geolocation. But today, technology can deliver insights from more varied data formats to circumvent inherent constraint in online transactional processing (OLTP), such as ERP, with star schema type databases that enable queries across multiple dimensions (a kind of metadata) in online analytic processing (OLAP). OLAP creates its databases to serve analytics by extracting, transforming, loading (ETL) data from OLTP databases.

Unstructured data, on the other hand, does not conform to a pre-defined data model or is not organized in a pre-defined norm or common understanding in the data structure industry. It is often text heavy, qualitative in nature and stored in data lakes or non-relational databases. Examples include e-commerce scraped data such as comments about a product's functions on Amazon or, social media data, and review data such as a product post on an official WeChat account, a restaurant review on Yelp, or a book review on Good Reads, as well rich media video or audio data and surveillance image data. A Zoom video recording is digitized data, but incorporating unstructured fields.

Between structured and unstructured data exists semi-structured data. A kind of structured data that does not follow the relational structure of typical databases, but instead contains tags or other markers that indicate hierarchies of records and fields within the data. Examples of semi-structured data include emails, zipped files, HTML, text messages in smartphones, Whatsapp and Wechat, where the text fields are structured, but content is not. Hence, a common and growing field is human interactive tagging (HIT) which processes unstructured data such as a Tiktok video or a product photo by tagging or flagging specific words images and patterns and enabling ML to mine for insights as part of data analytics.

According to Gartner, much work is taking place to enrich databases with semantics, i.e., the meaning in and relationship among data points. The data management field is developing a data structure layer (analogous to a middleware) to serve more diverse use cases with less or even no intervention by data professionals. This data management layer will append data with metadata the functions of which include the description, organization, integration, sharing, governance and implementation of the data. A resulting data "fabric" can incorporate existing and new databases that better serve data collection and insight connection. Hence, data can be discovered to fit the purpose of the search with more efficacy and efficiency, with little to no data "cleaning." Erstwhile data exhaust, metadata is the "new black" and the further refined products in the new oil's petrol value chain. This development will take place in part in standardizing and systemizing the metadata, especially the data fields that have not yet made their way to common tabular data and relational databases formats.[16]

The image below visualizes that much of the unstructured and external data sits below the surface and out of sight from the traditional view of



data in the organization above the water. Only the tip of the iceberg, structured data, can serve as inputs to analytic engines. What lies below the surface presents opportunities and risks but is often neglected. An enterprising business leader, supported by data analytics, can navigate the ice underneath with tools and intelligence, break it down and monetize the cubes.

(Image of Iceberg on page 20: Advanced Analytic Services)

Technology to Sort Out the Data Conundrum

The good news is that the fourth industrial revolution has already started interaction with massive amounts of data and sparked another enterprise technology innovation. The business software industry is hard at work creating the proverbial picks and axes for the data gold rush, and Silicon Valley's VCs are allocating record amounts of capital to early stage companies developing breakthrough data technology. In data infrastructure, the journey has progressed from corporate computers and servers to data marts, data warehouses, enterprise data warehouses, Hadoop clusters, data lakes and cloud based data warehouses and lakes to meet the challenges and costs of analyzing data and the growing demand for real time data streaming. The past few years have witnessed the advent and rapidly growing adoption of cloud databases and cloud data models. With analytics closely following data to the clouds, cloud data services can provide instantly available and scalable infrastructure for handling data and which can be purchased on an a la carte basis.

The database evolution reflects the overall expansion of data in volumes and complexity. Much is simply about speed and processing power. Today a business expects a query to return an answer in seconds. That would be after searching a billion records, executing complex calculations, formatting the output and returning that in three seconds. And even that is not enough. The data needs to "stream" real time into the database and real time to update reporting processes, something once used only at the stock markets and banks. Now consumer goods companies are deploying this real time, benefiting from the precipitous drop in cost of the technology.

Smart databases, a combination of database and data processing engines, can recognize different types of data, even certain unstructured data and automatically classify, tag and relate data, thus eliminating the need for classic database design. These databases are now generally distributed in tens or hundreds of thousands of servers, on the cloud, or a combination, enabled by frameworks like Hadoop and MapReduce acting like they are on one single server. This increasingly enables the system to stream data in or near real time as the underlying analytics engines advance and become ever more powerful.

Automatic pattern and anomaly detection of data sets can now automatically point data scientists to areas of interest, opportunity or concern without humans needing to build detection models. Today, data science tools run data sets through dozens or hundreds of models and suggest the best options to the growing population of citizen data scientists, anyone in a business, often analysts, with some coding skills to query data.

ML has been formulating algorithms to cut through massive volume of repetitive and focused tasks and practically automate much of the mid-stream data processing like data mapping and cleaning. ML models and data sets to validate simulation or predictive models are rapidly replacing tedious to write and difficult to maintain data mapping and regular expressions such as completing missing data fields with the average of the available data in relevant fields and classification and categorization of data. In recent years the rollout of ML enabled robotic process automation (RPA) software triggered the automation of multi-variate routines and nonlinear tasks that were formerly performed by white collar labor in finance, legal, and operations. Some may recall gratefully the application of finance RPA to reconcile account balances between systems, a critical but often tedious function. This advancement has both freed humans from executing these tasks and created new sources of data learning for the enterprise.

Data visualization tools now suggest the best way to present data and recommend page and chart layouts. Interactive graphs and charts are produced simply as drag-and-drop and can automatically filter and update data for users. Graphics represents a major advance as capturing relationship among data points, enhances answers to queries and in turn accelerates ML's contribution to data analytics. The recent trend towards augmented analytics techniques is showing how technology is able to automatically generate stories, which are starting to replace the dashboards. For example, market research firms or reporters can now use data storytelling software to produce reports and actionable insights including sharing, annotation, and drill-in story functions using a no-code framework.

On top of all of this technology sit business intelligence (BI) tools that have evolved from benchmarking reports, to drill-down and click-through, combing data from data warehouses and data marts through multi-dimensional filters. Traditionally BI technology enabled an analyst to use query technology and follow a data exploration process to analyze and discover business insights. Meanwhile, data scientists would focus on building model and developing custom algorithms for specific applications that require tailored data configurations to produce predictive models. Today we are seeing the convergence of BI and data science platforms through ML and natural language processing (NLP) resulting in Augmented Analytics.

However, the dependence of effective data analytics on complex mapping integration and customized query, data prep and reporting can be tedious and expensive. So, the BI software market is reacting by leveraging ML technologies in order to increase efficiencies in the data infrastructure and analytics process. The wrangling of data and report building that is currently done by developers is being automated and performed by the software itself. Enterprise analytics providers using ML and other artificial intelligence (AI) tools are building more elasticity in the data pipeline and infrastructure to enhance efficiency in the data-to-insight process.

With hope, evolving data formats will standardize more metadata to bridge the gap from unstructured to structured data or largely circumvent this bottleneck. One structural approach is to de-couple data from specific- and narrow- purpose applications so that data can be deployed and reused across multiple applications and processes for use cases other than at original collection. Along the way, it will reduce a significant part of resources today that has been pouring into cleaning or applying “feature engineering” to format data to enable ML. The resulting Common Data Models and Knowledge Graphs will enable citizen data scientists to produce output that was once the domain of experienced data scientists, eliminating a major bottleneck while also further reducing time and cost to insights.

Illustrating this trend, a new technology or company seems to surface in this field every month. Some names are even well known such as Amazon, Microsoft Azure, Oracle, Google, Alibaba, and Tencent, not surprisingly some of the top providers of cloud service. A massive and growing amount of activities in the cloud is driving innovation in this part of the data business. A snapshot of the internet bandwidth of data traffic across the two coasts in the U.S. now shows less traffic than that coursing through a couple hundred of servers inside a single data center.[17] To serve this demand going vertical, and leveraging the emergence of programmable chips, these cloud-as-a-service (CaaS) providers not only increase performance, but innovate throughout the cloud pipeline, from operating systems to networking interface to user applications.

Other examples include Databricks or Snowflake, cloud based data lakes and warehouses; Splunk, software for searching, monitoring, and analyzing machine-generated data; Alteryx, a software that makes advanced analytics accessible to any data worker; Spark, an open-source unified analytics engine for large-scale data processing; and Kafka, a framework implementation of a software bus (software architecture model with a shared communication channel) using stream-processing. Better known may be Tableau or Power BI, popular BI tools from Salesforce.com and Microsoft, and a lesser known startup called ThoughtSpot, bringing AI to BI that has raised US$664 million dollars with a valuation of US$4.2 billion.

Enabling and integrating these applications, tools and data are a large industry of service providers exemplified by Deloitte, Cap Gemini, Accenture, Cognizant, Wipro, Infosys, and also specialty analytics service companies such as Fractal, Mu Sigma plus thousands of specialized local service firms.

"Software Is Eating the World."[18] The data sphere is no exception to this adage, especially with ML starting to drive algorithms to perform much of the data manipulation, analysis and insights process. Big Data analytics generally requires large amount of fresh data to model, validate and test a thesis for further application. Failing validation often means repeating the process but with new data. After a successful launch of a model, data stream continues to feed the need to maintain the algorithm's efficacy. And alongside from project to maintenance is all the data infrastructure from capture to processing, storage, governance, and security.

These tools and technologies increase the feasibility of managing a corporation's vast, growing trove of data and access to third party data while maintaining a view to a path for complying with expanding and increasingly local data regulations. They enable data management by providing the core capabilities necessary to store, process, query and analyze data, enabling better decisions and harnessing business insights. Also, they support compliance by organizing data, providing access and authorization functionality, masking capabilities, and cataloging and creating traceability of data.

There Are No Solutions. There Are Only Trade-offs. [19]

As in the case in countless spheres, technology may usher in disruptions, but often cannot by itself solve for choices.

Companies are realizing the strategic importance of data and desire to enhance the level of control of critical data. However, they are finding that the majority of the technical assets able to capture this data are located outside the corporate boundaries. While this may relieve them of the risk of swimming in massive flows of PII, it can also result in a weakened position across product development, go-to-market and other key business processes. The seemingly unstoppable shift in data structure and services to the cloud and everything-as-a-service (EaaS) may lower the upfront costs of big data analytics capability, but at the same time it is creating an often under-appreciated dependence on emerging players with the potential to extract economic rent in the data-to-insight supply chain.

Given the reality of the consumer internet marketplace today, where the most important and valuable data properties, including Facebook and Google, are well beyond the reach of all brands and product firms, what are firms to do?

One emerging trend, especially in the West, is that brands and retailers setting up their own e-commerce sites and focusing on selling online direct to consumer (DTC) as a complement to their presence in the major e-commerce and social media platforms. This provides a direct contact point with consumers, and in turn enables direct data collection and an interactive channel.

Brands, product companies and retailers are also increasingly leveraging their traditional assets for data collection including (x) existing omni-channel assets which often provide a mixed physical and digital connection point, (y) customer service call centers and websites that establish a direct conversation with product owners and users, and (z) the growing product digital activation or registration processes that may provide home and IP address and other valuable data. These and other channels are increasingly being digitized and optimized for data collection and triangulation with externally purchased demographic data to weave a widening net of 360-degree end customer profiles.

The benefits of directly accessing and owning this consumer data run across the economic, operational and strategic (e.g., appreciating the customers' JTBD). They include (a) minimizing payments to the platforms and market places for demographic information, (b) enabling marketing to existing customers without paying third party advertising or traffic fees, (c) the ability to run A/B tests (for product development, website engagement, click through maximization, etc.) directly with customers, and (d) direct feedback on product features and JTBD, quality, customer service, and the ability to survey product users and owners.

Still, companies need to invest in the data and analytics infrastructure, databases, applications, and tools to enable the effective use of DTC data flows. This is in addition to the significant costs of buying or building and running the business's own e-commerce shopping website and the multitude of payments and logistics integrations that are required to process orders. Not to be neglected are the challenge and cost of attracting and retaining a highly skilled technical team in one of the tightest segments of the labor market. Last but not least, there is the increasingly responsibility of touching, let alone owning, reams of PII and liability for its management, governance, and protection.

Indeed, building systems and solutions for digital interactions and data collection comes at a cost, and given the current rapid evolution of technology, the risk of an even shorter shelf life.   The time gap between standing up a data solution and the business realizing a return on this investment calls for a thorough understanding of short-term value propositions and foresight to recognize the future value of certain data supply chains and the possible strategic role that they can play in an industry. Put simply, another trade-off. A company must devise and operate a business case to derive value from the data and insight today and position itself for the future, while balancing budgets and meeting demands for current period from stakeholders such as investors. However, the hurdle rate exercise will need to incorporate strategic orientations and an infinite-game mind frame in a competitive world to avoid missing objectives for short-term zero-sum outcomes.

There is a broad array of solution scale, size, scope and corporate standards and polices that determines the cost of delivering a project and then operating a production data pipeline and analytics system. Generally, each project consists of software and service costs. Albeit being highly variable depending on the specific project and technology, they tend to be in a 1:2 ratio - that is two dollars of initial implementation service for every dollar of software. Most corporate data analytics projects are comprised of (a) a database layer for distributed computing and storage, (b) a software layer for data extraction, transformation and loading (ETL), and (c) a data analysis and consumption layer, often consisting of one or more BI tools.

For a large corporation deploying a solution in a large country such as China or India for a function with 75 users, the reporting software, e.g., Tableau, a full-service BI worldwide license described in (c) above can cost anywhere from US$500-$800/year/name user or about US$50,000/year. Adding an ETL software license described in (b) for another US$100,000/year, plus a database or cloud data warehouse license described in (a) for a similar cost level (US$100,000) will bring a total software cost to US$250,000 for a mid-size functional solution. Assuming the software-to-service ratio of 1:2, the entire project will cost US$750,000 in the first year. In a subsequent year, annual operating costs will be around 25 percent of the implementation costs in support and maintenance, therefore, in the range of US$375,000 (250k+125k). Note that software and services are generally 5 times more expensive in North America and Europe, and perpetual software would cost a great deal more up front.

Looking ahead, the biggest shift is from ETL to ELT. Until very recently data engineers would extract the data, transform it (cleaning and mapping), then load it into the analytics database. A cutting edge process today is to extract the data, load into a smart cloud data warehouse like Databricks or Snowflake, and the database will transform the data or that transformation takes place inside the cloud database. The value just moves to a different software vendor, with enhanced productivity for the customer.

Juxtapose this with how much easier and cheaper it is simply to rely on external data and pay for access to traffic and insights from third parties. This alternative presents an asset light approach for data and technology and eliminates much of the cost of building and running and the risk of owning data systems with PII. For larger firms this is a business and capital allocation decision, integral to the corporate competitive strategy process and dependent on the competitive, industrial and broader ecosystem landscape and trends. For smaller businesses, directly owning and managing their end customers data may be an unaffordable luxury.

Cloud technology providers and subscriptions license models help to address the software and infrastructure costs with products and tools that are immediately available and license models that often allow for scalability. These providers and the pay-as-you-go model enable enterprises to get started on their data journey without large up-front investments, reducing both cost and risks especially where technology can help to resolve some data governance issues. Their cloud offering enables them to learn best practices from thousands of customers and build that capability into future versions. They also invest in solving complex technical or compliance issues and spread the cost of that investment across a large number of customers and volume of business. The opening up in opportunities and reduction of threats while balancing the costs and risks of creating, maintaining and owning business data is particularly attractive to smaller enterprises and lower level data demand.

Other Drivers in Formulating Approaches to Data Analytics

The trend of software subscription swept North America around 2015 and is now spreading around the world. In 2006, AWS began offering IT infrastructure services to businesses in the form of web services and pioneered the concept of corporations storing their business data at third party data centers, with third party owned servers and operating systems and databases. Salesforce.com then normalized the previously unthinkable operation of a company's customer relationship management (CRM) system as SaaS, kicking off the movement of applications into the cloud. Along the way, this also opens up the option to finance information technology as an operating expense rather than a capital cost to be amortized.

Perpetual software licenses are becoming rare, replaced by 1- to 3-year enterprise subscriptions to database, data analytics application, and tools. On the one hand, this is a major benefit for corporations in reducing the up-front investment and shifting capital expenditures to current period expenses. Subscriptions also offer flexibility to add and subtract users on the fly and payment only for usage. On the other hand, if a system is stable and would last more than three years, a subscription's Total Lifetime Cost likely will exceed the cost of a perpetual license. This explains why Wall Street rewards the valuation of data software companies which have shifted away from perpetual licensing.

Today many U.S. enterprise software companies and virtually all VC funded SaaS and cloud applications providers no longer offer perpetual licenses. They are only available as a subscription. The providers are now taking it to the next level and offer their software only over subscription on their own private clouds. The result is that data analytics software vendors are increasingly controlling both the storage of customers' critical business data and the tools to analyze the data.

Yet, relying on today's vendor or partner to be available tomorrow can be a risky bet. In the fast-moving software and technology industry, mergers are a constant, often resulting in support for products being cut off, prices raised, and new data service business models invented and reinvented. Other unexpected circumstances can disrupt - witness Tableau, now part of Salesforce, announcing an abrupt exit from China and giving existing subscribers 60 days' notice.

However, another consideration will be the growing trend of data and analytics cost. The continuing climb in data volume, the increasing complexity of data as sources and formats multiply and the places in which it is stored and analyzed expand across clouds and traditional internal systems, and increasing need to receive data from and share data with external parties, up and down the data analytics and business supply chains, and in real or near real time all contribute to rising data related expenses. This outlook will add to the burden of companies however they will allocate their financial resources between buy vs. lease. An unfortunate rule of thumb is that the party with less negotiation power often pay for the integration.

The pace of change presents another consideration weighing in on the side of leveraging external resources. Ten years ago, a project or system with a multimillion-dollar build was expected to last for a decade, but today the lifetime of systems is often planned for no more than three years. Thus, the cost of maintaining big data systems is often underestimated. Unlike the water pipes laid in the ground that lasted a hundred years with no maintenance, data pipes obsolescence needs constant attention. As businesses move and change at an ever faster pace, the never-ending stream of business and corporate reorganizations, product line adaptations, and new rules and regulations will keep data systems integrator vendors busy and result in ongoing costs to users and owners of data.

To avoid being unduly locked into a software and service arrangement, businesses need no reminders to engage IT and Purchasing to plan better for expansion, reduction, or contingency and follow through in negotiations. The good news is competition in the service sector and among software companies remains dynamic. AWS and MS Azure are battling it out, and in China competing with AliCloud, Tencent and UCloud.

A darker scenario will be if the data-to-insight industry will develop in a similar way as the computing industry before the turn of the millennium which ended in domination by the Wintel monopoly, leveraging the network effect through the PC operating system, CPUs and peripherals. This analogy presents a threat for everyone else in the data value chain. However, unlike the computation industry, the winners in the data analytics sphere will present a threat also to the consumers of data, especially those occupying previously lucrative industries like food or healthcare. While much of the early ML algorithms started as open sourced, the data platforms are racing to own the access to the growing data streams, and the silicon and the software to operate and navigate in this data infrastructure. To the extent that an equivalent of GitHub is scaling up to supply packages in the data and analytics industry, it may also be acquired one day by one of the platform players, as in GitHub's case its acquisition by Microsoft in 2018.

While capital investment, development cost, and regulatory concerns are key considerations, a factor often overlooked today is the constraint in the supply of skilled labor. Many companies will be simply unable to recruit and retain the kind of data talent necessary to handle everything in-house. Most companies, especially industrial and consumer firms, will need to contend with moderate internal data capabilities and leverage third party systems simply because a vast shortage of data engineering talent worldwide has not abated. Even if the business can afford the technology, it may not be able to attract enough human resource.

******

IV. Bringing It Together (see Data Protection - When Legal Meets Data Analytics Part 3)

Notes

1. The quote is attributed to T.A. Callaghan, Jr., How Much Is Not Enough, The Non-nuclear Air Battle in NATO's Central Region, Allied Interdependence Newsletter No.13, Center for Strategic and International Studies, Jun.21, 1979, Vol.33 Mar.-Apr. (1980), https://www.quora.com/Who-said-Quantity-has-a-quality-all-its-own. This quote is also attributed to Joseph Stalin.

2. G. Press, A Very Short History Of Big Data, Forbes, May 9, 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/?sh=1f0ec3ba65a1. An exabyte is followed by 18 zeros, or 10 to the 18th power. This will be in the ball park of 500 trillion pages of standard printed text, or a billion feature films. For a sanity check, recall that YouTube was founded only in 2005, and Facebook was founded in 2004. Most people only changed to digital cameras after 2003, and the first iPhones hit the markets in June 2007. *See also* M. Vopson, The World's Data Explained: How Much We're Producing and Where It's All Stored, May 4, 2021, https://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964; M. Roser, The internet's history has just begun, Oct.3, 2018, https://ourworldindata.org/internet-history-just-begun.

3. T. Pohl, How Big Data Keeps Planes in The Air, Forbes, Feb.19, 2015, https://www.forbes.com/sites/sap/2015/02/19/how-big-data-keeps-planes-in-the-air/?sh=15b13b5638a7.

4. K. Naughton, Driverless Cars' Need for Data Is Sparking a New Space Race, Bloomberg Businessweek, Sep.17, 2021, https://www.bloomberg.com/news/articles/2021-09-17/carmakers-look-to-satellites-for-future-of-self-driving-vehicles. For a bit of further context, an "iPhone 6's clock is 32,600 times faster than the best Apollo era computers and could perform instructions 120,000,000 times faster". T. Puiu, Your Smartphone Is Millions of Times More Powerful Than the Apollo 11 Guidance Computers, ZME Science, May 13, 2021, https://www.zmescience.com/science/news-science/smartphone-power-compared-to-apollo-432/.

5. F. Lambert, Tesla Unveils Dojo Supercomputer: World's New Most Powerful AI Training Machine, Aug.20, 2021, https://electrek.co/2021/08/20/tesla-dojo-supercomputer-worlds-new-most-powerful-ai-training-machine/.

6. A. Bryant, Ask a Neuroscientist! What Is the Synaptic Firing Rate of the Human Brain?, Aug.27, 2013, http://www.neuwritewest.org/blog/4541.

7. M. Eckler, Digital Payments in China Are Cheap and Convenient, Feb. 2, 2021,https://www.practicalecommerce.com/digital-payments-in-china-are-cheap-and-convenient.

8. https://www.statista.com/statistics/275806/programmatic-spending-worldwide/.

9. S. Carr, How Many Ads Do We See a Day in 2021?, PPC Protect, Feb. 15, 2021, https://ppcprotect.com/blog/strategy/how-many-ads-do-we-see-a-day/.

10. M. Prater, 25 Google Search Statistics to Bookmark ASAP, Jun.9, 2021, https://blog.hubspot.com/marketing/google-search-statistics#.

11. https://www.statista.com/statistics/871513/worldwide-data-created/.

12. MG Siegler, Eric Schmidt: Every Two Days We Created As Much Information As We Did up to 2003, Aug.5, 2010, https://techcrunch.com/2010/08/04/schmidt-data/.

13. J. Sun, Capture the Real Potential of China's Data and Analytics Market Growth, Gartner Local Briefing, Dec.6, 2020, p.3.

14. 25 Biggest GDPR Fines So Far (2019-2022), Email DPL Compliance, Jan.27, 2022, https://www.tessian.com/blog/biggest-gdpr-fines-2020/.

15. S. Lebow, Google, Facebook, and Amazon to Account for 64% of US Digital Ad Spending This Year, Nov.3, 2021, https://www.emarketer.com/content/google-facebook-amazon-account-over-70-of-us-digital-ad-spending/.

16. J. Sun, Gartner Local Briefing, see *supra* note 13, pp.18, 42-44.

17. How "Hyperscalers" Are Innovating – and Competing – In the Data Center, The a16z Podcast, Dec.11, 2021; comment by Nick McKeown, a Stanford professor of computer science, recently appointed Senior Vice President and General Manager of a new Intel organization, the Network and Edge Group.

18. A quote attributable to Marc Andreessen; see also T. Simonette, Nvidia CEO: Software Is Eating the World, But AI Is Going to Eat

Software, MIT Technology Review, May 12, 2017, https://www.technologyreview.com/2017/05/12/151722/nvidia-ceo-software-is-eating-the-world-but-ai-is-going-to-eat-software/.

19, T. Sowell, A Conflict of Visions: Ideological Origins of Political Struggles, (2002) Basic Books.

******

About the Authors

Kenny Tung is General Counsel at Lex Sigma Ltd., where he served as the China advisor to a top U.S. PE fund, and the Asia Pacific advisor to one of the world's top auto components companies and continuing to advise companies in strategic projects and transactions in the region. Kenny also co-founded In-Gear Legalytics Ltd. which helps legal departments of world class companies and law firms to explore and design strategy and process optimization to facilitate transformation of legal services. Projects cover consulting, capability assessment, workshops to address longer term issues, but a common stream concerns the design and implementation of corporate legal strategies. Recently, Kenny undertook an additional role as the Senior Advisor to SSQ in Asia Pacific, facilitating business development and alliance for law firms and management of legal departments, focusing more on people aspects of the people-process-technology spectrum.

Previously Mr Tung served as the Chief Legal Counsel of Geely Holding (during which time the department received the top award for Best Asian & South Pacific Legal Department 2014 by International Legal Alliance Summit) and before that as general counsel in the region at PepsiCo, Goodyear, Honeywell and Kodak where he fielded a vast variety of issues and projects and drove efficiency projects/practices. In 1994, he came to China as a lawyer with Coudert Brothers and led major projects such as the Shanghai GM JV negotiation.

Glenn McCarthy is an entrepreneur and private investor with a primary focus on software and big data analytics companies in China & SE Asia. Glenn is currently the CEO, of Early Data Market Intelligence Co. Ltd., based in Shanghai, and sits on the boards of a number of early and growth stage companies. A native of Boston, Glenn has been living in Shanghai, China for 23 of the past 26 years. During the 1990's Glenn worked for General Electric, responsible for World-Wide Sourcing Programs in Europe, China, Asia, and then globally for the entire corporation.

#KennethTung #GlennMcCarthy #data #legal #inspection #AI #BigData #datamining #DataProtection #DatarotectionLaw #Compliance #DataCompliance #DataAnalytics #DataStrategy #LegalProfession #strategy #lawyers #BigData #BigDataAnalytics #ArtificialIntelligence #MachineLearning #legal #LegalProfessionals #LegalDepartment #LegalDesignThinkiing #LegalManagement #LawTechnology #LawAndTchnology #LegalFunctionTransformation #DataStrategy #DigitalAdvertising #ProgrammaticAdvertizing

More by Kenny Tung

Kenny Tung regularly shares his expertise on Legal Business World. Click on one of these articles and you will be redirected to his publications.

Going In-house: A Changing Marketplace

“These are the times that try men’s soul.” : Thomas Paine, The Crisis, Dec.23, 1776...

If “Software Is Eating the World,”* Is Legal Service on the Menu?

* “Why Software Is Eating the World”, Marc Andreessen, The Wall Street Journal, August 2...

How the challenges of cybersecurity reflect those in the legal...

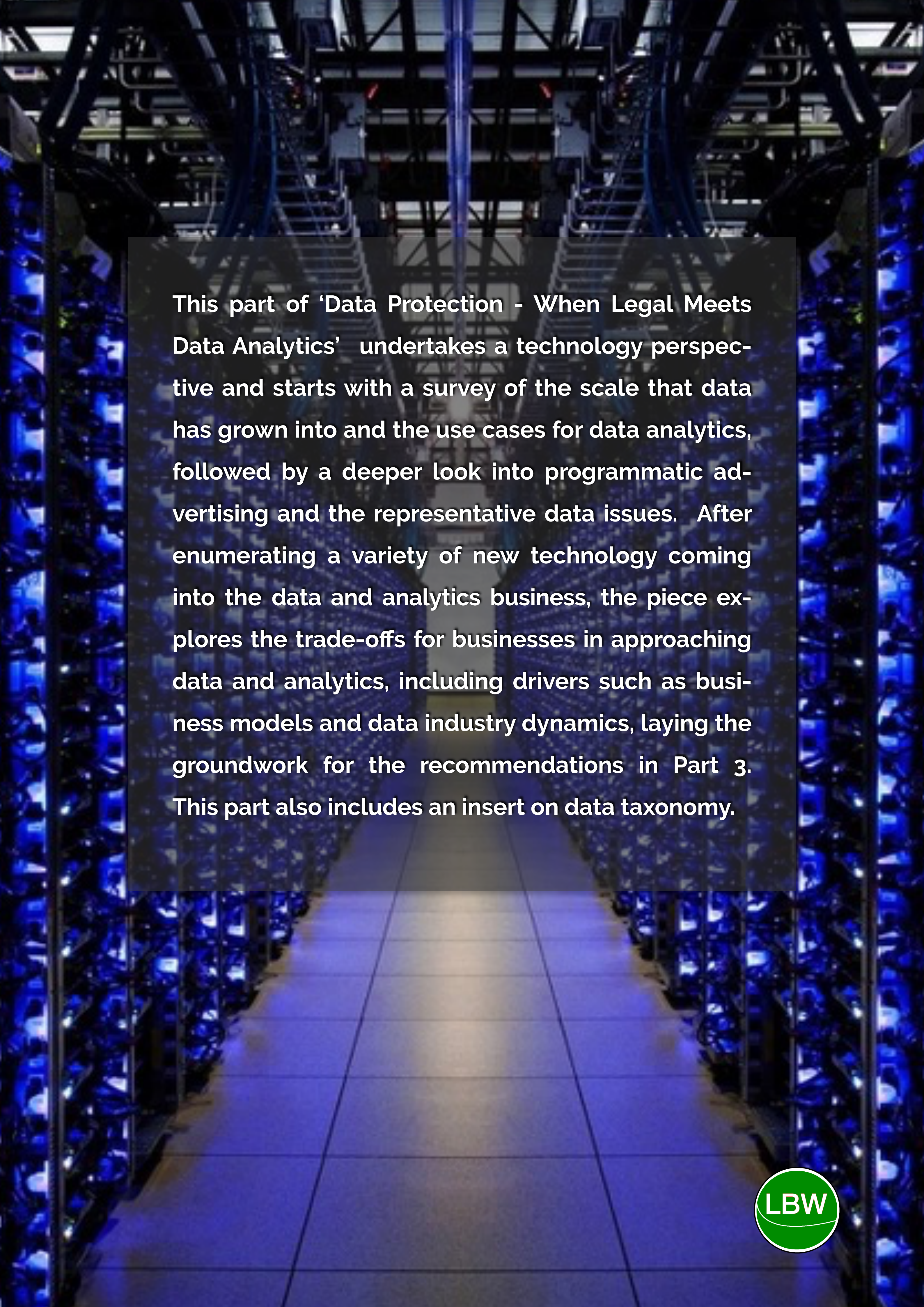
Cybersecurity used to be viewed as black magic. From a non-technical, user or...

Ecce Advocate - Reflections from Conversations in the...

The backward-looking, risk averse approach to the law, which is so common in...

Establishing a Legal Function Beyond Home Jurisdiction

Albeit less newsworthy today, news of surging and sustaining investments from...



This part of ‘Data Protection - When Legal Meets Data Analytics’ undertakes a technology perspective and starts with a survey of the scale that data has grown into and the use cases for data analytics, followed by a deeper look into programmatic advertising and the representative data issues. After enumerating a variety of new technology coming into the data and analytics business, the piece explores the trade-offs for businesses in approaching data and analytics, including drivers such as business models and data industry dynamics, laying the groundwork for the recommendations in Part 3. This part also includes an insert on data taxonomy.