



ARTICLE

Data Overload? A Beginner's Guide to Handling Big Data in Construction and Engineering

Construction projects are rich in data. It is common to find schedules, spreadsheets, letters, reports and photographs on even the smallest projects. This is set to increase with the wider adoption of mobile and cloud-based technologies on construction projects. For organisations, this presents a challenge to effectively manage the ever-increasing volumes of data, and an opportunity to gain valuable insights from it.

What is Big Data?

Big data has become something of a buzzword in recent years. This is likely to have been fueled by the explosive growth in data generated globally. The World Economic Forum¹ reports that we generate 294 billion emails, 65 billion WhatsApp² messages and 720,000 hours of new content on YouTube³ every day.⁴

As the global 'datasphere' continues to grow, we will see a similar increase in the number and size of datasets we encounter with on construction projects. But is this "big data", and if so, how do we analyse it?

Data scientists do not define big data by reference to a certain number of bytes or records, but rather in terms of

the technology used to analyse it. Big datasets are often too large to analyse using conventional databases on a single computer, so the data is distributed across several networked computers. Using the collective resources of these computers allows for greater efficacy when performing computations. Big data is therefore not a synonym for a large database or excel sheet; it is a field of its own with specialised tools that rely on advanced programming skills.

Whilst we may not all deal with big data in a strict technical sense, the frequency with which we interact with large and complex datasets is likely to increase. The effective management and analysis of this data will determine the extent to which we are able to benefit from it.

¹ The World Economic Forum is an international non-governmental organization that brings together political, business, cultural leaders to shape global, regional and industry agendas. It is headquartered in Geneva, Switzerland (https://en.wikipedia.org/wiki/World_Economic_Forum).

² WhatsApp LLC

³ YouTube is an online video sharing and social media platform headquartered in San Bruno, California (<https://en.wikipedia.org/wiki/YouTube>).

⁴ Melvin M. Vopson, "The world's data explained: how much we're producing and where it's all stored," World Economic Forum (May 7, 2021), <https://www.weforum.org/agenda/2021/05/world-data-produced-stored-global-gb-tb-zb/>.

⁵ The Economist is a weekly newspaper focusing on current affairs, international business, politics, technology, and culture (https://en.wikipedia.org/wiki/The_Economist).

Data = Information

The value of data is found in its ability to reveal patterns, trends and risks, which can be used to inform strategic operational decisions. An article from The Economist⁵ stated in 2017 that “the world’s most valuable resource is no longer oil, but data”.⁶ Oil is an appropriate analogy for data given that both undergo a process of extraction and refinement before they reach their maximum value. For data, this process transforms it from its raw format into something useful: actionable information

The steps involved in this transformation are described in greater detail in the following paragraphs.

Collecting the Data

Data exists in a variety of formats and the method of collection can differ for each.

Structured data is information that has been organised into a set format or hierarchy. Examples include relational databases and excel spreadsheets. These are usually straightforward to sort, filter and analyse. Collection of these datasets can include an export from planning or accounting software, or compiling a set of excel files.

Unstructured data is the opposite; it has no clearly defined format and can be difficult to filter and organise. Examples include visual records such as photographs and videos, or text documents such as monthly reports. If recorded on a GPS-enabled device (such as a mobile phone or tablet), information relating to the exact location and timing may be embedded within the record. These files are often referred to as ‘semi-structured’ because this metadata is structured but the recording is not.

Whilst unstructured files are often difficult to analyse, tools such as Aconex⁷ can simplify this process by structuring files and allowing efficient content searches of a large corpus of documents. Image recognition tools that rely on sophisticated machine learning algorithms are becoming increasingly powerful too, albeit that these often need a ‘training set’ of images to teach them how to recognise and label certain objects.

Cleaning the Data

Most people are painfully familiar with the adage “rubbish

in, rubbish out”. We can use the most sophisticated tools and data-mining techniques but if errors and inconsistencies are not identified and corrected, the analysis will produce spurious results. This is why data cleaning can be the most important but time-consuming step.

Typical errors include incorrectly formatted data (for example, dates stored as text), duplicate or missing values, or multiple variants of the same value. Distilling the data to a common set of groupings will allow the analyst to summarise the data effectively, ultimately leading to more robust conclusions.

The process of sifting for inconsistencies and errors is nearly impossible to undertake with the naked eye or even using filters. Using an Access⁸ database or a pivot table in Excel⁹ will allow a more thorough examination of the data.

Data Exploration and Analysis

The data analysis stage should be relatively straightforward if the data is properly prepared. Ironically, the bulk of the work is done before the data is even analysed.

Exploratory data analysis (EDA) is a technique for interrogating the data and will vary in accordance with the desired outcome of the analysis. EDA can involve a numerical analysis of the data, for example computing distributions or averages, or a graphical approach to reveal trends or interrelationships in the data.

Power BI¹⁰ is free data visualization software that can be used for exploring data. This was developed by Microsoft¹¹ so integrates easily with Office¹² applications and allows for visual inspection of the relationships between variables in a dataset.

Presenting the Data

The analysis findings should be communicated as clearly and efficiently as possible. This may entail the use of summary tables, charts or other data visualisations.

Information design is both an art and a science, but often little thought is given to either aspect to the detriment of the audience. As a rule, less is often more when presenting information. In other words, graphics that are cluttered and heavily annotated require more time to understand and interpret. This diminishes the impact of that graphic

⁶ “The world’s most valuable resource is no longer oil, but data,” The Economist (May 6, 2017), <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>.

⁷ Aconex is a mobile and web-based collaboration platform for project information and process management on a software as a service (SaaS) basis, to clients in the construction, infrastructure, power, mining, and oil and gas sectors (<https://en.wikipedia.org/wiki/Aconex>).

⁸ Microsoft Access

⁹ Microsoft Excel

¹⁰ Microsoft Power BI

¹¹ Microsoft Corporation

¹² Microsoft Office suite of applications, including Word, Excel, and Access.

and clarity of any insights gained from it.

A better practice is to minimise the graphical furniture (e.g. lines, dashes, axes etc.) which can be distracting to the human eye. This will ensure the salient data points are the focus of the audience.

Conclusion

Tradesmen are trained to use their tools, teachers are taught how to teach, but there is often little training in how to manage and analyse large datasets, despite the frequency with which we encounter them. Construction organisations have historically been slow to embrace new technologies, which often results in overloaded spreadsheets and poorly structured data.

By training employees to proficiently manage and analyse data, companies can proactively adapt to the growth of data volumes and even benefit from it. This may even assist in improving the delivery of projects and retention of key records in the event of a dispute.

Organisations may also wish to consider hiring professional data analysts. A LinkedIn¹³ survey recently found that data scientist was the fastest growing job in Singapore.¹⁴ This reflects a wider trend of companies recognising that they are data rich but information poor.

Either way, doing things the way they have always been done is no longer a viable option for construction firms.

¹³ LinkedIn Corporation

¹⁴ A commentary on the survey's key findings is available at: Claudia Chong, "Here are the 5 fastest-growing jobs in Singapore, says a LinkedIn survey," The Straits Times (September 7, 2018), <https://www.straitstimes.com/business/economy/here-are-the-5-fastest-growing-jobs-in-singapore-and-why-many-are-filled-by-foreign>. This is based on LinkedIn's Emerging Jobs Report: Singapore (2018) (<https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/webinars/2018/images/infographics/sg-emerging-jobs-report.pdf>)

The views expressed herein are those of the author(s) and not necessarily the views of FTI Consulting, Inc., its management, its subsidiaries, its affiliates, or its other professionals.

FTI Consulting, Inc., including its subsidiaries and affiliates, is a consulting firm and is not a certified public accounting firm or a law firm.

MARK CURTIES

Associate Director
Construction Solutions
+65 6831 7894
mark.curties@fticonsulting.com