BREAKING DATA SILOS FOR EFFECTIVE DECISION MAKING

K. Fordham & N. Brown (Auckland Council), S. Greenwood & N. Tomasi (Mott MacDonald New Zealand Limited)

Abstract

Reality is complicated and messy. Key decisions that might allocate resources, communicate to the public, or direct rapid responses to events, all strongly depend on the full context of the situation. While the amount of data collected every day is growing at an ever-increasing rate, the interpretation of these datasets for decision making often present challenges. Complex problems can rarely be resolved by the analysis of a single data source. The integration of multiple datasets into a unified system is crucial to inhibit data silos, enable big-data analysis, and deliver effective outcomes. Auckland Council and Mott MacDonald have spent several years employing such unified systems, which provide real-time data analysis from a wide range of sources to deliver targeted insights.

The Auckland Council Rainfall Analysis System is one example of an integrated system. This cloud solution analyses data such as rain radar, rain gauges, water flow/level sensors, public requests for service, cameras, weather warning and GIS layers. It supports the Auckland Council's flood response by consolidating datasets and specialist data services (such as rain radar processing), providing information in an easy to access format for the organisation and their partners. Crucially, the analysis of these datasets is visualised and reported to operators and decision-makers to quickly and reliably identify areas impacted by storm events a prioritizing resources efficiently, without the need to access and compare information from multiple proprietary software packages.

Another example of an integrated system is Safeswim, the water quality digital twin for public beaches. This solution allows the public to make informed decisions around the health risks of swimming and reflects a wide range of integrated datasets and models the Auckland Council, Mott MacDonald and other partners such as Watercare, Puhoi Stour, DHI and Surf Life Saving New Zealand have collaborated on. The collective datasets provide real-time accurate forecasts of water quality and beach safety, in one location, providing a quality source of information that can confidently be shared with the public.

These systems operate upon a strong platform specially built to enable data integration and analysis, owing its success to a series of key development plans and design architecture choices. Early and continued investment has allowed the platform to be appropriately planned in collaboration with organisations with a shared vision, of continuous improvement to deliver improved outcomes. The platform is scalable, to meet the challenges of working with large datasets such as the rain radar, which provides >20,000 virtual rain gauge sites across the Auckland region with 60 second temporal resolution. A robust API enables the ability to build automated workflows for any third party who can access and add to these datasets for further analysis.

To reach the intended outcomes, clear communication is necessary between all teams involved, from software developers working on the data ingestion through to end-users interpreting data and interacting with the platform. Each step in the pipeline from acquiring and storing the data, processing them in conjunction with other data sources, and visualising it to communicate information about the physical world need to complement each other.

Introduction

A data or information silo is a repository of information existing within an organisation that is cut off from other systems within the same organisation. Analogous to the idea of a grain silo, any single grain is freely available to all other grains within the silo but is entirely unknowable outside. Silos can form for a number of reasons such as a siloed organisational structure or the use of legacy technology not built to easily share data. Most commonly, different departments, or even teams within departments, collect and store their data in a fashion specific to an activity, without prior consideration of how that data may be made accessible to the rest of the organisation. The concept of breaking data silos is hence about implementing technologies and practises that enable interconnections across these information repositories. Doing so breaks down the metaphorical walls of the silos, promoting collaboration and freeing up data analytics to take advantage of the wider scope of information.

Data analytics that exist within any particular silo are usually constrained and not able to take advantage of data from others; so the more silos that exist, the more limiting the capability of the analytics to inform data driven decisions becomes. For information to be considered siloed, it does not have to be entirely inaccessible. Sometimes the speed at which that data is accessible can be sufficiently slow that it may practically be unusable. This is particularly true in time-pressured situations, such as following a large rainfall event, when decision-making relies upon up-to-date rapid analytics. If decisions need to be made within a timescale of hours to days, but siloed information takes a week or more to be obtained and processed, then the information loses all value in that moment and it is already too late to address the silo.

A crucial piece of the puzzle is to have members of the organisation all on board and driven to remove data silos. If only a few members are pushing for the removal of silos, they will often come up against insurmountable obstacles, be them budget or time constraints or a lack of technical support to implement change. On the other hand, when everyone is in agreement, the culture of the organisation lends itself to providing the motivation for putting plans into action. Every organisation is held back to some scale by data silos, they are not a unique problem, nor will the potential for their creation ever disappear. As quickly as we can produce new data, we create the risk of starting to accumulate this information into a disconnected system.

To adjust a common saying about saving money for a pension: the best day to start breaking a data silo is yesterday and the second-best day is today. The scale of work required to identify, plan, and finally take action to break a data silo is directly related to the size of the silo, in that, the more data that is isolated from the rest of the organisation, the more difficult it is to make it available. Undeniably, there is often a higher upfront cost with breaking down a silo than continuing with 'business as usual'. However, the technical debt associated with a disconnected data environment constantly accumulates and over time it will significantly outgrow the upfront cost of dealing with the isolated information early on. Understandably, some silos will be considered more valuable to address than others but at the very least being aware is the first step to knowing how to deal with them. Identifying silos sooner rather than later is therefore critical. Additionally, as progress is made to break down silos, additional benefits and value can be identified, compounding the impact and utility of the data silo as new ways to access, analyse and publish data are realised through time.

In this paper, we discuss examples of Auckland Council led programmes in collaboration with Mott MacDonald and others, highlighting how breaking data silos has led to producing effective outcomes by enabling data-driven decision making. Therefore, we will provide brief descriptions of the projects' intricacies and focus on a high-level overview of the systems in place that have contributed to their success. Additionally, we aim to draw attention to the supplementary benefits that the programmes have realised during their continued development that have complemented the original vision.

Case Studies

The Auckland hydrometric data set, was the first silo identified, where additional benefits could be realised within Healthy Waters. Historically this data was stored and maintained in a corporate repository, requiring specialised skills and knowledge to access, analyse and interpret data, during and post rainfall events. One of the key objectives of the project was to utilise real-time rainfall time series data to produce Annual Recurrence Interval (ARI) statistics to quantify how large a particular rainfall event was (e.g., a 50-year event) for the various stakeholders in impacted areas. Calculations for these statistics across numerous rain events every year, represented a significant investment of time to produce.

In order to generate these reports more efficiently, the rain gauge data was setup to automatically be ingested into a real-time data analysis platform (in the case for the projects in this paper, the platform is Moata from Mott MacDonald). The key point here is that now, the architecture of the platform enables analysis on the live data, producing ARI statistics that are instantly accessible. Employing this data pipeline took the timescale for producing these reports on rainfall events from 3 days down to 30 minutes, giving stakeholders much faster access to the information and reducing workload/costs. Both the time series data and the ARI statistics are now

democratised, allowing anyone with an internet connection to able to access them via the cloud based central platform. Figure 1 demonstrates the high-level overview of the system.

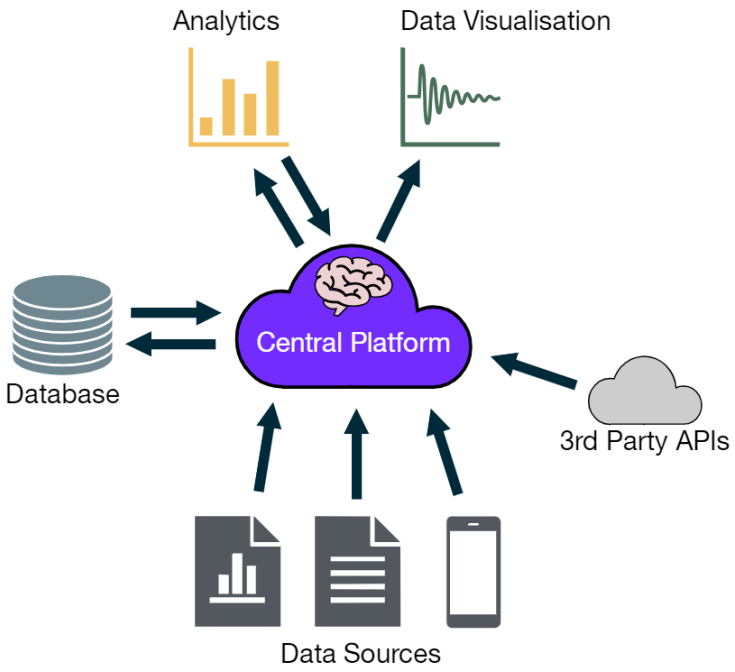

Figure 1. Data from different sources, including 3rd party data via Application Programming Interfaces (APIs) is fed to the central platform. The platform is then able to provide intelligent delivery of analytics and data visualisation.

One of the core ideas behind breaking down the data silos is having a solution that is data agnostic. For example, the time series rainfall data is processed by the central platform in the same way as the tide predictions, where the rainfall data are uploaded from the council but the tide predictions are obtained via a third party API. Obviously, both are being analysed in different ways but this is an important feature that ensures interoperability with many different types of data coming from many different sources. For any new or improving programmes, not being limited by necessitating bespoke systems for each of them rapidly accelerates their development and capabilities.

With the foundation for Auckland Rainfall in place, additional datasets were added to complement our understanding of rainfall events, evolving and improving the project over time. Stream flow/level gauges, tide data, soil moisture probes, MetService warnings, images from remote cameras monitoring storm drains and GIS layers (e.g. flood plains) could all be integrated into the platform and be drawn upon to contextualise real life decisions. Datasets will never be complete and perfect but being able to move forward with available information and comparing these datasets in a common environment enables users to have a more comprehensive understanding of the situation on the ground. Rain radar data now complements the rain gauge data to provide improved spatial representations of rainfall events, trajectories of weather systems, and reliable high-quality input to hydraulic flood models.

The expected trend that emerged is that it becomes exponentially easier to bring more and more data together to drive decision making in the absence of data silos. This trend continued with the delivery of the Operational Dashboard, a project similar to Auckland Rainfall but with key operational requirements over how the data is presented and reported, particularly given it may be needed in time pressured situations. Rather than starting from scratch, select data from the Auckland Rainfall project can be shared and supplemented by Request For Service (RFS) data originating from calls to the council from the public. Prior to the Operational Dashboard, the RFS data was contained within its own data silo that required breaking to allow it to be shared across departments. Following the initial challenge of breaking the silo, the RFS dataset is now central to the decision-making process and is now used for other activities across council business.

The Closed-Circuit Television (CCTV) pipe inspection programme is a proactive (and reactive) data collection programme to assess the condition of the piped stormwater network. In Auckland there are upwards of 6500km of pipe assets, and a large volume of data to manage. Pipe inspections are carried out by many different contractors,

therefore, to manage increasing large data sets associated with ongoing CCTV inspections it was critical to develop a centralised data system rather than having multiple datasets from each contractor with potentially incompatible data, A significant amount of effort was required to design the system architecture, processes and workflow to collect the data from different contractors and educate users on how to use it (and inevitably deal with issues and bugs from a new workflow). Years later however, this initial investment to standardise the data collection continues to pay off and has ensured data silos did not have a chance to form in the first place.

When data silos are prevented from being created, opportunities not originally envisaged can reveal themselves. A realised benefit from the stormwater pipe inspection dataset is that it now opens avenues for data analysis not previously feasible. Using the consolidated regional dataset, statistical modelling to predict which pipes are most at risk of failing can be realised, which can focus intervention and resource allocation. Since the standardised collection of inspection data has been carried out over the last decade, and prevented data from being siloed away, the construction of models to predict pipe condition utilising this dataset has been practical to achieve.

Improving the utility of data and removing data silos can enable wider benefits that would otherwise not be realised if silos existed. The rainfall datasets (gauges and radar) are widely used datasets for the aforementioned Auckland Rainfall/Operational Dashboard projects and now also for Safeswim and its associated water quality sampling programmes. All projects are able to tap into the same dataset as opposed to maintaining separate copies. If the later were true, inevitably, inconsistencies would arise and trust in the dataset would decrease. In the current scenario where all users of the data are able to effectively point back towards the single source of data, then downstream users are able to instantly benefit from improvements made to the upstream dataset. This continuity benefits Safeswim, which is able to have high confidence in the rainfall data which is an important driver of water quality predictions and hence an important driver of decisions around communicating the risks of swimming to the public.

In summary, breaking down data silos early and unifying data access across an organisation has enabled Auckland Council to be confident in its decisions as they are backed by robust analysis of multiple datasets. There is an initial cost with breaking down data silos but the technical debt it cancels, which can accumulate for years, can vastly outweigh the initial investment. Key datasets, such as the rainfall data, have been made easily accessible across the council and relevant 3rd parties and are used across multiple programmes, maximising the value from the information and avoiding costs in maintaining unnecessary duplicates by keeping a single source of truth. With a good data agnostic foundation, opportunities for improvements are made easier, permitting and leveraging additional data as it becomes available. Finally, the early investment into breaking data silos has meant that the datasets collected by the various programmes are now reaching a high level of maturity. This depth of information is starting to unlock much more sophisticated analysis around complex physical systems such as flooding, which require clear data communication and a trusted, decisive decision-making process. Had the original implementation of the initiatives required this later level of analysis immediately, the projects would have quickly been deemed a failure. Only by building upon a long-term vision, with consistent incremental improvements, was it practical to get to the current capabilities.