Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 119 (2015) 458 – 463

Procedia Engineering

www.elsevier.com/locate/procedia

13th Computer Control for Water Industry Conference, CCWI 2015

Waternomics: a cross-site data collection to support the development of a water information platform

Peter O'Donovan[ab*], Daniel Coakley[ab], Jan Mink[f], Edward Curry[de], Eoghan Clifford[abc]

[a] College of Engineering and Informatics, NUI Galway
[b] IRUSE, NUI Galway, Ireland.
[c] The Ryan Institute, NUI Galway, Ireland.
[d] DERI, NUI Galway, Ireland.
[e] INSIGHT Centre for Data Analytics, NUI Galway, Ireland.
[f] VTEC Engineering, Netherlands.

Abstract

In Europe, 20 to 40% of water is being wasted due to poor infrastructure, consumer negligence and lack of proper resource management. Effective and efficient management of water resources requires a holistic approach considering all the stages of water usage, which includes the application of ICT technologies that are capable of transmitting and analysing water data to deliver actionable insights to end-users. A primary requirement for ICT-based water innovations is the ability to collect and process water data from disparate locations. In this paper, we present a cloud-based cross-site data collection system, which is being developed as part of an EU funded research project named Waternomics.

Peer-review under responsibility of the Scientific Committee of CCWI 2015

Keywords: water, ict, information systems, decision support, data analytics, data integration

1 Introduction

In Europe, 20 to 40% of water is being wasted due to poor infrastructure, consumer negligence and lack of proper resource management [1]. Effective and efficient management of water resources requires a holistic approach considering all the stages of water usage . Given the emergence of pervasive and ubiquitous computing in recent years, ICT can play an important role in different aspects of water monitoring and management [2]–[5], as well as providing the basis fault detection and diagnostics to promote corrective and predictive maintenance strategies [6],

* Corresponding author. Tel: +353 91 524 411
E-mail address: peter.odonovan@nuigalway.ie

Peer-review under responsibility of the Scientific Committee of CCWI 2015
doi:10.1016/j.proeng.2015.08.861

[7]. This paper centres on one such research project, namely Waternomics. In simple terms, Waternomics is an EU funded research project that aims to promote awareness of water usage by integrating, contextualising and disseminating information that can be used to realise water efficiencies. The main component of Waternomics is the development of a water information platform that provides decision-makers with an easy and intuitive means of accessing key metrics and analytics relating to their water usage, leakage and maintenance. In turn, it is envisaged that these real-time insights can be employed by end-users to inform better decision-making, and gain a better perspective on actual water usage, demands and patterns.

A unique aspect of the Waternomics platform is its ability to personalise and customise the metrics and analytics presented to different groups of users. More specifically, the platform focuses on delivering pertinent water information to three different groups of users, namely domestic, municipal and corporate. By delivering insightful and targeted information to users that is aligned with their water usage context, the platform can serve highly relevant information that has the best chance of promoting behavioural changes for that particular target group. In turn, these behavior changes can have a positive effect on water usage, operational efficiency and maintenance, as well as being a primary source of information for future initiatives regarding water policy.

To deliver water information to end-users a process for collecting data from disparate and remote pilot sites must be established, as well as the development of a protocol that will enable development partners to easily access that raw water data. However, defining such a data integration strategy can be challenging as development partners may be using different technologies, platforms and frameworks. Given these differences, reaching an agreement on a strict set of data access protocols and policies for routine data ingestion, is likely to be a time-consuming endeavor, which may only serve to reduce overall productivity and output. Therefore, in this paper, we present the requirements, strategy and architecture for a cloud-based intermediate staging area for water data. This serves as an open and accessible cross-site data repository of log files relating to water usage, which development partners can ingest using standard HTTP.

This paper is organised as follows – (1) methodology describing our approach to this aspect of the research, (2) results in the form of requirements and architecture for a cross-site data collection system, and (3) conclusions and summary of this research.

2 Methodology

Our methodology for this research followed a four step process, which involved agreeing on the main research questions and objectives, collaborating with relevant research partners and pilot sites to determine the data requirements and characteristics that are applicable to them, and finally prioritise and filter these requirements so that the proposed cross-site data collection solution can address the most common and meaningful requirements.

2.1 Research questions

RQ1 – what characteristics and attributes should pilot site data collection exhibit to simplify and unify data access for development partners?

Given the need to service the data requirements of multiple development partners, the purpose of our initial research question was to identify the needs of these stakeholders, and highlight those requirements that were essential to the successful operation of a cross-site data collection system.

RQ2 – what type of information system architecture can be used to meet the data requirements of development partners?

The purpose of this research question was to produce an information system architecture and workflow that could enable development partners to access data in accordance with the findings of RQ1.

2.2 Stakeholder collaboration

As the Waternomics project includes stakeholders from different EU countries, we used virtual meetings to discuss data requirements and suggestions from each stakeholder. These real-time discussions were supplemented with asynchronous messaging in the form of an internal mailing list, which enabled us to refine and prune post-

meeting requirements.

2.3 Pilot site evaluations

We conducted an evaluation of two pilot sites to ascertain data and communication requirements for each site. These pilot sites chosen were the NEB and Colaiste Ni Corba, which are both situated in Galway, Ireland. These sites were chosen because of their geographical proximity to us, in addition to the fact that we had direct access to operational and technical staff that could support or evaluation of these sites.

2.4 Filtering and prioritisation

After gathering requirements through discussions with stakeholders, and pilot site evaluations, we carried out a filtering and prioritisation process to highlight important and common requirements across both sets of data. We approached the prioritisation process from the perspective of the development partners, as these stakeholders are the primary data consumers of pilot site data.

3 Results

3.1 RQ1 - Requirements and objectives

Based on the output from meetings and discussions with development partners and stakeholders, coupled with our own internal prioritisation and filtering process, the following items represent the key requirements for the data collection system.

Consistent data structure

A common issue with data access and integration scenarios, such as the case with BMS databases, is that the representation and structure of the data varies from site-to-site – ranging from vendor-specific, to bespoke, data models. This type of heterogeneity in the underlying data model makes integration time-consuming and expensive, as consumers of that data are required to implement custom scripts and code to wrangle the data in to a form that is suitable for their application. Furthermore, in scenarios where multiple development partners are working on several integration scenarios, there is potential for inefficiencies and waste through the implementation of duplicate scripts and code.

Therefore, a key requirement for serving multiple development partners is the ability to provide a single and consistent structure to the data that is presented to their applications and platforms.

Lightweight and geographically accessible

Historically, data integration scenarios were considered database integration, warehousing or management problems. These types of database solutions are governed by a schema, which enforces a particular structure on the data to which all consumers of that data must subscribe. Further to this, consumers of data from a database must utilise specific drivers to communicate with the database, as well as ensuring that the appropriate communication ports are open on both the application and database layers. While schemas are not necessarily always a bad way of structuring data, the existence of multiple proprietary data models (e.g. multiple BMS's) is a significant factor in reducing data accessibility, and greatly reduces the opportunity for code reuse, or indeed, interoperability between applications.

Therefore, a key objective of the data collection system should simplify data access by making the process for consuming the data lightweight, in terms of its connection and querying protocol, and be highly accessible to development partners from across the EU, without any significant network or communication constraints.

Archived and real-time data streams

Another challenge is the amalgamation of water data at different latencies and resolutions. In general, BMS's that are used to monitor water data export log files, or initiate database dumps every 24 hours. This is classified as archived data as it contains historical data regarding sites state. In contrast to this, newer and internet-aware sensors are capable of transmitting real-time water data at sub-second intervals. However, while IoT smart sensors are likely to become the de facto standard for embedded monitoring of water usage in the future, it is still important to

incorporate traditional database systems and log files in the analysis of water data.

Therefore, a key requirement of the data collection system is to manage sources of archived and real-time data in a transparent fashion, so that data consumers can access water data in an indiscriminate manner.

3.2 RQ2 - Cross-site data collection architecture

illustrates the cross-site data architecture that we tailored to meet the requirements identified by RQ1. The diagram is decomposed in to three layers. At the bottom layer two pilot sites are represented along with their local resources for monitoring water data. In the middle layer is the Amazon S3 cloud-based file repository, which offers a highly distributed and fault tolerant means of serving files to end-users. Finally, the top layer shows the array of data consumers, which are essentially our development partners in the Waternomics project. The alignment of the data collection system architecture with that of the requirements from RQ1, are discussed to in the following sections.

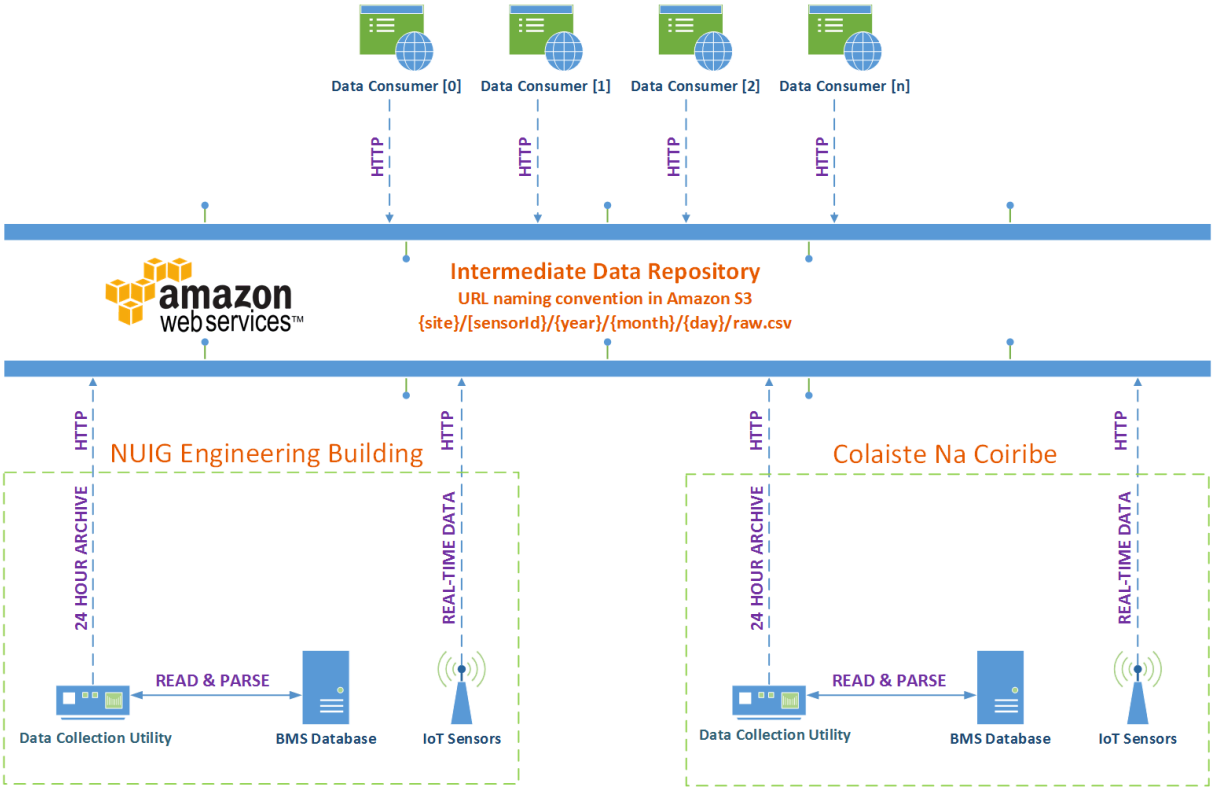

Figure 1 Cross-site data collection architecture

Data collection utility

To collect data from existing database systems, or indeed, other static data repositories residing on these pilot sites, a data collection utility is used. This utility is deployed to a location where it has direct access to the relevant data sources. In both of the pilot sites presented in the data collection utility is used to read and parse measurements from the BMS database every 24 hours, before transmitting the data to the cloud repository over HTTP. The parsing routine ensures that all data is transformed to a consistent format consisting of key/value pairs (i.e. timestamp/measurement), and that each 24 hour log file is placed in the appropriate directory according to the naming convention shown in .

Building Management System (BMS) database

BMS's operating on the pilot sites capture measurements taken from water meters and sensors located in the respective buildings. The measurements table contains a SensorId field, which is used by the data collection utility to filter, group and organise the measurements before uploading the logs to the cloud.

IoT sensors and devices

While the data collection utility caters for the integration of water data using legacy information system and data archives, IoT sensors on the pilot sites enable real-time high-resolution measurements of water data, such as flow rates in the buildings. These sensors implement the same protocol/interface as the data collection utility, and transfer data to the cloud over HTTP. In a similar manner to the data collection utility, IoT Sensors ensure that data is in a consistent format before uploading to the cloud, and distils its readings in to the appropriate directory by following the naming convention shown in .

Intermediate repository

The intermediate repository is a highly accessible and scalable staging area for pilot site data. Its purpose is to provide an open and standard approach for development partners to access site data, without enforcing proprietary technologies, policies or processes. Accessing the data can be undertaken using simple HTTP GET requests from a wide-variety of applications, tools and services. In particular, the intermediate data repository abstracts development partners from low-level integration, and provides all partners with a consistent and predictable method for consuming all pilot site data. This consistency is realised through the use of a descriptive and contextualised URL, as well as the structure of each pre-processed log file (i.e. key/value pairs of timestamp/measurement). As a result of this uniform approach, it is reasonable to expect a reasonable level of reusability between applications, as ingestion and processing is coded against the same structure and protocol.

Data consumers

Given the data stored in the intermediate repository is in a raw and highly granular state, it must be consumed, processed and analysed to deliver value. In the context of this research, the intermediate data repository is the primary source for streaming data from pilot sites, which is ingested by intelligent and insightful analytics applications that comprise the Waternomics water information platform. By abstracting development partners from time-consuming and expensive integration scenarios, and removing the possibility of duplicate integration routines across different sites, the data collection system enables partners to focus on delivering high-value and high-impact applications.

4 Conclusions

The cross-site data collection architecture presented in our results illustrates how an intermediate staging area for pilot site data, could be used to deliver data to development partners and stakeholders from across the EU. The architecture is based on the idea of non-obtrusive data management, which is void of proprietary technologies or protocols. By unifying the structure of the data and enforcing a naming convention, which are accessible via open standards and protocols (e.g. HTTP), development partners do not need to manage the modalities of distributed data integration, and can therefore exert more effort on delivering high value applications. Furthermore, as development partners build tools and applications that use the intermediate data repository to build new and interesting data sets, the potential for reuse between partners will also increase, and if this is realised, the level of redundant and duplicated code across the project could also be controlled.

In conclusion, while data collection in itself does not provide value for end-users directly, modern data-driven systems, especially those of a distributed or large-scale nature, would be unable to operate in a timely, reliable and efficient manner without cohesive and flexible data integration processes that can serve higher functioning applications. Furthermore, as the size, diversity and scale of information systems increase, the introduction of multi-stage data storage and processing, such as the intermediate data repository that was mentioned in this paper, can be used to service the needs of different data consumers without employing restrictive or proprietary technologies.

Acknowledgements

The research leading to these results has received funding under the European Commission's Seventh Framework Programme from ICT grant agreement WATERNOMICS no. 619660. It is supported in part by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

[1] World Business Council for Sustainable Develeopment, "Business in the world of water: WBCSD Water Scenarios to 2025," 2006.

[2] E. Curry, V. Degeler, E. Clifford, D. Coakley, A. Costa, S. J. van Andel, N. van de Giesen, C. Kouroupetroglou, T. Messervey, J. Mink, and S. Smit, "Linked Water Data for Water Information Management," in *11th International Conference on Hydroinformatics (HIC)*, 2014.

[3] E. Curry, J. O'Donnell, E. Corry, S. Hasan, M. Keane, and S. O'Riain, "Linking building data in the cloud: Integrating cross-domain building data using linked data," *Adv. Eng. Informatics*, vol. 27, no. 2, pp. 206–219, Apr. 2013.

[4] E. Curry, S. Hasan, and S. O. Riain, "Enterprise Energy Management using a Linked Dataspace for Energy Intelligence."

[5] D. Hartley, "Acoustics Paper," in *Proc. of 5th IWA Water Loss Reduction Specialist Conference*, 2009, pp. 115–123.

[6] K. Bruton, P. Raftery, P. O'Donovan, N. Aughney, M. M. Keane, and D. T. J. O'Sullivan, "Development and alpha testing of a cloud based automated fault detection and diagnosis tool for Air Handling Units," *Autom. Constr.*, vol. 39, pp. 70–83, Apr. 2014.

[7] K. Bruton, D. Coakley, P. O'Donovan, M. M. Keane, and D. O'Sullivan, "Development of an Online Expert Rule based Automated Fault Detection and Diagnostic (AFDD) tool for Air Handling Units: Beta Test Results," in *ICEBO - International Conference for Enhanced Building Operations*, 2013.