Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 2250–2259

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Towards an operational database for real-time environmental monitoring and early warning systems

Bartosz Balis[1], Marian Bubak[1], Daniel Harezlak[1], Piotr Nowakowski[1], Maciej Pawlik[1], and Bartosz Wilk[1]

AGH University of Science and Technology, Department of Computer Science, Krakow, Poland

Abstract

Real-time environmental monitoring, early warning and decision support systems (EMEWD) require advanced management of operational data, i.e. recent sensor data needed for the assessment of the current situation. In this paper we evaluate the suitability of four data models and corresponding database technologies – MongoDB document database, PostgreSQL relational database, Redis dictionary data server and InfluxDB time series database – to serve as an operational database for EMEWD systems. For each of the evaluated databases, we design alternative data models to represent time series data, and experimentally evaluate each of them. We also perform comparative performance evaluation of all databases, using the best model in each case. We have designed performance tests which reflect realistic conditions, using mixed workloads (simultaneous read and write operations) and queries typical for a smart levee monitoring and flood decision support system. Overall the results of the experiments allow us to answer interesting questions, such as: (1) how best to implement time series in a given data model? (2) What are the reasonable operational database volume limits? (3) What are the performance limits for different types of databases?

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Environmental monitoring, time series data, operational database, smart levee monitoring

1 Introduction

Large-scale environmental monitoring, early warning and decision support systems (EMEWD) need to process massive sensor data streams in real time [9, 12]. In such systems it is useful to clearly distinguish between operational data which is needed for ongoing assessment and decision-making, and archive data which is used for research and analytical purposes. While the distinction between operational and analytical databases is common in business applications [6, page 9], EMEWD systems have specific characteristics in terms of the nature of operational data (the bulk of which consists of time series records), and the way it is used.

In this paper we evaluate the suitability of four different data models and representative databases to serve as an operational time series database for EMEWD systems: (1) MongoDB document database, (2) PostgreSQL relational database, (3) Redis dictionary data server, and

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.193

(4) InfluxDB time series database. Our objective is to conduct experimental performance evaluation in order to find out how to best represent sensor data and what are the limits of the investigated databases in terms of the storage volume and performance.

Related work evaluating performance of sensor storage platforms includes [13], where authors compare MongoDB, PostgreSQL and Cassandra, and [11], evaluating MongoDB, Redis, and CouchDB. However, these publications do not present alternative sensor data models (instead focusing on a single, simplistic model) or representative queries, and do not perform mixed workload tests (simultaneous reads and writes). In [8] three native time series databases are compared (OpenTSDB, KairosDB and Databus), but the main focus of experiments is to investigate their scalability in the cloud. Aydin et al. [1] present a sensor data storage solution, but their main focus is data analytics leveraging big data techniques. In [7], the authors propose and evaluate a data management solution for smart environments and robotic applications. Experimental results are provided to compare the performance of Cassandra and MongoDB. Some test scenarios, however, are rather specific to the application.

The paper is organized as follows. Section 2 describes research context and methodology. Section 3 introduces the design of alternative time series data models for all databases. Section 4 presents experiments and discusses their results. Finally, section 5 concludes the paper.

2 Context and methodology

2.1 Research context

The context of the presented research is an operational database for an environmental monitoring, early warning and decision support system, whose generic architecture is illustrated in Fig. 1. We have implemented such an architecture in the smart levee monitoring [3] and flood decision support project ISMOP [2, 4].

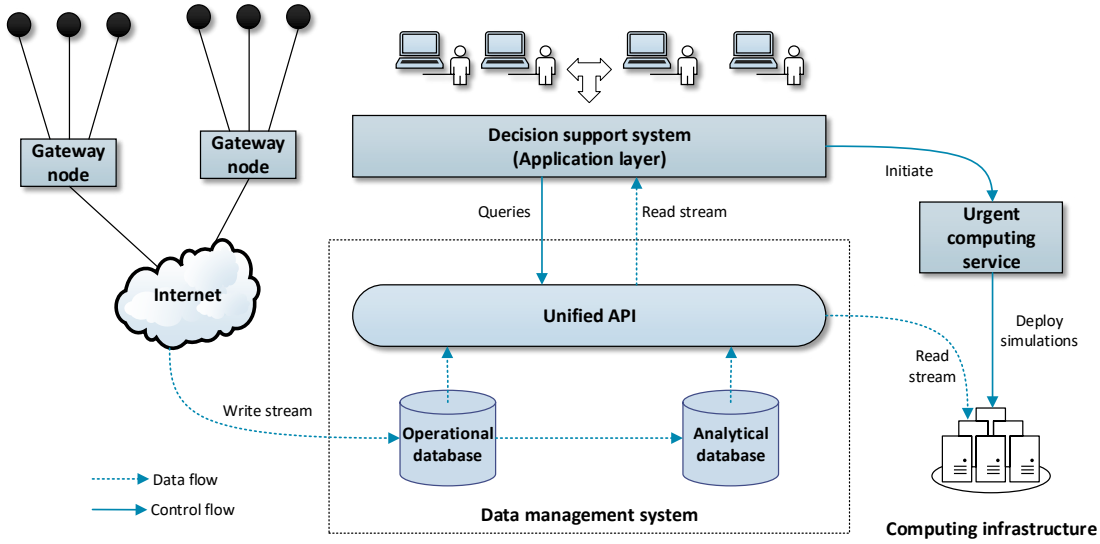

Figure 1: Simplified reference architecture of an environmental monitoring, early warning and decision support system.

The architecture focuses on the Data management system which needs to be designed to