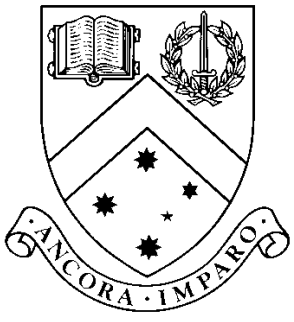

Monash University

Faculty of Information Technology

Computer Science Honours Reading Unit

Case Study of Inconsistent Railway Data With MongoDB

Author:
Jonathan Poltak Samosir

Supervisor:
Dr. Maria Indrawan-Santiago

May 13, 2015

Contents

1 Introduction

While relational database management systems (RDBMS) have been somewhat of a "go-to" solution for for a number of years for general data storage and management problems that many application developers face, we have noted a recent rise in the use of non-relational data management tools [6]. Most of these tools have traditionally fallen into the domain of big data analytics, with platforms such as the Hadoop ecosystem [1] being notably popular. However, outside of the domain of big data, looking more at general purpose data storage and management, what are now commonly referred to as "NoSQL" solutions are proving to be a popular solution.

NoSQL databases refer to those databases that are not built on top of the relational algebraic concepts, as laid out by Codd in 1970 [1], unlike the more commonly used RDBMS technologies, such as MySQL [2]. Being free of the strictness the relational model enforces on its data allows NoSQL databases to focus less on the overall structure of data, and more on factors such as scalability and performance [5].

While the relational model is a good fit for many data problems, its strictness in terms of flexibility of managing data eventually led to the introduction of the NoSQL model. The following characteristics can be given as a starting point for NoSQL databases in comparison to relational databases [4]:

- **Unstructured data support:** While the relational model would often force data to be stored in tabular formats, the NoSQL model does not force any kind of data schema.
- **Designed with distributed processing and horizontal scalability in mind:** Given the commoditisation of computer hardware in the last decade, support for horizontal scaling and processing among clusters is an important factor for adoption.
- **Less strict adherence to ACID principles:** While the relational model attempted to very much adhere to the transactional principles of data atomicity, consistency, isolation, and durability (ACID), this very much impacts performance in terms of distributed computing. Relaxing the strictness of adherence to these principles, allow many NoSQL databases to make the trade-off for higher performance.

Of course, these differences between the NoSQL model and relational model vary between each individual database technology's design, and trade-offs are often made depending on the goals and aims for that given database.

In this paper, we will look at the use of MongoDB [3], a popular NoSQL database solution, as a solution for a case study based upon a railway data problem using data from Monash University's Institute of Railway Technology (IRT). An overview of the case study in question will be given in §2. A small overview of the MongoDB database will be given in §3. Implementation and evaluation details will be given in §4 and §5, respectively, before concluding in §6.

[1] https://hadoop.apache.org/
[2] https://www.mysql.com/
[3] https://www.mongodb.org/

2 Case Study Overview

The case study that is being looked at involves a project that has been worked at at the Monash University Institute of Railway technology (IRT). The project involves trains the operate in the Pilbara region of Western Australia, taking ore and minerals from loading points at mines to a specified unloading port. The data that is being dealt with comes from numerous categories of sensors being placed on certain specialised train cars to record data monitoring track and train car conditions [2, 3].

Data currently gets unloaded at sent back in large batches to remote servers once the car completes a trip and pulls back into port [7]. The project currently makes use of a RDBMS solution to manage data storage, however this solution is faced with many problems that currently require painful work-arounds. The most obvious of which involves unreliable data being received from sensors. For example, in the case of a damaged or failed sensor, reliable data cannot be guaranteed to be returned from such a sensor. Hence, data received by the remote server is often inconsistently structured, and thus the data has to go through a series of preprocessing work arounds to fit in within the strict schema that the RDBMS expects. As the only guaranteed consistency in the data that gets received is the timestamp and geocoordinates, the strictly structured relational model is not an appropriate solution.

The Monash IRT team are currently investigating further solutions in the NoSQL big data space, where they intend to replace the current system with an appropriately tested solution. While the scope of that project is much larger than what will be covered in this paper, we will look at the differences in what is possible when attempting to use a NoSQL database for the data storage and management.

For this paper, we will be looking at the possibility of using MongoDB, a popular NoSQL database that does away with the concept of the schemas and tables so commonly found in databases following the relational model. This is done with the expectation of better handling of the given inconsistent data.

3 MongoDB Overview

4 Implementation

5 Evaluation and Discussion

6 Conclusion

References

[1] Codd, E. F. A relational model of data for large shared data banks. *Communications of the ACM 13*, 6 (1970), 377–387.

[2] Darby, M., Alvarez, E., McLeod, J., Tew, G., and Crew, G. The development of an instrumented wagon for continuously monitoring track condition. In *AusRAIL PLUS 2003, 17-19 November 2003, Sydney, NSW, Australia* (2003).

[3] Darby, M., Alvarez, E., McLeod, J., Tew, G., Crew, G., et al. Track condition monitoring: the next generation. In *Proceedings of 9th International Heavy Haul Association Conference* (2005), vol. 1, pp. 1–1.

[4] Indrawan-Santiago, M. Database research: Are we at a crossroad? reflection on nosql. In *Network-Based Information Systems (NBiS), 2012 15th International Conference on* (2012), IEEE, pp. 45–51.

[5] Leavitt, N. Will nosql databases live up to their promise? *Computer 43*, 2 (2010), 12–14.

[6] Padhy, R. P., Patra, M. R., and Satapathy, S. C. Rdbms to nosql: reviewing some next-generation non-relational databases. *International Journal of Advanced Engineering Science and Technologies 11*, 1 (2011), 15–30.

[7] Thomas, S., Hardie, G., and Thompson, C. Taking the guesswork out of speed restriction. In *CORE 2012: Global Perspectives; Conference on railway engineering, 10-12 September 2012, Brisbane, Australia* (2012), Engineers Australia, p. 707.