Cloud Based Sensor Big Data Management: A Literature Review

Anthony J. Christe

December 10, 2016

Abstract

1 Introduction (abstract?)

The exponential increase in volume, variety, velocity, veracity, and value of data has caused us to rethink traditional server architectures when it comes to data acquisition, storage, analysis, quality of data, and governance of data. With the emergence of Internet of Things (IoT) and increasing numbers of ubiquitous mobile sensors such as mobile phones, distributed sensor networks are growing at an unprecedented pace and producing an unprecedented amount of streaming data. It's predicted by the European Comission that IoT devices will number between 50 to 100 billion devices by 2020[10].

The size of sensor networks is quickly growing. BBC Research provides figures that the market share for sensor networks in 2010 was $56 billion and was predicted to be closer to $91 billion by the end of 2016 [14]. Data generated from the IoT are surpassing the compute and memory resources of existing IT infrastructures. [2]. Not only is the size of data rapidly exploding, but data is also becoming more complex. Data from sensor networks is often semi-structured on unstructured with data quality issues.

Sensor networks can benefit from the generally "unlimited resources" of the cloud, namely processing, storage, and network resources. We believe that by leveraging cloud computing, distributed persistence models, and distributed anayltics, its possible to provide a platform that is able to meet the demands of the increasing distributed sensor market and the increasing volume, velocity, variety, and value of data that comes along with that.

This review hopes to summarize the current state of the art surrounding distributed sensor networks and the use of cloud computing as a means for big sensor data acquisition and analysis. We will also briefly discuss privacy concerns relating to sensor data in the cloud.

The rest of this review is as follows: The rest of chapter 1 will review the concepts cloud computing, big data, and sensor networks with motivation as to why these technologies are intertwined. Chapter 2 will compare and contrast various Big Data persistence models. Chapter 3 will provide a review of the current state of the art analytics for Big Data. Chapter 4 will discuss privacy in relation to sensor networks and the cloud. Chapter 5 compares reference implementations of sensor frameworks in the cloud. Chapter 6 discusses future directions and open research questions.

1.1 Big Sensor Data

Sensor data is generally different from Big Data in several ways [2]. One of the most common characteristics of sensor data is the amount of noise in the data. Environmental sensing will always include noise because the data is born analog[8]. Often sensor networks provide data in a large variety of unstructured formats with missing, partial, or conflicting meta-data. Sensor data can also contain a large amount of data redundancy from multiple sensors in similar locations. The problem very quickly becomes a needle-in-the-haystack problem or more aptly put, finding the signal in the noise.

1.2 Cloud Computing

NIST[7] defines cloud computing as “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”

Cloud computing frameworks can provide on-demand availability and scaling of virtual computing resources for storage, processing, and analyzing of very large data sets in real-time or near real-time. This model makes it possible to build applications in the cloud for dealing with Big Data sets such as those produced from large distributed sensor networks.

By using the cloud as a central sink of data for our devices within a sensor network, it’s possible to take advantage of central repositories of information, localized dynamic computing resources, and parallel computations. With the advent of cheap and ubiquitous network connections, it’s becoming easier to do less processing within sensor networks and to offload the work to a distributed set of servers and processes in the cloud[4].

Cloud computing includes both technical and economical advantages as discussed in [1].

On the economical side, computing resources are pay-per-use. Businesses can dynamically increase or decrease the computing resources they are currently leasing. This makes it possible to utilize massive amounts of computing power for short amounts of time and then scale back resources when demand isn’t at its peak. Before cloud computing these same businesses would be required to manage and maintain their own hardware for peak load without the ability to dynamically scale their hardware if the peak load were to increase.

On the technical side, the localization of computing resources provides for a wide variety of benefits including energy efficiency, hardware optimizations, software optimizations, and performance isolation.

1.2.1 Cloud Computing Service Models

When discussing cloud computing, it’s useful to understand the service layers that cloud computing provide. The three service models are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

At the lowest level is IaaS which provides virtual machines that users have the ability to manage. Users can install operating systems on these virtual machines and interact with this virtual machines as if they were local servers. Users do not need to worry about the underlying hardware or network infrastructures at this level. This level of interaction is lower-level than we would like to examine for our survey.

At the next lowest level is the PaaS layer which provides infrastructure for users to deploy custom software to. At this level, users do not have control over their operating systems or virtual machines and these are managed by cloud.

At the highest level is the SaaS layer. This layer provides for pre-built applications running in the cloud that users interact with.

This review aims to look at software and architecture somewhere between and including the PaaS and SaaS layers. There are many examples in the literature at optimizing data centers for cloud architecture and the IaaS service level (i.e. [5], [13], [11], [9], [6]). These discussions are outside of the scope of this review as we want to focus on software framework for sensor data collection in the cloud.

1.3 Mobile Cloud Computing

Mobile devices such as smartphones make fantastic distributed sensors for temporalspatial data. Not only do they carry a wide array of sensors on-board (microphones, barometers, accelerometers, GPS, compasses, cameras, clocks), but they generally have multiple modes of offloading data (WiFi, bluetooth, cellular, SD cards), and support some pre-processing on-board.

1.4 Scientific Data Infrastructure

1.5 Applications of Distributed Sensor Networks

Zaslaveky et al. [14] cites several examples of distributed sensor networks in-the-wild including: a real-time greenhouse gas detection network deployed across California, real-time structural monitoring such as the St. Anthony Falls Bridge sensor network in Minneapolis, distributed radiation detection in Fukushima, real-time parking space inventory in San Francisco

Software for cloud environments include distributed fault-tolerant databases and distributed parallel algorithms for computer clusters.

One area that shows a lot of promise for distributed sensor networks with centralized management is smart grids. The smart grid is an collection of technologies aiming to advance the electrical grid into the future with respect to intelligent energy distribution and integration of renewables. Electrical grids can benefit by using a large distributed sensor network to collect power consumption, production, and quality information and use that information to control power production and consumption in real-time.

Many times, the sensor nodes in smart grids lack powerful local computation abilities, but generally have network connections and sensing capabilities. This makes the cloud a perfect sink of information for analyzing complex power trends from a large scale distributed sensor network for smart grids[1].

2 Big Data Persistence Models

Traditional storage methods for meta-data and related products has traditionally made use of the filesystem and relational database systems (RDMS).

Big Sensor Data by its nature can be structured, unstructured, large, diverse, noisy, etc. Many of the properties of BSD do not fit nicely into the structured world of traditional RDMSs.

In-order to meet the needs of Big Sensor Data and distributed sensor networks, we look to the ever growing field of NoSQL (not only SQL) and related Big Data storage models. There are multiple types of data models with different use cases. We will review the current players in this field and with a focus on how these technologies could aid in managing Big Sensor Data in the cloud.

According to Song et al.[3] an ideal NoSQL data model strives for "high concurrency, low latency, efficient storage, high scalability, high availability, reduced management and operation costs." The challenges of realizing an ideal NoSQL data model however are lie in three main areas[2]: consistency, availability, and partition tolerance.

Consistency issues occur when data is stored in a distributed manner with multiple copies. In situations of server failure (or with systems that support different consistency models), situations can arise where multiple copies of the same resource contain different contents.

Vogels and Wener[12] explain the main forms of consistency. Assume a record is being updated across multiple servers. With strong consistency, any access of that resource after the update will return the updated result. With weak consistency, subsequent access of that resource is not guaranteed to return the updated result if that access is within a certain "inconsistency window".

We also believe, ease of use, maturity of the product, and community (or commercial) support should also factor into the comparisons.

With the above factors in mind, we big categorizing and analyzing.

2.1 Data Models

2.1.1 Key-Value

2.1.2 Document

2.1.3 Graph

2.1.4 Wide Column Store

2.2 Indexing

3 Big Data Analytics

4 Privacy

5 Reference Implementations

6 Open Research Question

References

[1] Alessio Botta, Walter de Donato, Valerio Persico, and Antonio Pescap. Integration of cloud computing and internet of things: A survey. 56:684–700.

[2] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. 19(2):171–209.

[3] A. Jin, C. Cheng, F. Ren, and S. Song. An index model of global subdivision in cloud computing environment. In *2011 19th International Conference on Geoinformatics*, pages 1–5, June 2011.

[4] Supun Kamburugamuve, Leif Christiansen, and Geoffrey Fox. A framework for real time processing of sensor data in the cloud. 2015:1–11.

[5] Niloofar Khanghahi, Ramin Nasiri, and Mahsa Razavi Davoudi. A new approach towards integrated cloud computing architecture. 4(1):24–34.

[6] Praveenkumar Khethavath, Johnson P. Thomas, and Eric Chan-tin. Towards an efficient distributed cloud computing architecture.

[7] Peter Mell and Tim Grance. The nist definition of cloud computing. 2011.

[8] President's Council of Advisors on Science and author Technology (U.S.). *Report to the President, big data and privacy : a technology perspective.* Washington, District of Columbia : Executive Office of the President, President's Council of Advisors on Science and Technology, 2014. Includes bibliographical references.

[9] Han Qi, Muhammad Shiraz, Abdullah Gani, Md Whaiduzzaman, and Suleman Khan. Sierpinski triangle based data center architecture in cloud computing. 69(2):887–907.

[10] D. Reed, J. R. Larus, and D. Gannon. Imagining the future: Thoughts on computing. *Computer*, 45(1):25–30, Jan 2012.

[11] Zhiming Shen, Sethuraman Subbiah, Xiaohui Gu, and John Wilkes. Cloudscale: elastic resource scaling for multi-tenant cloud systems. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, page 5. ACM.

[12] Werner Vogels. Eventually consistent. *Queue*, 6(6):14–19, 2008.

[13] Bin Wang, Zhengwei Qi, Ruhui Ma, Haibing Guan, and Athanasios V. Vasilakos. A survey on data center networking for cloud computing. 91:528–547.

[14] Arkady Zaslavsky, Charith Perera, and Dimitrios Georgakopoulos. Sensing as a service and big data.