2014

Commercial Data Intensive Cloud Computing Architecture: A Decision Support Framework

Shane Archiquette
Hitachi Data Systems Colorado Technical University, s.archiquette1@my.cs.coloradotech.edu

D. Lance Revenaugh
Montana Tech University Colorado Technical University, lrevenaugh@mtech.edu

13R. Commercial Data Intensive Cloud Computing Architecture: A Decision Support Framework

Shane Archiquette
Hitachi Data Systems
Colorado Technical University
s.archiquette1@my.cs.coloradotech.edu

D. Lance Revenaugh, PhD
Montana Tech University
Colorado Technical University
LRevenaugh@mtech.edu

Abstract

Scientific and commercial computing is undergoing an immense change with increasing demands for processing of large-scale datasets for a variety of needs such as simulation, modeling, and calculations of multivariate equations. Computation has largely been used as a method for achieving the results and is now shifting to a data intensive computing model to accomplish some of the most demanding scientific challenges in existence. Current data and storage architectures are not sufficient to provide for Petascale and Exascale data processing and analysis which will require new ways of data access at multi-terabyte per second speeds. An architectural framework is proposed for data intensive cloud computing called Datalanx.

Keywords

Data Intensive, Cloud Computing, Architecture, High Performance Computing (HPC), Big Data Analytics, Data Science, Communications, Media & Entertainment, Healthcare & Life Sciences, Oil & Gas Exploration

1. Introduction

The motivation behind this research is to investigate the creation of a cloud based data intensive architectural framework to serve various applications and industry verticals that currently all must use separate and completely different architectures to facilitate proper data processing of their information that is interesting and useful for the mission of research or business. Some examples of these data intensive industries are communications, media & entertainment, healthcare and life sciences, and oil and gas exploration. The emergence of Big Data has helped create new methods and architectures of dealing with large datasets primarily to perform different levels of business intelligence and analytics on. These analytic architectures are not well suited to actual data intensive computing (Hazelhurst 2008) workloads for multiple reasons; however, the basis of the architectural principles can be useful. As this research is focused on building a foundation for developing a cloud based data intensive architectural framework with greater bandwidth to storage resources, the following questions:

> **Research Question 1:** *Can the creation of a guiding architecture provide a framework for cloud architects to construct dynamically scalable environments to handle large scale fixed and variable datasets?*

Research Question 2: *What is the impact of a scalable Data Intensive architecture in a Cloud computing framework for commercial data intensive applications?*

Research Question 3: *What are the benefits and drawbacks of a data intensive architectural framework for scientific and commercial communities relative to decision support?*

Research Question 4: *As the requirements for higher resolution output of various datasets increase, can a common platform provide an elastically scalable solution?*

2. Data Intensive Computing Defined

When a supercomputer is unable to get its data fast enough, there is a need to provide faster access to storage than what has been previously defined. There are new and emerging scientific applications (Soltesz 2007) that require large scale datasets to be pushed in and out supercomputers to allow for the analysis to be completed with success. Data Intensive Computing provides the ability for this requirement to function; however, it is not without many problems in storage bandwidth to the supercomputers that are architected for this type of computing. Data Intensive Computing provides an architecture where access to storage is a primary consideration and the computational power is a secondary consideration. Common Cloud Computing platforms have very similar I/O patterns that allow Horizontal IT application uses to be largely shared on the same computing platform infrastructure. The five types of I/O:

1. Small block random ← horizontal IT primary type
2. Medium block random
3. Large block sequential
4. Very large block sequential
5. Mixed block random and sequential

Vertical Industries require different types of I/O patterns, which also place specialized requirements on the infrastructure that is implemented in separate and optimized datacenters that are primarily single purpose and single organization.
Figure 1 depicts some of the common commercial data intensive applications and the associated I/O profiles.

Data intensive computing is a branch discipline of High Performance Computing as a result of greater dependence on larger datasets and faster access to data storage resource infrastructure. Scientific and commercial computing are actively expanding in to emerging technology areas such as hybrid SMP/parallel cluster computing, high performance cloud computing, and data intensive cloud computing, new architectures need to be considered and implemented for these new areas to flourish (Osamu 2002). Currently commercial data intensive applications are run on separate infrastructure architectures due to differential block input/output requirements.

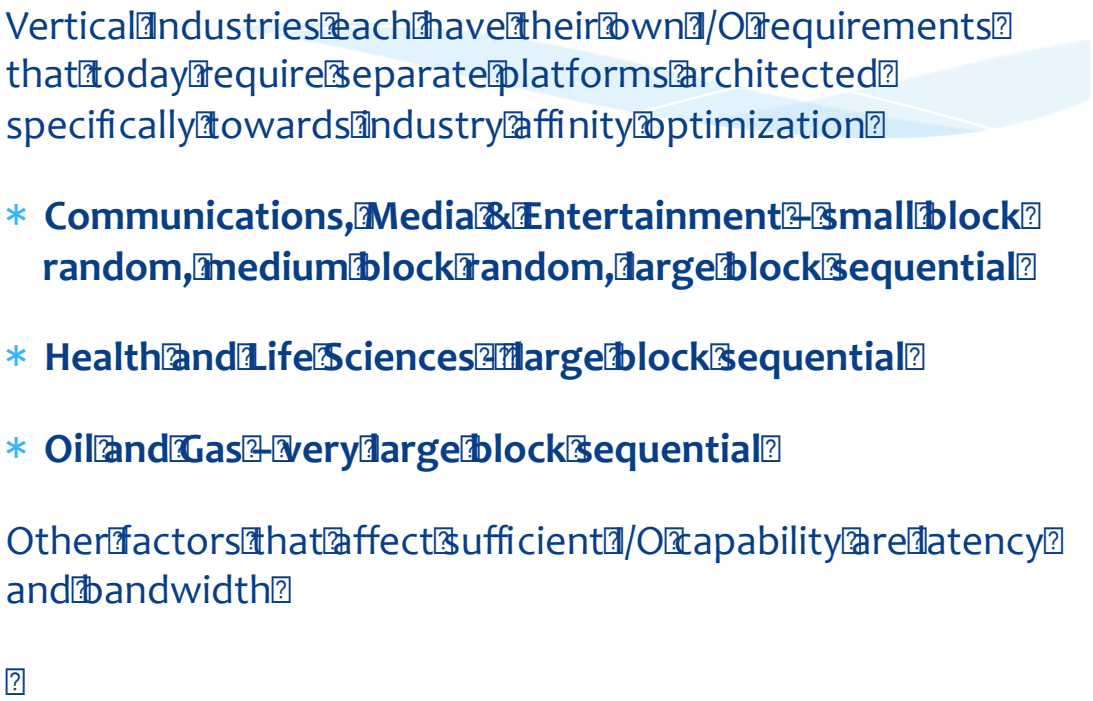

Figure 1 – Data Intensive Computing Applications

3. Big Data as an Emerging Science Discipline

Every industry in existence has ever-increasing demands for data storage with no end in sight. Some industries have such significant data stores they lack the tools to be able to analyze and make any decisions on what to do with the data. This has created a post cloud computing area called Big Data, which is a variant of cloud computing due to the type of architecture that is designed to process large (Terabyte) to extremely large (Petabyte) datasets of information. Every organization sits at the intersection of big data, predictive analytics trying to leverage multiple data sources. For example, a large department store handles data from click stream receive from the tens of millions of consumers every month, data from all of its distribution channels, search, Twitter, Facebook, and data from the millions of products that offers. Handling big data requires machine learning, data mining and predictive analytics algorithms to determine, in close to real time, how to best acquire customers across all channels. Some label these petabyte scale, big data, and web scale. Big data does not need to be defined in terms of TB, PB; it should be defined, as the data is too big to fit in main memory. Big data has created revolutionary breakthroughs in commerce, science, and society. Figure 2 displays 5 important attributes of big data, which are volume, variety, and velocity value, veracity and a sixth could be

variability, especially with the combination of vertically disparate datasets on the same infrastructure.

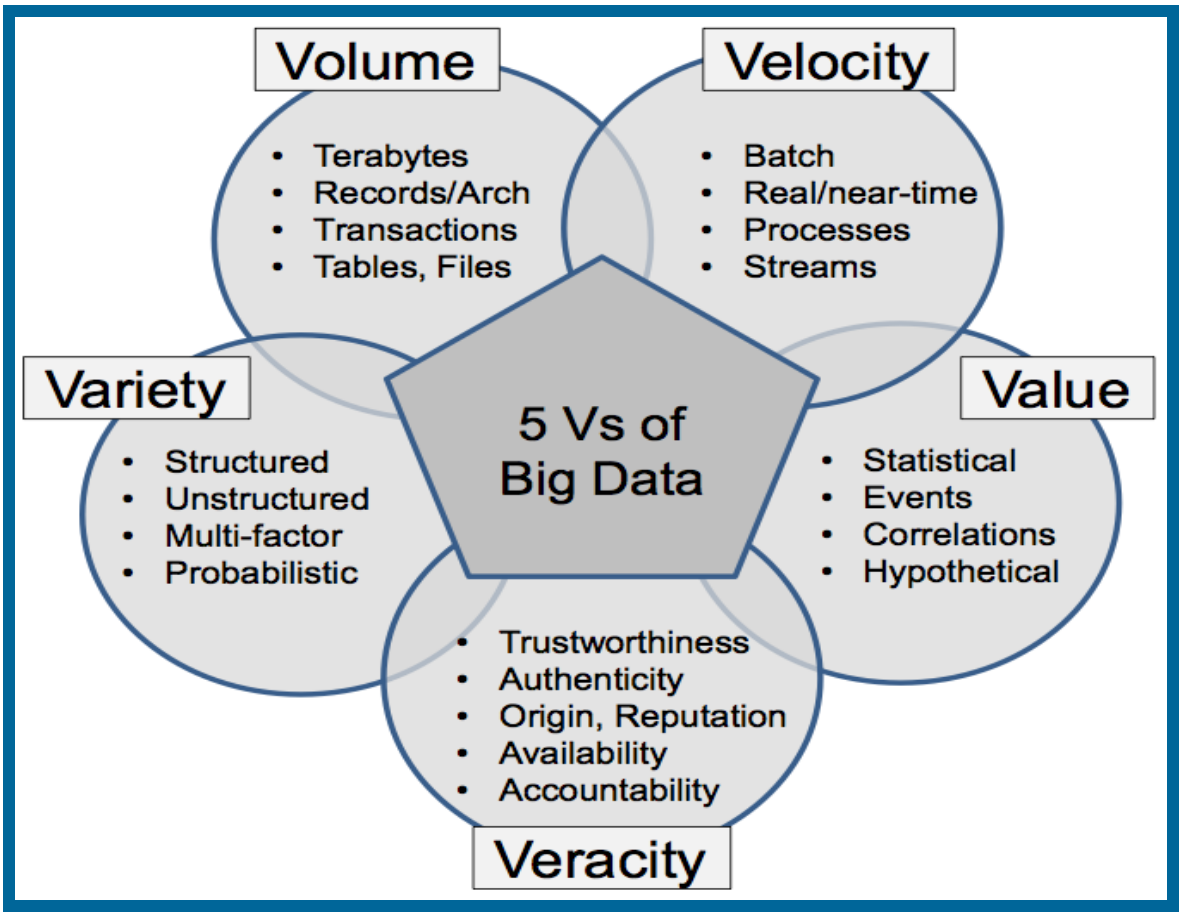

Figure 2: Five-Vs of Big Data

4. Big Data Analytics

Big Data Analytics requires the ability to leverage hundreds to thousands of compute nodes scanning metadata intensive indexes for correlations of valuable data to individuals or companies doing research or identifying service or product areas that can be monetized. Many companies are literally sitting on mountains of valuable information that could be very difficult to use and access the information properly (Vinayak 2011). Understanding the environment requires collecting and analyzing data from data sources everywhere, from Web 2.0 and user-generated content to large scientific projects, and other data points. Data analysis is the process of inspecting data in order to extract meaningful information. Figure 3 indicates the framework for Big Data and Big Data Analytics.

Figure 3: Big Data Analytics Framework

5. Datalanx Framework

The organization of an architectural framework for three distinct forms of data intensive computing (media & entertainment, oil & gas, healthcare and life sciences in to a useful working model is the goal of the research and presents several challenges in being able to validate and construct. The basic premise of the Datalanx framework is a distributed parallel hardware architecture that leverages a block level optimization software file-system layer. Figure 4 presents a high level diagram of the proposed framework.

6. Methodology

Due to the large-scale nature of the proposed framework and architecture, field-testing and validation is both time and cost prohibitive. To establish validity and agreement of the research, leveraging an Inter-Rater Reliability (IRR) study will be necessary with survey techniques and data collection, and statistical analysis (Li 2012). The study will cover a series of experts in the media & entertainment, oil and gas, and healthcare and life sciences industries and use a series of well-formed survey questions to determine validity of the research constructs and utilize IRR coefficients and statistical inference for determining the findings and ultimately the anticipated validity of the research.

7. Current Research Status and Summary

The majority of the question sets formation and industry expert identification has been done for the inter-rater reliability experiment and study. There will be a six to 10 rater inter-rater reliability experiment two question sets per major industry (communications, media & entertainment, oil and gas, and healthcare and life sciences) with assessment of various levels of agreement and validity of the study. This research covers real world architectural issues at Hitachi Data Systems and other storage companies in dealing with disparate industry storage performance requirements across different platform configurations.

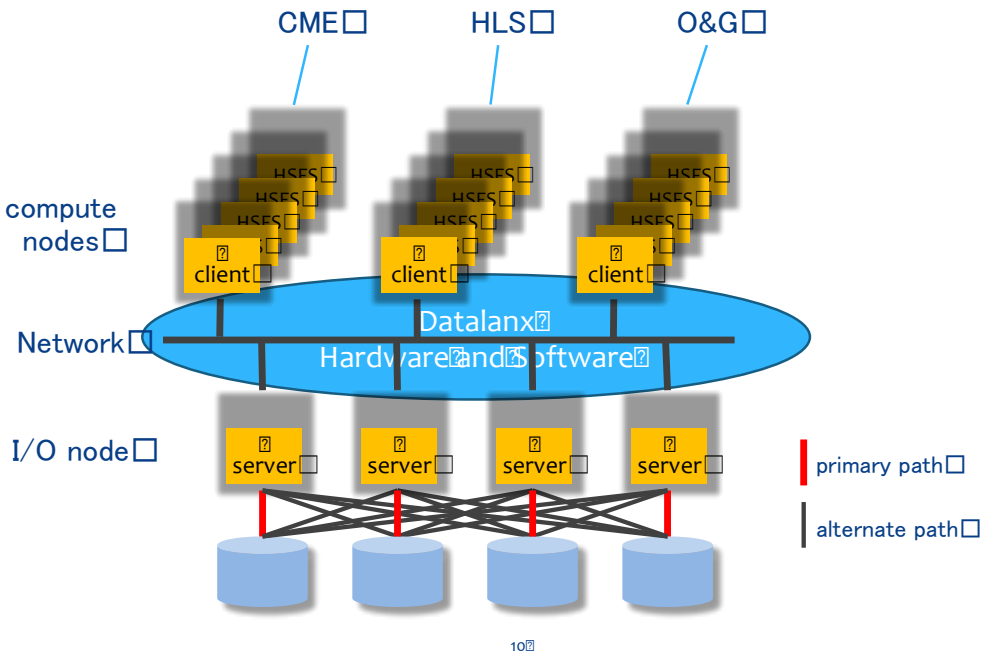

Figure 4: DIC Architecture Proposal

References

Bader, D. A. (2008). Petascale Computing: Algorithms and Applications, Chapman & Hall/CRC.

Borkar, Vinayak (2011) "Hyracks: A Flexible and Extensible Foundation for Data-intensive Computing." Data Engineering (ICDE), 2011 IEEE 27th International Conference on IEEE.

Buyya, R., Abramson, D., & Giddy, J. (2000). An Economy-driven Resource Management Architecture for Global Computational Power Grids. Proceedings of the 2000 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000).

Ekanayake, Judy Qiu Saliya (2013) "Data Intensive Computing for Bioinformatics." Bioinformatics: Concepts, Methodologies, Tools, and Applications.

Gwet, Kilem Li. (2012) Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters. Advanced Analytics Press.

Hazelhurst, S. (2008). Scientific Computing Using Virtual High-performance Computing: A Case Study Using the Amazon Elastic Computing Cloud, ACM New York, NY, USA.

Kouzes, Richard T. (2009) "The Changing Paradigm of Data-intensive Computing." Computer 42.1, 26-34.

Lysne, O., S. A. Reinemo (2008). "Interconnection Networks: Architectural Challenges for Utility Computing Data Centers." Computer **41**(9): 62-69.

Soltesz, S., H. Pötzl (2007). Container-based Operating System Virtualization: A Scalable, High-performance Alternative to Hypervisors, ACM New York, NY, USA.

Tatebe, Osamu (2002) "Grid Datafarm Architecture for Petascale Data Intensive Computing." Cluster Computing and the Grid, 2nd IEEE/ACM International Symposium on IEEE.

Youseff, L., M. Butrico (2008). Toward a Unified Ontology of Cloud Computing.

Zhao, Dongfang, and Ioan Raicu (2012) "Distributed File Systems for Exascale Computing." Doctoral Showcase, SC 12.

Zhou, S., Van Aartsen, B. H. (2008). "A Lightweight Scalable I/O utility for Optimizing High-End Computing Applications" Proceedings of the 22nd IEEE Parallel and Distributed Processing Symposium, Miami, April 14-18, 2008.