International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 5 Issue 2 March 2015
International Manuscript ID : 2249054XV5I2032015-05

BIG DATA ANALYSIS ON THE CLOUD USING IN MEMORY COMPUTING

Neha Saxena
Research Scholar
IITT College of Engineering
Pojewal, Punjab, India

Dr. Pawan Kumar
Head, Computer Science and Engineering
IITT College of Engineering
Pojewal, Punjab, India

ABSTRACT

Big data analytics and cloud computing are two technology initiatives based on which many technological advancements are happening around the globe. Big Data is often seen as providing valuable insights that can create competitive edge and make way for further innovations. Cloud computing has the means to enhance business agility and productivity enabling greater efficiency in big data analytics while reducing the cost associated with it. As these technologies continue to evolve, organizations are moving

beyond questions of what and how to store big data to addressing how to derive meaningful analytics that respond to real business needs. True potential of big data analytics can only be achieved if the analytics is based on real time data, which can be achieved effortlessly using In Memory Computing. By using in-memory computing model on the cloud space one can bring all the advantages of cloud computing to data exploration, analysis and sharing. This paper outlines a review of big data analytics using in memory computing on a cloud network and provides a solution implementation using JAVA along with its results and analysis.

INTRODUCTION

Big data refers to huge data sets generated over time by a connected ecosystem, it is heterogeneous and can be made available in many formats. The real value of big data is in the insights it produces when analyzed. Big data analytics is a set of advanced technology frameworks designed to work with large volumes of heterogeneous data. It uses sophisticated quantitative methods such as artificial intelligence to explore the data and to discover interrelationships and patterns.

With the potential for so much data to reveal insights that can boost competitiveness, there is a need to find new approaches to processing, managing, and analyzing data. Meanwhile the technology for processing real-time is growing steadily into a high degree of maturity supporting predictive analytics.

The hybrid cloud model enables the use of on-demand storage space and computing power via public cloud services for analytics processes also providing for scaling up when needed. Using cloud infrastructure to provide ***D***ata ***A***nalytics as a ***S***ervice (DAaaS), organizations can address user needs across the full range of analytics requirements from data delivery and management to data usage. By developing an

end-to-end cloud-based big data strategy, one can define an insight framework and optimize the total value of enterprise data.

At a basic level, almost all Big Data problems are about time. The bigger the data volume and the faster the data streams into the enterprise application, the longer it takes for traditional analytics and data management software to turn this data into actionable information. In-memory computing can be used to greatly solve this problem. The main architecture of in-memory computing moves the data as close as possible to processors. Traditionally, to analyze any data, a look-up must be performed on the appropriate database, which is bounded by the speed of the disk drive it is stored on. This process is further slowed by any latency involved in transferring this data through an input / output (I/O) connection between the storage device and server.

By using applications designed to take full advantage of this technology in available computing power, organizations can now realize the dream of real-time or near-future-time analysis of core business data.

LITERATURE SURVEY

Because of significant business interest on both cloud and in memory computing, both the topics have been feature of latest research in computer science and its related technology.

A study by (*TATA Consultancy Services, 2012*) [1] found that telecommunication, travel, and finance spent the most on Big Data investment, while energy and utility industries expect the highest return from Big Data investments. Studies indicate that almost all corporations can benefit from Big Data and Analytics, as it can revolutionize business operations across the board.

(*IBM, 2012*) [2] released a study concluding that Big Data represents an enormous opportunity for marketers. Big Data can drive decisions by accurately delivering the right message to the right person at the right time for the right price. The travel industry has recognized the importance of Big Data and Analytics in transforming its industry. (*Amadeus IT Group, 2013*) [3] concluded in its study that Big Data and Analytics provides significant benefits for travel companies by offering better decision support, new products and services, and better customer relationships.

Another dimension of the Big Data definition involves technology. Big Data is not only large and complex, but it requires innovative technology to analyze and process. (*NSIT, 2012*) [4], the National Institute of Standard and Technology (NIST) Big Data workgroup proposed the following definition of Big Data that emphasizes application of new technology:

"***Big Data exceed the capacity or capability of current or conventional methods and systems, and enable novel approaches to frontier questions previously inaccessible or impractical using current or conventional methods.***"

In order to overcome the volume issue, big data requires technologies that store vast amount of data in a scalable fashion and provide distributed approaches to querying or finding that data. The cloud computing model is a perfect match for big data since it provides unlimited resources on-demand. With cloud, organizations no longer need to purchase hardware or high cost software.

On the other hand generally data intensive or data driven applications generate and process massive data sets usually stored in the cloud. These applications have large

data processing requirements and are engineered with custom algorithms to run on scalable infrastructure as presented by (*Venugopal, Desikan and Ganesan, 2011*) [5].

The characteristics of cloud computing has provided for both big data acquisition, and software data processing strategies. (*Gartner, 2012*) [6] has estimated that 50% of data will be stored on the cloud by 2016 (*Schouten, 2012*) [7]. The availability of cloud based solutions has dramatically lowered the cost of storage, amplified by the use of commodity hardware even on a "pay as-you-go" basis that is directed to effectively and timely processing large data sets.

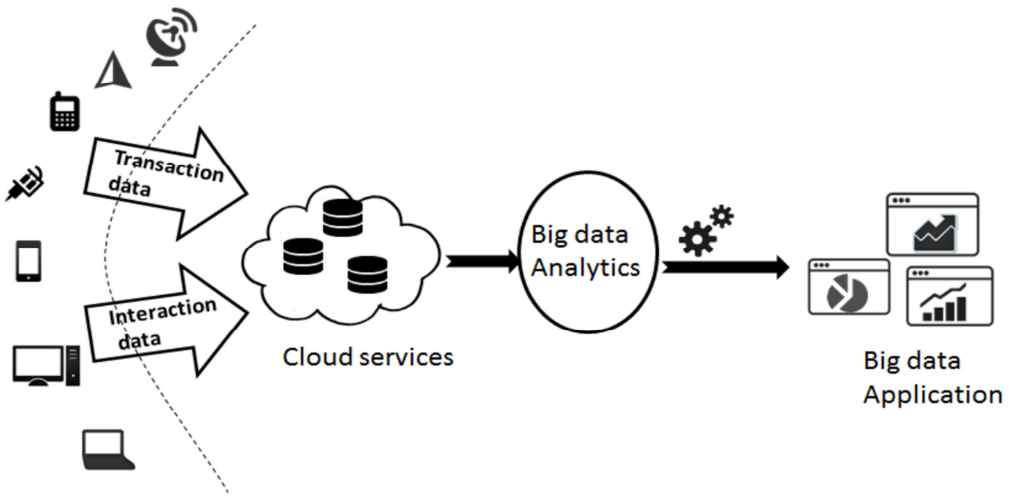


Figure 1 - Big Data visualization on the cloud

Big data and its analytics is a fairly broad and complex subject. So is Cloud Computing, as an enabler for the analytics delivery. Existing literature covers drivers that point out to Cloud being the 'next generation' option for in-sighting and analyzing data. A ringside overview of these drivers is provided in this section.

Data storage and its associated costs in cloud are discussed by (*Todd Weller, 2013*) [8] in an analyst interview with Wall Street. Demand for hosted storage and its implications are dealt with (*Clive Longbottom, 2012*) [9] in his insightful article on how to make sense of the Big Data Universe.

Storage as is on the cloud has undergone transformation over the years and has seen regular advancements in its functioning, which have helped in applications making their data available on a cloud network. The Storage Network Industry Association (SNIA) proposes Cloud Storage Initiative (CSI) to adopt Cloud Data Management Interface (CDMI) standard as cloud service standard.

There have been several researches in the areas of improving the performance of storage infrastructure on a cloud network and have come up with various strategies to efficiently manage storage.

One such strategy suggested by (Yunhong Gu, 2010) [10] in his research paper "***A high performance wide area community data storage and sharing system***" formulates the idea of sectors which enables application users to work with large datasets stored over multiple distributed nodes as if the files were on their local disk. Users do not need to locate data, manage data across multiple nodes, back up data, and manage the addition of new nodes or the deletion of existing nodes to the system.

Another strategy suggested in the white paper Harnessing 'Storage Clouds' for high performance content delivery by (*James Broberg. 2009*) [11] introduces MetaCDN, a system that exploits 'Storage Cloud' resources, creating an integrated overlay network that provides a low cost, high performance CDN for content creators. MetaCDN removes the complexity of dealing with multiple storage providers, by intelligently

matching and placing users' content onto one or many storage providers based on their quality of service, coverage and budget preferences. MetaCDN makes it trivial for content creators and consumers to harness the performance and coverage of numerous 'Storage Clouds' by providing a single united namespace that makes it easy to integrate into origin websites, and is transparent for end-users.

With the advantages provided by IMDB's in terms of cost and performance, (*Francesco Pagano, 2012*) [12] in his paper ***Using In-Memory Encrypted Databases on the Cloud*** has used traditional encryption on in memory databases for creating a sense of data security on a cloud storage there by offering In Memory Databases as a safe and secure choice for applications to move on to cloud network and use In Memory databases as a backend for themselves. Thereby giving them faster response times and enabling the end users to get a hold of real time analytical data.

As it can be seen, the value adds that applications get by using in memory computing is huge and this advantage can be easily made use of by leveraging this technology for the data on a cloud network.

There is not yet a global standard specification and general architecture to cloud computing and cloud storage. The main idea is to integrated and improvement the current architecture, distribution mode, application area, etc. to construct a low cost, fault torrent, reliable, scalable, high performance and fair cloud storage alliance system

OVERVIEW: IN-MEMORY COMPUTING

Increasingly complex business decision models are dependent on a very speedy access to and computation of massive data stores. Analysis into business operations often demands data sizes that are beyond the capabilities of traditional disk-based

systems to process in real time, which can limit access to benefits such as the following:

- Efficiency provided by the ability to respond in real time to the changing needs of the business
- Flexibility based on insight that accurately directs quick action
- Empowerment of business users to make and act on smart decisions

In-memory computing is, as the name indicates, a way of keeping data close to computation, instead of the standard method of storing data in another machine and sending it back and forth to the computer servers. In-memory theoretically means that computation of big data sets can take place a lot faster. It matters for an increasingly large number of use cases, as more business takes place on the Internet.

In-memory computing technology is an evolution through various hardware and software technology innovations. Hardware innovations include blade servers and CPU with multi core architecture and memory capacities measured in terabytes for massive parallel scaling. Software innovations include an in-memory database with highly compressible row and column storage specifically designed to maximize in-memory computing technology. Parallel processing takes place in the database layer rather than in the application layer as we know it from the client-server architecture.

In-Memory databases are very cost efficient as compared to traditional relational database technology due to:

- ***leaner hardware and less system space required***, as mixed workloads of analytics, operations, and performance management are handled within a single system, which also reduces redundant data storage
- ***Reduced extract, transform, and load (etl) processes*** between systems and fewer prebuilt reports, reducing the support effort required to run the software

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 5 Issue 2 March 2015
International Manuscript ID : 2249054XV5I2032015-05

TECHNOLOGY STACK OVERVIEW

To achieve our goal we have utilized the following open platform/technologies to build and system and highlight the power of Data Analytics as a Service (***DAaaS***) and focus on its future.

To fulfil our platform requirements we have using an EC2 windows based instance on Amazon Web Services (AWS) public cloud. Public cloud is a service, which is provided by a third party. The analytics application is built upon JAVA, which is a general-purpose computer programming language that is concurrent, class-based, object-oriented and specifically designed to have as few implementation dependencies as possible.

As a web based framework we have used Spring MVC upon which the application is built. The Spring Web model-view-controller (MVC) framework is designed around a Dispatcher Servlet that dispatches requests to handlers, with configurable handler mappings, view resolution, locale, time zone and theme resolution as well as support for uploading files.

In Spring Web MVC you can use any object as a command or form-backing object; you do not need to implement a framework-specific interface or base class. Spring's data binding is highly flexible: for example, it treats type mismatches as validation errors that can be evaluated by the application, not as system errors. Thus you need not duplicate your business objects' properties as simple, un-typed strings in your form objects simply to handle invalid submissions, or to convert the Strings properly. Instead, it is often preferable to bind directly to your business objects.

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 5 Issue 2 March 2015
International Manuscript ID : 2249054XV5I2032015-05

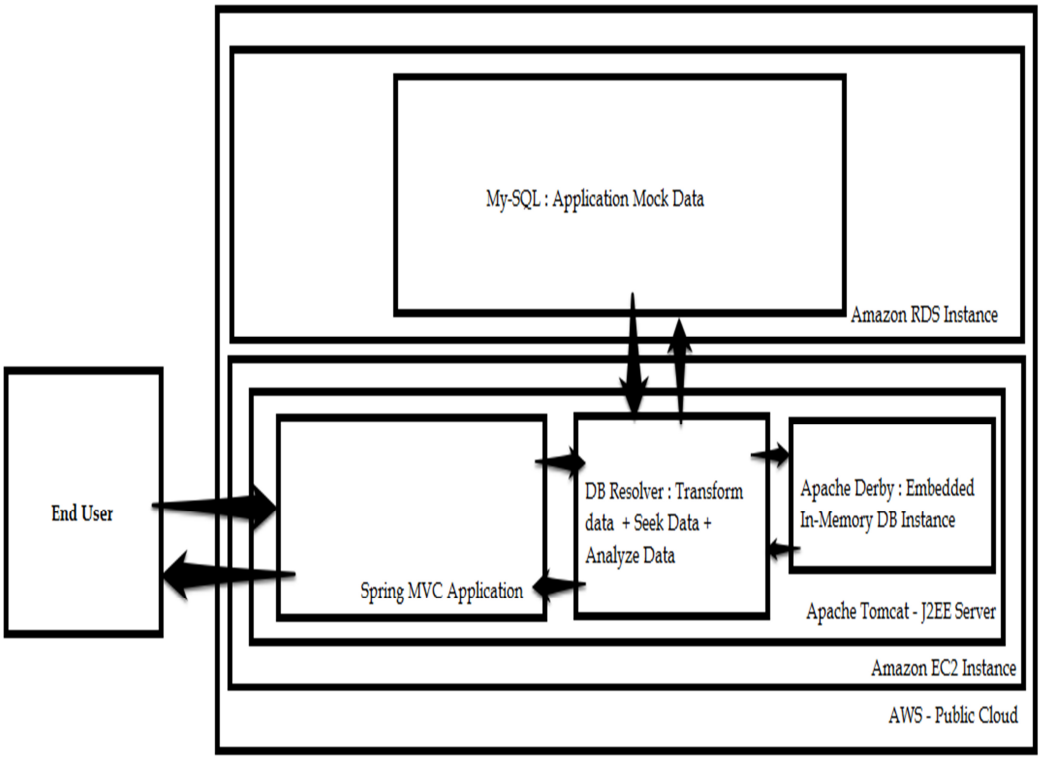


Figure 2 - System Architecture

The database layer forms the core of the analytics application built as part of this thesis. The main aim of the application here is to

- ***Build a transformation layer*** where-in data from any database can be migrated to any other database, which in our case is an in-memory database.
- Show from an analytics point of view the ***comparative analysis*** of the fetch records from a traditional database and an in-memory database

For this purpose we have used the services of two databases

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 5 Issue 2 March 2015
International Manuscript ID : 2249054XV5I2032015-05

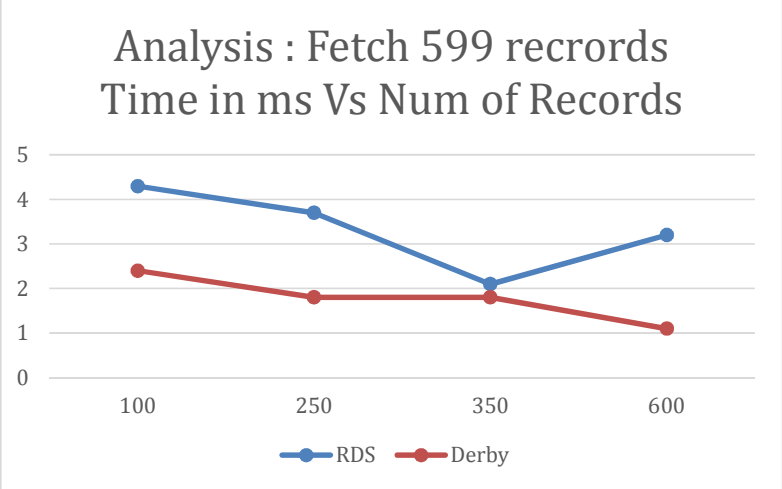
MySQL as a traditional RDBMS system to act as a data store for traditional application data. The Amazon public cloud provides the services of such databases using the Amazon RDS interface. The application can access this database to fetch/save data using a remote connection provide by AWS public cloud

Derby as a testing In-Memory database system to provide for analysis on how the use of in-memory database in embedded on a cloud server can greatly increase the speed at which analytics can be done on big-data.

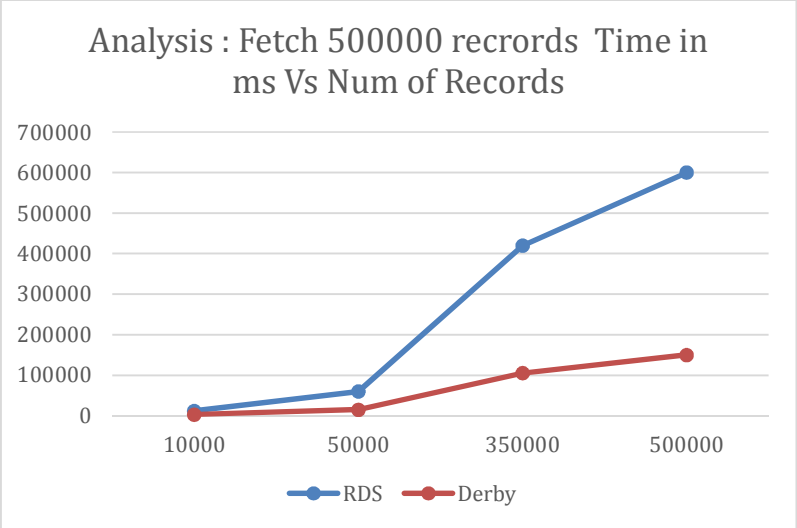
RESULT EVALUATION

Response time can be defined as a time that is need for calculations of underlying data and partly including the time needed for visualization of data. For instance, in case when data analysis graph loads on the front end, the response time includes time required by the java script to load the graph on the html page.

Result Summary : Graphical Analysis on AWS Public Cloud

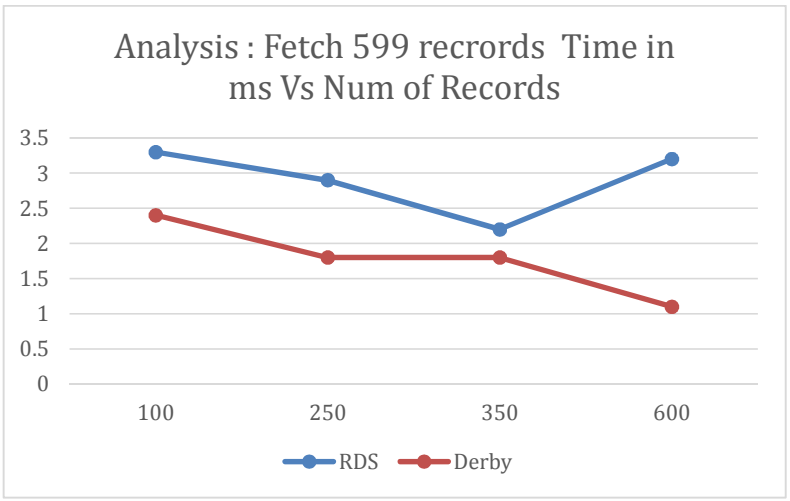


International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 5 Issue 2 March 2015
International Manuscript ID : 2249054XV5I2032015-05

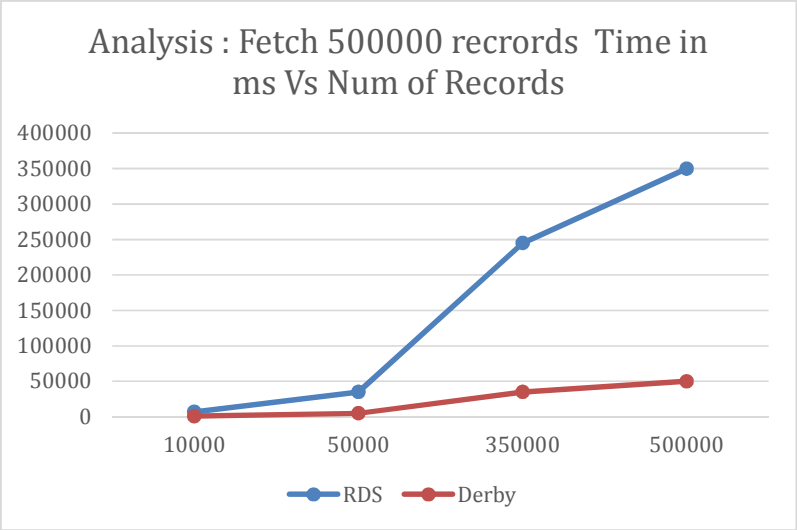


The values of response time in the results table imply that basic calculation (fetching, evaluation of data) have been processed within three second for all datasets. This data can be treated as base for comparison with other groups. Such as, explorative analysis, or better multidimensional analysis that typically involves the on-the-fly calculation of aggregated values is fast enough to compete with basic calculation.

Result Summary : Graphical Analysis on local setup



International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 5 Issue 2 March 2015
International Manuscript ID : 2249054XV5I2032015-05



A slightly bigger response time is noticeable with the largest dataset containing billion of records. Although the final response times with aggregation and computational analysis may be much longer for the largest dataset due to the complexity, but for user the analytical operation is still reasonably fast.

CONCLUSION

However hyped it may be currently Big Data is certainly a business changing trend, as the facts are evident: the data explosion is real and some companies have shown clear competitive advantage by creating and implementing new analytic capabilities over previously unused data. But getting this kind of capability may be not easy for some companies. Here the flexibility that Cloud delivery models bring can simplify adoption for some companies and even those that could have the resources to implement it internally can obtain significant cost advantages with DAaaS.

Data Analytics as a Service, the model proposed in this research can be applied to multiple use cases and industries even as the analytic approaches to different scenarios may vary considerably. Beyond that DAaaS puts analytics as a first-level element component in a new vision of Enterprise Computing, which makes extensive usage of the advantages of Cloud technologies.

REFERENCES

1. TATA Consultancy Services. (2013), "**The Emerging Big Returns on Big Data**" http://www.tcs.com/big-data-study/Pages/default.aspx.
2. IBM. (2012), "**Moving up the digital marketing maturity with big data analysis. A Thought Leadership White Paper**" http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/Moving_up_digital_marketing_maturity-IMW14658USEN.pdf.
3. Thomas H. Davenport. (2013), "**At the Big Data Crossroads: turning towards a smarter travel experience.**" http://www.amadeus.com/web/binaries/blobs/60/112/Amadeus_Big_Data.pdf.
4. NSIT. (2012), http://bigdatawg.nist.gov/home.php
5. Venugopal, S., Desikan, S. and Ganesan, K. (2011), "**Effective Migration of Enterprise Applications in Multicore Cloud. 2011 Fourth IEEE International Conference on In Utility and Cloud Computing**"
6. Bu M.A. Beyer and D. Laney, Gartner. (2012 - 2013), "**The Importance of big data: A Definition**"
7. Schouten, E. (2012), "**Big Data 'as a Service'. The Atlantic.**" http://www.theatlantic.com/sponsored/ibm-cloud-rescue/archive/2012/09/big-data-as-aservice/262461/
8. Todd Weller. (2013), "**Cloud Computing a Positive Driver for Data Centre Companies**".

9. Clive Longbottom. (2012), "**How to make sense of the Big Data Universe**" http://www.computerweekly.com/feature/How-to-make-sense-of-the-big-data-universe
10. Yunhong Gu. (2010), "**Sector: A high performance wide area community data storage and sharing system**"
11. James Broberg and Rajkumar Buyya. (2009), "**MetaCDN: Harnessing 'Storage Clouds' for high performance content delivery**"
12. Francesco Pagano. (2012), "**Using In-Memory Encrypted Databases on the Cloud**"