

Distributed Processing and Cloud Computing in Business Analytics

Processing data faster and cheaper

Monday, February 15, 2010

ClicData SARL
Telmo Silva



54/58 rue Nationale
F-59800 Lille
France

www.clicdata.com

This document contains confidential material proprietary to ClicData SARL. This information and the ideas herein may not be disclosed to anyone outside of the recipient(s) of the document.

Ce document contient des informations confidentielles qui appartiennent à ClicData SARL. Les informations et idées qu'il contient ne doivent être divulguées à l'extérieur des personnes auxquelles le document est distribué.

TABLE OF CONTENTS

Preface.....	2
Purpose of This Paper	3
Distributed Processing.....	3
Infinis and Distributed Processing	4
Risk Factors	5
Cloud Computing.....	6
Cloud Computing in Infinis.....	6
Risk Factors	7
Summary	7
References	8

PREFACE

In January 1996, George Woltman along with several other collaborators, launched one of the first public distributed applications onto 50 computers (equipped with the now defunct 386 processors) in order to find new Mersenne prime numbers (a Mersenne prime is a positive integer that is one less than a power of two). Sufficient to say that finding a Mersenne prime number is a complex and time consuming task that gets more difficult as the lower prime numbers are found and the exponent increases. Within a few months, on the 13th of November 1996 a prime number with 420,921 digits was found. Approximately every year after that, a new prime number has been found even though it should take longer each time to find new numbers. In fact, it should be almost exponentially harder.

By April 12, 2009, the largest to date, a Mersenne prime number with 12,837,064 decimal digits was found.

Although there was a \$100 000 prize for the person/institution that discovered a prime with more than 10 million decimal digits by the Electronic Frontier Foundation, the entire project had been running for over 13 years using nothing but "spare computing power" available in everyone's home and offices.

More recently, Stanford's University chemistry department has launched a distributed processing project aimed at understanding protein folding which is linked to diseases, such as Alzheimer's, ALS (Amyotrophic Lateral Sclerosis), Huntington's, Parkinson's disease, and many cancers. Although the project has been running for over 10 years in home PCs, Sony recently contributed to the project by adding code to their gaming machine Playstation PS3 which participates in the computing grid while their owners listen to music. In addition, users can continue to contribute to the project while surfing the internet.



Figure 1 - The PlayStation 3's Life with PlayStation client.

This project, known as Folding@home is the most distributed processing cluster in the world and is used by researchers worldwide. It uses over 400,000 machines concurrently. However, it has run in over 4.5 million devices since it first started.

To date there have been over 70 research papers based on the Folding@home project completed with visualizations of how the proteins react to temperature variations, mutations depending on

diseases, sizing of proteins and trajectories and many other useful chemical simulations and experiments.

Purpose of This Paper

The main idea behind this paper is to address the usability of distributed processing and cloud computing concepts in the area of Business Intelligence.

Typically, Business Intelligence relies on the concept of centralization of large volumes of data into a single location for processing and analysis. In contrast, distributed processing, as the name suggests, it is about distribution of processing effort and cloud computing is about remote and partitioned location of data, application and infrastructure.

We investigated how these latter two concepts can be applied to support Business Intelligence processes and achieve better results. We developed Infinis to address the issues that we found in current technology and will explore with you the risks and advantages of our new approach.

DISTRIBUTED PROCESSING

Distributed processing is a method to assign work to be done to different devices (processor, computer, gaming console, telephone, etc.) so that they can work on their own while contributing to a greater end result.

It is a complicated system since if a device fails or is slower than other devices in completing their assigned task; the end result should not suffer, but rather be failure tolerant and re-assign that task to another device. Because the individual nodes are typically low power devices, the tasks must be simple and short in duration.

When examining distributed processing for data, an additional issue that arises is that many of the assigned tasks rely on data and data is harder to distribute. Each device typically has its own memory and the central server needs to take this into account when assigning jobs. Distributed Processing is ideal for scientific and mathematical research (as discussed above) where it is calculating possible outcomes with little need for data as input, but insufficient for enterprise data processing and warehousing since each device needs access to the same data.

This is where Parallel Processing comes in, a branch of distributed processing, where the memory is shared between the devices. It can be seen in large scale applications such as Google where there is a master index of all web sites replicated to several memory areas which is then shared by all the different servers capable of answering the queries from users.

Until recently, distributed processing was directed at large, safety/time critical applications and as shown above, scientific applications, but with the decrease of hardware cost and increase in connectivity speeds, this same concept can now be used for personal use or by small and medium sized companies as well.



Figure 2 - How many PCs are on and connected to the network at night at your office?

Next time you leave the office late at night take a stroll around and see how many desktops and laptops are available to be used. Many of them are still on only with their monitor turned off! All of them connected to your company's network ready for distributed processing. Imagine the power of the available processing power at no additional cost!

Infinis and Distributed Processing

When gathering the requirements for the design of Infinis, it was found that it typically takes a very long time to bring large amounts of data from internal and external data sources into a state where analysis can be made against that same data. The key issue Infinis sought to address to resolve this issue was in the time and processing power it takes to copy data into a format understood by the application with the objective to show the data quickly to the user.

Typically bringing millions of records across to the data warehouse is a process that can take hours and/or even days. As we continually try to make the data warehouse more relevant and accurate we purchase or create more data sources each containing large amounts of records and columns. As we add more and larger data sets, (to the same hardware system) it means that the time window available for processing the various data sources is reduced, forcing these data load jobs to take place on weekends or outside of intensive usage periods such as month end or quarter end. However, it is usually during the weekday and month end that the information is needed the most which makes the data warehouse at risk of becoming a system that nobody uses, simply because it does not have up-to-date information. So the data warehouse risks becoming an obsolete tool, or the cost of the solution continues to rise as it is used, since it will require new and updated hardware to continue to manage increase data processing and peak usage needs.

A secondary design requirement was that Infinis would be useful for both small and large companies. Smaller companies typically do not have a large infrastructure due to the high cost of purchase and maintenance and as such are not capable of handling; the volume or the speed requirements.

The concept was to use distributed processing in Infinis to load a data file faster.

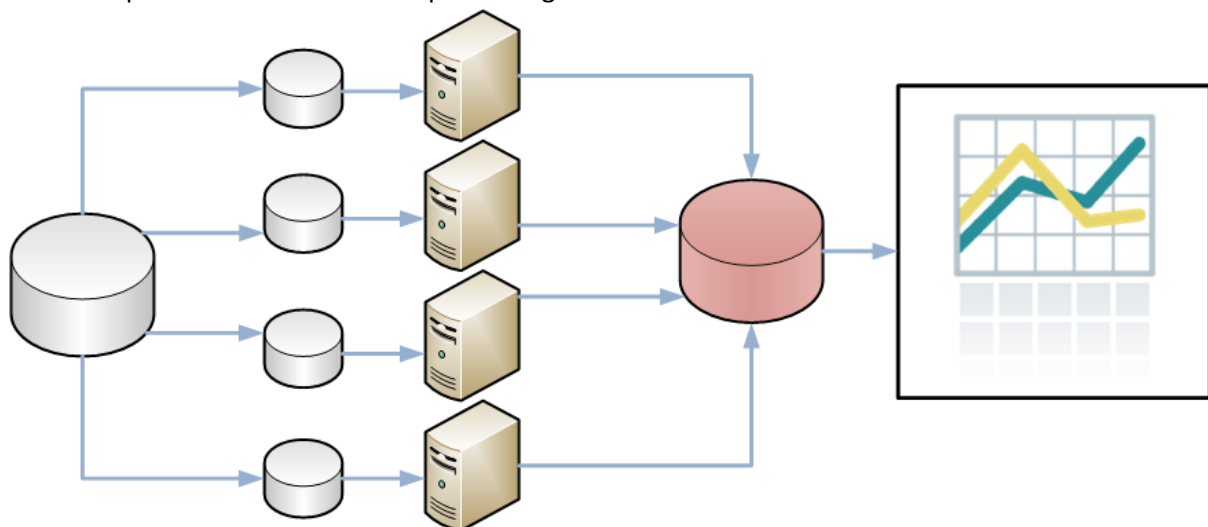


Figure 3 - Parallel loading concept of Infinis

When loading data from a text file, for example, a typical approach is to first copy the text file into a "staging area" in the database native format. This is usually done using a bulk loading mechanism where the data is basically copied row by row from the beginning to the end of the file. This process is typically sequential nature and this means that it can take hours or days.

In Infinis, the same process splits the text file (or database table) into manageable smaller chunks and instructs available *workers* (computers with access to the database that run a special Infinis application) to process and load the data. The availability of the servers/desktops/laptops on the same network means that there is very little delay in transferring the data to the machines with processing capacity, and the Infinis worker application manages the issue of dealing with memory, processing and job related information. In this way the limitations of distributed processing can be overcome by Infinis and unleash the power of your existing infrastructure.

Infinis creates a shared memory pool for its processing tasks by its data management methods and its ability to create and manage discrete data processing tasks. Previously parallel processing was only available to very large and complex infrastructure installations, but now, Infinis has simplified the process and management of parallel processing so that it is available as part of the Infinis solution.

The processed data is restored in exactly the same order (if required) and ready for further processing and analysis. The total time required for this to take place is dependent on how many workers are available; however as a general rule, for each worker, the time required for processing is reduced by 45% as compared to traditional processing methods.

The same principle is to be used in other processes in Infinis. But it does not end with just improving the time of processing data but can also be used to offset infrastructure costs such as those required for large databases - a new concept called Cloud Computing can also be used with Infinis.

Risk Factors

There are several risks in moving towards a distributed process but they are controlled risks. In essence, by trimming distributed processing to a single device we revert back to a typical architecture existing in most companies for data processing. The main risks for distributed processing are how well the management application can handle the parallel processing of tasks and more importantly how it can recover from individual node failures.

A secondary risk is associated with your data security. The devices need access to the data repositories to be able to process the data (or at least their assigned chunks of data) and put the processed data back into a centralized, or at least accessible, area. This represents a security risk as the passwords need to be encoded in each worker device to be able to access the data to be processed. An associated security risk, which can be mitigated, is the usage of web services. Web services expose particular methods for use by distributed computers, and as such could be a point of access for unauthorized users, however, there are steps that can be taken to secure web services and protect your network and your data.

CLOUD COMPUTING

In assessing our database and processing needs, cloud computing is an obvious contender to assist Infinis in its end result of providing access to more data and faster data access.

Cloud computing is a term used by many of the large technology companies such as Amazon, Google, Microsoft, among many others where they put at the disposal of consumers and other companies their technical infrastructure.

This has some benefits since it avoids capital expenditure on the hardware, database and software. It minimizes the issues that arise from growth (increase in hardware) or shrinkage (high maintenance costs for unused infrastructure).

As an example, Amazon's Elastic Computer Cloud (EC2) is a service that virtually anyone can use to take advantage of disk space, database, and processing power.

In other words, even if your company does not have numerous devices available for distributed processing, there are external providers that can provide a scalable, secure and large infrastructure for your needs.

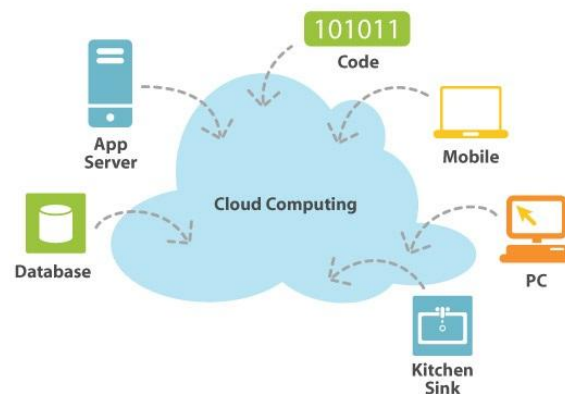


Figure 4 - Cloud Computing concept

Cloud Computing can be seen as a single point of access for your data, application logic, and the infrastructure required to maintain both.

Cloud Computing in Infinis

Both distributed processing concepts and the emerging trend of hosting data and application in a server infrastructure remotely located and accessed via the internet were both part of the design strategy of Infinis. However, the initial release of Infinis is focused around distributed processing only.

Mainly because there needs to be a change of thought on the usage of cloud computing. It is still not widely trusted as it means less control for companies to control their own data, application and infrastructure.

There was a second issue which will delay the native connectivity of Infinis to Cloud services such as EC2 (Amazon) and Azure (Microsoft) among others. Infinis relies on very fast data access from and

to SQL Server 2008. At this time there are no publicly available cloud service companies guaranteeing the required connectivity speeds as an option.

Risk Factors

Security and reliability is the number one concern of Cloud Computer users. It has to be said that lack of control augments this perception but there are many cases of connectivity outages which can cause large lack of productivity and financial losses. Security risks can also be minimized by having cloud computing in a smaller cloud, perhaps in a cloud within your company (which may negate to many extents the cost and scalability effects).

Taking into account the risks, it still is a growing trend which is slowly allowing companies to move to a zero install computer for their employees. This means that once an employee is hired, a single user is created in the network with immediate access from any computer (at home, at work, on the road) to a wide breadth of applications (word processing, spreadsheet, internal company applications, etc.) simply by accessing a portal.

More interesting is the idea of not even needing an operating system at all, where the operating system is also a cloud service.

Summary

Distributed processing, and more specifically parallel processing, is a key industry trend especially when coupled with existing cloud services readily available at a low cost.

Future enterprise ready applications need to be designed and built with these concepts, which although not easy to implement into a Business Intelligence/Data Warehouse application, are critical. Furthermore as higher volumes of data and more complex data sets are available to the enterprise via their R&D, ERP, HR, CRM, Financial, Forecasting and external market data providers, the greater the need to process data faster and distributed processing is the only scalable and cost efficient way to accomplish this.

Infinis is a new approach to business intelligence software and incorporates many new technology and user-centric changes to business intelligence software. We have looked at many of the challenges facing businesses today; with the continuing rapid change in technology, the globalization of competition and business, the continuing growth of data generation and data consumption by businesses, increased competition as well as the ever increasing demands of customers. Infinis has been designed to provide you with faster and better access to your data, so that you can meet these challenges, and expand your business.

Visit us at InfinisWorld.com to see and understand how Infinis is changing the approach to business intelligence applications and how Infinis can help you achieve better results for your business.

References

1. *Folding@home distributed computing*. Stanford University. <http://folding.stanford.edu/>
2. *Folding@home diseases studied FAQ*. Stanford University. <http://folding.stanford.edu/FAQ-diseases.html>.
3. *Great Internet Mersenne Prime Search*. <http://www.mersenne.org/>
4. Woltman, George (February 24, 1996). "The Mersenne Newsletter, issue #1" (txt). Great Internet Mersenne Prime Search (GIMPS). <http://www.mersenne.org/newsletters/news1.txt>. Retrieved 2009-06-16.
5. *The Theory of Database Concurrency Control*. CS Press, 1986. Christos Papadimitriou
6. *Computational Complexity*. Addison Wesley, 1994. Christos Papadimitriou
7. *Cloud Computing Savings - Real or Imaginary?* Balakrishna Narasimhan. (April 16, 2009) <http://blog.appirio.com/2009/04/cloud-computing-savings-real-or.html>
8. Amazon Elastic Compute Cloud (Amazon EC2). <http://aws.amazon.com/ec2/>
9. *Cloud Computing Image, Source*: <http://infreemation.net/cloud-computing-linear-utility-or-complex-ecosystem/>