A Systems Thinking Approach to Business Intelligence Solutions Based on Cloud Computing

by

Eumir P. Reyes

B.S. Information Systems Engineering (2001)
Instituto Tecnológico y de Estudios Superiores de Monterrey

Submitted to the System Design and Management Program
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering and Management

at the

Massachusetts Institute of Technology

February 2010

Signature of Author _____

Eumir P. Reyes
System Design and Management Program
February 2010

Certified by _____

John R. Williams
Thesis Supervisor
Associate Professor of Civil and Environmental Engineering

Accepted by _____

Patrick Hale
Director
System Design and Management Program

This page intentionally left blank

A Systems Thinking Approach to Business Intelligence Solutions Based on Cloud Computing

by

Eumir P. Reyes

Submitted to the System Design and Management Program in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering and Management

Abstract

Business intelligence is the set of tools, processes, practices and people that are used to take advantage of information to support decision making in the organizations. Cloud computing is a new paradigm for offering computing resources that work on demand, are scalable and are charged by the time they are used. Organizations can save large amounts of money and effort using this approach.

This document identifies the main challenges companies encounter while working on business intelligence applications in the cloud, such as security, availability, performance, integration, regulatory issues, and constraints on network bandwidth. All these challenges are addressed with a systems thinking approach, and several solutions are offered that can be applied according to the organization's needs.

An evaluations of the main vendors of cloud computing technology is presented, so that business intelligence developers identify the available tools and companies they can depend on to migrate or build applications in the cloud.

It is demonstrated how business intelligence applications can increase their availability with a cloud computing approach, by decreasing the mean time to recovery (handled by the cloud service provider) and increasing the mean time to failure (achieved by the introduction of more redundancy on the hardware).

Innovative mechanisms are discussed in order to improve cloud applications, such as private, public and hybrid clouds, column-oriented databases, in-memory databases and the Data Warehouse 2.0 architecture.

Finally, it is shown how the project management for a business intelligence application can be facilitated with a cloud computing approach. Design structure matrixes are dramatically simplified by avoiding unnecessary iterations while sizing, validating, and testing hardware and software resources.

Thesis Supervisor: John R. Williams
Title: Associate Professor of Civil and Environmental Engineering

Acknowledgments

I would like to acknowledge everyone who supported me during my experience at MIT and my work on this thesis. I extend my thanks to:

My thesis advisor, John Williams, who introduced me to the world of cloud computing during his lectures and helped me to materialize my ideas on this thesis, combining two important knowledge areas on information technology.

Pat Hale and the SDM staff, who gave me the opportunity to study the MIT's exceptional SDM program, and provided me the aid and flexibility required as a distance, international student.

My SDM colleges, who helped me on several educational challenges, let me learn with them, and with whom I shared priceless moments.

Raúl Livas, an MIT Ph.D. in Economics, who encouraged me to study at the MIT and showed me a complete new perspective on graduate education.

Eduardo Graniello, my manager, and Intellego, the company I work for, for allowing me to study for my master's degree in combination with my professional responsibilities, and providing me with required resources.

And my family, who stayed with me and supported me all the time while I made this dream come true.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

1.1 Motivation

Business Intelligence is one of the information technology areas that has grown and evolved the most in the last few years. According a study by AMR Research, the total spending on business intelligence and performance management by 2008 was of 57.1 billion dollars, in a market growing 4.2% per year[1].

Through business intelligence it is possible to improve the decision making process in virtually any department, organization or industry. Organizations have seen how business intelligence has changed over this time, how the tools have evolved offering more functionality to the analysts, and at the same time, providing solutions for more users. Information requirements have grown exponentially: while only a few gigabytes of data were needed some years ago, now data warehouses are populated with terabytes of data and rapidly moving to the petabytes range. Also, many companies are now exposing their information to external users, like suppliers and customers, in order to share data and improve their operational process.

With this scenario, traditional technology strategies were not fast enough to satisfy the business needs most of the time, under situations in which data warehouses and application servers reached their limits just months after having released the applications to final users. Thus, a new flexible technological approach was needed to address this challenge.

Since the recession began, companies have been forced to optimize their budgets and reduce costs. Many of these companies have seen cloud computing as an option to achieve this goal. They have started acquiring services like Salesforce.com or migrating essential productivity tools such as e-mail from Microsoft Exchange to Google Apps[2]. The next natural step is to migrate business intelligence solutions to the cloud.

Besides all these events, Gartner, the information technology research company, has identified the top 10 strategic technologies for 2009[3]: Cloud Computing and Business Intelligence were two of them. In the forecast for the top 10 strategic technologies for 2010[4], the company repeated Cloud Computing and added Advanced Analytics, which is one of the axes of business intelligence solutions.

According to Gartner, the cloud computing market will increase from $56.3 billion in 2009 to $150 billion by 2013[5], requiring organizations to pay attention to the benefits they can get by changing the way they deliver applications to their internal and external customers, and how CIOs can save money by moving their IT efforts to the cloud.

Cloud computing also helps to create a greener IT. According to a study by Greenspace, companies that run Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) applications on NetSuite, a company that offers cloud computing services, reduced up to 423,000 metric tons of CO2 per year[6]. This decrease was possible because the number of servers, power supplies and cooling systems were reduced dramatically, creating a more sustainable architecture.

Although much has been written about cloud computing and its benefits, some challenges still remain that have to be addressed in order to effectively leverage the organizations' information and new technology. The objective of this thesis is to assess these challenges and try to solve them through a systems thinking approach.

1.2 Current Scenario

It is estimated that companies spend an average of 80% of the time gathering data for decision making and only 20% analyzing it. This 80% is formed by many activities, such as:

1. Discovering the data sources
2. Mapping required data
3. Extracting the identified data
4. Transforming the identified data
5. Modeling the data warehouse structure
6. Loading the data into the data warehouse
7. Creating the front end
8. Validating the information
9. Fine tuning the solution
10. Managing the required hardware
11. Managing the required software

The first nine activities can be solved with a robust business intelligence solution, which if well designed, built, and automated, may be able to avoid user involvement. But the two last activities require enormous amounts of time, money and human effort. Usually, the IT departments in the organizations are in charge of dealing with these expenses.

Part of these efforts can be optimized by taking this type of solution to a cloud computing paradigm. Some CIOs and CTOs have started to act by running some applications under the modality of Software as a Service (SaaS). However, it is important to note that SaaS is not the same as cloud computing. SaaS is one of cloud computing service models, and not all SaaS offerings run on a cloud computing basis.

According to IDC, the global information provider, 43.2% of the companies found business intelligence to be the best suited for SaaS delivery among other 11 technologies[7]. In spite of this, many challenges remain for a successful implementation, mainly thinking about large-scale systems and complex data warehouses.

Software as a Service (SaaS) has been one of the first offerings based on Cloud Computing, but one limitation is that SaaS consists mainly of packaged analytic applications[8]. That is why a robust business intelligence cloud computing solution has to be customized, and for this, a previous consulting and tailoring effort is needed. Although some ERP solutions are prepackaged and are part of the data sources that feed BI applications, most of them are customized during the implementation. These customizations also need to be applied to the BI packages, creating the need of more customer-oriented solutions. This is the main reason why each case has to be analyzed separately.

2 Technology Concepts

In order to understand how business intelligence and cloud computing can work as a solution for organizations, it is important to clarify the meaning of both concepts and identify their main elements. This section will describe both technologies.

2.1 Business Intelligence

There are several definitions of what business intelligence means, and many of them have evolved over time and have acquired a larger scope. For purposes of this document, we must understand business intelligence as the set of tools and processes that gather data from several sources, organize them, process, store and present them to end users in order to improve the decision making in the organization and to generate value through information and knowledge.

The business intelligence process can be simple or complex depending on the company, its data sources, and its analysis needs. Figure 1 illustrates the typical sections of a complete business intelligence solution:

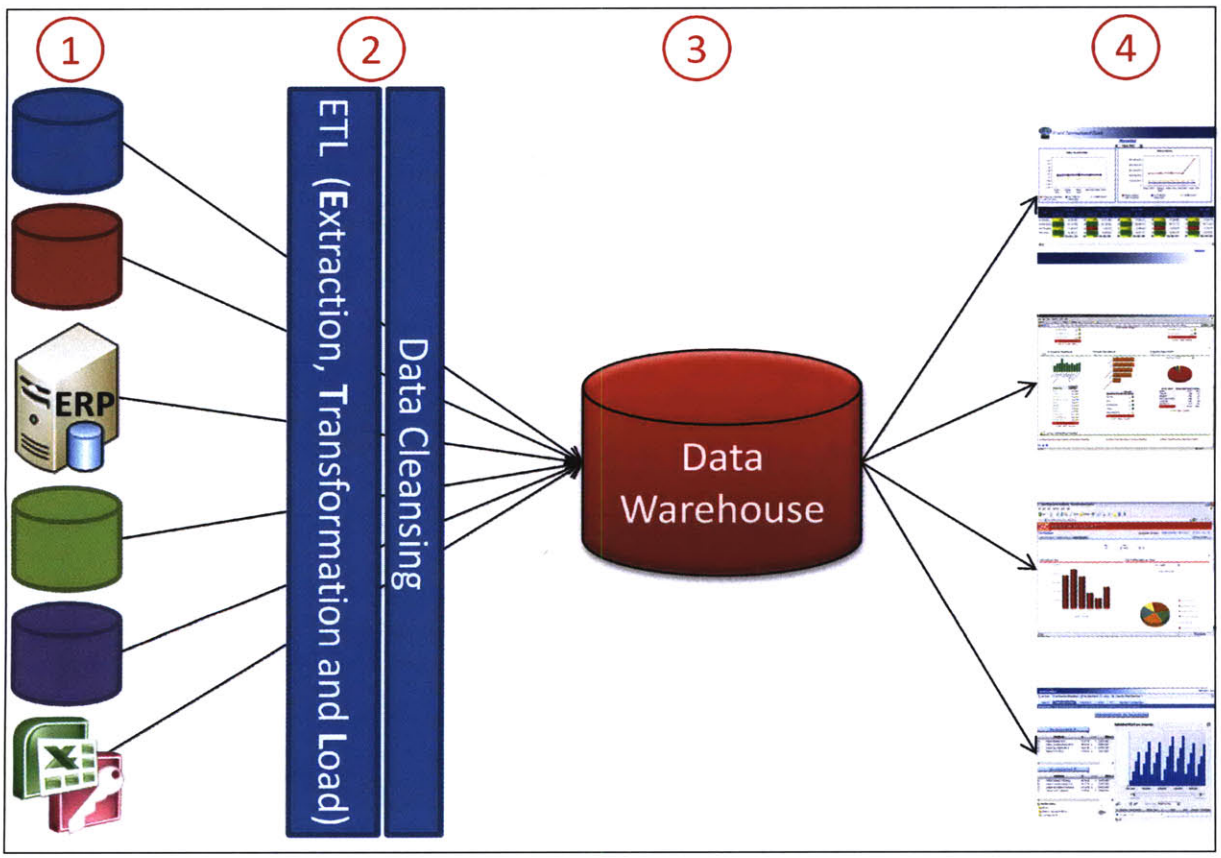


Figure 1. Typical Business Intelligence Solution Sections

It is important to recognize all the elements that form a business intelligence solution, since each of them will play different roles in a cloud computing approach and will deal with different challenges. The four main sections will be explained below.

2.1.1 1- Data sources

Most organizations have grown without an initial strategy for information technology. This lack of perspective has caused them to have different systems and data sources, and even, to generate isolated departmental applications. When decision makers need information, analysts have to go over all these data sources to get the required data. The most common source of data is ERP systems. This is usually the most robust and easy way to get data. Also, many business intelligence suites contain predefined structures to obtain data coming from these systems, like the SAP NetWeaver BI platform does with its Business Content, which contains specific objects that help to rapidly obtain information from mySAP systems.

But not everything is as simple as obtaining data from an ERP's. Several solutions need to connect to relational databases like Oracle, DB2, SQL Server or MySQL, or multidimensional databases like Essbase or Microsoft Analysis Services. Another kind of source is unstructured data sources, such as spreadsheets and flat files. Here is where actual challenges arise, since valuable data can be found on these unsystemized sources.

2.1.2 2 - ETL and Data Cleansing

The Extraction, Transformation and Load (ETL) process is the phase that usually takes the longest time in business intelligence solutions. Its objective is to obtain the required data from specific data sources in the organization. This demands a detailed task of mapping data from its origin, which needs to be done with the people who own or administer the system or unstructured data.

The next step is to transform the data into specific formats and structures needed by the application. As explained above, data is obtained from several data sources, and obviously, each one is heterogeneous and data is so, then they has to be homogenized in order to satisfy further analysis needs. Some example could be to homogenize catalog keys for items coming from several systems, or to convert all the measures to metric units supposing that some systems store them using imperial units.

In some solutions, data cleansing is also required. Considering that data is not always captured properly on transactional systems, information has to be cleaned before loading it into the data warehouse. Examples of data cleansing are completing or correcting mail addresses, correcting amounts (e.g. person ages greater than 120 years) or deduplicating records. Although simple data cleansing can be done during the ETL process, best practices recommend creating a new project just for data cleansing before initiating a business intelligence initiative.

2.1.3 3 - Data Warehouse

The data warehouse is considered the core of any business intelligence solution. According to Bill Inmon, recognized as the "father of the data warehouse", a data warehouse "is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process"[9]. This element is responsible for containing all the data coming from data sources after they have been organized and homogenized. These data are later accessed by analysts and decision makers through front end tools, such as dashboards or executive reports.

A good data warehouse design must consider the current and future user needs of information, and at the same time, be optimized for fast and interactive access. Also, the data warehouse has to be flexible enough to support the rapid data growth and changes in the organization.

2.1.4 4 - Front End

Front end tools allow users to access data contained in the data warehouse. Since several roles and profiles can be found in the organizations, not all the front end tools have the same functionality, and they have to be chosen according to the final user's goal. Some tools are focused on more operative analysts, who need to go deep into the data to find specific metrics or patterns. Other applications, like executive dashboards, are aimed to managerial levels that need a general and consolidated view of the company's current status.

Wayne Eckerson, of The Data Warehouse Institute, identified five dimensions of business intelligence, which can help us to categorize the myriad of tools and applications that vendors have created to match users' needs. Figure 2 shows these dimensions and the interaction among them:

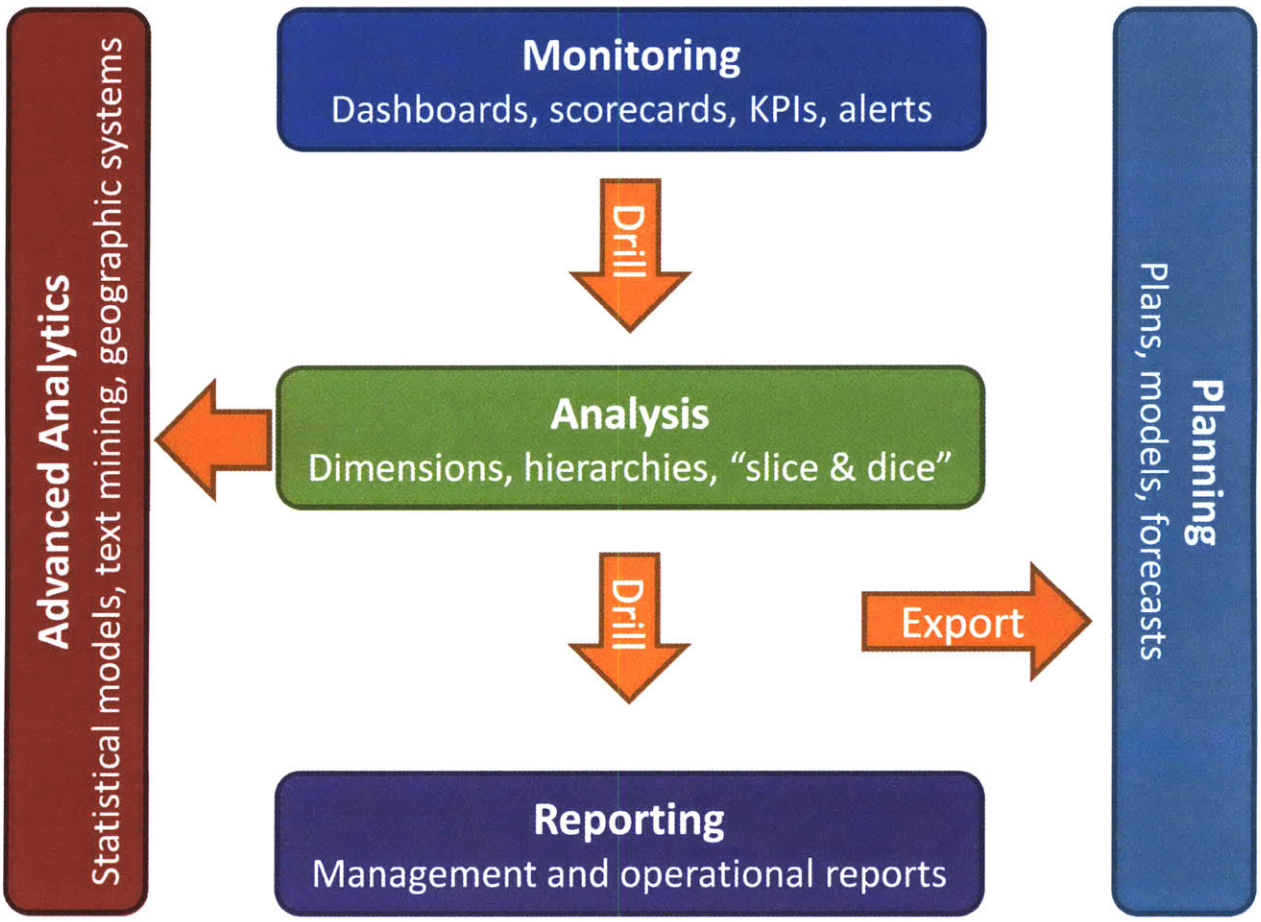


Figure 2. Eckerson's Five Dimensions of Business Intelligence[10]

The newest dimension in this model is Advanced Analytics, much focused on data mining. As mentioned above, it is now considered as strategic by Gartner, and needs to have all the

business intelligence phases mentioned earlier as its foundation. That is why all the elements have to be taken into account for a cloud computing approach.

The front end layer is the one that should be the least disrupted by the cloud computing approach, since most of these tools are available via web interfaces, and it is irrelevant for the user if the application server is running on his own computer, in his office or in a data center in another city or even another country. But still, there are several applications that work on a client/server paradigm, demanding access to high volumes of information.

2.2 Cloud Computing

There are several definitions for cloud computing, but all of them concur in that it is a new paradigm based on a pay-per-use model to flexibly access hardware and software resources through Internet, allowing companies to reduce costs and increase performance. The definition that seems to have all these elements is the one created by the National Institute of Standards and Technology (NIST), which is reproduced below:

Cloud computing is a model for enabling 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.[11]

Regarding cloud computing interaction with business intelligence, the next elements have to be emphasized:

- On-demand. Also related to pay-per-use. Under this model, users would only pay for what they need and use according to the required capacity. Under a traditional model, the organizations have to acquire hardware and software, even though these resources are not used at their full capacity. Business intelligence applications requirements are not constant. For example, ETL processes require high processing levels at night, when the information is extracted from data sources. Also, application servers need more memory and data warehouses need more I/O (Input / Output) operations in the mornings, when analysts seek information for daily operations and decision making.
- Network access. All the resources to be accessed and modified are located outside of the company, that is why data uploads and data requests need to travel on the Internet. This is different to on-premises solutions, in which almost all the information is obtained within the same local network.
- Pool of resources. The cloud computing provider may have a set of servers and storage devices to serve several customers. Under this paradigm, the information contained in a data warehouse may be distributed physically in several places and the application instances may be shared with different companies.
- Rapid provision. One of the main concerns for business intelligence solutions is how to scale the data warehouse when it has reached its maximum storage and performance capabilities, or when the front end application servers start to delay users' responses. Under the cloud computing paradigm, the escalation and release of resources is done automatically and is transparent to the users.

It is also important to identify two types of cloud computing models as defined by Gartner: Public Cloud Computing and Private Cloud Computing. According to this company's definition we have[12]:

Public cloud computing is a style of computing where scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies.

and

Private cloud computing is a style of computing where scalable and elastic IT-enabled capabilities are delivered as a service to internal customers using Internet technologies

Both definitions vary only in the type of customers: one delivers service to external users and the other to internal users. It is relevant to understand both approaches since many organizations will opt to have private cloud computing services for BI, because, firstly, the information must only be exposed to internal users or users with "memberships", and secondly, because they need to maintain the control and ownership of the resources and information. In general, private cloud computing offers more security and privacy to the organizations, but at the same time, do not take advantage of low prices and economies of scale that public clouds may have.

3 Benefits of Cloud Computing for Business Intelligence Solutions

According to a recent Gartner survey, the most important drivers for investing in business intelligence are related to providing faster response to user's needs for data, decreasing costs, and enhancing the user's methods for data sharing and self service[13].

Cloud computing makes sense to business intelligence solutions only if it offers benefits to customers. In this section those benefits will be described[14].

Lower costs. Under a cloud computing paradigm, companies do not need to invest large amounts of money to acquire hardware, software, licenses and knowledge to put the business intelligence infrastructure up and running. They would only have to contract a cloud computing provider and pay for the resources they need. It is unnecessary to pay for the time in which no user accesses the application and computing resources remain dormant.

Multiple redundant sites. One of the main concerns of business intelligence professionals is to keep the solution available the longest time possible. One way to achieve this is to have multiple sites that offer redundancy. Since most of cloud computing providers have sites geographically dispersed, this characteristic is achieved.

Scalable provisioning of resources. Business intelligence solutions do not have the same load work during the day. This means that at certain points in time, some servers could be idle while others may be reaching their peaks on processing, memory usage or I/O operations. With cloud computing, resources can automatically and rapidly scale in and scale out. For example, during the ETL process at night, the solution could use processing power from application servers, and in the afternoon, analytical processes could use memory that is not in use by the ETL processes.

On-demand performance improvements. One of the most recurrent problems that are seen on business intelligence applications is when the customers need to expand their data warehousing capabilities without affecting daily operations. This task is usually very complex because it requires high investments in new hardware, storage, licenses, and human effort to perform the migration to the new environment. Under a cloud paradigm, this problem would be addressed almost instantaneously and transparently for users, by taking advantage of existing hardware and software resources.

Usage billing. By using cloud computing, companies pay for a service as they go or pay on a monthly or yearly basis. With this policy, the expenses move from Capital Expenditure (CapEx) to Operational Expenditure (OpEx). CFOs will greatly appreciate this characteristic.

Fast deployment. Instead of spending long time preparing and installing required hardware and software, the platforms can be up and running in just minutes, ready to configure applications and start populating the data warehouse.

Easy maintenance. Most of the maintenance needed for hardware and software, like firmware, updates and upgrades, are done by the cloud computing provider. Also, since these applications are accessed through internet browsers, maintenance on client computers is reduced dramatically.

4 Cloud Computing Requirements for Business Intelligence Solutions

One of the objectives that originated business intelligence and data warehousing was to develop tasks that were not able to be executed on transactional systems. This forced a rethinking of different hardware and software architectures, like databases optimized for reading instead of writing and high interactive and visual applications.

Is the cloud computing paradigm able to satisfy these requirements? What are the technological requirements? This section will discuss these requirements and show how different they are compared to the on-premises approach.

4.1.1 The Basic Architecture

The basic architecture needed to run business intelligence solutions in the cloud is depicted in Figure 3:

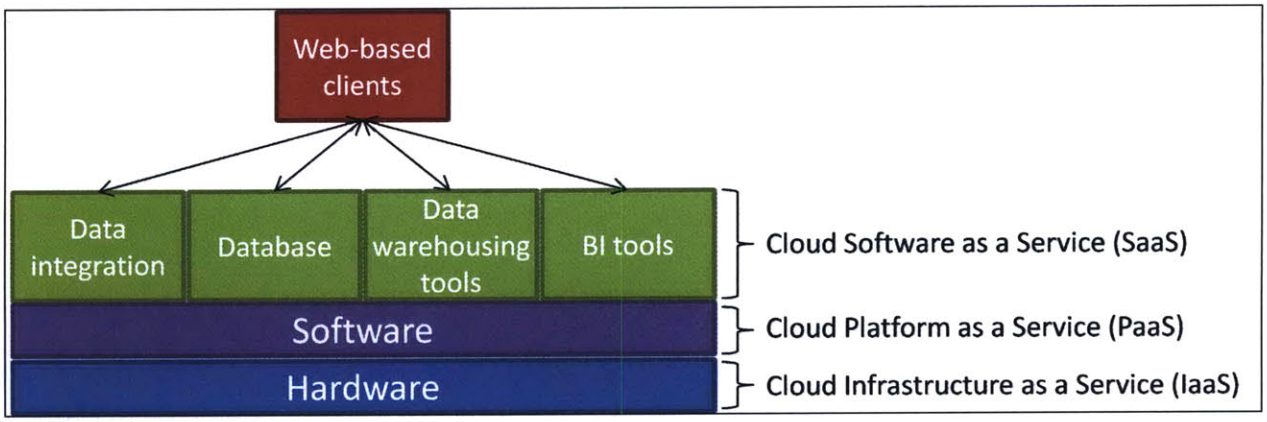


Figure 3. Basic Business Intelligence Architecture in Cloud Computing

The lower layers are formed by hardware and software systems. These are the minimum elements that have to be offered by the cloud computing provider. Hardware refers to processing, storage, and networks, while software refers to the operating systems and drivers required to handle the hardware.

The Data integration box refers to the tools needed to perform the ETL and data cleansing processes. The database box refers to the relational or multidimensional database systems that administer the information. It is important to note that there are new devices called "data warehouse appliances", which integrate hardware, software and databases elements in just one box. However, they should be considered as an integrated part of the architecture.

Data warehousing tools are the set of applications that allow the creation and maintenance of the data warehouse. BI tools are the set of front-end applications that enable the final users to access and analyze the data.

Finally, since all the architecture is going to be accessed through the Internet, there is no need for thick clients or preinstalled applications, because all the content and configuration can be reached through traditional internet browsers.

4.1.2 Available Technologies

While this thesis was being written, several vendors in the market announced products aimed to satisfy companies' needs in the cloud. This section discusses some of the most relevant technologies and specifies which layer they are focused on based on Figure 3.

4.1.2.1 IBM Smart Analytics Cloud

Vendor: IBM

Product/Service name: Smart Analytics Cloud

Offered layers: IaaS, PaaS, SaaS

Description: Based on its System Z mainframe, IBM launched this product intended to provide a private cloud to satisfy its internal needs for data analysis. The solution provides the required hardware, Cognos 8 as the business intelligence platform, and required services for the adoption of cloud computing in the organizations.

Advantages:

- Integration of hardware, software, business intelligence applications, and services in one solution.
- Good option for private clouds.
- Customer control over all the resources of the solution.

Disadvantages:

- New product.
- Product originally intended only for private clouds.

Observations: IBM initially launched this product to satisfy its internal business intelligence needs, trying to reduce costs and effort needed for data analysis.

4.1.2.2 Amazon EC2

Vendor: Amazon

Product/Service name: Amazon Elastic Compute Cloud

Offered Layers: IaaS, PaaS

Description: Taking advantage of its successful platform, Amazon is offering Cloud Platform as a Service. This solution provides all the characteristics defined for cloud computing, including scalability and pay-per-use. The customers customize the service according to their needs, based on memory, storage, processing, operating system, database, and other applications. Since this platform still does not offer business intelligence tools, we cannot categorize it as a BI SaaS provider.

Advantages:

- Proven technology and solutions

- Completely customizable by the customers
- Flexible pricing
- Customer control over the operating system layer

Disadvantages:

- Lack of customer control over the platform
- Inability to solve any problem when it fails.
- Some service outages have been registered, for example: October 7 2007, February 2 2008[15] and December 9 2009[16].

Observations: Some companies have already successfully installed business intelligence applications in the Amazon Cloud, having open source applications as the set of tools enabling the solution on this platform.

4.1.2.3 Informatica 9

Vendor: Informatica

Product/Service name: Informatica 9

Offered Layers: SaaS (not including database)

Description: Informatica was one of the first business intelligence tools providers to offer solutions focused on the cloud. Its solution enables customers to develop all the business intelligence cycle, from data access to data delivery, offering compatibility with several data sources and databases as data warehouse.

Advantages:

- Unified set of tools for business intelligence
- Openness, allowing integration of data in any IT environment

Disadvantages:

- Required integration into an existing IaaS, PaaS and database since the product only offers BI software.

Observations: It is expected that other business intelligence providers enable their suites to take advantage of the cloud as Informatica 9 has done.

4.1.2.4 Greenplum

Vendor: Greenplum

Product/Service name: Greenplum Database 3.3

Offered Layers: Database (part of SaaS)

Description: Greenplum is a database created specifically for data warehouse environments. It utilizes shared-nothing Massively Parallel Processing to improve the performance during data

access. The company has announced its cloud enabled version of the database with characteristics such as self-service provisioning, elastic scale and massively parallel operations.

Advantages:

- Designed for analytic processing and cloud computing
- MapReduce support

Disadvantages:

- Required integration into an existing IaaS, PaaS and database since this product only offers the database.

Observations: It is the first database oriented to data warehousing that is officially focused to the cloud.

4.1.2.5 Windows Azure

Vendor: Microsoft

Product/Service name: Windows Azure

Offered Layers: IaaS, PaaS

Description: Windows Azure is the Microsoft cloud computing offering. It includes the platform required to run applications in the cloud, operating system, and database on Microsoft's data centers. Microsoft is also including ERP services as part of its cloud computing offerings. Business intelligence is not yet part of the service; this is why it is not catalogued as BI SaaS.

Advantages:

- Integration with existing software applications
- Includes all the platform required for cloud computing

Disadvantages:

- Uncertainty about how to integrate data warehouse and business intelligence software into its SQL Azure database system.

Observations: It is a good option for continuing using Windows platform and make it interact with other architectures.

4.1.2.6 The Rackspace Cloud

Vendor: The Rackspace

Product/Service name: CloudServers / CloudSites

Offered Layers: IaaS, PaaS

Description: Rackspace offers on-demand servers with root access that can be scaled up and down at any time (IaaS). In another modality of its services, customers obtain PaaS, consisting of servers with Linux or Windows with automatic scalability.

Advantages:

- Flexibility regarding the level of control the customer needs: from direct access to the servers to application control.

Disadvantages:

- Some outages have been registered, such as the one on December 18, 2009[17].

Observations: Rackspace claims to be 60% larger than Google.

4.1.2.7 Force.com

Vendor: Salesforce.com

Product/Service name: Force.com

Offered Layers: IaaS, PaaS

Description: Force.com provides the required platform to build applications in the cloud.

Advantages:

- Proven platform
- Encapsulation of lower layers, like operating system, applications and database.

Disadvantages:

- Low flexibility regarding the set of operating systems and databases available.

Observations: After its successful CRM SaaS, Salesforce.com has launched its cloud computing services to allow organizations create customized applications.

5 Challenges of Business Intelligence solutions based on Cloud Computing

Although the benefits that cloud computing brings to the industry have been mentioned above and companies have taken advantage of them, there are several challenges that have been identified by several authors. This section will present the main challenges faced by business intelligence solutions in the cloud.

Cyrus Golkar of B-eye-Network identifies these CIO concerns[18]:

- Security
- Performance
- Availability
- Integration
- Customization

Mukund Deshpande and Shreekanth Joshi of B-eye-Network identify these issues with cloud computing[19]:

- Moving data to the cloud
- Storing data in the cloud
- BI components as a service
- Integration with on-premises data

Wayne Eckerson of The Data Warehouse Institute identifies these constraints[20]:

- Data transfers
- Data security
- Due diligence

Recombinant Data Corp identifies one main challenge[21]:

- Security of sensitive information

Stephen Dine from Datasource Consulting, LLC identifies these challenges[22]:

- The ability to scale-up is limited
- Difficult to quell security concerns
- Viability of moving large amounts of data
- Performance of physical data access
- Reliability of service concerns
- Pricing is variable and complex

Dave Wells of B-eye Network explains these factors as the downside of cloud computing[23]:

- Security
- Privacy

- Compliance
- Control
- Governance

Tom Lounibos of Eclipse Developer's Journal mentioned the next challenges according to SOASTA, a company based on the cloud that offers products for web application testing[24]:

- Security
- Governance
- Performance

As we can see from the sources cited above, there are several challenges. Synthesizing those ideas, we have that the more recurrent concerns are, in order of importance: security, moving large amounts of data, performance, integration, and reliability. Each of these challenges or constraints will be explained below.

5.1 Security

Security is clearly the most important concern when dealing with business intelligence solutions based on cloud computing. This concern is completely understandable, because companies would have to move valuable information out of their own servers and data centers and rely on a third party who would store the information somewhere in the world. With this change, several factors arise. First, if the connection between the data source and the target data warehouse or staging area is not secure, some intruders could get the data and get sensitive information. This is possible, because instead of having data travelling just inside the company's local area network (LAN) now it would have to travel over the Internet.

Second, sensitive data would be out of the control of its owners and would reside in the storage equipment of the service provider. Under this situation, the customer would not be able to know if some unauthorized individual were to access the data at the provider's facilities, and use the information for prohibited actions.

The third important concern is that there are some laws that prohibit mixing sensitive data with another organization's one. This is the case of Health Insurance Portability and Accounting Act (HIPPA). Also, regarding regulatory issues, there are laws that prohibit maintaining information in a country different to the one where it was originated.

Fourth, customers feel threatened by the possibility that external hackers could attack the provider's information systems and interfere with the information, such as illegally accessing it, taking it for commercial reasons or corrupting it. This danger is mainly relevant if the cloud computing provider does not have the protective software and hardware that the company had in-house before using cloud computing.

5.2 Moving large amounts of information

Organizations tend constantly to generate more information with their transactional systems. Modern practices suggest monitoring detailed data in several steps of the daily operation. If all this data is thought to be stored in the data warehouse, then large amounts of data will be

required to move to the cloud. If we think of data warehousing from a Bill Inmon's approach, then maximum data granularity would have to be transported to the cloud through Internet. It is understandable that network throughput over the Internet is much less than the existing in local area networks, thus, requiring more time to transfer information.

With this scenario, transferring data through the ETL process to the cloud could take longer, and if the organization does not count with a reliable internet connection, the night extraction processes would not be ready on time, resulting in the generation of incorrect data in the data warehouse, and thus, incorrect decision making.

Another point of concern is that several service providers, based on the characteristic of pay-as-you-go, charge for the amount of data that is transferred to the cloud. If large volumes are going to be exchanged, then the cost could be very high, and the total expenditure would not be, in this case, one of the main benefits of cloud computing.

5.3 Performance

Performance is related to two main factors: data processing and data transferring. Data processing refers to the required computational power to handle information in each business intelligence step, that is, ETL, data warehouse or data analysis, and consists of CPU, memory, and I/O operations. This concern comes when an organization evaluates if the provider will be able to offer hardware and optimized software at least as powerful as the one the customer has on-premises. If the provider fails to have this kind of architecture, then performance may be hindered.

Data transfer is related to the data sent and received over the Internet. Just as stated in the previous case, it is not the same to transfer data in the same local area network, where gigabit connections are available and data has to only go over a few meters, than to transfer data over the Internet, with slower connections and several kilometers to travel between the final user and the data center.

5.4 Integration

The first kind of integration that has to be considered in a business intelligence scenario is between the data sources and the data warehouse. This integration is handled by ETL tools. Although modern ETL suites have considerably expanded their connection capabilities to sources and targets of data, and some even claim to interact with virtually any system, under a cloud computing approach the number and kind of sources and targets increase, mainly affected by the type of data bases that are going to be used as data warehouses.

An example of new technologies that would have to be integrated to the cloud is MapReduce. MapReduce is a new software framework created by Google that supports processing of large datasets in distributed systems. MapReduce capabilities make it a very good choice for handling large volumes of information in cloud environments. Then, it this technology were going to be considered as the data management system for data warehousing purposes, all the access procedures would have to be rewritten in order to be handled by the system, that is, traditional SQL statements could not be used[25].

Another point of integration is drill-through. Although business intelligence literature offers different definitions of what "drill through" means, the general understanding is that doing a drill-through means to go from a less detailed piece of data to a more granular data, including existing filters and usually interacting between two different types of business intelligence structures. An example could be to find that a customer has bought $1,000. This data is stored in the data warehouse or in a data mart (an information cube), but if we wanted to know if these $1,000 consist of only one order of $1,000 or 10 orders of $100 each, we could automatically send a query to the original relational database and discover the actual buying structure. Of course, this kind of integration could be very complex, mainly considering that the data warehouse and multidimensional model would be in the cloud and the transactional source system would be on-premises.

Unstructured data also needs to be integrated to business intelligence solutions. It is estimated that 80% of the data in organizations is unstructured[26]. A relatively new tendency is to try to unify both types of data in order to get a complete view of the organization. Having business intelligence in the cloud could make this integration more difficult.

5.5 Availability

This challenge refers to the fact that the organization can be sure that the solution will be available and up and running more time than it would be on-premises. In cloud computing more issues have to be considered in order to have the system ready, like the network internet connection and the distributed devices that keep the system running. In addition, if something fails on the platform, there is little that the organization can do to get it running again, besides calling the service provider and waiting for them to solve the problem.

6 Addressing the Challenges

After having clear the challenges that affect a business intelligence solution in the cloud, this section will address those challenges and propose solutions based on a systems thinking approach.

6.1 Architecture of Business Intelligence in the Cloud

The first step is to understand how the system as a whole is architected, what its subsystems are, and how they interact among them. This section will show the general architecture that a business intelligence solution needs to have in order to run in the cloud. It is important to mention that according to different vendors and requirements, the architecture can change. For example, there are several data warehouse appliances in the market, which join in only one device hardware (CPU, memory), storage, operating system and database software. Still with this kind of differences, the general architecture must prevail.

6.1.1 Decompostional View

Figure 4 shows the system decomposition for this solution using OPM (Object-Process Methodology). The architecture is taken to level 3 for relevant elements. There are 4 general elements that compose the cloud: Infrastructure, Platform, BI Applications and Clients.

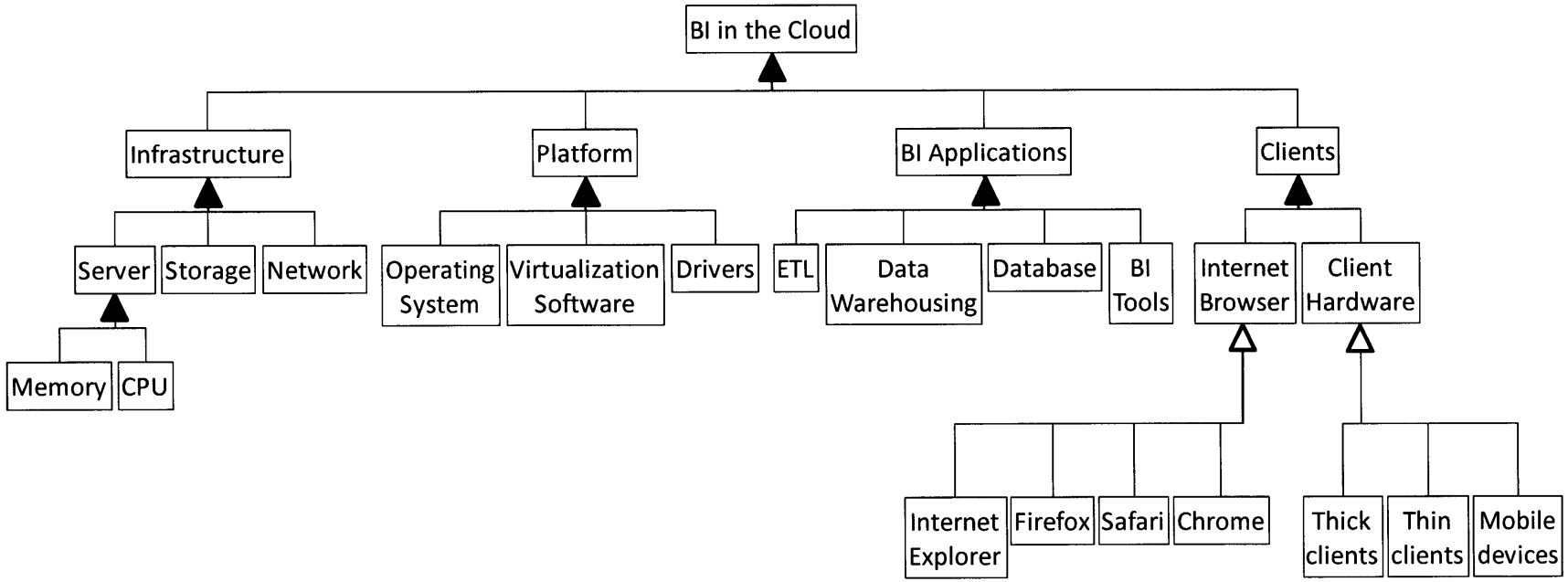


Figure 4. BI in the Cloud system decomposition

The Infrastructure component refers to the set of hardware needed to run the solution. This hardware must have the required characteristics in order to run in the cloud, like supporting virtualization and redundancy. Three main elements integrate the hardware: servers, storage and network. Servers are the actual boxes that process data. For them, two main elements have been decomposed: memory and CPU. It is important to consider these two elements because most of the system performance depends on them. Some cloud computing providers calculate the pricing based on the requirements the customers have of these two parts.

The next element of the infrastructure is storage. Storage is the hardware that will be in charge of storing the produced information on the system. It is generally formed by hard disks, disk controllers, network interfaces, and required hardware to keep them running. The last element of the infrastructure is network, which is in charge of transferring data among devices in the cloud, and taking them in and out in order to reach the users.

Platform refers to the set of software needed to run the hardware in the cloud. The most important element is the operating system, which must be able to support cloud essential operations, like scaling in and out and control the required hardware. The most used operating systems for cloud computing are Linux, OpenSolaris and Windows. Platform also contains virtualization software, like VMWare or Microsoft Hyper-V, and the needed drivers to operate hardware.

BI Applications refers to the set of tools that will enable the operation of the business intelligence solution. This layer is the one that differentiates business intelligence solutions with transactional ones. The main element for this layer is the database, which is the responsible of handling the data to be stored in the data warehouse. There are several databases in the market: Oracle, DB2, SQL Server, and MySQL are among the most popular. Some of these databases have already started working in versions enhanced for the cloud. Another type of database is the multidimensional databases, which do not store data in relational, but in proprietary form. Examples of these databases are Microsoft Analysis Services and Hyperion Essbase. These products should be seen as datamarts handlers instead of data warehouse. A datamart is defined as a set of data that contains information focused to one area or department in an organization. They are built to improve access performance or to generate very specific analysis and calculations.

ETL, as stated previously, is the tool in charge of extracting data from data sources, transforming data, and loading them to the data warehouse. Data warehousing tools are used to design, create, monitor and fine tune the data warehouse. Finally, BI tools are the set of applications that distribute information to final users, such as decisions support systems, dashboards, reports, or data mining applications.

The last element in the architecture is the client. Since almost all the processes needed in this approach are done by the servers in the cloud, there are little requirements demanded on client computers, and all the interaction is done through internet browsers on the user's equipment. For this reason is that many organizations have opted to use thin clients, that is, computers with reduced CPU and memory capabilities. Some thin clients don't even have hard disks.

A popular kind of client is mobile devices, such as iPhone, Blackberry, smart phones, and Amazon Kindle. Since cloud computing requires low processing and memory capabilities, these kind of device make a very good option.

6.1.2 Structural View

Figure 5 shows the system structural view, which includes the connections among the elements and recognize those subsystems that exist out of the system's boundaries.

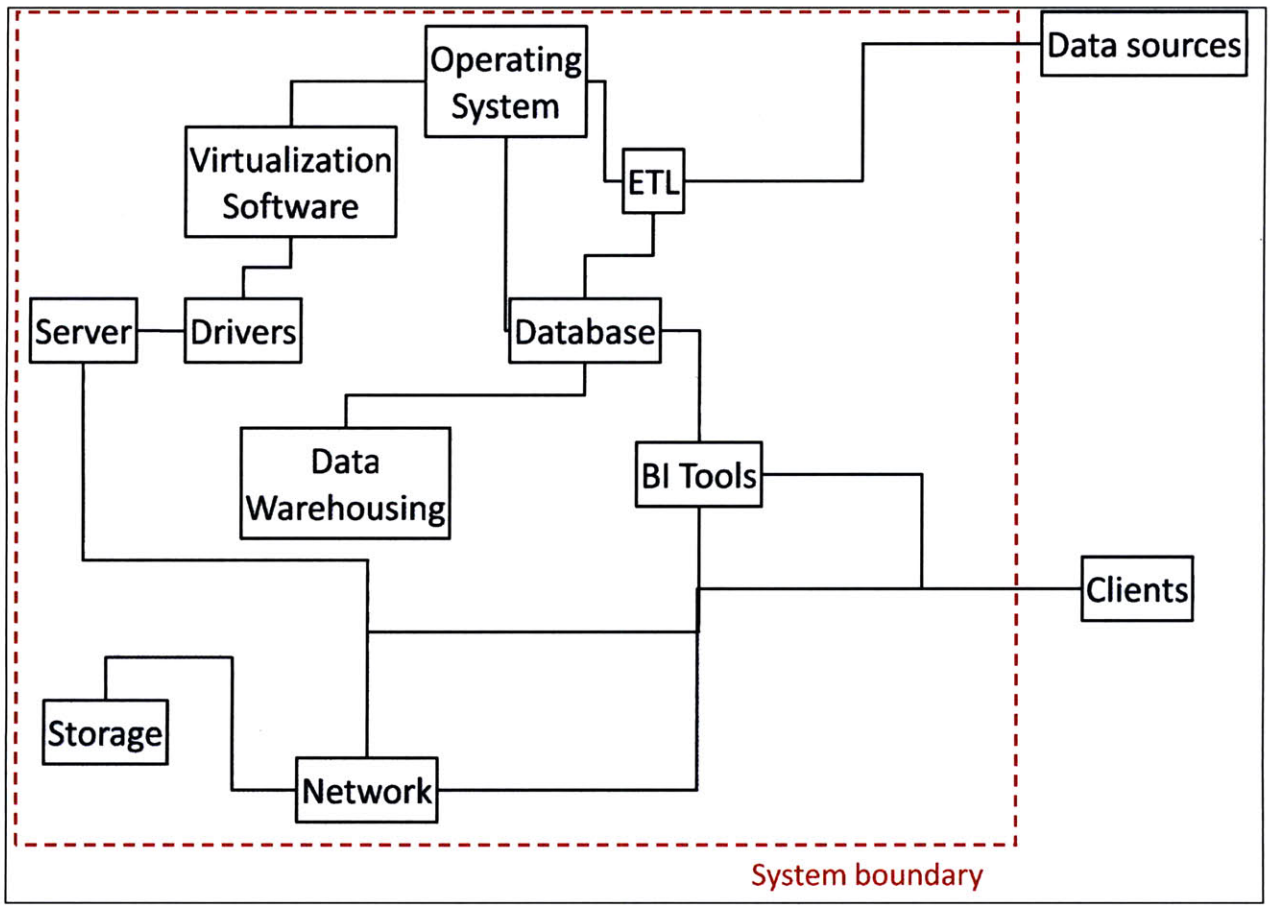


Figure 5. BI in the Cloud Structural View

As can be seen, the red dotted line separates subsystems that are inside from those that are outside of the system. Client computers and data sources are out of the system boundaries. Clients are out of the cloud, connected through internet connections and enabled by an internet browser. Although they are out of the system, cloud computing based solutions must make sure browsers are compatible with the applications and that they support all the features needed to execute the solution.

Data sources are also outside of the system. They are subsystems that contain the data to be extracted and may be found in several formats and geographically dispersed. Still, although they are out of the boundaries, the solution must make sure the source is reachable, data can be accessed in a promptly manner, and that is compatible with the ETL tools.

It is very important to identify these two elements that are outside of the cloud system, because they represent security concerns that need to be handled in order to provide a robust and secure cloud computing solution.

6.1.3 Architecture Matrix

Taking the main elements of a BI in the cloud solution, an architecture matrix was created in order to understand the interrelation among them.

Table 1 shows the architecture matrix for all these elements, identifying if they exchange information or are controlled by others.

	Server	Storage	Drivers	Virtualization software	Operating system	Data warehousing tools	Network	Database	ETL tools	BI tools	Data sources	Clients
Server	X	t	t	t	t		t					
Storage	t	X	t	t	t		t	t				
Drivers	c	c	X		t		c					
Virtualization software	c	c		X	c		c					
Operating system	c	c	c	c	X	t	c	t	t	t		
Data warehousing tools					t	X	t	c	t			
Network	t	t	t	t	t		X		t		t	t
Database		t			t	t		X	t	t		
ETL tools					t	t	t	t	X		t	
BI tools					t		t	t		X		t
Data sources					t		t		t		X	
Clients							t		t			X

Table 1. BI in the Cloud architecture matrix

t = Transfers information
c = Controls

Based on this matrix, we can see that operating system, ETL tools, and database (in this order) have the most interactions in the system. Special attention has to be taken considering security and reliability concerns.

6.2 House of Quality

The House of Quality tool will be used in order to define the relationship between the customer desires, in this case, the concerns or challenges, and the capabilities of current cloud computing solutions based on existing technologies.

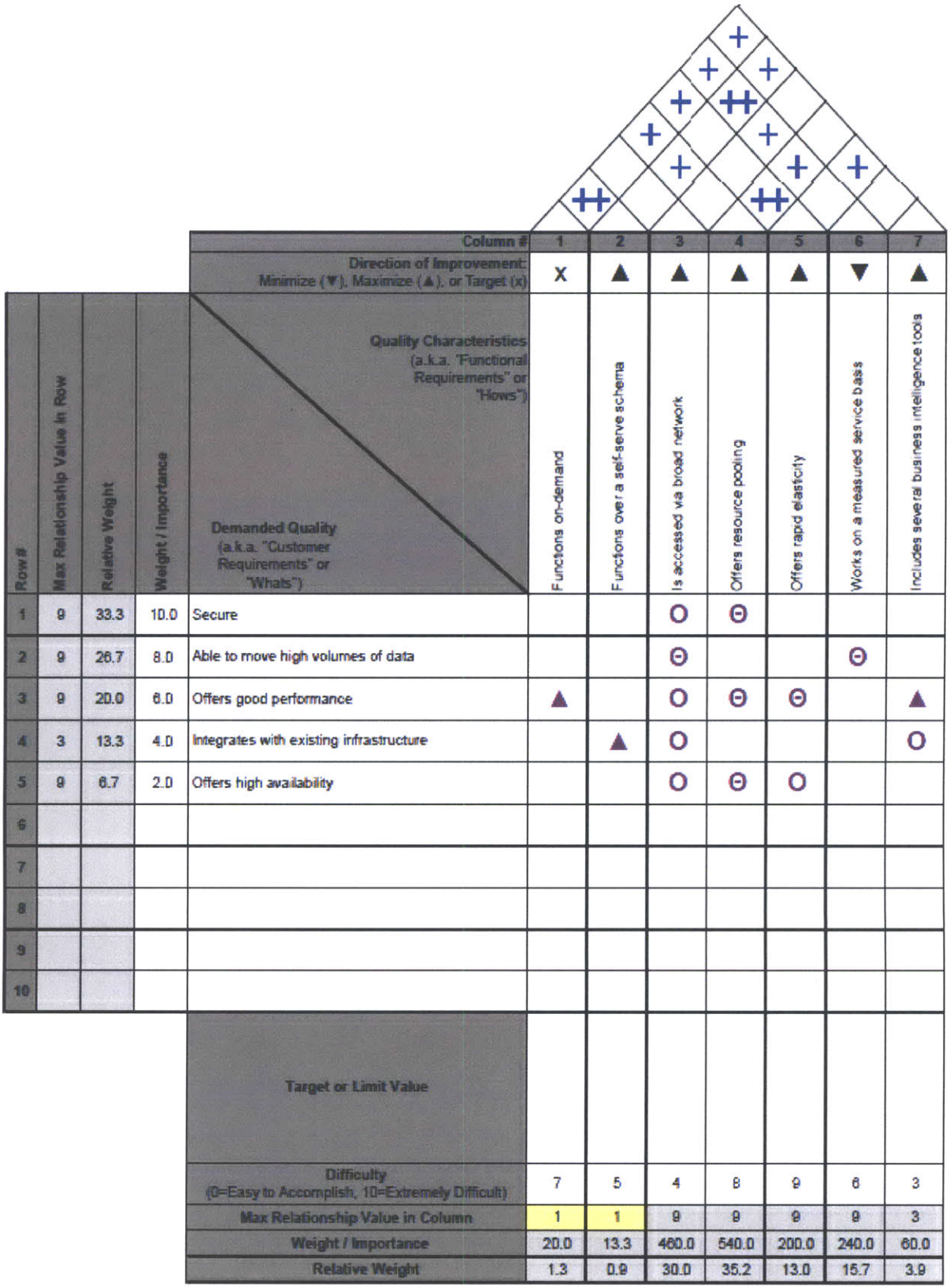
Figure 6 shows the House of Quality for business intelligence solutions based on cloud computing. The demanded quality elements are:

- Is secure
- Is able to move high volumes of data
- Offers good performance
- Integrates with existing infrastructure
- Offers high availability

And the quality characteristics are:

- Functions on-demand
- Functions over a self-serve schema
- Is accessed via broad network
- Offers resource pooling
- Offers rapid elasticity
- Works on a measured service basis
- Includes several business intelligence tools

Based on the results obtained in this house of quality, we can see that the two cloud computing characteristics that have to be addressed carefully to deal with the previously presented challenges are "Is accessed via broad network" and "Offers resource pooling". In the next sections, these two items will be mostly analyzed in order to find a solution for business intelligence applications working over cloud computing environments.



Row #	Max Relationship Value in Row	Relative Weight	Weight / Importance	Demanded Quality (a.k.a. "Customer Requirements" or "Whats")	Functions on-demand	Functions over a self-serve schema	Is accessed via broad network	Offers resource pooling	Offers rapid elasticity	Works on a measured service basis	Includes several business intelligence tools
				Column #	1	2	3	4	5	6	7
				Direction of Improvement: Minimize (▼), Maximize (▲), or Target (x)	x	▲	▲	▲	▲	▼	▲
1	9	33.3	10.0	Secure			O	Θ			
2	9	26.7	8.0	Able to move high volumes of data			Θ			Θ	
3	9	20.0	6.0	Offers good performance	▲		O	Θ	Θ		▲
4	3	13.3	4.0	Integrates with existing infrastructure		▲	O				O
5	9	6.7	2.0	Offers high availability			O	Θ	O		
6											
7											
8											
9											
10											
				Target or Limit Value							
				Difficulty (0=Easy to Accomplish, 10=Extremely Difficult)	7	5	4	8	9	6	3
				Max Relationship Value in Column	1	1	9	9	9	9	3
				Weight / Importance	20.0	13.3	460.0	540.0	200.0	240.0	60.0
				Relative Weight	1.3	0.9	30.0	35.2	13.0	15.7	3.9

Figure 6. House of Quality for Business Intelligence Solutions Based on Cloud Computing

6.3 Addressing Security

The main concern for cloud computing applications is security. There are some actions that can be taken in order to make sure that the information is safe and protected in the cloud. Those actions will be discussed below.

6.3.1 Security by moving applications to the cloud

The first idea that has to be considered is that in many cases the information that is moved to the cloud is more secure than in on-premises sites. Although some mature companies have state-of-the-art data centers with all the required elements to have applications up and running in high security environments, some others do not. Especially organizations that have grown without an IT strategy are more likely to have security problems and their information can be vulnerable to internal and external attacks.

Data centers offering cloud computing services are most likely to have advanced peripheral solutions to handle security, which in most cases, would be much better than the security implemented in traditional medium sized companies. For example, the security mechanisms implemented for Amazon EC^2 comply with all the set of certifications and accreditations required for robust data centers, as well as physical and logical network security[27]. By far, these procedures would keep data more secure than in traditional sites at the companies.

But still with this approach, there are some concerns that prevail, like possible attacks when information is transferred on the network to and from the cloud, and possible accesses to the data in the storage devices at the data center.

6.3.2 Network security

As a general rule, all information traveling through networks out of the organization (like Internet broadband) must be encrypted. SSL (Secure Socket Layer), and its successor, TLS (Transport Layer Security) are the standards in the industry. Figure 7 shows the flow that encrypted data must follow.

Talking about uploading information from the organization to the cloud (or from data sources to the cloud), data has to be encrypted. This job has to be done by the ETL tools, which are in charge of extracting data from data sources, encrypting it, sending it to the cloud, de-encrypting it, transforming it and then loading it into the data warehouse.

An alternative to encryption in the upload is opening a VPN (Virtual Private Network) connection from the organization to the cloud. In this way data could safely travel on the network. VPNs are mechanisms that are commonly offered by the cloud service providers and consist of a secure tunnel through which information can be transferred.

On the download flow, the process is easier because the client will always be an internet browser. Secure connections between applications servers and internet browsers are widely used via SSL connections, so it would not represent a big challenge. Most of SSL connections used nowadays encrypt data with 128-bit keys. It is said that with current traditional computing technologies, "it would take significantly longer than the age of the universe to crack a 128-bit key"[28].

The objective of these security tasks is that no intruder could get the information while is transferred on the network. Even by getting data via mechanisms like sniffers, the obtained information would not make sense since it would be encrypted.

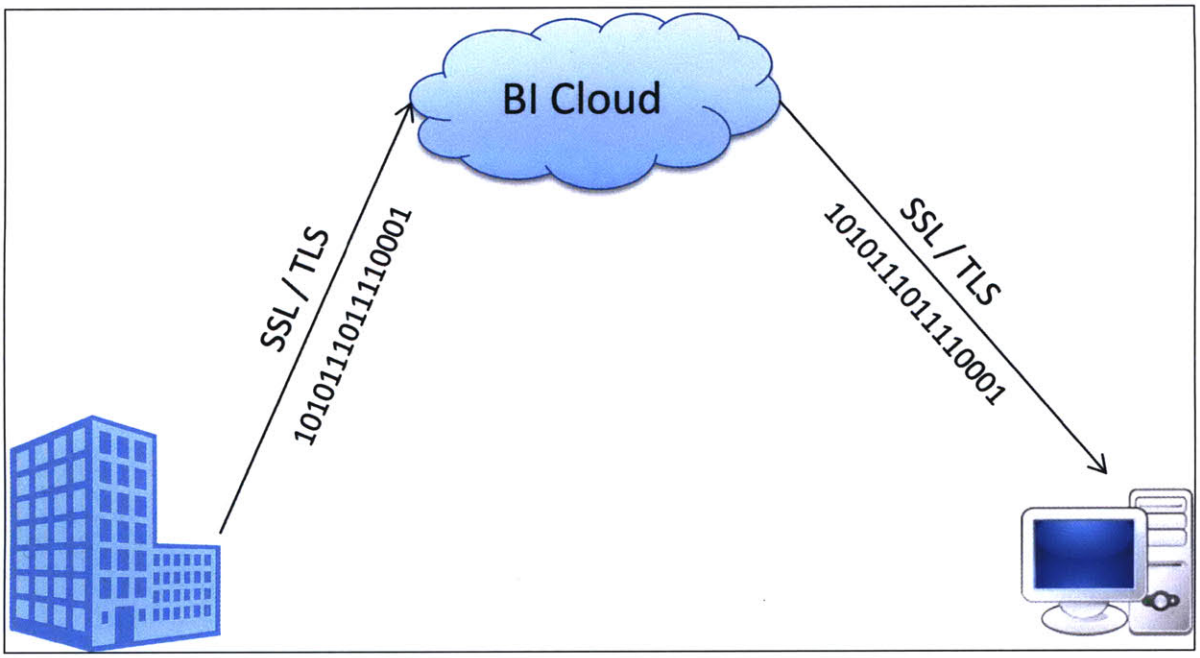


Figure 7. Flow of Encrypted Data.

6.3.3 Encrypting data in the database

A way to keep the data secure once it has reached the cloud is by encrypting the content in the database. In this way, even if an attacker is able to penetrate the security mechanisms and access the data warehouse database, he would not be able to understand the data, because he would need the key in order to de-encrypt the information. A disadvantage to this method is that database performance may be hindered since encryption tasks are CPU intensive. An example of this is encryption on Microsoft SQL Server 2008, on which the impact has been estimated to be between 4 and 5%[29].

6.3.4 More data can contribute to information security

This approach can take advantage of proper cloud computing characteristics. Platforms in cloud computing may distribute data in several places, starting from another server, another rack, or even another data center geographically dispersed. So, after the question: where the data are? The answer is uncertain, because nobody would exactly know where they are, having distributed systems like those described above. Although this dispersion could be seen as a disadvantage, in reality it is an advantage for data security, because in the case of a physical attack, the hacker would not know where the information is stored[30]. Also, if the information is partitioned and dispersed among several storage systems, getting small chunks of data would not make sense by themselves and the attack purpose would fail. Figure 8 shows the reinforcing loop that supports this theory, based on System Dynamics.

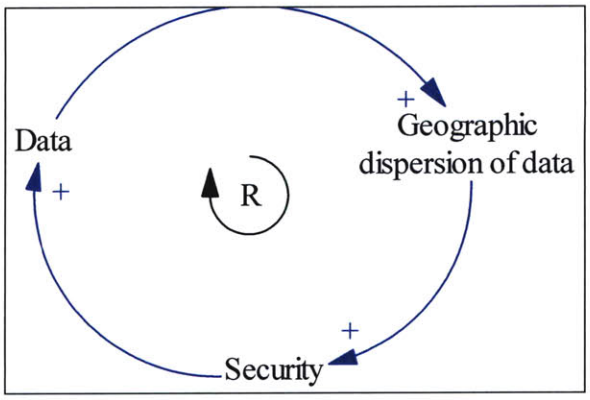


Figure 8. Data Volumes and Security Reinforcing Loop

The model indicates that as more data is uploaded to the cloud, more likely the data is going to be geographically dispersed. This phenomenon is based on the resource pooling characteristic of cloud applications. If not enough computational resources are available in the same data center, the cloud would automatically pool resources located in several places. Having data geographically dispersed, less successful attacks would be registered, contributing to increase data security. If customers identify that their data is secure with their service provider, then customers would be willing to upload more data to the data warehouse and continue using cloud computing. This all together forms a reinforcing loop in benefit of data security.

6.3.5 Regulatory issues

There are some situations in which even the application of the mechanisms described above would not be enough to keep data in the cloud. Those cases are mainly related to regulatory concerns. There are some laws that prohibit keeping sensitive information in places different to the owner's facilities, like health information. Other laws prohibit keeping data in servers outside of the country of origin. In these cases, different approaches have to be taken if organizations want to take advantage of cloud computing. Some alternative are private clouds or hybrid solutions. Both approaches will be discussed in further chapters.

6.3.6 Components Isolation

Some cloud computing providers have implemented isolation mechanisms in order to increase security and privacy for their customers. An example for this is RightScale, who gives to its customers a private virtual disk that is not shared with any other customer. With this functionality, low-level encryption can be done on the disks, besides that information is kept segregated.

Another component of the cloud infrastructure that can be isolated is the network. Through a private virtual network all the traffic can be segregated, so that only information from one customer would be traveling in the same virtual channel, so that no other individual could have access to the data.

6.3.7 Intelligent workload management

Under an intelligence workload management approach, companies can have certainty on where the information is stored. In this way, that information would be in devices residing in a country,

avoiding infringing laws that prohibit having information outside the country where it is generated. An example of this approach is offered by Amazon, with its Amazon S3 (Amazon Simple Storage Service. Customers can choose in which region their information will be stored. Figure 9 shows Amazon S3's regions and rates. Customers may choose if their information is going to be kept in US Standard, US North Carolina or in the European Union, in Ireland. This segmentation, besides incrementing performance and security, helps to comply with regulatory restrictions.

US – Standard | US – N. California | EU – Ireland

Storage		Data Transfer		Requests	
Tier	**Pricing**	**Tier**	**Pricing**	**Tier**	**Pricing**
First 50 TB / Month of Storage Used	$0.150 per GB	All Data Transfer In	Free until June 30th, 2010*	PUT, COPY, POST, or LIST	$0.01 per 1,000 Requests
Next 50 TB / Month of Storage Used	$0.140 per GB	First 10 TB / Month Data Transfer Out	$0.170 per GB	GET and All Other Requests*	$0.01 per 10,000 Requests
Next 400 TB / Month of Storage Used	$0.130 per GB	Next 40 TB / Month Data Transfer Out	$0.130 per GB	* No charge for delete requests	
Next 500 TB Storage Used	$0.105 per GB	Next 100 TB / Month Data Transfer Out	$0.110 per GB		
Next 4000 TB Storage Used	$.080 per GB	Data Transfer Out / Month Over 150 TB	$0.100 per GB		
Storage used over 5000 TB	$.055 per GB	* Data Transfer In will be $0.100 per GB after June 30th, 2010			

Figure 9. Amazon S3 Regions and Rates

6.3.8 Private Cloud

Private clouds are a reduced version of public clouds that increase security by providing services only to an organization or department. The computing equipment is operated by the organization and can be on-premises or off-premises. The advantage of this approach is that the information is only shared within a local area network and is restricted to the organization and its departments. A disadvantage is that the companies have to invest still in infrastructure to support the private cloud and may not get the benefits of economies of scale that service providers achieve.

Two versions of private cloud have been created in the market: an actual private cloud, which has been promoted mainly by IBM through its Smart Analytics Cloud, offering an all-in-one solution for internal clouds. The other version of private cloud is promoted by Amazon via its Virtual Private Cloud, which increases security by creating a bridge between an organization IT infrastructure and the Amazon cloud services by the use of a VPN.

Although both approaches increase security in the cloud, Amazon's Virtual Private Cloud would still be restricted by regulatory issues and may not be suitable for every organization. On the other hand, IBM's Smart Analytics Cloud offers regulatory compliance and security, but requires high investment of money to have it running on-premises.

6.3.9 Hybrid Solution

Another approach to address security in cloud computing is having a hybrid solution, composed of a public cloud and a private cloud, or a public cloud and traditional computing on-premises computing infrastructure and applications. The main objective of this combination is to keep sensitive data in the organization and only move public or semi public data to the cloud. This approach will be discusses later since hybrid solutions help to address several challenges for the business intelligence solutions on cloud computing.

6.4 Addressing High Volumes of Data

The second most important challenge for BI in the Cloud is the volumes of data that have to be transferred to the cloud. Although there are not straight solutions to address this concern, some general ideas and current trends will help to overcome the data movement from the organizations.

There are two constraints that affect data volumes in the cloud: the price to be paid for transferring data and the availability of enough bandwidth. One of the characteristics of cloud computing is that it is a measured service, and one of the elements that are metered is the number of gigabytes transferred. For example, Amazon charges $0.10 per GB transferred in, and up to $0.17 per GB transferred out. Business intelligence and analytic applications tend to be more data expensive than transactional systems, then this fee could increase the monthly price considerably.

From the customer side, there is little that can be done in order to avoid these fees, but from the provider side, fees could be reduced or even eliminated for applications running business intelligence in order to promote their services. Some companies have started reducing fees as data volume increases. Table 2 shows the pricing per GB transferred from Amazon EC2. These amounts tend to reduce based on economies of scale.

Data Transfer Out	Price
First 10 TB per Month	$0.17 per GB
Next 40 TB per Month	$0.13 per GB
Next 100TB per Month	$0.11 per GB
Over 150 TB per Month	$0.10 per GB

Table 2. Amazon EC2 Pricing for Data Transferred "In"[31].

The trend on communications indicate that more bandwidth will be available by less cost, so as cloud solutions penetrate the market, the amount of data that organizations will be able to transfer through broadband networks will be higher and with lower cost.

6.4.1 Data Compression

Data compression is a technique that can be followed in order to reduce the amount of information to be transferred on the network. Plain data has shown to be highly compressible without the danger of losing exactness. It has been demonstrated that compressing ratios can be as high as 80%[32]. In this way, data could be compressed just after extracted from the data source and just before transferring it. Once data reaches the cloud, it would have to be decompressed, then transformed and loaded to the data warehouse. Figure 10 shows the data flow of compressed data.

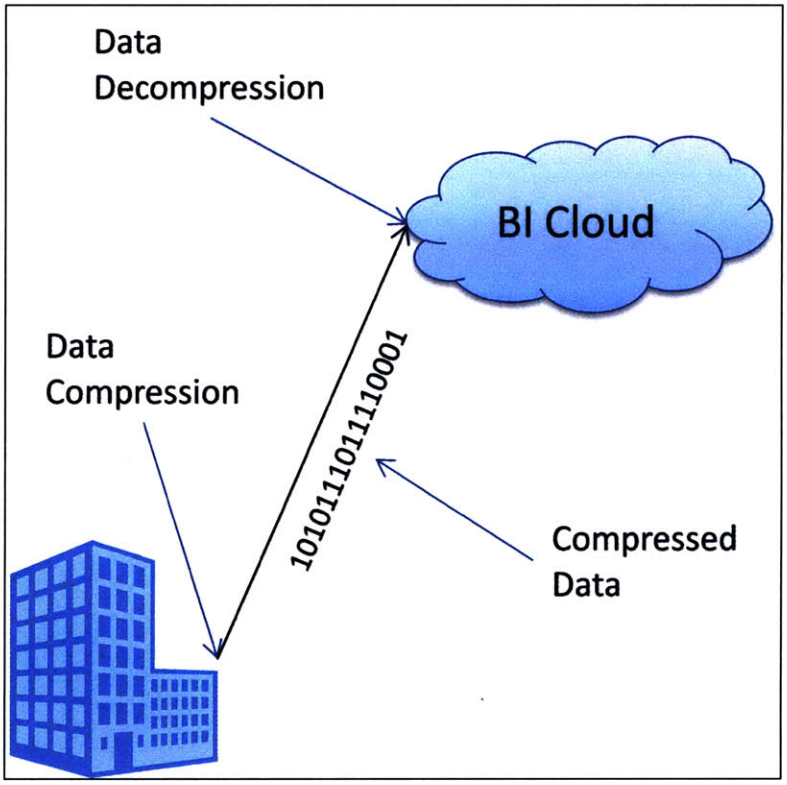


Figure 10. Compressing Transferred Data

With this idea, the amount of transferred information would be reduced as low as 20% of the actual size, thus decreasing the transfer times during data extraction and load. The challenge of this approach is to get the ETL tools working properly and compressing and decompressing data at the adequate times. Currently, not many ETL tools do these tasks natively, meaning that additional programming has to be done. Fortunately, most of the new tools are including web services as part of their integrated functionality, which will help to build the compression algorithms accordingly. Another option for data compression is that some VPN solutions compress data on transit. In this way, the ETL tools would have much less overhead during load times.

6.4.2 Data Sources in the Cloud

The best way to reduce data latency and transfer to the cloud is by having the data sources, like ERP systems, also in the cloud. With this approach, it is not tried to say that all data sources should be moved to the cloud just because the organization has a business intelligence

solution, but if the organization already has part of its data sources in the cloud, and with the same service provider, then data transfers would not be a concern while building BI in the cloud.

6.4.3 Avoid Querying the Data Warehouse

Many analysts are accustomed to send large queries to the data warehouse as part of their daily analysis tasks. This practice, besides overloading data warehouse, can consume large bandwidth while the results are transferred to the analyst computer. In local area networks this behavior may be almost unnoticeable, but with a constrained broadband connection to the cloud it can be painful. A solution to this kind of query is to keep access to the in-cloud data warehouse only through light clients, like internet browsers, and keeping all the filters and aggregations in the cloud. Analysts would not be allowed to download large amounts of information to their computers, but would receive the same quality of results. The response time would be even greater, because processing and I/O operations would take place in the cloud, not in a local machine.

Another approach to reduce data upload to the cloud is by having a hybrid solution, in which only aggregated data would be uploaded to data marts. This architecture will be explained later in this document.

6.4.4 Accelerated Internet

Organizations can take advantage of special network appliances that improve the connection between two distance points. One of the points is the cloud provider and the other the organization's data center. There are several solutions on the market that support data compression and encryption. The final result is that the connections are faster and more reliable. An example of technology companies offering this type of appliances is F5, with its WANJet[33].

6.5 Assessing Performance

By definition, performance of business intelligence solutions on the cloud should be better than on-premises solutions, because applications can take advantage of rapid elasticity and resource pooling. These two characteristics are difficult to achieve with conventional computing architectures. However, there have not been large scale data warehouses or business intelligence solutions so far in the cloud. The largest data warehouses have been built on-premises and have required large investments of money.

The world's largest data warehouse is supported by Sun data warehousing architecture and Sybase IQ columnar data base. Before these two companies joined efforts, the largest data warehouse was held by Wal-Mart on a Teradata machine.

To understand business intelligence solutions performance, the general architecture will be divided in three parts.

Figure 11 shows the parts that form the basic architecture: ETL, Data Warehouse and BI Applications. Data Warehouse has shown to present the highest performance challenge in this kind of solutions. The three elements will be assessed below in order of performance concern, from the least to the most: BI Applications, ETL and at the end, the data warehouse.

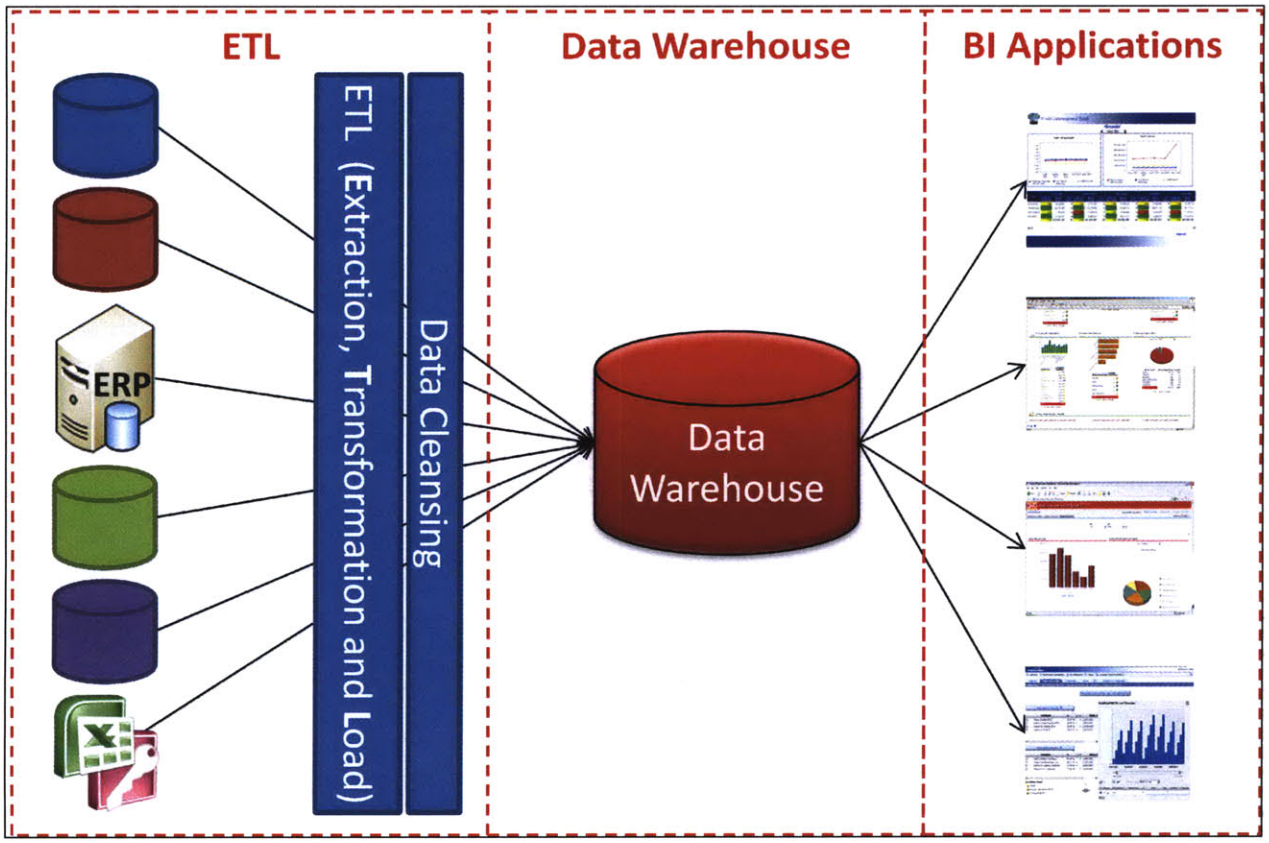


Figure 11. Business Intelligence Solutions' Three Main Elements

6.5.1 BI Tools Performance

BI tools are by far the part that can be more easily integrated to the cloud, and at the same time, take advantage of its benefits to improve performance of the overall solution. Although at the beginning most of the BI tools were intended to run with thick clients, nowadays almost all of the clients are based on internet browsers, offering great compatibility and flexibility. Organizations have run business applications from data centers for some time and have gotten good results.

Most of the BI tools require an application server to run. This application server works as a bridge between the data warehouse and the clients by parsing the queries and encapsulating the data that is sent to the final users. Since the application servers work based on layers and isolate the operating system from the users, the elasticity and scale up and down operations done by the cloud platform will be practically transparent for the users. The general aspect to keep in mind is licensing, because many of the applications servers are licensed by number of processors, but this concern can be easily solved by the service provider with the help of virtualization.

In conclusion, the BI Tools section should not really be affected by performance taking it to the cloud.

6.5.2 ETL Performance

ETL tools are perhaps the part that has been thought the least as a candidate for the cloud. Conventional developments of ETL tools require high customization, deep understanding of data sources, and connection in the same local area networks. There are four sub-elements that may hinder ETL performance:

- Access to data sources
- Data transfer
- Data transformation
- Insertion into the data warehouse

Access to data sources is an issue that would be present also in on-premises development. Every business intelligence project has to deal with it and is not exclusive for the cloud. The performance depends completely on the data sources. Although there are best practices to query information more efficiently from databases, and source systems can be tuned to get better performance, those techniques are out of the scope of this document since they are not limited to cloud computing solutions.

Data transfer problems, although in much less quantity, are also found in on-premises paradigm, mainly caused by high traffic or inappropriate configuration of the communications devices. The two approaches previously discussed in the "Addressing High Volumes of Data" section can help greatly to reduce latency on data transfer in the ETL process: data compression and data sources in the cloud.

Data transformation is the most important issue when talking about ETL performance. This process is very CPU and memory intensive. Conventional solutions usually assign one big box just to manage ETL tasks. Under cloud computing, performance should not be a problem anymore thanks to automatic and rapid elasticity. ETL process is one of the best examples in the computing industry that takes advantage of elasticity: during night ETL processes require high computing power, so resources can scale out in order to provide enough CPU and memory; as soon as ETL processes finish, resources scale in and release capabilities so that other processes can take advantage of it.

Insertion into the data warehouse is the second most important issue related to ETL performance. Since the insertion speed relies the most on the data warehouse per se, this is a matter of data warehousing performance, which will be discussed in the next section.

Under this understanding, we can conclude that performance is not an issue that directly affects ETL processes in the cloud, but may improve them if all the recommended tasks are followed and if the data warehouse solution is effectively taking advantage of the cloud benefits.

6.5.3 Data Warehouse Performance

Data warehouse industry has had important improvement over time. Technology has even offered several kinds of architectures in order to handle higher volumes of information, load data faster, and resolve queries quicker. Some vendors now have started to push their offering toward the cloud, like Greenplum or Teradata.

Data warehouse architectures can be divided in three groups: traditional, appliance, and columnar databases. The three of them try to maximize results while minimizing cost. Let's examine each one in order to understand which architecture would fit the best for cloud computing.

6.5.3.1 Traditional Architecture

The traditional architecture has been in the market for more time, has come as a direct evolution from OLTP (Online Transaction Processing) systems and has been adapted to work with data warehouse and analytic environments, that is, to improve data read rather than write. This architecture is formed by hardware (CPU, memory, etc.), software (operating systems), RDBMS (Relational Database Management System), and storage. These three parts are usually each supplied by different vendors.

The main advantage of this approach is the openness of its elements. Each element can be changed by another from other vendor if the rest of the parts are compatible, offering a great combination of alternatives.

Figure 12 shows the theoretical possible combination of these elements. Note that not all the possible vendors were included and not all the combinations are possible due to compatibility reasons.

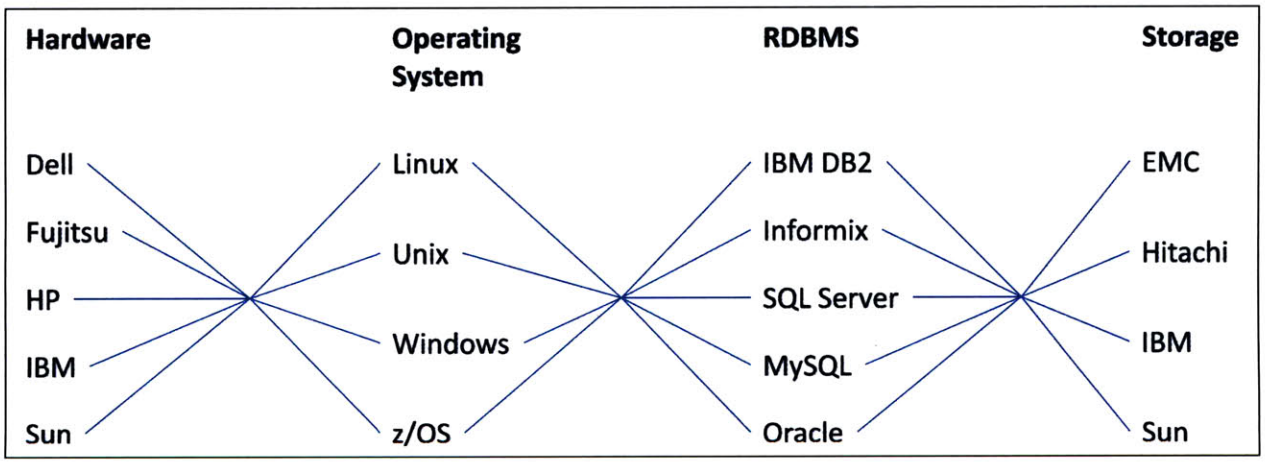


Figure 12. Combination of Elements for Traditional Data Warehousing Architectures

The main disadvantage of this approach is that costs may be higher because each of the elements has to be bought separately and then integrated. Also, specialized professional services are needed. Companies need experts on hardware, operating systems, RDBMS and storage to have the system running, making the administration even more complex.

Expanding capacities is another challenge. When the data warehouse has to be expanded it may require changing all the elements in order to get better performance. In this scenario, more resources are needed and downtime may affect the organization's operations.

Thinking of the cloud, however, this architecture may be the best suitable. As far as the service providers transparently administer the cloud platform (including hardware, operating system,

and storage), the elasticity mechanisms could be irrelevant for final users. The service provider can change all elements except the RDBMS. A change on the RDBMS would not be transparent because it could affect compatibility with the rest of the business intelligence architecture. For example, ETL tools may not have the proper connector to the new RDBMS containing the data warehouse, or BI tools could not be compatible with the new database.

This issue can bring a new challenge to BI in the cloud. Since traditional databases were not designed to run in the cloud and to take advantage of its characteristics, new database management systems need to be integrated. Examples of cloud oriented database systems are:

- Greenplum
- Amazon SimpleDB
- Microsoft SQL Azure

Of these databases, the only one completely focused to analytic environments is Greenplum. This company has moved rapidly into the cloud and has the advantage of its compatibility with some of the mayor players in the business intelligence industry, like BusinessObjects, SAS, Pentaho, AbInitio and Informatica[34].

On the other hand, Amazon SimpleDB is still a Beta product featuring applications like logging, online games and metadata indexing[35]. So far no business intelligence tools exist that are compatible with this data base.

Microsoft SQL Azure is very new in the market. It was recently launched by Microsoft, and although no business intelligence vendor has announced to be compatible, integration could be achieved by the use of its native ODBC (Open Database Connectivity) connector.

6.5.3.2 Data Warehouse Appliances

Data warehouse appliances are devices that contain in only one box all the elements to run a data warehouse: hardware, operating system, RDBMS and storage. In theory, less knowledge and administration effort are needed to operate these devices. Examples of data warehouse appliances in the market are:

- DATAllegro
- HP Neoview
- Netezza
- Teradata

Data warehouse appliances were created under the approach of shared-nothing and high parallelism. Some appliances have also focused on scalability by connecting new nodes without having to shut down the system.

The main disadvantage of data warehouse appliances is that they may not be completely ready to run on cloud computing environments, inhibiting fast elasticity. Also, some of them have proprietary components that may be difficult or impossible to replace while in the cloud.

6.5.3.3 Column-Oriented Databases

Column-oriented databases are a new paradigm in database technology that improves data access speed. Data are organized in columns instead of rows. This approach reduces the time required to get data from a query, because the database only reads those columns that are relevant, instead of reading the entire row. Besides improving data access, columnar databases reduce the disk space needed to store information, because data is compressed when is loaded to the system.

The architecture needed for columnar databases is similar to the traditional solution, but in this case, the relational database is changed by a columnar database. Also, vendors claim that required computing power is less compared to conventional relational databases. Some of the columnar databases in the market are:

- Sybase IQ
- Vertica
- ParAccel

The use of column-oriented databases is relatively new in the industry. Their use can also help to address integration with ERPs. This approach will be discussed later in this document.

6.6 Assessing Integration with Existing Infrastructure

The key for a good integration of business intelligence solutions in the cloud with existing infrastructures in the organizations resides on ETL tools. Three main challenges have to be dealt with in order to guarantee a correct integration:

- Data extraction from data sources
- Data transferring
- Data loading into the data warehouse

Data extraction from data sources depends much on the connectors the ETL tools have. As mentioned previously, modern ETL suites are compatible with most of relational database systems and unstructured sources of data. Also, through the use of ODBC and JDBC, connections can be achieved to new databases as far as they support these standards.

If the ETL has the capacity to connect to a data source, then there would not be any problem while trying to access it from the cloud, as far as the network connection and ports are available during the connection. In this sense, extracting data from cloud computing would be the same as on-premises.

Informatica is an example of company offering business intelligence suites. They have announced the new version of their product named Informatica 9, which is focused to the cloud. Table 3 shows some of the databases that are supported by this platform. This table represents only a subset of data sources, since according to the company, its tools enable the access to "virtually any and all enterprise data types"[36].

Enterprise Data Stores	Source	Target
Adabas for UNIX, Windows	X	X
Adabas for z/OS	X	X
Binary Flat Files	X	X
C-ISAM	X	
Datacom	X	X
DB2 for i5/OS	X	X
DB2 for Linux,	X	X
UNIX and Windows	X	X
DB2 for z/OS	X	X
Essbase	X	X
IDMS	X	
IMS DB	X	X
Informix	X	X
Dynamic Server	X	X
ODBC	X	X
Oracle	X	X
SQL Server	X	X
Sybase	X	X
VSAM	X	X

Table 3. Informatica 9 Supported Databases and Interfaces

The data transferring concern on infrastructure integration has been discussed previously. The same approach has to be used in order to extract data from data sources in the organizations and loading into the data warehouse.

Finally, regarding integration of data loaded to the data warehouse, the same principle as in extracting data is applied here. If the ETL tools have the connectors to the database containing the data warehouse, the integration can be natural. Table 3 also shows the target data stores that are supported by Informatica 9. It is important to mention that Informatica 9 has been used as an example in this section, but all major ETL suites offer similar connection capabilities.

6.7 Assessing Availability

Availability is the fifth concern IT professionals have regarding cloud computing. Business intelligence applications have become mission critical systems for several organizations. That is why it is vital to keep these solutions up and running as much as possible; otherwise, daily operations and decision making can be affected.

It is important to differentiate two terms related with operational hardware and software, that is, reliability and availability. Sometimes they are used interchangeably, which is incorrect. Although they are highly related, they refer to different things. According to the IEEE, reliability means that a system will perform its intended function for the required duration within a given environment[37].

Availability is defined as the probability that a system is operating properly when it is requested for use[38]. This probability also considers the time in which the system is under a repair action. Then, availability is a function of reliability and maintenance operations applied to the system.

Since cloud computing suggests that the applications are used under demand, that is, when the users need them, then availability is the characteristic that has to be evaluated in order to measure is the system is able to work properly when the customers require it. In general, a cloud computing paradigm will increase availability by offering virtualized systems running on more than one physical server.

The general formula for defining the availability in a system is:

Where is the unavailability of the system at time .

When we say that cloud computing increase availability is based on the fact that most of the services offer redundancy on physical servers, obtaining a system that can be expressed on Figure 13.

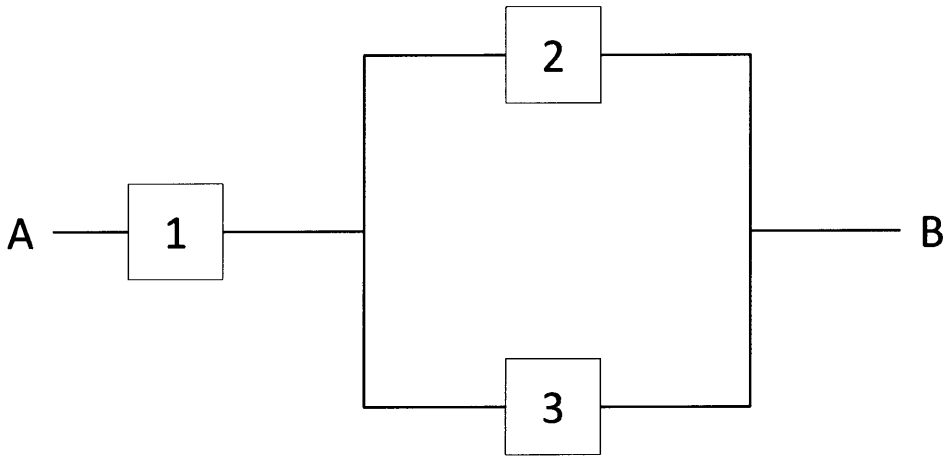


Figure 13. Representation of a system in the Cloud with redundancy

Figure 13 can be explained in the next way: an output B is expected to be solved by the system given an input A. Input A is processed by the software "1". Software "1" is run on server "2" OR server "3". With this architecture, output B can be obtained even if server "2" or server "3" is down, but not both.

With this understanding, the minimal path sets for success are:

And the structure function for success is:

Supposing that the documented availability of the software (box "1") and hardware (boxes "2" and "3") are equal to 90%, we would have a general availability of 89.1% by substituting all the "Ys" on the structure function of success.

Now let's compare the same systems but without redundancy, that is, without more servers available in the cloud. Figure 14 shows the representation of the new system.

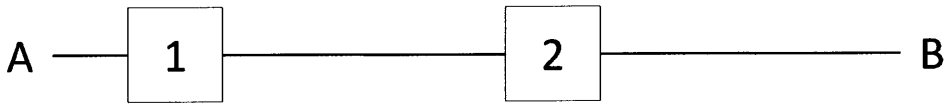


Figure 14. Representation of a system on-premises without redundancy.

In this case, there is not a server "3" providing redundancy, so, there is only one minimal path set for success: , resulting in the next structure function for success: . Finally, substituting the same documented availability by both "Ys", we get a general availability of 81%, which is 8.1 points lower than the cloud version with redundancy.

As mentioned before, maintenance is also part of the availability on systems. When something fails in any system (which could be hardware or software), there is a specific time to fix the error. This is known as Mean Time To Recovery (MTTR). The lower time the MTTR takes, the higher the availability. It is understandable that MTTR is lower in a cloud computing environment. For example, if a physical resource fails while operating (for example, a hard disk is broken), the service provider has specialized staff working 7 x 24, and may have available spare parts ready to be installed. This scenario would be difficult to have in on-premises IT scenario.

A cloud computing application can be defined as a continuously monitored repairable component, which shows a behavior similar to the depicted on Figure 15. The X axis represents the time, and the Y axis represents if the system is up (0) or if it had a failure and is under repair (1).

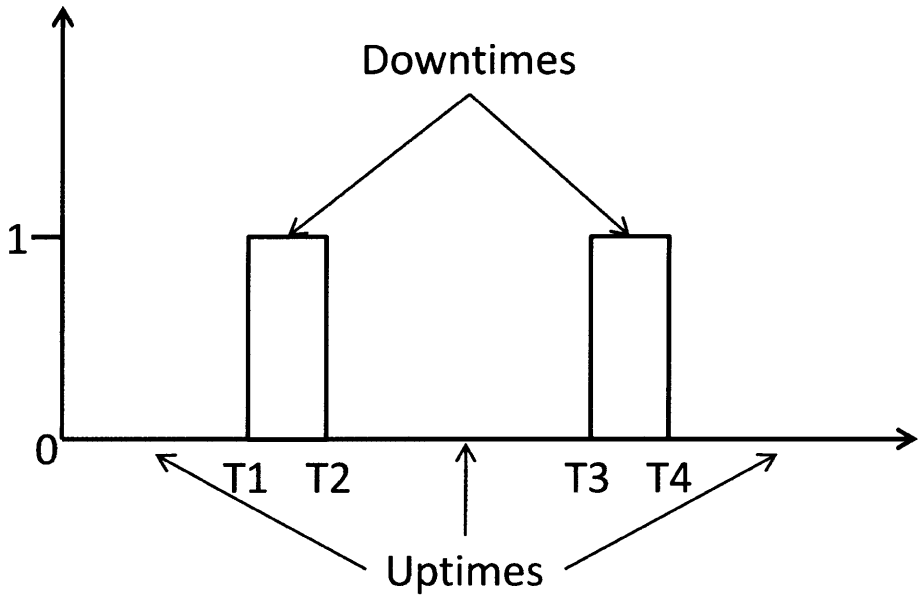


Figure 15. Behavior of Continuously Monitored Repairable Components

For a higher availability, downtimes have to be the least as possible, that is, the rectangles shown in Figure 15 have to be as thin as possible. Based on previous explanations, this is something that can be effectively achieved on cloud computing systems compared to those running on-premises. With these elements, the average availability can be calculated by taking the Mean Time To Failure (MTTF) and the Mean Time To Recovery in the next equation:

$$a = 1 - \frac{MTTR}{MTTF + MTTR}$$

Based on the previous equation, we can see that by reducing the mean time to recovery and increasing the mean time to failure, the general availability increases. Both variables can be successfully improved with business intelligence running on the cloud.

As shown above, it can be mathematically proved that the availability of a system in the cloud increases by the use of this new approach's characteristics. The next sections will show how availability is increased viewed from different perspectives.

6.7.1 Increasing availability by moving applications to the cloud

One of the objectives of cloud computing is to increase the availability of the applications it supports. This benefit is achieved by resource pooling, elasticity, and redundancy. Under this approach, it is logical that many organizations will increase systems availability just by migrating their applications to the cloud. This case is very similar to the security issue discussed earlier.

Many medium sizes companies do not have the required infrastructure to operate state-of-the-art data centers to host their systems, but cloud computer providers do. According to the Uptime Institute, there are four levels of availability depending on the infrastructure existing in the site[39].

Table 4 shows the types of tiers, the technology description they have, and the percentage of availability according to the UpTime Institute.

Tier	Description	Availability
Tier I	Basic Data Center Site Infrastructure	99.67%
Tier II	Redundant Site Infrastructure Capacity Components	99.75%
Tier III	Concurrently Maintainable Site Infrastructure	99.98%
Tier IV	Fault Tolerant Site Infrastructure	99.99%

Table 4. Tier Standards According to the UpTime Institute

To mention an example of a cloud computing provider, Amazon EC2 offers an availability of 99.95%, which is between tier II and III. This availability is by far higher than the offered by conventional on-premises sites for most of the organizations.

Under this understanding, we can conclude that for most of the customers, cloud computing per se offers better availability. The only availability constraint could be for migrating applications to the cloud when current on-premises systems have a higher availability rate than those offered by the chosen cloud computing provider.

6.7.2 Redundancy

One of the most common solutions for improving availability is redundancy. Cloud services providers are more likely to have redundancy for all the components required to run the applications, such as servers, storage devices, network equipment, connections, and power supplies. Again, many of mid-sized companies do not count with redundancy in their traditional data centers.

6.7.3 Synchronization and local backups

Another alternative to increase availability is to have multiple redundant systems containing the information. That is, keeping local backups of the information that is contained in the cloud. Local and remote data would be synchronized periodically. In the remote case that the cloud service is not available, the organization can have an emergency backup of the information. It is important to note that this alternative should only be used in extreme situations, but is worthier to keep in mind as a safe measure.

7 Other Approaches

In this section three new approaches will be discussed to address the challenges mentioned before. It was decided to address these approaches in a different chapter because they present solutions for more than one concern at a time. The approaches to be discussed are Hybrid, Column-Oriented Databases and Data Warehouse 2.0 (DW 2.0®).

7.1 Hybrid Approach

The hybrid approach joins two types of architectures that could be:

- Public Cloud with Private Cloud Computing
- Public Cloud with On-Premises Conventional Architecture

The main idea of this approach is to keep detailed, massive and sensitive information on-premises and only load into the cloud aggregated information. The detailed data warehouse would be kept on-premises with a relational schema, and the aggregated data would be uploaded to the cloud in form of data marts or cubes. On this way, only anonymized information would be outside of the organization's facilities. The application servers would also be located in the cloud in order to take advantage of elasticity and high availability. Figure 16 depicts this approach.

Due to this approach, the next challenges would be addressed:

Data security. Only anonymized or public information would be exposed to the cloud, and thus, an attack may be less dangerous or could take information that can be shared without any risk. Details like names, addresses, and individual amounts would remain in the company's site.

Massive data upload. Since only aggregated information would be uploaded, the volumes of data would be reduced dramatically. Depending of the level of detail that is going to be kept, the ratio of reduction could be of up to 250:1.

Integration. Data extraction, transformation and load would be done on-premises, just as conventional business intelligence projects do. In this part of the process there would no need of cloud sophisticated mechanisms.

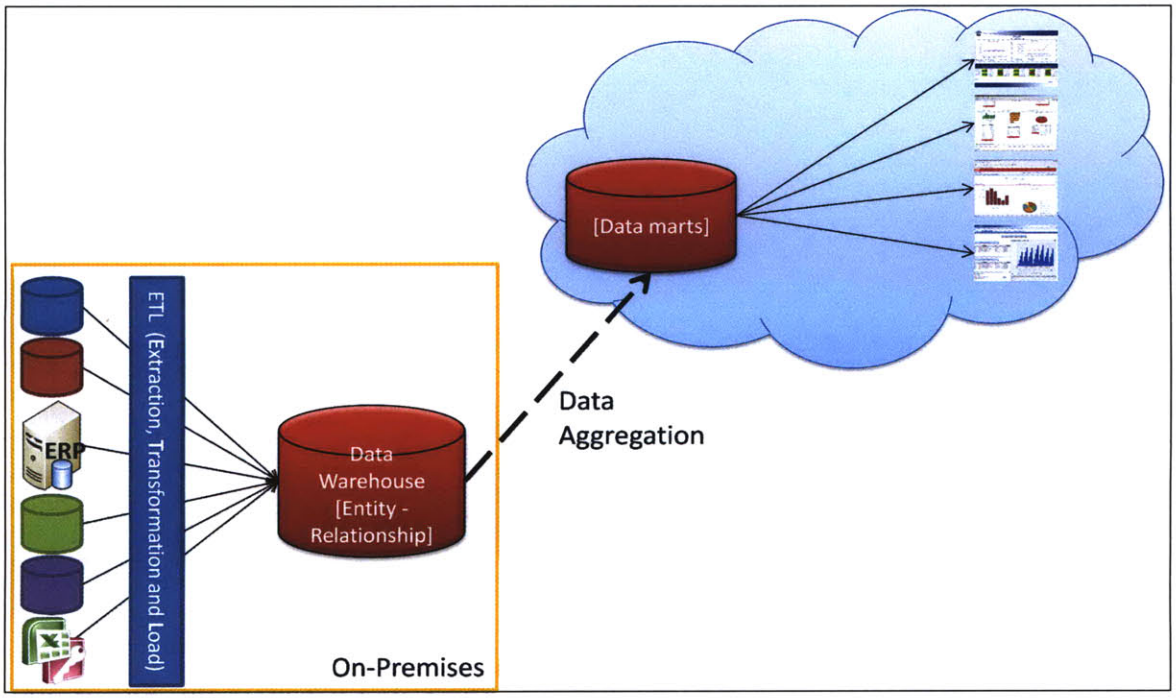


Figure 16. Hybrid Approach for Business Intelligence

However, some disadvantages arise with this approach:

- This architecture would not take advantage of all the cloud computing benefits.
- If the architecture considers a private cloud or on-premises architecture, large amounts of money and effort would be invested.
- Hardware, software and administration effort will be needed in order to maintain the data warehouse up and running.

7.1.1 Example for hybrid approach

To understand how the hybrid approach can be accomplished by maintaining detailed data on-premises, and aggregated anonymized data on the cloud, this section will show a case using sample data.

The sample data used for this case is part of the Microsoft SQL Server Community Projects & Product Samples. The chosen database is AdventureWorks, which contains information about a fictitious bicycle store[40]. The AdventureWorks database besides being fictitious, offers a good set of data that has been used for educative and tests purposes, and for the interests of this document, the sample shows how data can be anonymized.

Table 5 shows a sample of 20 rows from a query joining a sales table with eight dimensions (date, territory group, country, product category, product subcategory, product name, customer name and postal code) and two metrics (quantity and amount). This table has been already denormalized taking information from an OLTP system. As seen on the table, some sensitive information could be exposed to the cloud, such as the customer name and his or her postal

code. This information could be used for unauthorized purposes if it were intercepted by someone not related to the organization.

Besides the possible confidentiality attack, this detailed table represents a large amount of data. The total number of rows obtained with this query is 60,391. For this example, that number of rows would not mean a large amount, but considering a real organization, the number of rows of actual tables could reach several millions. Also, this table was reduced on the number of dimensions and metrics in order to present a coherent number of columns. Actual tables could have several hundreds of columns. Appendix 1 shows the SQL code utilized to generate the detailed table.

In order to avoid the problems explained above, an aggregated version of the data has to be created, and the results would be uploaded to the cloud. Only the detailed version would be kept on-premises. With this approach, confidential data would not be exposed to the network and less data would be transferred to the cloud.

Table 6 shows the table with the aggregated data sample. As can be seen, the sensitive fields have been removed, that is, the columns with the customer name and postal code. In this way, only summarized information is available in the cloud, and if detailed analysis is required, users could access the information with tools on-premises or with the drill-through mechanism.

Another advantage of this approach is that data volume decreases. This table has been reduced to 43,089 records, that is, it was reduced by 29%. This reduction was possible because several records were summarized and compressed. The filed "Quantity" on Table 6 shows higher numbers, meaning that the same product was bought in the same store in the same day. The products were bought by different customers, but since this table is an aggregated version, we do not carry the customer name. Depending on every organization, the percentage of reduction in summarized tables may vary considerably. Appendix 2 contains the SQL code utilized to generate the aggregated table.

7.1.1.1 Detailed Data

Date	Territory group	Country	Product Category	Product Subcategory	Product Name	Customer Name	Postal code	Quantity	Amount
7/1/2001	Europe	France	Bikes	Mountain Bikes	Mountain-100 Silver, 44	Rachael M Martinez	93500	1	3399.99
7/1/2001	North America	Canada	Bikes	Road Bikes	Road-150 Red, 62	Cole A Watson	V9	1	3578.27
7/1/2001	North America	United States	Bikes	Mountain Bikes	Mountain-100 Silver, 44	Sydney S Wright	97355	1	3399.99
7/1/2001	Pacific	Australia	Bikes	Mountain Bikes	Mountain-100 Silver, 44	Christy Zhu	2113	1	3399.99
7/1/2003	Europe	France	Accessories	Bike Stands	All-Purpose Bike Stand	Brad S Chande	59223	1	159
7/1/2003	Europe	France	Accessories	Helmets	Sport-100 Helmet, Red	Kaitlyn J Henderson	93290	1	34.99
7/1/2003	Europe	France	Accessories	Tires and Tubes	HL Road Tire	Brad S Chande	59223	1	32.6
7/1/2003	Europe	Germany	Accessories	Bottles and Cages	Mountain Bottle Cage	Ricky D Vazquez	22001	1	9.99
7/1/2003	Europe	Germany	Clothing	Jerseys	Long-Sleeve Logo Jersey, M	Hailey P Russell	33098	1	49.99
7/1/2003	North America	Canada	Bikes	Road Bikes	Road-350-W Yellow, 42	Lauren R Washington	V9	1	1700.99
7/1/2003	North America	Canada	Clothing	Jerseys	Long-Sleeve Logo Jersey, L	Lauren R Washington	V9	1	49.99
7/1/2003	North America	United States	Bikes	Road Bikes	Road-750 Black, 52	Isaac J Allen	94704	1	539.99
7/10/2003	North America	United States	Bikes	Mountain Bikes	Mountain-200 Silver, 38	Hunter J Robinson	91502	1	2319.99
7/10/2003	North America	United States	Bikes	Touring Bikes	Touring-1000 Yellow, 60	Gabriel Wright	94947	1	2384.07
7/14/2003	North America	United States	Bikes	Touring Bikes	Touring-3000 Blue, 58	Devin T Williams	97330	1	742.35
7/15/2003	Europe	France	Accessories	Bike Racks	Hitch Rack - 4-Bike	Lacey C Zheng	93500	1	120
7/15/2003	Europe	Germany	Accessories	Fenders	Fender Set - Mountain	Samantha R Lewis	64283	1	21.98
7/15/2003	Europe	Germany	Bikes	Mountain Bikes	Mountain-500 Black, 52	Samantha R Lewis	64283	1	539.99
7/15/2003	Europe	United Kingdom	Accessories	Bottles and Cages	Mountain Bottle Cage	Latoya C Goel	SW1P 2NU	1	9.99
7/15/2003	Europe	United Kingdom	Accessories	Bottles and Cages	Water Bottle - 30 oz.	Latoya C Goel	SW1P 2NU	1	4.99

Table 5. Detailed sample data from AdentureWorks database

7.1.1.2 Aggregated Data

Date	Territory group	Country	Product Category	Product Subcategory	Product name	Quantity	Amount
12/8/2001	Pacific	Australia	Bikes	Road Bikes	Road-150 Red, 62	4	14313.08
2/26/2002	North America	Canada	Bikes	Road Bikes	Road-150 Red, 56	4	14313.08
12/7/2002	Europe	United Kingdom	Bikes	Road Bikes	Road-250 Red, 58	3	6544.6875
1/4/2003	Pacific	Australia	Bikes	Road Bikes	Road-250 Red, 48	3	7330.05
1/9/2003	North America	United States	Bikes	Mountain Bikes	Mountain-200 Silver, 38	3	6214.2588
8/7/2003	Pacific	Australia	Accessories	Bottles and Cages	Water Bottle - 30 oz.	8	39.92
12/18/2003	Europe	United Kingdom	Clothing	Caps	AWC Logo Cap	4	35.96
1/11/2004	Europe	United Kingdom	Clothing	Caps	AWC Logo Cap	4	35.96
5/12/2004	North America	United States	Accessories	Tires and Tubes	Mountain Tire Tube	11	54.89
5/21/2004	North America	United States	Bikes	Mountain Bikes	Mountain-200 Black, 38	5	11474.95
5/23/2004	North America	United States	Accessories	Helmets	Sport-100 Helmet, Black	5	174.95
5/23/2004	Pacific	Australia	Accessories	Bottles and Cages	Road Bottle Cage	5	44.95
5/24/2004	North America	Canada	Accessories	Fenders	Fender Set - Mountain	5	109.9
5/25/2004	Pacific	Australia	Accessories	Bottles and Cages	Road Bottle Cage	5	44.95
5/26/2004	Europe	United Kingdom	Accessories	Tires and Tubes	Touring Tire Tube	5	24.95
5/28/2004	Europe	Germany	Accessories	Tires and Tubes	Patch Kit/8 Patches	5	11.45
5/31/2004	Pacific	Australia	Bikes	Road Bikes	Road-750 Black, 44	4	2159.96
5/31/2004	Pacific	Australia	Clothing	Caps	AWC Logo Cap	4	35.96
6/6/2004	North America	United States	Clothing	Caps	AWC Logo Cap	5	44.95
6/7/2004	North America	United States	Accessories	Fenders	Fender Set - Mountain	11	241.78

Table 6. Aggregated sample data from AdventureWorks database

7.2 Column-Oriented Databases

A new approach for utilizing column-oriented databases to handle OLTP and OLAP data has been proposed by the Hasso Plattner Institute. This idea comes after analyzing conventional relational database systems, which become very slow after loading high amounts of information. Relational databases are a good option for OLTP systems, where update and insert are common, but they are not designed for responding after complex queries.

A solution is the use of column-oriented data bases. The technology that these databases offer is best suited for data reads, that is, OLAP environments, but they have poor performance on data updates. A work around for this limitation is to do inserts instead of updates.

The main principle of this approach is that no updates would be done, but only inserts. In order to differentiate two rows (the original and the recently inserted row), a timestamp would be added to every record. If a query is sent to the database, the most recent record would be returned, and the user would obtain information up to date.

With this method, both environments would coexist in just one database and analytic processing could be done in real time, with data recently inserted as product of daily transactional operations in the organizations. The next challenges would be addressed with this approach:

Integration: Virtually, no ETL tool would be needed since BI tools would be applied directly to the operational database. Also, no insertion to a data warehouse would be needed.

Performance: Column-oriented databases have shown better performance and lower cost than conventional relational databases. BI tools would produce results faster for the final users.

Data transfers: Under this approach, both, transactional system and analytical system would be in the cloud, and thus, no data transfer would be needed: everything resided in the same database.

This style of data warehousing currently presents some disadvantages. The first disadvantage is that columnar databases are still a technology that has to be tested under great demand of data access. They have been successful on analytical environments, but as the OLTP approach is new, they still have to show their performance gains.

The second disadvantage is that most transactional systems (like SAP) need to be modified in order to work on an insert and time stamp basis instead of the traditional insert and update approach. Also, analytical BI applications would need to be adjusted to work under this new paradigm.

7.3 Data Warehouse 2.0®

Data Warehouse 2.0 is the new term coined by Bill Inmon to describe the new generation of data warehouse. Figure 17 shows the Data Warehouse 2.0 general architecture.

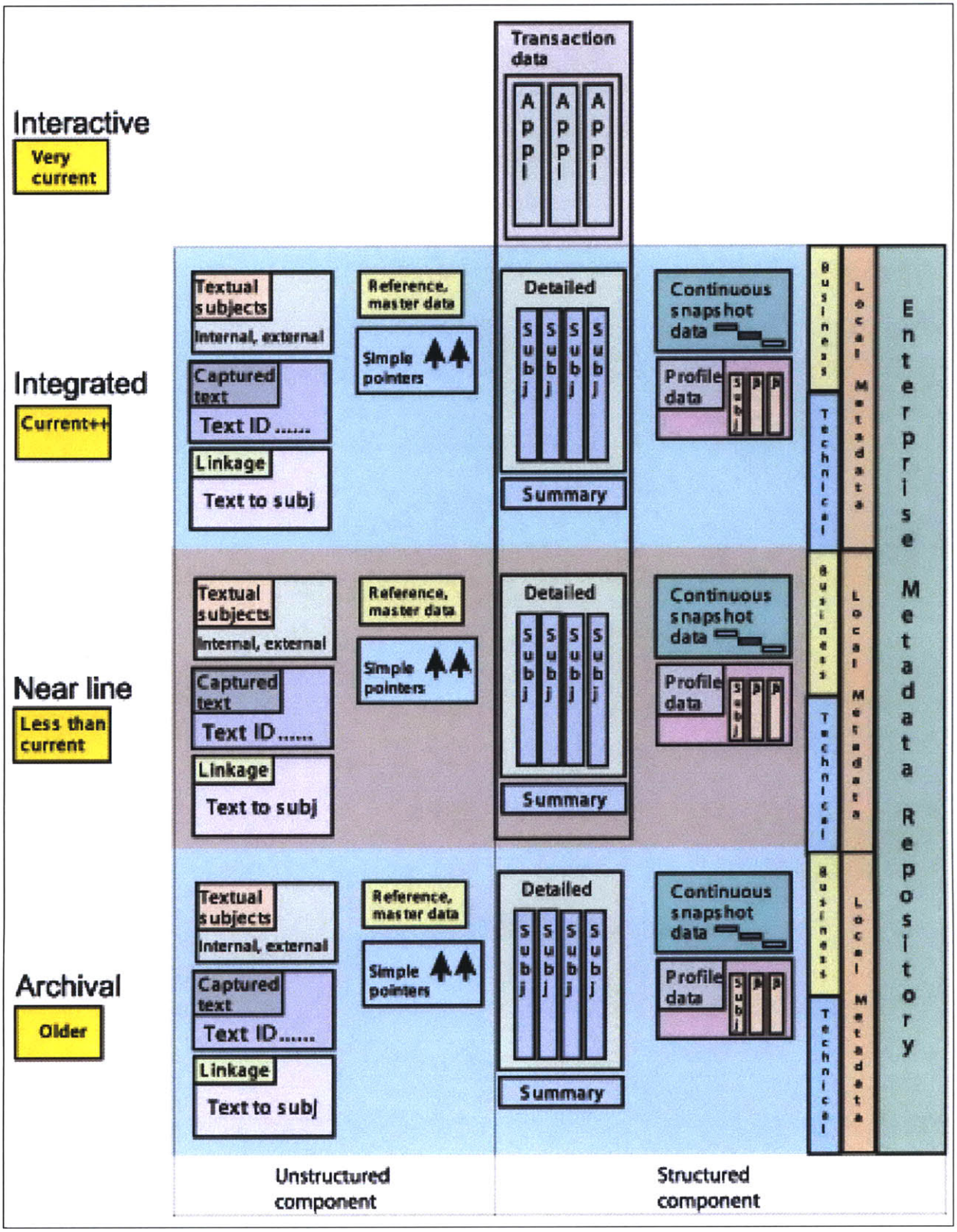


Figure 17. DW 2.0® Architecture

There are three important elements that differentiate this architecture with the traditional data warehousing architectures: information life cycle, metadata, and unstructured data.

7.3.1 Information Life Cycle

The information that is loaded into the data warehouse has a life cycle. That is, not all the information has the same relevance for the users. For example, 90% of the queries may require information only from the last 12 years, while 9% need information from the last two years and only 1% of previous years. If this is the reality in the organization, there is no sense to have all the information in the same data warehouse, because besides of requiring more storage, queries start to become slower every time.

In order to avoid this, four sectors have been introduced: The interactive sector, which is aimed for real time data access, that is, for data that has up to two seconds of age. In this case, real time data warehousing would be applied.

The next sector is the integrated. Information relatively current (no older than one or two years) would be stored in fast disks, so that queries can be responded in a short time with information not older than one year.

The third sector is near line. In this case, information with life in the range of two to five years would be stored in near line storage, which consists in special devices with slower but inexpensive disks. Response times of information coming from near line storage are in the order of some minutes to one hour.

The last sector is archival, for which old information would be stored in archiving devices. Data with life of more than five years would persist in this sector and response times could be between one week and one month. Of course the number of queries soliciting this information is rare in common organizations.

7.3.2 Metadata

Data Warehouse 2.0 promotes extensive use of metadata. Metadata has to be used to describe every characteristic in the data warehouse and business intelligence solutions. Only in this way integration, maintenance and development will be consistent in the application life-time, regardless of the developers that build the solution.

7.3.3 Unstructured data

This new architecture considers unstructured data as part of the core of data warehouse. The idea is to join and integrate unstructured and structured data in one repository. In this way, organizations would be able to get the most of their information for decision making.

The integration of unstructured data is achieved by the use of textual ETLs, which extract information from unstructured data sources, interpret the information, transform it, and load it in into the data warehouse. It is important to apply several business rules in order to grant the proper integration and to make sure that the links among data make sense to the final user.

By the use of Data Warehouse 2.0 in the cloud, two challenges are addressed:

Performance. By maintaining in the data warehouse only recent information, query development can improve considerably. In this way, less scale out would be needed while the data warehouse expands as part of the business intelligence solution. Less current sectors could also remain on-premises.

Integration. Integration with unstructured information is feasible thanks to this architecture and the integration of textual ETLs. Also, the use of metadata facilitates the understanding of data sources and target repositories.

8 Project Development

An evaluation of cloud computing would be incomplete without a clear understanding of how a project like this has to be managed. Managing a business intelligence project in the cloud is different to a project with the on-premises approach, mainly because some of the iterations needed are not needed in the cloud. Also, some tasks will be easier and some other will not be needed, because complex hardware and software decisions and configurations are done by the cloud computing provider.

There are several methodologies for building business intelligence projects. The methodologies vary according to the customers' needs, consultants' best practices, and project complexity. The methodology that is going to be used in this document is the one proposed by Larissa T. Moss and Shaku Atre in their book "Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications"[41]. This methodology has been adopted in several projects and is recognized as a best practice in the industry.

8.1 Critical Path Method Analysis

Moss and Atres' methodology is divided on 16 steps organized in 6 stages. Figure 18 shows a generic project plan for business intelligence solutions based on this methodology. The plan is organized on a Critical Path Method format in order to identify tasks dependencies and possible parallelism. A cloud computing approach affects the way some tasks are developed (those that are shown in orange on the CPM chart). Each affected task and how they are modified are explained below.

8.1.1 Business Case Assessment

The business case assessment is a very important task because it is used as the foundation to justify why cloud computing makes sense when dealing with business intelligence solutions. The business case includes the cost justification and a projection of the costs related to the project, not only during its development but also for its further maintenance. As stated above, cloud computing offers important savings of money, and taking in consideration this paradigm, it may help to ease the project justification during the business case, and at the same time, explaining the business benefits that the project may bring to the organization.

8.1.2 Enterprise Infrastructure Evaluation

There are two types of infrastructures that have to be evaluated for a business intelligence solution: the technical infrastructure and the nontechnical infrastructure. The first one is related to hardware, software, middleware, network, and so on. The second one is related to standards, methodologies and naming conventions. The technical infrastructure evaluation is the task that is affected the most by a cloud computing paradigm, because the organizations do not have to worry about the set of hardware and software required to keep the applications up and running. All these infrastructure requirements are handled by the service provider. This change helps to save time and effort during the evaluation of enterprise infrastructure.

8.1.3 Project Planning

The project planning is also different compared to an on-premises solution, because projects now can be shorter since there is no need to assign resources to the hardware and software

installation and configuration. Instead, the project manager can use more time and effort to create value on the core of the business intelligence project.

8.1.4 Application Prototyping

Application prototyping is also one of the tasks that can be greatly affected by a cloud approach. On a traditional project, the team would have to acquire temporal hardware and install and configure operating systems and business intelligence applications. These tasks could take long time taking in mind that a procurement process would have to be completed before actually starting to build the prototype.

On a cloud based project, the team does not have to wait to acquire the required equipment, they just activate the CPU, memory and storage parameters with the service provider and can start working on the prototype just some minutes after that. Another benefit is that the project team does not have to worry about the temporal hardware disposal, because they just have to deactivate the shared resources with the service provider and would only pay for the time they used the equipment.

8.1.5 ETL Development

ETL development is the only task that is not positively affected by a cloud approach compared to an on-premises approach. Developing an ETL capable of running on cloud computing and including the required mechanisms mentioned before regarding encryption and compression, may take longer time. The main constraint is that the ETL processes have to gather information from geographically dispersed locations, which also adds more time while extracting the first samples of data. Another constraint is that the project team has to be sure that all the required drivers and data connectors are available to get information from several sources outside of the cloud.

8.1.6 Application Development

Application development is modified in the sense that some business intelligence applications may be different to the ones used on-premises. Examples of these differences are special database management systems required to support scale mechanisms and demand adaptation. At the same time, some tools capable of metering the use may be introduced.

8.1.7 Implementation

The last task that is modified by a cloud computing approach is implementation. The project can save long time during the phase of getting up and running the required final infrastructure. Since the cloud service provider has the infrastructure ready all the time, the implementation can begin as soon as the development has finished, without the need of adjusting the required computing power or sizing the applications.

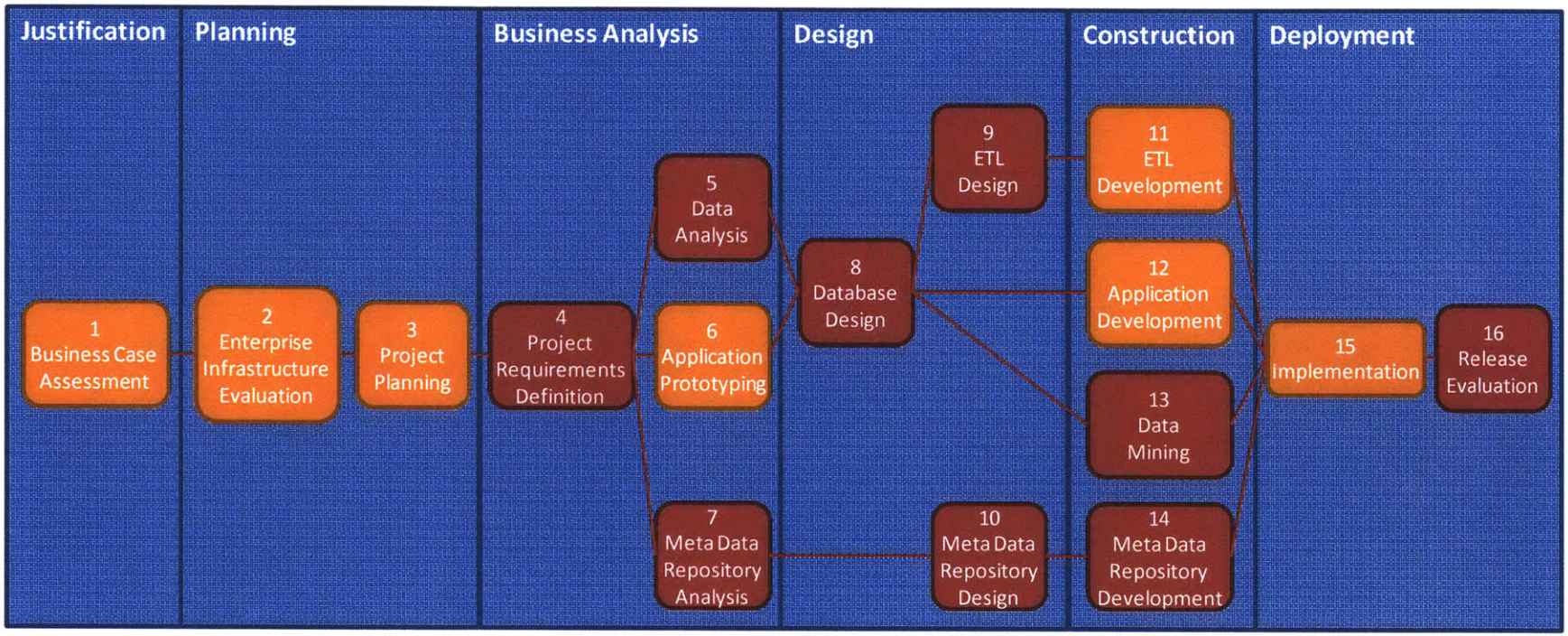


Figure 18. Generic Business Intelligence Project Plan based on Moss and Atre's methodology.

8.2 Design Structure Matrix Analysis

Business Intelligence projects are characterized by the use of several iterations along all their development. These iterations are based on the need of continuous validations and verifications with the customers. Some users like to work under the premise of "I will know exactly what I want as soon as I can see my data represented on a dashboard", which creates the need of a spiral project methodology in some parts of the development.

Based on this is why the use of Design Structure Matrixes (DSM) helps much to have greater control on the project. In this section, a general design structure matrix will be analyzed and will be shown how it can be greatly simplified by the use of the cloud computing paradigm.

Figure 19 shows an abridged version of an DSM applied to a generic on-premises business intelligence solutions. This DSM contains two sets of tasks that are influenced by the use of cloud computing: prototype creation and solution deployment. It is important to note that this DSM has been oversimplified to facilitate the analysis and application of cloud computing benefits. An actual DSM may contain much more task and iterations.

The DSM shown on Figure 19 contains two set of coupled tasks, which represent iterations. The blue square indicates the cluster of tasks related to the prototype development. Most of the tasks have natural dependencies that are found on the lower left triangle in the DSM. For example, in order to acquire the correct hardware, the estimation or required hardware has to be done first, because information ("I") regarding the number and speed of the CPUs, the memory and storage specifications are needed. Also, after the servers have been acquired, there is a physical (marked with the letter "P") dependency to the task of install and configure hardware. The dependencies continue on the DSM, but those included in the triangle mentioned above do not require iteration.

However, in the same blue square, we can find four intersections marked with the letter "I" that represent iterations. It may happen that during the prototype generation, the project team realizes that the estimated hardware is not enough and does not have good performance. This mismatch between the original estimation and the actual needs may arise after identifying the data existing in the organization. There could be complex data structures or demanding data transformations, which require more intensive CPU cycles, more memory or even more storage capacity. When this mismatch is found, the team has to estimate again the required hardware, and after that is done, continue with the appropriate acquisition of resources, such as additional CPUs, memory or hard disks. Long time is usually needed to complete all the procurement process, meaning a delay in the project.

The second cluster of activities on the DSM is identified by the green square. This square represents the tasks related to the project deployment. After the design and development have been finished, the applications have to reside on a dedicated server, which similar to the prototype hardware has to be sized and estimated. This sizing is more important because the hardware will remain permanently at the organization. It may happen that after the application is deployed, again the hardware resources are not enough and the performance is hindered after a few days or months. Then, a new sizing is required followed by the extra resources

acquisition. In this case, this iteration does not only affect the project, but also the final users' operations after the project has finished.

	Estimate required prototype hardware	Acquire hardware	Install and configure hardware	Install platform software	Install application software	Generate prototype	Validate prototype with customer	.	.	.	Estimate required deployment hardware	Acquire hardware	Install and configure hardware	Install platform software	Install application software	Implement applications	Deploy solution
Estimate required prototype hardware						I	I										
Acquire hardware	I					I	I										
Install and configure hardware		P															
Install platform software			P														
Install application software				I													
Generate prototype					I												
Validate prototype with customer						I											
.																	
.																	
.																	
Estimate required deployment hardware																I	I
Acquire hardware											I					I	I
Install and configure hardware												P					
Install platform software													P				
Install application software														I			
Implement applications															I		
Deploy solution																I	

Figure 19. Generic Design Structure Matrix for on-premises Business Intelligence Solutions

Under a cloud computing paradigm this situation changes considerably. Figure 20 shows the simplified DSM for a business intelligence project developed with this approach. As can be seen in the figure, there is no need for iterations for these two sets of tasks. Iterations can disappear thanks to the ability of the infrastructure to automatically scale out and scale in during normal operations, requiring only some minutes to adjust to the users demand.

A big difference in this new DSM is that no dependencies are found on the upper right triangle. There is no need to do a sophisticated sizing and the beginning of the prototype or the deployment. The project team only has to estimate the initial values required to start with the development, activate them with the service provider and start working. If after the works have started, there is need of more resources, the cloud infrastructure will automatically assign

additional resources without spending more time or effort acquiring and configuring extra CPU, memory or storage. This entire job is done automatically by the cloud-enabled platform.

	Activate initial prototype resources	Install application software	Generate prototype	Validate prototype with customer	.	.	.	Activate initial deployment resources	Install application software	Implement applications	Deploy solution
Activate initial prototype resources	■										
Install application software	I	■									
Generate prototype		I	■								
Validate prototype with customer			I	■							
.					■						
.						■					
.							■				
Activate initial deployment resources							I	■			
Install application software								I	■		
Implement applications									I	■	
Deploy solution										I	■

Figure 20. Generic Design Structure Matrix for cloud computing based Business Intelligence Solutions

As seen on these different DSMs, the business intelligence project can save important amounts of time and effort in order to continue with the project, not to mention the overall costs that can be waived by applying the cloud paradigm.

9 Conclusion

Business Intelligence applications have been less developed on the cloud than transactional operational applications. There are indeed some challenges that have to be solved in order to take advantage of cloud computing benefits for handling high volumes of information for decision making in the organizations.

The benefits that cloud computing offers to business intelligence applications are lower costs, multiple redundant sites, scalable provisioning of resources, on-demand performance improvements, usage billing, fast deployment and easier maintenance.

Cloud computing is offered in three modalities: Infrastructure as a Service, which consists basically of hardware; Platform as a Service, formed by the previous layer plus operating system and virtualization programs; and Software as a Service, integrated by the previous two layers plus the business intelligence applications required to manage databases and decision support systems.

The most important challenges that affect business intelligence applications in the cloud are security, moving large amounts of information to the cloud, performance, integration, and availability.

Security is usually increased by the sole fact of moving applications to the cloud, since more mature and prepared environments are found in service providers' data centers than in on-premises facilities. Mechanisms like data transfer encryption, in-database encryption, data distribution, component isolation in the data center, and intelligent workload management can be used to increase security in the cloud.

Moving high volumes of data is improved by adopting larger bandwidth connections to the Internet, compressing data before sending it to the cloud, extracting information from transactional systems in the cloud and by using acceleration devices.

Performance is increased by moving applications to the cloud, since cloud service providers offer more computing power at lower cost than on-premises data centers. These computing services benefit the performance of ETLs, data warehouses, and BI applications. The use of new architectures on the cloud like data warehouse appliances and new software able to scale on demand, are also options to improve performance.

Integration of cloud computing resources with on-premises data sources is achieved by the use of the newest ETL tools, which support the connection to a large variety of applications and databases for both, data sources and data warehouse repositories.

Availability is increased also by moving applications to the cloud. Cloud service providers have among their offering, redundant sites and specialized hardware, software, and professional services. With these elements, the availability is increased by decreasing the mean time to recovery (the time the cloud staff needs to repair the system if something goes wrong) and increasing the mean time to failure (the time when the system works properly). Most of best know service providers offer at least 99.67% of availability.

There are some new approaches that can be applied to improve the users' experience with cloud computing solutions, such as hybrid approaches, column-oriented databases, and the Data Warehouse 2.0 architecture.

Column-oriented databases improve performance on data reading, and offer a new technique for integrating transactional systems with decision support systems in just one database. Since both applications reside in the same environment, then data transfers would be practically unnecessary.

The Data Warehouse 2.0 architecture offers extended metadata for a better integration with transactional systems, recognition of unstructured metadata for a better decision making, and use of information life cycle, which offers a better performance by separating data that are more used from those that are less used, and storing them in different storage devices.

The hybrid approach is the one that suits the best with business intelligence solutions in the cloud. It can be also mixed with the approaches mentioned before. Although this technique requires larger investment, it can take advantage of cloud computing benefits, and avoids most of the challenges. The main principle of the hybrid approach is to have detailed sensitive information on-premises and only transfer aggregated anonymized information to the cloud. This solution is the most recommended. Products and services such as Amazon's EC2 Private Cloud or IBM Analytics Cloud facilitate this duty to the organizations.

Finally, project development is also greatly benefited by cloud computing. Business intelligence projects usually require several iterations to satisfy customers' requirements. Iterations usually involve validations and adjustments. Computing power estimations are usually exceeded by the actual prototypes and final project development. In a cloud computing environment extra computing power is easily and rapidly obtained taking advantage of the service providers' architecture. Hardware is scaled out to satisfy the required power, and extra iterations are avoided on the Design Structure Matrix used to control the project.

As proved in this document, business intelligence applications can effectively be improved by the use of a cloud computing approach. Some techniques and special considerations have to be implemented, but in the end, organizations will obtain greater benefits and will be able to leverage their information for better decision making.

10 Abbreviations

Amazon S3: Amazon Simple Storage Service

Amazon EC2: Amazon Elastic Compute Cloud

BI: Business Intelligence

CapEx: Capital Expenditure

CFO: Chief Financial Officer

CIO: Chief Information Officer

CPM: Critical Path Method

CTO: Chief Technology Officer

DSM: Design Structure Matrix

ERP: Enterprise Resource Planning

ETL: Extraction, Transformation and Load

I/O: Input / Output

IaaS: Infrastructure as a Service

IT: Information Technology

JDBC: Java Database Connectivity

LAN: Local Area Network

MTTF: Mean Time To Failure

MTTR: Mean Time To Recovery

ODBC: Open Database Connectivity

OLAP: Online Analytical Processing

OLTP: Online Transaction Processing

OpEx: Operational Expenditure

OPM: Object-Process Methodology

PaaS: Platform as a Service

RDBMS: Relational Database Management System

SaaS: Software as a Service

SSL: Secure Socket Layer

TSL: Transport Layer Security

VPN: Virtual Private Network

11 Appendix 1: AdventureWorks Detailed Data Query

```
SELECT DimTime.FullDateAlternateKey,
       DimSalesTerritory.SalesTerritoryGroup,
       DimSalesTerritory.SalesTerritoryCountry,
       DimProductCategory.EnglishProductCategoryName,
       DimProductSubcategory.EnglishProductSubcategoryName,
       DimProduct.EnglishProductName,
       DimCustomer.FirstName + ' ' + ISNULL(DimCustomer.MiddleName + ' ', '')
+ DimCustomer.LastName AS CustomerName,
       DimGeography.PostalCode,
       SUM(FactInternetSales.OrderQuantity) AS Quantity,
       SUM(FactInternetSales.SalesAmount) AS Amount

FROM   DimProductSubcategory INNER JOIN
       DimProduct ON DimProductSubcategory.ProductSubcategoryKey =
DimProduct.ProductSubcategoryKey INNER JOIN
       DimProductCategory ON DimProductSubcategory.ProductCategoryKey =
DimProductCategory.ProductCategoryKey INNER JOIN
       FactInternetSales ON DimProduct.ProductKey =
FactInternetSales.ProductKey INNER JOIN
       DimCustomer ON FactInternetSales.CustomerKey = DimCustomer.CustomerKey
INNER JOIN
       DimSalesTerritory ON FactInternetSales.SalesTerritoryKey =
DimSalesTerritory.SalesTerritoryKey INNER JOIN
       DimTime ON FactInternetSales.OrderDateKey = DimTime.TimeKey INNER JOIN
       DimGeography ON DimCustomer.GeographyKey = DimGeography.GeographyKey
AND
       DimSalesTerritory.SalesTerritoryKey = DimGeography.SalesTerritoryKey

GROUP BY DimTime.FullDateAlternateKey,
         DimSalesTerritory.SalesTerritoryGroup,
         DimSalesTerritory.SalesTerritoryCountry,
         DimProductCategory.EnglishProductCategoryName,
         DimProductSubcategory.EnglishProductSubcategoryName,
         DimProduct.EnglishProductName,
         DimCustomer.FirstName,
         DimCustomer.MiddleName,
         DimCustomer.LastName,
         DimGeography.PostalCode
```

12 Appendix 2: AdventureWorks Detailed Data Query

```
SELECT DimTime.FullDateAlternateKey,
       DimSalesTerritory.SalesTerritoryGroup,
       DimSalesTerritory.SalesTerritoryCountry,
       DimProductCategory.EnglishProductCategoryName,
       DimProductSubcategory.EnglishProductSubcategoryName,
       DimProduct.EnglishProductName,
       SUM(FactInternetSales.OrderQuantity) AS Quantity,
       SUM(FactInternetSales.SalesAmount) AS Amount

FROM   DimProductSubcategory INNER JOIN
       DimProduct ON DimProductSubcategory.ProductSubcategoryKey =
DimProduct.ProductSubcategoryKey INNER JOIN
       DimProductCategory ON DimProductSubcategory.ProductCategoryKey =
DimProductCategory.ProductCategoryKey INNER JOIN
       FactInternetSales ON DimProduct.ProductKey =
FactInternetSales.ProductKey INNER JOIN
       DimSalesTerritory ON FactInternetSales.SalesTerritoryKey =
DimSalesTerritory.SalesTerritoryKey INNER JOIN
       DimTime ON FactInternetSales.OrderDateKey = DimTime.TimeKey

GROUP BY DimTime.FullDateAlternateKey,
         DimSalesTerritory.SalesTerritoryGroup,
         DimSalesTerritory.SalesTerritoryCountry,
         DimProductCategory.EnglishProductCategoryName,
         DimProductSubcategory.EnglishProductSubcategoryName,
         DimProduct.EnglishProductName
```

13 References

[1] McGreevy, Maura. Spending on Business Intelligence and Performance Management to Top $57.1B in 2008. May, 2008. http://www.amrresearch.com/content/view.aspx?compURI=tcm:7-37602&title=Spending+on+Business+Intelligence+and+Performance+Management+to+Top+$57.1B+in+2008

[2] BusinessWeek. Turbulence on the Way to the Cloud. November, 2009. http://www.businessweek.com/globalbiz/content/nov2009/gb2009116_988428.htm.

[3] Gartner. Gartner Identifies the Top 10 Strategic Technologies for 2009. October, 2008. http://www.gartner.com/it/page.jsp?id=777212.

[4] Gartner. Gartner Identifies the Top 10 Strategic Technologies for 2010. October, 2009. http://www.gartner.com/it/page.jsp?id=1210613.

[5] Gartner. Gartner Says Worldwide Cloud Services Revenue Will Grow 21.3 Percent in 2009. March, 2009. http://www.gartner.com/it/page.jsp?id=920712

[6] GreenerComputing. Cloud Computing Highlighted as an Emissions-Reduction Strategy. July, 2009. http://www.greenercomputing.com/news/2009/07/15/cloud-computing-highlighted-emissions-reduction-strategy

[7] IDC. Everyone's a Genius: SaaS-Delivered Business Intelligence Tools Put Decision Making in the Hands of Decision Makers. April, 2009.

[8] Eckerson, Wayne. Implementing BI in the Cloud. June, 2009.

[9] Inmon, William H. Building the Data Warehouse. 2005.

[10] Eckerson, Wayne W. The Five Dimensions of Business Intelligence. 2005.

[11] National Institute of Standards and Technology. NIST Definition of Cloud Computing v15. October, 2009.

[12] Gartner. Five Refining Attributes of Public and Private Cloud Computing. May, 2009. http://my.gartner.com/portal/server.pt?open=512&objID=260&mode=2&PageID=3460702&resId=965212.

[13] McMurchy, Neil. Survey of BI Purchase Drivers Shows Need for New Approach to Business Intelligence. July, 2008. http://my.gartner.com/portal/server.pt?open=512&objID=260&mode=2&PageID=3460702&resId=714209.

[14] Potter, Randolph & Bezuidenhout, Brendon. Matching Business Intelligence with Cloud Computing. October, 2009. http://xqrx.com/writing/a_cloud.php.

[15] Dine, Stephen. B.I. In The Cloud. August, 2009.

[16] Thibodeau, Patrick. Amazon's data center outage reads like a thriller.December, 2009. http://www.computerworld.com/s/article/9142154/Amazon_s_data_center_outage_reads_like_a_thriller

[17] Bradley, Tony. Rackspace Outage Has Limited Impact. December, 2009. http://www.pcworld.com/businesscenter/article/185171/rackspace_outage_has_limited_impact.html

[18] Golkar, Cyrus. Top Business and Technology Questions in Cloud Computing. July, 2009. http://www.b-eye-network.com/channels/1550/view/10905.

[19] Deshpande, Mukund & Joshi, Shreekanth. Incorporating Business Intelligence in the Cloud. August, 2009. http://www.b-eye-network.com/channels/1550/view/11143.

[20] Eckerson, Wayne. Implementing BI in the Cloud. June, 2009. http://portals.tdwi.org/blogs/wayneeckerson/2009/06/implementing-bi-in-the-cloud.aspx.

[21] Recombinant Data Corp. Cloud Computing for Healthcare and Life Sciences Data Warehousing. 2009. http://www.b-eye-network.com/files/CloudComputing_WP.pdf.

[22] Dine, Stephen. B.I. in the Cloud. August, 2009.

[23] Wells, Dave. What's Up with Cloud Analytics. December, 2009.

[24] Lounibos, Tom. SOASTA's 10,000 Hours in the Cloud. December, 2009. http://eclipse.sys-con.com/node/1150203.

[25] Swoyer, Stephen. The Mainstreaming of MapReduce. November, 2009.

[26] Austin, Tom, et. al. Introducing the High-Performance Workplace: Improving Competitive Advantage and Employee Impact. Gartner. May, 2005. http://my.gartner.com/portal/server.pt?open=512&objID=260&mode=2&PageID=3460702&resId=481145

[27] Amazon. Amazon Web Services: Overview of Security Processes. November, 2009. http://awsmedia.s3.amazonaws.com/pdf/AWS_Security_Whitepaper.pdf.

[28] Inet2000. How secure is the encryption used by SSL?. Retrieved: January 4, 2010. http://www.inet2000.com/public/encryption.htm
[29] Microsoft. Database Encryption in SQL Server 2008 Enterprise Edition. February, 2008.
[30] Reese, George. Cloud Application Architectures. April, 2009.
[31] Amazon. Amazon Elastic Compute Cloud. 2009. http://aws.amazon.com/ec2/#pricing.
[32] Lelewer, Debra A. and Hirschberg, Daniel S. Data Compression. http://www.ics.uci.edu/~dan/pubs/DataCompression.html.
[33] F5. WANJet 500 Series Datasheet.2008. http://www.f5.com/pdf/products/wanjet-ds.pdf
[34] Greenplum. Our Partners. 2009. http://www.greenplum.com/partners/our-partners/.
[35] Amazon. Amazon Simple DB. 2009. http://aws.amazon.com/simpledb/.
[36] Informatica. Informatica Power Center Features. 2009. http://www.informatica.com/products_services/powercenter/Pages/powercenter_features.aspx
[37] IEEE. IEEE Reliability Society - Reliability Engineering. 2009. http://www.ieee.org/portal/site/relsoc/menuitem.e3d19081e6eb2578fb2275875bac26c8/index.jsp?&pName=relsoc_level1&path=relsoc/Reliability_Engineering&file=index.xml&xsl=generic.xsl
[38] Weibull.com. Reliability Basics. April, 2003. http://www.weibull.com/hotwire/issue26/relbasics26.htm
[39] Uptime Institute. Data Center Site Infrastructure Tier Standard: Topology. 2009. http://professionalservices.uptimeinstitute.com/UIPS_PDF/TierStandard.pdf.
[40] Microsoft. SQL Server 2005 SP2a. May 2007. http://www.codeplex.com/MSFTDBProdSamples/Release/ProjectReleases.aspx?ReleaseId=4004
[41] Moss, Larissa T.; Atre, Shaku. Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. March, 2003.