A Research Report for Data Warehouse in Cloud

Embracing Data Warehouse as a Service

Mao Chuan Li
Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
mao.chuan.li@gmail.com

***Abstract*—Data warehouse has been around for decades to store the most important business data and help knowledge workers to reveal the hidden values and make the wisest decisions. However, with the more volume of data accumulated, the pain on managing and maintaining the data is getting worse. Buying, installing and setting up a super computer is never the right answer, instead, cloud computing should be the ultimate solution for any size of business organizations. This report reviews the basic concepts of data warehouse, and the related challenges in traditional data ware houses development, and examine the benefits of current cloud computing. Lastly a few of current data warehouse service providers in cloud are reviewed.**

***Index Terms*—data warehouse, data warehousing, cloud computing, Database as a Service, Data Warehouse as a service.**

I. Introduction

Back in 1969, IBM published the world’s first commercial database management system (DBMS) called “Information Control System and Data Language/Interface (ICS/DL/I)” [13]. Since then DBMS has experienced a great development through navigational systems, to relational database systems, and up till now the NoSQL database systems. With the rapid development of both computer hardware and software, the business data scale has grown from the ‘mega’ megabytes to the contemporary ‘big’ terabytes, even petabytes.

Saving history data is not a product of computers, but the human natures. People believes that the history data could provide us insight of the future events. As in our history, we also believe that the business history data could help our business organizations better understand our environments, and support our decision makings with sufficient evidences hidden from the data. To deal with the enormous raw data generated in our daily transactions, the IBM researcher Barry Devlin and Paul Murphy developed the concept "business data warehouse" to formally address this requirements in the late 1980s. With such a data warehouse, based on the modern database systems, most of the valuable business data finally find a place to reside. A rich set of business intelligent tools, such as On Line Analytics and Data Mining are created to better mine the hidden values in the big data.

Nevertheless, creating such a data warehouse is not a trivial work for any size of company, which includes a huge amount of capital investment of the hardware devices and software licenses, a full range of data warehouse architects, developers and maintenance teams, etc. Not every company could afford such a beautiful looking data machine and solution, especially for the small size companies. For most of startup companies, such a data warehouse could only be an extravagant hope.

Among all the challenges in realizing such a data warehouse for a company, three of them are the most prominent:

A. The upfront hardware and software investment

There are 2 flavours for the system configuration, using commodity hardware and commercial special purpose database systems, and using specialized data warehouse appliances provided by the companies like IBM, Oracle or Teradata, which has bound the hardware with acceleration technologies and their database systems optimised specially for data warehousing. The second flavour is obviously more expensive than the first one, but it provides an integration solution for companies with easy configuration management. Although the first flavour spent less on hardware procurement, the extra maintenance cost as well as the employment requirement for those highly skilled specialist cannot be overlooked. In either way, the payment for the high performance database management system is inevitable.

B. System Maintenance and Tuning

Comparatively, querying the data is far more comfortable than working with the CPUs, disks, and networks. A healthy running data warehouse is supposed to respond to business users’ queries and help them generate reports within a reasonable time, normally measured by a quality of service. To keep a data warehouse running smoothly with periodical operational data coming in, a team of skilled system administrators, database administrators, storage administrators and network administrators are a must to maintain and tune the system.

C. Scale up or out when data gets bigger

As computer hardware is too ‘hard’ to expand the capacity, such as CPU and memory, and storages, it is unavoidable to run out of disk spaces eventually when more and more transactional data flows in data warehouse. Even it has not reached the ceiling of the storages, the performance of the data warehouse may be downgraded gradually. One common solution to this issue is to delete or archive the part of history data, however, this obviously deviates from the original data warehouse design principle – to keep the Single Version of the Truth of the company, because of loss of the data.

This paper examines the state of the art of latest and poplar technology - cloud computing and looks to utilize the cloud services to solve the aforementioned issues.

II. Background and Motivation

As the data volume in the data warehouse getting bigger and bigger, the performance of query of the data, and the processing of source data becomes slower and slower. The most straightforward solution to solve or mitigate the problematic issue is to increase the computer's power with more and faster CPUs, more memory, and more disk storages. But any single hardware system has a maximum capacity due to the current computing technology, and all the hardware systems can only scale up to a point where the system bus gets overloaded as the following Figure 1 shows:

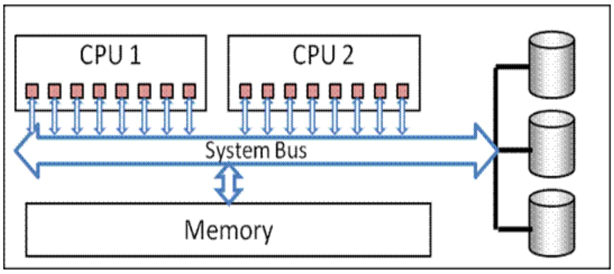

Figure 1- SMP Architecture [9]

To further solve the performance issue, a new architecture shown in Figure 2, called "Massive Parallel Process (MPP)" is innovatively created to dispatch the workload to different nodes, which are each a standalone computing node as a SMP system illustrated above. Such MPP systems could easily be extended horizontally with more computing nodes which has more CPUs, memory and storages to fulfill the system requirements. One of the biggest issue with this architecture is that most of the MPP data warehouse vendors package their proprietary hardware and software, and related business intelligent tools altogether for end users, which incurs a high cost for customers and becomes a major blocker for most of small to medium-sized companies to adopt such a solution.

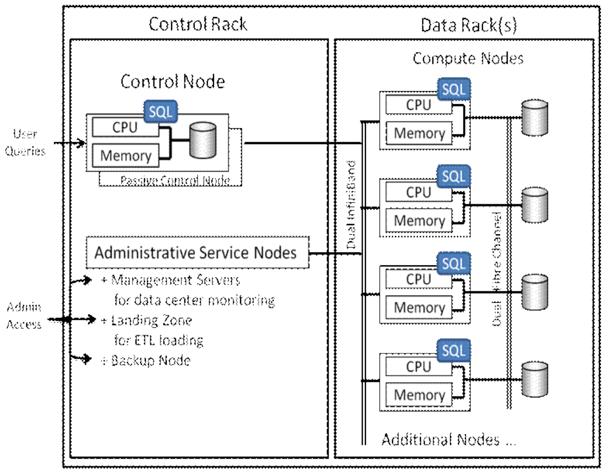

Figure 2 – MPP Architecture [9]

Another direction to solve the scaling problem was going to Grid computing, which utilizes the distributed systems thanks to the mature development of modern network. All the data warehouse data is deliberately distributed into the grid nodes, which are composed by an unlimited number of heterogeneous commodity computers each with their own CPUs, memory and storages. The Grid computing architecture is simply depicted in Figure 3:

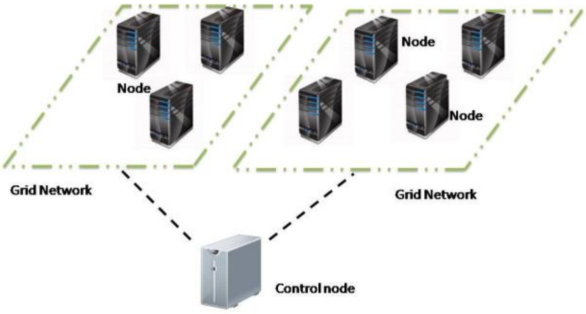

Figure 3 – Grid Architecture for Data Warehouse [8]

All these measures are trying to increase the data warehouse capacity with either single node vertical scale-up techniques or multiple nodes horizontal scale-out techniques, which may solve the third issue aforementioned. But none of them have faced the other two issues: the unaffordable upfront hardware and software investment and the endless system maintenance and tuning tasks.

Cloud computing has become a hype since around 2000. Although the concept Cloud Computing could be dated back in the 1960s, referred to as the "ability to provide and organize computation as a utility" by John McCarthy, the realization of it has just been true for around 10 years. The cloud architecture as shown in Figure 4 could completely solve the previous mentioned issues with the following unbeatable major features [4]:

A. Full Managed by Cloud Provider

All hardware and software are provided to end users by Cloud providers in a form of web service, for example, the most popular REST service. End users do not need to procure any hardware or software on their premises any more, but access the computing power provided by the Cloud and pay for the service. Using the computing power is just like turning on the tap and enjoying the flowing water. Moreover, end users could turn off the tap at any time.

B. *"Infinite" computing resources*

Cloud providers host a few huge size data centres to give end users an illusion that the computing resources are not capped, instead, are elastic to scale up or down on their demands.

C. *Access from anywhere in the world*

With a lightweight device, for example, a smart phone which has connection to the internet, end users could access the cloud services anywhere in office or out of office.

D. *Trivial cost to begin with*

Normally, Cloud providers charges the end users based on their usages of the cloud services. Any size of business users could afford to set up a data warehouse.

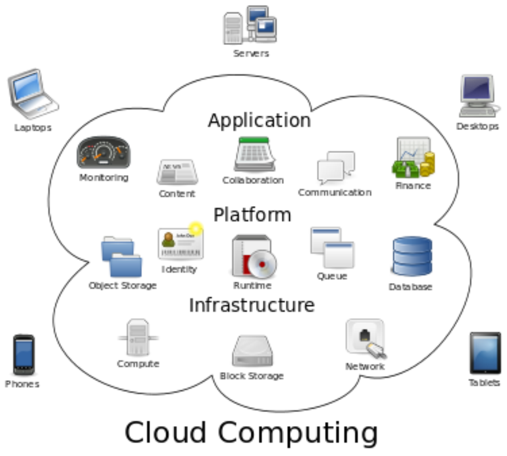

Figure 4 – Cloud Computing Architecture [10]

III. Literature Review

A. *Mary Breslin: Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models[1]*

This paper introduced the two main data warehouse development models and methods proposed by the two experts: Mr. William Inmon and Mr. Ralph Kimball.

In the field of Data Warehouse, there are two important figures who have influenced the development of data warehouses for more than two decades: Inmon – the Father of Data Warehousing and Kimball – the Father of Business Intelligence. Both Inmon and Kimball defined a model for creating a data warehouse with different philosophies and techniques, so that a debate on which one is better among the data warehouse practitioners has lasted for one decade.

Inmon proposed a top-down development model following strictly the relational database design and methods, such as ERD. So an organization should create a big centralized enterprise-wide data warehouse upfront to achieve the concept of "single version of truth", and then more satellite data marts could be derived from the big data warehouse.

On the contrary, Kimball suggested a bottom-up development model to build small sized departmental data marts for specific business requirements. Later with all data marts, organizations could merge all the data marts into a big data warehouse for the whole organization.

For the data structure, Inmon suggested to follow traditional proven database design and the relational model to ensure consistency of data. In contrast, Kimball abandoned such a mature data structure and innovatively create a brand-new multi-dimensional structure (star schema and snowflakes) to represent the data.

What they share is the data massaging operations, both defined an ETL (Extract, Transform and Load) process.

B. *Pedro Furtado: A Survey on Parallel and Distributed Data Warehouses[2]*

This paper reviewed the basic concepts of parallel computing in data warehouse systems, and then the works on distributed data warehouse systems. Using parallel system or distributed system are two important, main stream techniques to achieve the scalability of a data warehouse system.

For parallelization, there are three main architectures based on the computing architecture of processing unit (PU), the storage device (S) and memory (M).

1) *Shared Memory/Everything*

As the Figure 5 illustrates, the parallel system is working within a single system with multiple CPUs or multiple cores. All CPUs or cores share the same system memory and peripherals like the hard disks to process IO requests. The major drawback of such a SMP system is the limited scalability of such a single system, imposed by the limitation of ability of multiple processing of operating system, and the system bus bandwidth.

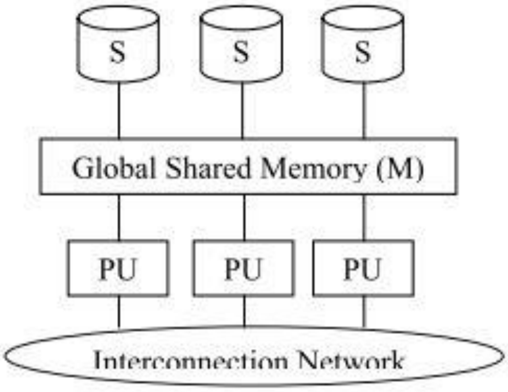

Figure 5 – SMP shared memory architecture [2]

2) *Shared Nothing*

Shared noting architecture as noted in Figure 6, dose not share the processors, memory nor storages, all the computing nodes are running separately, independently from each other. For cooperation, all of these nodes are connected by a high speed network. In such an architecture, the most important advantage of it is the unlimited scalability with more and more hardware joining in such a cluster system. On the other hand, the disadvantage is that the interconnection between each node may become a bottleneck because the cooperation of these independent nodes depend heavily on the data exchange to make decisions. Data allocation, query processing optimization and load balancing are the most relevant issues in such an architecture.

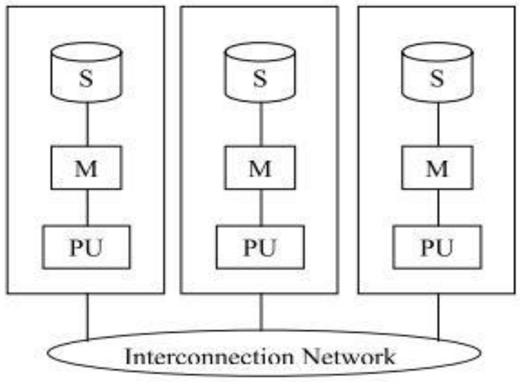

Figure 6 – Shared Nothing architecture [2]

3) *Shared Disk*

Similar to shared nothing architecture, all the computing nodes shared the whole array of disks as illustrated in Figure 7. In nature, such an architecture allows expanding the computing nodes to improve the whole system's performance, however the performance of storage system and the interconnection system might be a bottleneck.

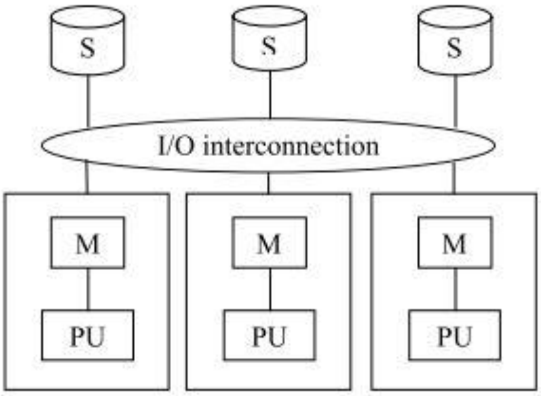

Figure 7 – Shared Disk architecture [2]

C. Mohammad Hamdaqa, Ladan Tahvildari: Cloud computing uncovered: a research landscape[3]

This paper gives a general introduction of cloud computing, and explained all the cloud attributes around the NIST definition of cloud. Also it summarizes the basic deployment models and service models of general could computing. At last, they made a comparison between the Cloud computing with the other computing technologies, such as Grid computing.

There are six main characteristics of Cloud computing most people agree upon:

- On-demand self-service – which means all the computing resources in cloud is reachable and obtained on demand as we are using the tap water.
- Easy to access standardized mechanisms – all cloud services are accessible to end users who have network connections, with standard web services interfaces like REST, can access the computing resources hosted in Cloud.
- Resource pooling and multi-tenancy – a cloud is servicing all the customers at the same time with the same set of hardware and software resources. Virtualization plays an important role here to support this functionality.
- Rapid elasticity – With the help of virtualization, and sharing of the same hardware resources, the computing power could be fully utilized by every customer, who do not always 100% utilize the resources all the time. This fashion of utilization gives the end users an illusion of unlimited available resources. The automation of resource allocation and dispatch in the cloud enables users to increase or decrease the hardware resources elastically.
- Measured service – all the cloud services are measured in some way for customers to be charged and closely monitored by cloud provider.
- Auditability and certifiability – to comply with some regulations or rules, the services must support auditing and logging.

For cloud providers, the paper summarized 4 different deployment models for the cloud services:

- Public Cloud – is the most popular form of all four models, which is the cheapest solution for the cloud consumers. The cloud services are open to all public with a multi-tenant model. The only concern or obstacle of this model for end users is the lack of trust.
- Private Cloud – is appealing to big companies which has both financial resources and technical resources to set up one cloud for their own usage on their own data centres. The biggest challenge of it is the upfront cost of investment of the hardware and the manpower to maintain such a huge computing system. The big advantage of it is the company has full control of the cloud and

could enforce any polies, security controls to it to fulfil their specific requirements.

- Hybrid Cloud – is a compromise between the public cloud and the private cloud. For the less important data and applications, the company could choose to move them to public cloud, while keeping the sensitive confidential data and applications within their own private cloud.
- Community Cloud – such cloud is built by a group of organizations who share the same goals or interests, yet don’t trust the public cloud. They are not able or willing to set up a private cloud by themselves, but would like to share a ‘private cloud’ with others.

All the resources in cloud are represented by a form of service. All the services are categorized into three groups:

- SaaS – Software as a Service, the most typical example of this service is the Microsoft Office software. Without SaaS before, users have to purchase a copy of the software and install it on their local computers, which has to be windows operating system, and then use it. With SaaS, users could simply operate the online web browser based Office 365 software without all the hassles of installing, updating locally. Users can pay for what they use on a subscription basis.
- PaaS – Platform as a Service, the cloud provider provides a range of hardware and software platforms for customers to run their own applications, and help them manage and maintain the systems for customers. All customers will not care about the system underneath the application and they could focus on their applications.
- IaaS – Infrastructure as a service, the cloud provider provides the fundamental facilities, like computers, network and storages to customers as a service. End users could install whatever operating systems and applications based on their own needs. All the hardware management and maintenance work is done by cloud providers.

IV. Blossom of Data Warehouse in Cloud

Back in 2002, IBM has proposed the concept “database as a service (DBaaS)” [6] which was delivered and accessed with the internet connection, however the prevalence of DBaaS only started from 2009, when Amazon released their first Relational Database Service to the public. Although Data Warehouse implementation on a regular database is a simple and natural task as it seems, it has took Amazon 3 years to deliver the Data Warehouse in their cloud service. This may be mainly attribute to the special complex requirements of a data warehouse. These requirements could be categorized into two groups:

A. Big Data support and Fast Query Response Ttime

The data warehouse should be able to store and process petabytes level data, and meanwhile sustain the reasonable response time for the high volume of data queries for business users. This is the biggest challenge of the Data Warehouse service in the cloud. Traditional databases cannot fulfil this requirement easily without a deliberate design of the database system. Also the commodity hardware in cloud limit the performance of traditional database systems. To overcome this problem, Amazon redesigned the database system to specialize in data warehouse based on the PostgreSQL relational database and adopted the columnar storage concept to reduce the overall disk throughput so as to optimize the analytic query performance. Besides that, Amazon also introduced the massive parallel processing (MPP) architecture to support dynamic and automatic scaling as Figure 8 shows:

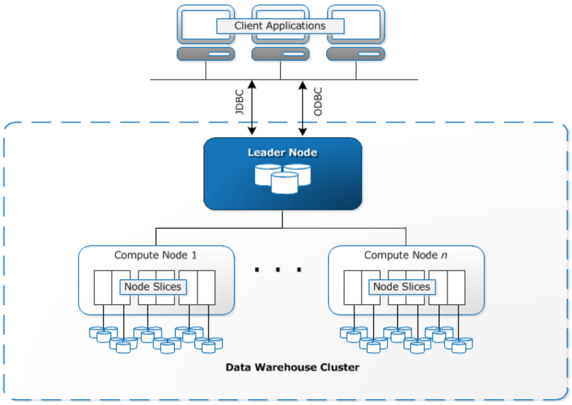

Figure 8 – Amazon Redshift Architecture [11]

B. Integration with surrounding Business Intelligent products

To get the most out of data warehouse, the data users need to extract the transactional data from multiple external sources; to analyse the data, users need business intelligent tools like OLAP, data mining to extract the hidden values, and generate the reports. Amazon has provided a range of tools to help customers to import the data from either customers’ site, or from the data resides in Amazon existing data stores, and integrated with a rich set of third party business intelligent tools, such as TIBCO, IBM Cognos.

With this whole solution for customers, every data warehouse users could either start using it from scratch (for those small to medium sized companies) or migrate their traditional data warehouses in house to Amazon cloud with the extreme low prices.

Although Amazon is the first cloud provider to support Data Warehouse service, it is not the only player in the market at the time of writing. Right after the release of Redshift, IBM, Teradata, and Microsoft have also released their editions of data warehouse in their clouds. Among them, Microsoft has made an innovation to separate the charging of the storage and the computing [12]. The SQL Data Warehouse service can elastically grow or shrink, even pause the computing power for the data warehouse. This would further save the cost of customers and provide more flexibility.

With the Data Warehouse service in the cloud, every business willing to have their own data warehouse could afford to create and maintain their business data and benefit from utilizing it. All the three aforementioned issues could be lightly solved with such a service in the cloud. Users do not need to worry about the procurement of the hardware and software, the installation and maintenance, even the site to place those behemoths. Furthermore, no more system management is needed as before. At last, the extremely low price on a subscription pay mode is affordable for any size of businesses. It is not hard to imagine that the data warehouse of future must fall into the cloud computing.

V. Conclusions

Although Cloud computing has developed only for about 10 years, and the Data Warehouse as a Service has just emerged in the cloud market, the advent of era of Big Data is the irrefutable fact. Every business organization, no matter how small or big, they are all accumulating important and valuable business data day by day, even minute by minute. Data Warehouse is the best solution for them to store their precious data, and provide them the ability to query and analyze the history data, and help them make the most important business decisions in such a fast-paced economy. Traditional data warehouses on users' premises must be replaced by the data warehouse services hosted in either private or public clouds, because of their exclusive advantages. Embracing the Cloud for business organizations shall be their only choice in future.

VI. Future Issues

Almost all cloud providers are utilizing commodity hardware to support their infrastructures. The holistic performance may be limited by the single point of the computing nodes. In such case, Data Warehouse users might experience a performance issue with the services in cloud. Also the internet connection delay between customer site and cloud provider may be another factor of causing the performance issue [5].

The cost of transfer data between customer site and cloud platform might be another issue. Since the data volume of external data sources loaded into the cloud platform is huge, the cost of transferring these data has to be considered in customer's budget [5].

The Data Warehouse saves the most important data of the business organizations, any violation of these data could cause a loss of the businesses. In public cloud, or hybrid cloud, this is always an unavoidable issue [5].

At last, the cloud platform may stop working and the data warehouse may become unavailable to users. The interoperability between cloud providers seems impossible, at least for now, so the migration from one cloud provider to another may incur unpredictable cost and issues [5].

References

[1] M. Breslin. (2004) Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models. Business Intelligence Journal. 6-20.

[2] P. Furtado, "A Survey of Parallel and Distributed Data Warehouses," International Journal of Data Warehousing and Mining, vol. 5, p. 57, 2009.

[3] M. Hamdaqa and L. Tahvildari, "Cloud computing uncovered: a research landscape," Advances in Computers, vol. 86, pp. 41-85, 2012.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, et al., "A view of cloud computing," Commun. ACM, vol. 53, pp. 50-58, 2010.

[5] I. Awoyelu, T. Omodunbi, and J. Udo, "Bridging the Gap in Modern Computing Infrastructures: Issues and Challenges of Data Warehousing and Cloud Computing," Computer and Information Science, vol. 7, pp. 33-40, 2014 2014.

[6] H. Hacigumus, B. Iyer, and S. Mehrotra, "Providing database as a service," in Data Engineering, 2002. Proceedings. 18th International Conference on, 2002, pp. 29-38.

[7] Salesforce. What is "the cloud"? Available: http://www.salesforce.com/eu/cloudcomputing/

[8] A. Roselló. (2012). Job submission to Grid Computing systems. Available: http://dana.i2cat.net/job-submission-to-grid-computing-systems/projects/

[9] G. Edmondson. (2012). Massively Parallel Processing and the Parallel Data Warehouse. Available: http://www.sqlservercentral.com/blogs/microsoft-business-intelligence-and-data-warehousing/2012/04/15/massively-parallel-processing-and-the-parallel-data-warehouse/

[10] Anonymous. (2015). Cloud Computing. Available: https://en.wikipedia.org/wiki/Cloud_computing

[11] Amazon. (2015). Data Warehouse System Architecture. Available: http://docs.aws.amazon.com/redshift/latest/dg/c_high_level_system_architecture.html

[12] Microsoft. SQL Data Warehouse. Available: http://azure.microsoft.com/en-us/services/sql-data-warehouse/

[13] IBM. Information Management System. Available: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/ibmims/