

# **StorHouse-- The eBusiness Warehouse (eBW)**

## **Concepts and Applications for eBusiness Infrastructure**

---

All rights reserved. No part of this publication may be reproduced, translated, stored in any electronic retrieval systems, or transmitted in any form or any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of FileTek, Inc.

©2000 by FileTek, Inc. Rockville, MD  
January, 2000

FileTek and StorHouse are registered trademarks of FileTek, Inc.  
All other products are trademarks of their respective owners.  
StorHouse patent pending.

---

### **FileTek, Inc.**

9400 Key West Avenue, Rockville, MD 20850  
<http://www.filetek.com> • 301/251-0600

## **Executive Summary**

Too much has been written about eBusiness. Volumes have been written regarding the application layers, web site ingenuity, and the wonders of the World Wide Web. What is missing is the infrastructure requirements needed to support these multiple applications and web site ingenuity. eBusiness can cause considerable strain on an enterprise's IT department. For example, the more an enterprise engages in eBusiness initiatives, the more information storage and retrieval requirements skyrocket.

eBusiness infrastructure is the foundation for successful eBusiness initiatives. Cornerstone to the foundation is access and availability of all of the highly granular information that is generated by eBusiness initiatives. Without all of this information, there is no significant advantage in the marketplace. To prove this point, this paper does five things:

- 1) Provides a common definition for eBusiness
- 2) Shows the dramatic increase in information storage for eBusiness enterprises
- 3) Identifies the types of highly granular information generated by eBusiness initiatives
- 4) Provides justification for storing & accessing all of the granular information over time
- 5) Suggests a means for reducing IT burden by using StorHouse as the eBusiness Warehouse (the cornerstone)

## **Introduction**

No matter how the industry pundits slice it, the phenomena is staggering...

- More than 142 million people using the Internet in 1998
- 502 million worldwide using the Internet by 2003
- 197,000 new users (a.k.a, customers) each day
- More than two new customers per second
- eBusiness revenue generation of \$48 billion in 1998 to \$1.3 trillion in 2003

The effects of eBusiness (electronic business) are hard to miss. At the macro-level, stocks have soared and the United States economy has been, and continues to be, bullish. At the micro-level, what it means to 'do business' has been redefined. 'Doing business' in today's economy means organizations need to embrace the Internet as part of their business strategy. More specifically, they must embrace the management philosophy and technical concepts necessary to support eBusiness infrastructure.

### **eBusiness Comes of Age**

eBusiness is the conduct of business on the Internet, not only buying and selling, but also servicing customers and collaborating with business partners. eBusiness can include, but is not limited to: eCommerce, Supply Chain Management, Customer Relationship Management (CRM) and Web-based Business Intelligence.

IBM pioneered the term, 'eBusiness' in October of 1997, when they launched their thematic campaign around the term. While today's business economy is defined by eBusiness initiatives, there have been other 'Internet Ages' that have brought us to the eBusiness world of today:

- 1) Age One: Information Content (1993)
  - Disseminate information
  - Reach 'eyeballs'
  - Create mindshare
  - The 'build' content phase
- 2) Age Two: Consumer eCommerce (1996)
  - Build new revenue channels
  - Revenue 'shift'
  - Remove geographic barriers to customers
  - Reach new/ more customers
  - Transaction processing/ Secure environments
- 3) Age Three: eBusiness Initiatives (2000)
  - Build more efficient business models
  - Create B2B marketplaces (trading communities)
  - Virtually integrate value chains
  - Accelerate intelligent information flow (bots)
  - Optimize customer Net Present Value (NPV)

This progression leading to eBusiness initiatives resulted in continual organizational learning and/or restructuring. Therefore it can be said that eBusiness involves the continuous optimization of an organization's value proposition and position in the value chain, using the Internet as a primary communication and trading medium.

### **Impact of eBusiness**

Pushing this continual evolution are several eBusiness objectives:

- Acquiring new customers
- Retaining existing customers
- Cross-sell/ Up-sell products
- Market Ownership
- New/ improved vendor supplier relationships

And while the eBusiness objectives may be familiar, the conditions are different:

- New, cost effective channel: the Internet
- ‘Smart-shoppers’
- Value-chain competition
- Real-time demands
- Immediacy to deliver content for market ownership
- New types and forms of value-chains

Furthermore, as organizations embrace eBusiness objectives and compete under new conditions, impact is felt throughout many different areas in an organization. For example, the objective of an eBusiness initiative may be to streamline parts purchasing and fulfillment via web-based systems using the Internet as the primary channel. In this example, impact will notably be felt in purchasing and receiving departments, and partners of the organization within their supply chain.

However, one area of the organization that is consistently affected by eBusiness initiatives is the information technology (IT) department. Although other departments (marketing, strategic planning, joint-committees, etc.) may own or sponsor the eBusiness initiative, IT has the lion’s share of the labor for implementation. (Even if the organization is attempting their eBusiness initiative by external acquisition, the IT department will still be responsible for integration). eBusiness has had significant impact with regard to an organization’s IT infrastructure. Not including labor issues and application integration, eBusiness infrastructure demands are found in two primary areas: bandwidth allocation and information storage and retrieval.

### **Bandwidth Allocation**

Bandwidth allocation has been, and continues to be, a challenge. Nielson’s law states that bandwidth will continue to lag behind demand. Moreover, Nielson’s Law<sup>1</sup> claims Internet bandwidth grows by 50% each year. There are three reasons for the lagging increase of Internet bandwidth:

- Telcos are conservative (in terms of investment and time)
- Users are reluctant to invest in bandwidth
- Internet user base is getting broader, not deeper (resulting in average bandwidth requirements shifting lower)

---

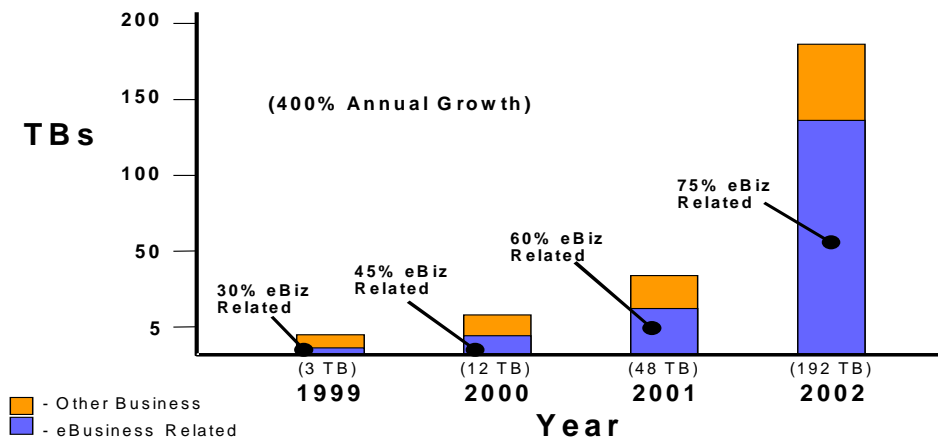
<sup>1</sup> Neilson’s Law: Jakob Nielson, Nielsen's Law of Internet Bandwidth: A high-end user's connection speed grows by 50% per year. Jakob Nielsen, Ph.D., is a User Advocate and principal of the Nielsen Norman Group. Dr. Nielsen was a Sun Microsystems Distinguished Engineer and the company's Web usability guru. Dr. Nielsen coined the term "discount usability engineering". He holds 45 United States patents, mainly on ways of making the Internet easier to use.

## Information Storage & Retrieval

Another, perhaps more challenging issue of eBusiness infrastructure is the information storage and retrieval requirements. In fact,

- eBusiness organizations' storage growth is exploding at a rate of 400%/ year
- By 2002, 75% of incremental storage demand will be 'e'Infrastructure related, with expected revenues of \$30 billion
- Within two years more than 75% of server sales will be to support eBusiness initiatives, with expected server revenues near \$80 billion.

As the chart below illustrates, for a typical business, the rapid growth in incremental storage demand will be fueled by eBusiness information:



Again, the chart above shows two important facts:

- 1) For a company engaged in eBusiness, the amount of information generated will increase by 400% each year.
- 2) By 2002, 75% of the company's information storage requirements will be eBusiness related.

Considering this increase in eBusiness generated information, the next logical question becomes "What exactly is this information, and where does it come from?"

**eBusiness Granularity: Business Intelligence Resources in eBusiness Systems**

In the most simplistic view, a client browser sends a request to a web server for resources. The web server accesses the information (resources) and sends the information back to the client browser. Although this interaction may seem simple, there is a lot of granular data and information within this transaction.

Building from the ground-up (meaning that each type of information is ‘in addition to’, not ‘in place of’ the previous data type), the most granular information, or data types within eBusiness systems are:

Network Information: This data is at the switch level (IP packets) in telecommunications systems. This may be referred to as call detail records, or IPDR (Internet Protocol Detail Records).

Atomic Data Example: Analogous to the Call Detail Record; See [www.ipdr.org](http://www.ipdr.org)

Applications: Users of this information may include web masters and network administrators. This information may be helpful in optimizing network resources, server capacity, network infrastructure planning and development, and new applications such as billing based on bandwidth usage and customer profiling from switch level information (e.g., Narus, [www.narus.com](http://www.narus.com)).

Information Amount Generated by eBusiness: Terabytes to Petabytes (Depending on the size of the network, the specific applications, and the number of users).

Log Information: This data comes from the network data within the TCP/IP packet. This data is recorded onto the web server, commonly in an access log.

Atomic Data Example:

User Host	Log-In Auth./ User	Date & Time	Request	Status	Bytes
gateway.iso.com	- -	[05/OCT/1998:00:00:49 + 000]	”GET/Index.htmlHTTP/1.0”	200	33362
gateway.iso.com	- -	[06/OCT/1998:00:00:51 + 000]	”GET/Index.htmlHTTP/5.0”	200	23581

Applications: Users of this information may include web masters and marketing managers. This information is what is conventionally thought of to be unique to eBusiness initiatives (a.k.a, clickstream, within a web site or clickthrough, to other web sites). From access logs, behavioral information of customers can be obtained, along with duration of site visit, information requested, and several other defined fields.

Information Amount Generated by eBusiness: Terabytes (Depending on the size of the network, the specific applications, the number of users and the number of web servers).

(e.g., DoubleClick, www.doubleclick.com, employs 320 web servers for web advertising analysis. DoubleClick generates 52 GB of log data per day, or 19 TB/ year—which will be 95 TB next year, and 475 TB, or nearly one half of a Petabyte by 2002).

Order Information: Data that is derived from a purchasing environment either on the web server or on another OLTP server, as directed by the web site. A customer specific ‘eTail’ market basket would be an example of order data.

Atomic Data Example:

XACTIONID	Date/ Time	SKU/QTY	Unit Price/ Ext. Price	Store ID	Tender	OnAD
1234	01/05/00	5471998/1	12.99/12.99	57	MC	N
1234	01/05/00	359187/2	4.50/8.00	57	MC	Y

Applications: User of this information may include marketing managers, supply chain partners, logisticians, and order fulfillment personnel. This information relates to what is found within each purchasing transaction (frequently within a secure environment), and is related to product, size, amount paid and other conventional market basket criteria.

Information Amount Generated by eBusiness: Gigabytes/ Terabytes  
 (Depending on the size of the network, the specific applications, number of users, number of web servers and specific purchase environments). (e.g., Dell Computers, www.dell.com, generates \$30 million/ day from their online store. This translates into 200 GB/ order data per day or 73TB year).

Considering these data types all start with the IP packet information, some may argue that these data types are all one in the same. However, an argument can be made that these data types are distinct because they each provide specific business intelligence in often very different areas as noted above. The next question becomes; “How much of the data being generated is actually useful? Do I need to store, manage, and provide access to all of this information?”

**All the Data. All the Time.**

When attempting to determine the appropriate level of detailed information or data from eBusiness initiative to keep available, opinions vary. Some feel that most information that is generated by eBusiness is of little practical value (e.g., all of the information that is generated within a web access log).

Others may argue that determining *in advance* which information from eBusiness activities should be kept, and which should not be kept is the best policy because it is not cost effective or practical to store and provide access to all of the information. For example, some argue that the appropriate grain of a fact table in a database should be one fact record for one visitor session. In other words, summarize up front-- keeping only the starting time, number of pages visited and ending time of each user visit to a web site.

Additionally, one could also argue that keeping all of the eBusiness-generated information would be difficult because of the inherent complexity of some of the data (e.g., access logs). The primary arguments for summarizing and/or *not* keeping all of the information generated by eBusiness initiatives:

- *The practical value of keeping all eBusiness information...*

Today, all eBusiness information (from access logs to credit card information from a processed transaction) has practical value. Why?

1) No one knows the effects of the legislation (i.e., tax) being proposed for the Internet. Keeping details of client's access and purchasing patterns may be the best steps for preparedness.

2) Complete Customer Information. By keeping the atomic detail of access logs and purchase details over time, more complete pictures of customer experiences will develop. This data must include all of the pages visited (e.g., all of the clickstream and clickthrough). Moreover, when this atomic data is combined with other sources of customer information (i.e., ERP, CRM, and external data) over time, a complete customer profile will be created.

- *Not cost effective for systems management/ development...*

The cost of summarizing and/or not keeping detailed eBusiness information could easily be more costly than keeping it. Why?

1) Current value of information versus future value of information. Although the current value of some of the detailed information obtained from eBusiness initiatives may be questioned, who knows what tomorrow's new analytical applications will be able to accomplish? Moreover, with a rich history of specific information, customer analysis, and even organizational analysis, becomes more accurate and poignant.

2) Statistically speaking, the more detailed information provided for decision support tools, the more accurate the response (e.g., standard deviation). In most cases, it becomes a question of marginal effectiveness for each reduction in standard deviation. For example, a real scenario may be "Will a 1% increase in the accuracy of my eBusiness channel forecasting return a significant savings or revenue return versus other channels?" The answer may be yes sometimes, and no other times.

The point is that the accurate answers to these questions are not possible without rich, (highly granular), deep (historic) readily available eBusiness information.

3) Information storage costs continue to decline. With the costs of storage media declining 35%/ year (IDC Research, [www.idc.com](http://www.idc.com)), the cost to store data has been reduced significantly. However, the demand for data storage is growing by 80%/ year.

- *Difficult because of the complexity of the data which is generated by web servers...*

‘Complexity’ implies voluminous, extremely detailed data is generated by eBusiness initiatives. While this may be true, it is important to review the granularity, probable volumes, and necessary system resources to manage eBusiness data, before making this statement.

This statement is a system resource issue. This is not a statement about the value of the granular eBusiness information. In other words, the systems used must be able to store, manage, and provide access to all of the voluminous granularity of eBusiness information. And as such, if your enterprise does not recognize the value of granular eBusiness information, your competitors more than likely will.

According to the Meta Group, ([www.metagroup.com](http://www.metagroup.com)) with ‘dot-com’ data farms, which will grow well above the 50TB range for most eBusiness companies, the challenge will not necessarily be storing the data, but keeping the stored information accessible and exploiting the stored information.

Therefore, even though the cost of storage is declining, there are other, perhaps more significant reasons why IT managers should plan for storing, managing and accessing all of their detailed eBusiness information:

- 1) Management Impact/ Effectiveness  
Data storage is a manageable issue. There is real opportunity for an IT department to impact the performance of an organization, simply by recognizing the value of all eBusiness information and implementing systems which support this information.
- 2) Regulatory/ Compliance  
Considering the increasing debates on eCommerce taxation, telco settlement processes, and content ownership, ensuring detailed eBusiness information is readily available is, to put in simply, a mandatory organizational safeguard.
- 3) ‘Routine’ Decision Support/ OLAP  
Ensuring detailed information from eBusiness systems is available for routine decision support (DSS) functions is essential for customer and product analysis, network and system resource management, etc.
- 4) Complete Customer Profile/ Management  
Perhaps one of the most important reasons to store, manage, and provide access to all of the detailed eBusiness information which is generated is to be able to provide a complete customer profile, including customer information from other sources (ERP, CRM, etc.)

- 5) **Hedge Against New Technologies and Applications**  
As new decision support, OLAP, and CRM applications come available, having complete information regarding customer's history, habits, and behavior will undoubtedly become increasingly important.
- 6) **Exploration (organizational, customer, etc.)**  
True information exploration comes being able to not only compare and analyze information in new ways, but having unlimited availability to all of the information which is needed now, or in the future.
- 7) **Information Security**  
When organizations engage in eBusiness, information security between all parties is a necessity. In the event of an unplanned outage or system failure, having all eBusiness information available and online is invaluable.
- 8) **New Business Models**  
Information is money. This can occur as a result of improved forecasting, new and improved DSS tools, or the more discreet sale of raw information. The idea is simple: harvest data reserves for new revenues by selling these data reserves (information sources) to customers. Many organizations have adopted this model. (e.g., NDC/Health Information Services, [www.ndchealth.com](http://www.ndchealth.com)).

### **StorHouse-- The eBusiness Warehouse (eBW)**

Each of these independent eBusiness information types have their own value to individuals within an organization (some even have value to individuals outside the organization, e.g., trading partners). As a result, an independent data mart can be constructed to support each one of these information types. Moreover, new 'e-centric' companies are booming as a result. For example, companies such as Accrue and net.Genesis are specific in their attempts to offer analytic solutions in the Business-to-Consumer eBusiness model.

However, attempting to compare and contrast these data types would pose a problem in this environment. What is required is a single source to be able to store, manage and provide access to each of these eBusiness information types over time—regardless of the eBusiness model. The ideal would provide scalability (in size) that is cost-effective where performance is maintained.

Moreover, keeping the most granular eBusiness information on-line and over time is necessary for numerous reasons (information exploration, current DSS applications, yet-to-be uncovered applications, regulatory, etc.).

There is nothing wrong with information summarization at the right time and for the right reasons (e.g., prior to building a cube for complex analysis, etc.). However, summarization should not occur simply as a means to be able to store, manage, and access your information.

In practical application, keeping all of the historic information on transaction processing systems or web servers is not practical, primarily for performance reasons. Moreover, the inherent current and future value of eBusiness information requires user access to this information beyond day-to-day analytical processing.

Since some of the applications for eBusiness information remain ‘untapped’, a key requirement (in addition to the routine analytical processing, and the other eBusiness applications outlined in prior sections) of the eBusiness information repository will be information exploration (e.g., ad hoc query processing).

Additionally, the Meta Group recently outlined their recommendations for a successful eBusiness infrastructure. According to Meta, “To construct robust, yet agile, eBusiness infrastructure, IT organizations must master mapping eBusiness models (B2B, B2C), onto reusable infrastructure patterns.... Bottom Line: Reusable patterns of integrated server and network components, skills, and organizational roles are the key to eBusiness infrastructure success.”

In other words, there are a number of requirements for an eBusiness Warehouse, which include but are not limited to:

- Massive scalability (In terms of data granularity and size of database)
- Information Security
- Extensibility among multiple platforms
- High data availability (Information storage and access)
- High and demonstrable ROI
- Reusable with multiple servers and infrastructure changes

These requirements are the primary reasons why StorHouse® ([www.storhouse.com](http://www.storhouse.com)) was developed. In fact, these are the inherent characteristics of StorHouse software. StorHouse software is uniquely designed to store, manage and access relational and non-relational, infinitely granular data. More specifically, StorHouse software is unique in two ways:

- 1) StorHouse software provides the most effective and efficient way to store, manage and access vast amounts of extremely granular data. StorHouse stores and provides continuous access to data, with unlimited scalability and infinite data granularity.
- 2) StorHouse software is storage media independent. StorHouse can reside on any variety or combination of storage media. It does not solely rely on disk. In fact, since StorHouse provides row-level data access directly from tape, many StorHouse customers use high-performance tape as their primary storage medium.

Together, these components make StorHouse software the ideal solution for storing, managing, and providing access to transaction-level data that is being generated by eBusiness, and other data-intensive applications (e.g., customer relationship management, business intelligence, etc).

Moreover, StorHouse was built ‘from the ground-up’ to enable eBusiness applications-- applications that require storage of very large amounts of data and Structured Query Language (SQL) retrieval for future analysis. Within these data-intensive application environments, StorHouse can be used in a variety of scenarios that include, but are not limited to: Hub & Spoke, Active Archive, and Database Extension.

*StorHouse as a Hub & Spoke*

Provides timely, shared access to enterprise-wide, infinitely granular data. As the hub, StorHouse enables data marts (spokes) to access data from a centralized, standardized single data store. StorHouse is ideally suited for the multiple applications of today, and tomorrow’s new applications. StorHouse enables data exploration and established query processing, reduces the necessity for ‘stovepipe’ systems, and increases enterprise ROI.

*StorHouse as an Active Archive*

Delivers database transparency and flexibility by extending the storage capacity of a merchant database. StorHouse off-loads data from the merchant database, yet ensures the transaction-detail is still active and readily available. StorHouse ensures access to infinitely granular data and supports multiple applications.

*StorHouse as a Database Extension*

Extends the storage capacity of a merchant, or specialized database. Like the Active Archive scenario, StorHouse off-loads data from the database and ensures the transaction detail is still active and readily available. StorHouse as a Database Extension cost-effectively increases primary database performance and enables extremely fast access to transaction-level data— without disk.

**Summary**

Industry experts are turning their attention to eBusiness infrastructure requirements. The staggering information storage requirements and need for on-line access of this information are coming to the forefront. Moreover, most of today’s databases and storage systems simply are not designed to handle the new requirements of eBusiness infrastructure (largely due to cost, performance, and overall scalability issues).

In fact, according to Meta, “eBusiness infrastructure requirements will expose brand vulnerability”. What is needed is a comprehensive approach for storing, accessing and managing the voluminous, granular information generated by eBusiness initiatives. What is needed is StorHouse.