



Meeting the Big Data Challenge

A Business and Technology White Paper

Brian Farrell
John Konya

April, 2002

Contents

THE BIG DATA CHALLENGE	1
DATA EXPLOSION	1
INTERNET TRANSACTION DATA.....	1
KNOWLEDGE IS KEY.....	2
THE CURRENT TECHNOLOGY SCENE.....	3
THE SOLUTION: CORWORKS	5
SIMPLE TO USE = SOPHISTICATED TECHNOLOGY.....	5
BEYOND COMPRESSION	7
CORDATA.....	7
RELATIONSHIP BUILDER.....	8
TIME SERIES	10
CORWORKS DATA EXCHANGE	12
CORWORKS IN ACTION	13
VIRTUAL INTEGRATED DATA WAREHOUSE.....	13
DAILY SCORING OF LARGE DATA BASES	13
CHANGING DATA SCHEMAS	13
ALL THE DATA CLOSE AT HAND.....	13
MODEL BUILDING MADE EASY	14
CONCLUSION	15
BUSINESS BENEFITS	15

Summary

BIG data is here. In fact, data doubles every eight months. If your company is using 25 terabytes of data today, it will likely need to manage one *petabyte* by 2005.

Across a wide range of industries and applications there is a search for a solution to allow the ever-growing amount of data to be converted to useful business information. One of the key components is the storage

and management of the raw data and the extraction of workable and meaningful subsets of that data for further analysis.

This paper analyzes the data explosion and current technology's inability to cope with it, and explains how the data compression solution can offer advantages above and beyond savings in just hardware expenditures. In addition, several examples are presented that demonstrate how compression technology can solve the BIG data challenges in real-world environments.

About Corworks

Corworks™ is a privately held company headquartered in Stamford, Connecticut. Corworks represents a revolution in big data management through three breakthrough products:

- 1** **Corworks Knowledge Server** is an open platform software solution that accesses files in compressed form to create speed and hardware savings.
- 2** **Corworks Data Exchange (CDX)** allows automatic data exchange to create unprecedented speed and flexibility in accessing and distributing disparate sources of data.
- 3** **Corworks Relationship Builder** is the most flexible householding product on the market, allowing you to create customer definitions in any way you choose, as often as you choose, using your existing data.

Corworks Knowledge Server was first released in 1996. Corworks Relationship Builder was released in June of 1998. And Corworks Data Exchange made its debut in March 2001. Among the accomplishments Corworks is most proud is that its customers have enjoyed enormous success and return on investment by implementing Corworks products.

Corworks set out to create software solutions to meet the coming big data explosion. Businesses around the globe can take comfort in the fact that as the demands of big data management grow, **Corworks is already there with a solution.**

The BIG Data Challenge

Data Explosion

The ability to gather information is far outpacing organizations' ability to use it effectively, a development that Corworks has identified as "Hitting the Data Wall" - when the amount of data needed to make optimal decisions and carry on operations overwhelms a firm's technology and infrastructure.

As companies grow and their data mounts (BIG Data), the complexity of the data grows exponentially. Here are just a few examples of the BIG Data challenge:

Gartner predicts "the average large enterprise will need to manage and integrate up to 500 terabytes of data by 2005. In 2004, enterprises will be managing 30 times more data than in 1999." In another study, according to the Yankee Group: "The demand for storage is growing so fast that if it takes a company one year to use one terabyte of storage today, it will take only 30 days to use the same capacity in 2002. In 2003, it will take one day to use one terabyte of storage; and in 2004, just one tenth of a day."

"At least 200 companies will need a petabyte hard drive two to three years from now," said Steve Duplessie, analyst at the Enterprise Storage Group, Milford, Mass.

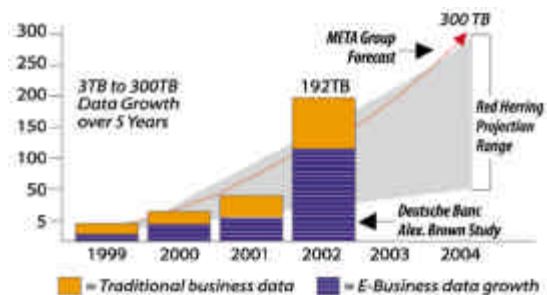
To drive business results and remain competitive, companies need to leverage their massive data stores (their most important asset) to generate greater business intelligence, data sharing, and to create new revenue streams.

What business processes are driving the creation of massive amounts of raw data?

Internet Transaction Data

The Internet is creating a torrent of customer information. In a recent *InformationWeek* article, Clinton Wilder states that "every customer action on a Web site, from a simple click-through to a complex buying transaction or product configuration generates data that can be captured and mined." Conventional technology cannot effectively manage or even capture these "customer clicks," thus ensuring that they will never be turned into useful business information.

Data Management Review indicates that E-Business data will comprise over half of an organization's data in 2002.



Source: "Surviving the Perfect Storm in Data Management" DM Review, January 2001

"The problem that IT managers now face is how to take the raw customer data obtained via the Internet and other sources and turn it into useful, strategic business knowledge that the appropriate people can access with ease. There's no magic bullet." (Michael Lattig, *InfoWorld*)

Some analysts predict that it will be at least five years before the software industry produces a product capable of effectively capturing, managing and analyzing all

potentially useful Internet clicks, rather than a very limited subset.

Historically, the IT consensus was that BIG data management could not be done effectively.

Knowledge is Key

BIG Data management goes beyond just managing large volumes of data – it should address the larger issue of Knowledge Management.

Knowledge Management is the combination of data and organizational processes that govern the creation, dissemination, and utilization of knowledge. In other words, the key is not just storing large amounts of data, but *how*, *where* and *why* the data is used.

Knowledge Management – either the work isn't connected to the knowledge or the knowledge isn't connected to the work. Companies cannot effectively manage all their customer data and often lose out on the information “gold nuggets”.

DATA HAS VALUE

Data is a valuable asset. The statement appears self-evident, but data is valuable only if it is converted to meaningful information. As the number of collection points and the amount of captured data increases, the trend is for the data to be dropped into the “bit bucket” and deleted or accumulated and then archived without close examination. In the areas of health and insurance, data is often kept simply to meet government imposed archival requirements. In the scientific field, automatic data collection devices create more data than can be studied. Busy Internet sites accumulate usage data from millions of hits per day.

CUSTOMER RELATIONSHIP MANAGEMENT

Customer Relationship Management (CRM) is the market area that deals with the issues of maintaining a 360-degree view of your customer. CRM is not only associated with customer touch points such as call centers or direct marketing campaigns, it has a more comprehensive strategic definition of encompassing the full customer interaction cycle. CRM is at the core and likely to be the preeminent area of Information Technology implementations over the next few years as organizations look to improving their customer-related activities, following a decade of concentrating on internal efficiencies through Enterprise Resource Planning (ERP) and similar endeavors.

The ability to maintain and manage large amounts of raw data, and to present that data in formats suitable for modeling, profitability analysis and customer interaction is the most critical component of the CRM infrastructure.

CUSTOMER DATA

The customer and the analysis of customer data are clearly essential to any business. Improving the performance of an organization can take place in three customer dimensions:

- Increase the number of customers
- Increase the profitability of each customer
- Retain customers over a longer period of time

Just one of these dimensions, retention, can have a major effect on a corporation.

Research analyst Alex Brown estimates that “US organizations lose one-half of their customers every five years, and a five percent incremental improvement in the customer retention rate could have the effect of doubling profits.” By being better able to use and analyze their customer data organizations can significantly improve their performance.

The Current Technology Scene

To achieve the business benefits outlined previously, the IT professional is challenged with the task of extracting useful, significant business information from a very large volume of heterogeneous data.

The challenge is sometimes overwhelming. Studies show that a very large percentage of effort, 80% according to Bill Inmon (90% according to others) is spent on collecting the data, with only 20% of effort spent on the real work – generating business intelligence. In addition, the business demand for this intelligence is rising. *Data Management Review* estimates that tens of thousands of people within large corporations have access to a wide range of corporate data.

DATA WAREHOUSING

The challenge of coping with these large amounts of data is not trivial. A body of knowledge and standard practice has evolved to meet that challenge. This revolves around the creation of data warehouses and data marts. This area is one of major growth in the IT industry – IDC estimates some \$29 billion in new data warehouse implementations through 2002.

Data warehouses are very expensive in terms of hardware to store the data and in administration to maintain them. Often the

compromise is to store only a subset of corporate data – either by restricting the types of information stored or by dropping historical data. The reason for the high cost is that the data warehouse typically uses the same basic technology used in production environments – on-line transaction processing (OLTP) systems. OLTP systems are designed for real-time updates, rapid point query response and fail-safe operation – all necessary features in their place, but cumbersome overheads for business intelligence needs.

Data marts grew to sidestep the limitations of data warehouses. The idea was to build a small collection of data, relevant to a small segment of the business. Many data mart implementations are successful and very valuable to their users. However, with proliferation of data marts, the overall picture of the organization’s business is lost. On a technical level, often the data definitions, identifying information and algorithmic logic differs between data marts.

The problems are well recognized and a continual source of debate in the industry. A number of ideas have been proposed – the use of a large data warehouse to feed “dependent” data marts is one approach. The rigorous use of standards for data and metadata is another major theme. And underlying all of these solutions is the idea of a rigorous design process within the rigid parameters of the basic relational database management system (RDBMS).

No approach entirely solves the problem, especially when trying to use legacy data stored on mainframe systems or data scattered across small departmental servers. The rigorous design ideal often fails because

The Current Technology Scene (*cont'd*)

by the time the design steps are carried out the business has often moved to another stage making the work irrelevant or outdated. As a result, IT managers admit that they handle only a fraction of the business requests for data analysis. Gartner Group estimates that 50% of all data warehouse implementations will fail or fall significantly behind schedule. In a 1999 *Data Management Review* survey, “only 24% of our readers rate their current data warehouse initiatives as successful.”

There are other complicating factors, including a shortage of skilled personnel, and rapid changes in the business environment, especially mergers and acquisitions which dump extra data formatted in unfamiliar ways into the pool.

The Solution: Corworks

Simple to Use = Sophisticated Technology

It is a truism that ease of use hides technological complexity: the electric outlet hides the power generation and distribution infrastructure, the “D” position on the shift lever hides the automatic transmission. In Information Technology, the browser hides the proxy servers and firewalls of the web, the graphical user interface covers the complexity of the operating system and layered applications. In the same manner, a user-friendly data manipulation system will be supported by an innovative technology.

Corworks has developed such a technology. The Corworks suite of products uses:

- Data compression technology
- Algorithms which allow the data to be manipulated in its compressed form
- Data Exchange
- Knowledge Center interface and Instant Queries
- The capacity of the latest generation of Very Large Memory hardware and operating systems.

What was inconceivable just a few years ago is being accomplished today by Corworks’ clients. Corworks is the first simple-to-use, comprehensive data management technology designed expressly for handling very large volumes of data for analysis purposes. The company’s core technology and customer-focused design philosophy drives every Corworks product. Each

product leverages the following core technological capabilities:

INDUSTRY-LEADING COMPRESSION

Corworks compresses data to 20% of the raw data size, on average, and allows users to work on data in its compressed state. Customers can keep all of their useful transaction, customer and business data, save on technology costs, and extend the life of existing technology investments. In fact, there is no limit to how much data can be stored using Corworks technology.

EFFICIENCY

The Corworks compression advantage is magnified by the fact that Corworks does not require the large amounts of work space, sort space, and overhead that cause traditional databases to “teraflate” to several times the raw data size and sap valuable resources for their own processes.

SPEED

Industry-leading compression translates into unparalleled processing speed. Because Corworks technology can process large amounts of compressed data in memory, Corworks products do not suffer the significant I/O exchange penalty that other products must pay to process large amounts of data. Compressed indexing and other innovations further enhance processing speed.

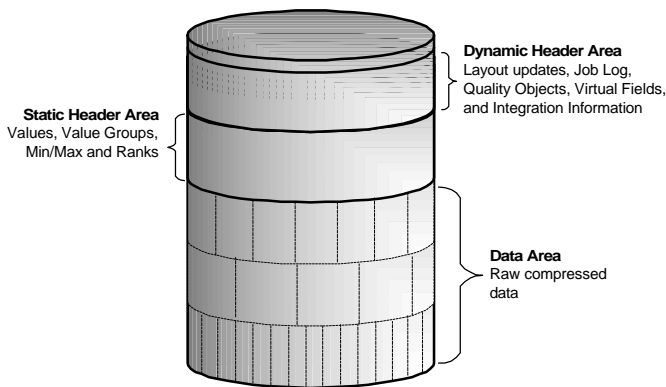
PERFORMANCE

Corworks built-in multi-threaded and parallel processing support achieves superior performance on both high-end MPP and SMP hardware platforms. IT staffs with an extensive software background in I/O intensive applications (such as OLTP) often

believe that "hardware is hardware": that there is only a few percent difference in performance between top line servers. That is sometimes a misperception. Corworks, a CPU and memory intensive application, leverages the fastest CPU's and a hardware/operating system combination that allows efficient access to very large memory spaces.

INTELLIGENCE AND FLEXIBILITY

Corworks Data Objects use "header areas" to store key data characteristics with the data itself, simplifying sophisticated data analysis.



Corworks Data Objects are malleable enough to permit end users to append new data, add new fields, and manipulate the overall data schema to suit changes in business needs and business rules. Corworks allows you to directly address and access individual records within the data object in near sub-second time. The Corworks Knowledge Server programming language (CKS), which manipulates compressed objects, allows end users the flexibility and functionality to manipulate and manage their entire set of data. Corworks product design philosophy ensures that end users have maximum flexibility in defining their own data input and output schemas and parameters.

UNIVERSALITY

Because Corworks technology automatically reads the ASCII or EBCDIC extract files of any database (together with associated data definitions), an organization can rapidly load large amounts of data into Corworks products from almost any source without intervening middleware. Thus, data from a variety of sources and formats (text, numeric, or unstructured) can be automatically standardized into Corworks format and output to a variety of previously incompatible platforms.

Corworks has effectively harnessed the power of compressed object technology to produce products that support large terabyte or *petabyte* sized applications in mission critical production environments.

Beyond Compression

Compressed data offers several advantages beyond the reduction in storage space, hardware costs and increased processing efficiency. Corworks compression delivers valuable business intelligence information (CorData), and the ability to create extracts from historical data (Time Series) and build relationships across the entire data store (Relationship Builder).

CorData

CorData is business intelligent data captured during compression, that is always there, up-to-date and accurate, and requires no additional processing or storage overhead.

The four types of CorData are:

- *Values* – the number of occurrences of each value in a field.

Rank	Count	Percentage	Value
1	1,109	44.68	REP
2	1,043	42.02	DEM
3	84	3.38	IND
4	63	2.54	LIB
5	34	1.37	REF
6	32	1.29	UNK
7	22	0.89	OTH
7	22	0.89	NNE
9	17	0.68	GRE
10	16	0.64	DFL
11	11	0.44	NLP
12	10	0.4	CRV
13	7	0.28	RTL
14	3	0.12	CST
15	2	0.08	SWP

- *Min/Max* – the minimum and maximum values in each Data Object.

Dataset	Min/Max	Count	Percentage	Value
20001201-20011201	Max	1	0	2057560
20001201-20011201	Min	15	0.01	0
19990101-19991231	Max	4	0	9350125
19990101-19991231	Min	2	0	0
19980101-19981231	Max	2	0	3000000
19980101-19981231	Min	3	0	0

- *Ranks* – similar to percentile ranking where values are separated in rank order. The count of values in each rank is determined by the number of records in the Data Object.

Rank	From Value	To Value	Avg. Value	Count	Unique Values
1	0	769	388	2,900	55
2	769	1867	1282	2,900	281
3	1867	2116	1965	2,900	93
4	2116	2186	2124	2,900	60
5	2186	2584	2289	2,900	105
6	2584	3833	3572	2,900	273
7	3833	6103	4053	2,900	433
8	6103	6830	6340	2,900	206
9	6830	7043	6931	2,900	76
10	7043	7632	7431	2,900	143
11	7632	8034	7815	2,900	161
12	8034	8884	8329	2,900	226
13	8884	10019	9058	2,900	43
14	10019	10022	10021	2,900	4
15	10022	10128	10057	2,900	46

- *Value Groups* – counts of the same and unique values in a field based on the similarity of the initial bytes.

Rank	Count	Percentage	Common Value	# of Unique Values
1	24	0.97	THO	23
1	24	0.97	MCC	23
3	23	0.93	CAR	22
4	22	0.89	MAR	19
5	21	0.85	WIL	21
5	21	0.85	SMI	21
5	21	0.85	DAV	20
8	19	0.77	SCH	19
8	19	0.77	JOH	19
8	19	0.77	COL	18
11	18	0.73	BRO	18
11	18	0.73	WAL	18
11	18	0.73	MOR	17
11	18	0.73	BAR	17
11	18	0.73	GRA	18

Relationship Builder

Corworks Relationship Builder (patent pending) easily synthesizes disparate elements of customer information from different sources by providing the ability to clearly define the business rules that identify a customer, household, or an unlimited number of user-defined identifiers.

For example, an individual might maintain a checking account, a credit card account and an insurance policy at different subsidiaries of the same financial institution. Each account record might be identified by a different account number and stored in a

different format (e.g., checking information in a relational database running on a UNIX server vs. insurance information on a mainframe). Although these accounts all belong to the same individual, the financial institution has no way to identify the relationship. Companies that cannot identify this relationship among accounts cannot effectively cross-sell complementary products to the same customer or view a total picture of the customer for CRM.

Data Source (ID = 001) - Checking Accounts

Acct #	Name	Address	Details...
357	Frank Able	Center Street	
892	Julia Able	Center Street	
314	Thomas Able	Center Street	
390	Felix Able	Railroad Ave.	
103	Ava Baird	Rural Lane	
432	James Baird	Rural Lane	
239	Evita Cross	Center Street	

Data Source (ID = 002) - Credit Card Accounts

Acct #	Name	Address	Details...
430	Frank Able	Center Street	
209	Julia Able	Center Street	
201	Felix Able	Railroad Ave.	
344	Ava Baird	Rural Lane	
556	Anne Bruin	Woods Hole	
879	Evita Cross	Cosmo Blvd.	

Data Source (ID = 003) - Insurance Accounts

Acct #	Name	Address	Details...
458	Frank Able	Center Stree	
972	Felix Able	Railroad Ave.	
304	James Baird	Rural Lane	

Relationship Builder - Anchor File

Name	Source ID #	Acct #	Customer #	Household #
Frank Able	001	357	1	1
Frank Able	002	430	1	1
Frank Able	003	458	1	1
Julia Able	001	892	2	1
Julia Able	002	209	2	1
Thomas Able	001	314	3	1
Felix Able	001	390	4	2
Felix Able	002	201	4	2
Ava Baird	001	103	5	3
Ava Baird	002	344	5	3
James Baird	001	432	6	3
James Baird	003	304	6	3
Anne Bruin	002	556	7	4
Evita Cross	001	239	8	5
Evita Cross	002	879	9	6

Relationship Builder (*cont'd*)

KEY FEATURES

- Customizes match parameters by allowing you to specify which data fields will be searched and which patterns within those fields will be accepted as a match.
- Specifies “fault tolerance” by allowing you to determine how close a match must be before it is acceptable and what type of “fuzzy logic” will be tolerated on individual fields.
- Cleanses data by comparing data input with comprehensive postal files to correct errors and eliminate duplicates.
- Identifies matches by comparing records by customer name, address and/or other user-defined variables.
- Alerts you to relationships among accounts or inserts a hard link to ensure that the relationships are always noted.
- Returns data relationship results to Corworks Knowledge Server or third party tools for further analysis and aggregation.

SUPERIOR PERFORMANCE AND FLEXIBILITY

Initial benchmarks indicate that Relationship Builder has the potential to process more than 5 million records an hour *per processor*, compared with 500,000 an hour on 4 processors for comparable products. Faster processing is made possible by compressed indexing and the opportunistic use of system memory to hold and compare compressed files.

Relationship Builder has unprecedented flexibility to specify the data “match” parameters for data cleansing and for linking accounts. In contrast with comparable products, this capability makes Corworks Relationship Builder marketable across countries and regions where naming and addressing conventions differ significantly.

Unprecedented control of “fault tolerance” allows a user to tailor searches to the data, correcting for suspected data deficiencies. For example, an end user could specify that if social security numbers in two records are identical, with the exception of one digit transposition, the record should be deemed a match despite the transposition – or checked further to determine whether phone numbers and addresses match.

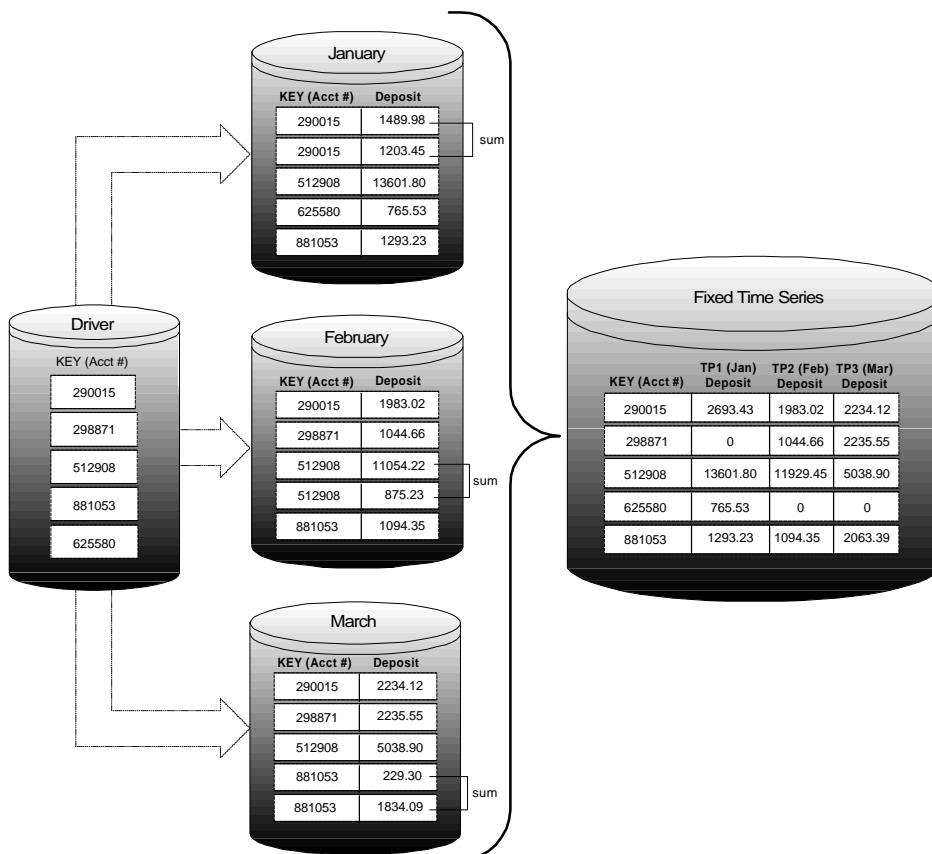
Time Series

Corworks Time Series is an internal reorganization of data from multiple files representing periods of time so that the data can be easily analyzed. For example, if a bank's customer deposit information is stored in several month files, the data can be extracted and organized in a single file where analysis can be easily performed.

Corworks can create a Time Series in one of two distinct types, Fixed or Relative, to help accomplish your data analysis goals.

FIXED TIME SERIES

A Fixed Time Series is characterized by static data. For example, if your customer deposit data is stored in several files, each file representing a month, a Fixed Time series reorganizes the data so that data from each file is placed in a single file's field representing the month. The key to the record, in this example a customer account number, is matched between each of the input month files so that deposit transactions are summed and "rolled up" to a single record. As an option, you can use another file, known as a "driver", so that only customer accounts in the driver file are processed in the Fixed Time Series.



RELATIVE TIME SERIES

A Relative Time Series is similar to a Fixed Time Series in that it is a reorganization of data from multiple files, each representing a period of time. A Relative Time Series, though, creates an internal representation that is based on a user-defined event in the time period. For example, you can determine if a credit limit increase resulted in an increase in credit spending even if the credit limit increase were granted for customer accounts in different months. The Relative Time Series internally assembles and delivers data so that analysis can be performed relative to the event (in this case, the credit limit increase), regardless of when the event occurred for each account.

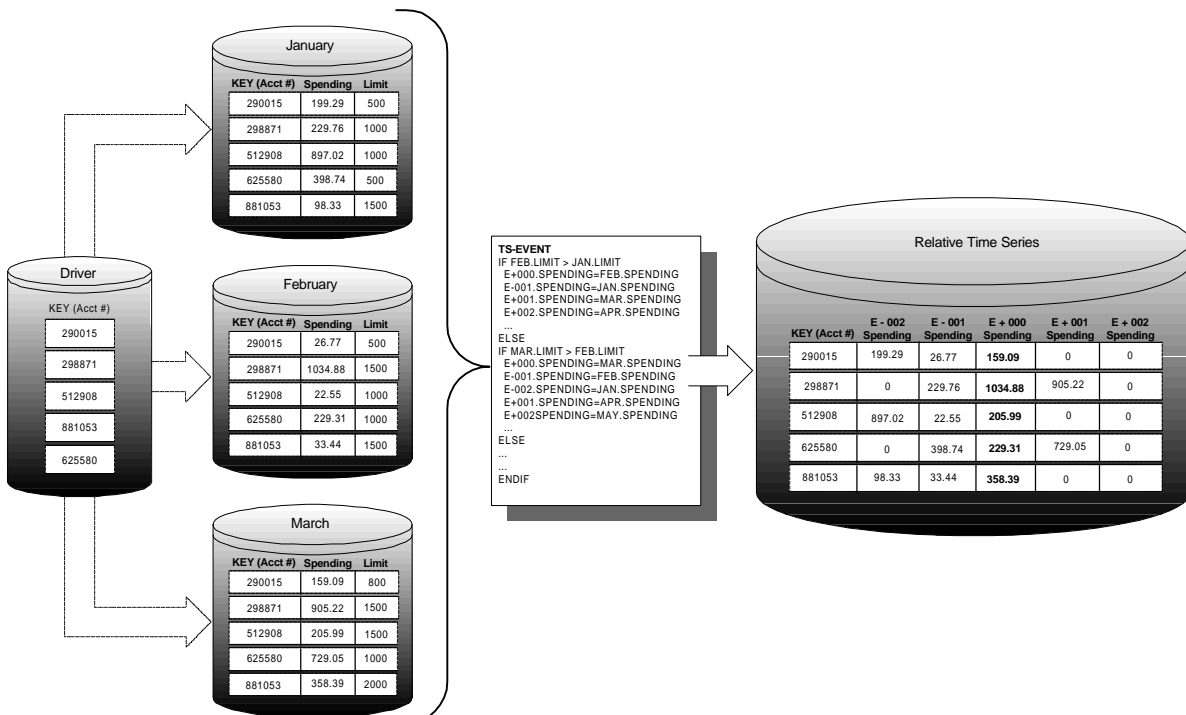
EASE OF IMPLEMENTATION

Typically, the reorganization of data from disparate historical data files could only be achieved by several days, if not weeks, of a highly skilled programmer's effort. Since the resulting program is likely to be large and complex, the probability of coding errors

increases and requires an additional test and debugging effort. Corworks simplifies the process of creating complex time series by empowering non-programmer business experts to build Time Series from their desktop, in a "drag and drop" wizard environment.

PERFORMANCE

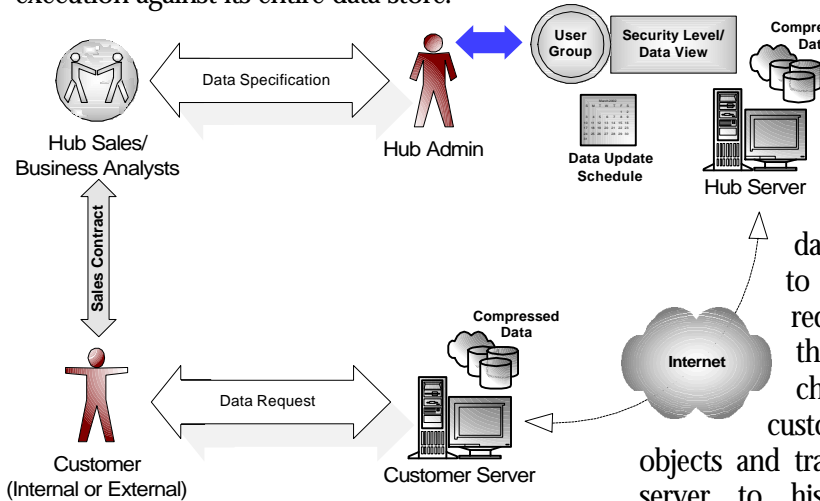
With other technologies, analyzing large volumes of data in a Time Series can decrease performance to the point that data analysts are not likely to bother with creating these kinds of analyses. The hardware may not be available, or the time required to process the entire data store may be so long as to make the project impractical. This is not the case with Corworks technology. A Corworks Time Series is created by accessing compressed files stored in memory, so that the process does not suffer the penalty of multiple I/O requests to disk storage. The result is performance that is efficient and, more importantly, *usable*.



Corworks Data Exchange

Corworks Data Exchange (CDX) is the third component of the basic Corworks suite of products that allows companies to build, market and sell information products to external customers or utilize an Internet connection to distribute data to internal customers within a company. CDX is used in a hub/client relationship where a client (a.k.a. customer) accesses and manipulates metadata to construct a customized request that is submitted to the hub for automated execution against its entire data store.

defined specifications via the Internet, the data hub provider can automate large portions of its data processing operation. Increased automation dramatically reduces data processing time and significantly improves response accuracy by eliminating the need to interpret customer specifications and standardize customer files. Because customers tailor their own electronic processing specifications, a data hub provider need not analyze and implement complex instructions from each customer.



For example, a customer wishing to submit a Corworks CDX request uses Corworks graphical user interface, Knowledge Center, to visit a data hub provider's web server to view the data elements, requests and analysis objects that the data hub provider chooses to make available. The customer selects from these

Customers equipped with CDX User (remote metadata access and compression/decompression capability) can submit their own data for processing in accordance with their metadata instructions and decompress the processed result set returned by the hub site.

objects and transfers them from the Web server to his PC. Knowledge Center integrates the objects from the data provider's web site with the customer's own objects. The customer then builds their own custom objects which contain all of the processing instructions for complete processing of the customer's request. The objects are submitted to the data hub using the CDX interface and the results are returned to the customer via the Internet or a secure network connection.

Corworks with CDX allows a data provider to create its own electronic specifications for a data request and then merge those specifications with standardized, compressed customer data before submitting the request for processing. Because it receives pre-standardized data merged with customer-

Because the process flow using CDX is automated, a data provider can process a customer request in a fraction of the time previously required to understand, standardize and code a customer data request.

Corworks in Action

Virtual Integrated Data Warehouse

The market for a product like Corworks BIG Data Management software suite has been validated by industry experts and by our own experience. Companies are seeking a way to capture, store and analyze customer data generated by operational systems, including the Internet, and third party data. They want to aggregate and analyze data across business units and platforms and organize it in a customer-centric fashion.

One prominent Corworks prospect coined the term “Virtual Integrated Data Warehouse” to describe a data store comprised of all of their available data, collected from a number of disparate source systems. They want all of their data to be readily accessible, compatible and transferable across company and external networks. Many of our other prospects are seeking the same power and flexibility. We believe that Corworks technology is the right platform for the Virtual Integrated Data Warehouse, and we have proven in live customer environments that Corworks can perform the necessary tasks.

Daily Scoring Of Large Data Bases

Our largest customers are generating enormous amounts of data, and they are seeking to use that data aggressively to generate profits and keep high-value customers. For example, one customer is using Corworks to manage more than fifteen terabytes of information, updated nightly, to identify high-risk customers and reduce bad debt write-offs. Another customer is using Corworks to replace multiple data reporting platforms with one platform that can handle a high volume of

heterogeneous data, leading to cost savings and greater consistency in reports provided to end users. We know of no other product that could have enabled these innovative business approaches.

Changing Data Schemas

One of the biggest issues facing IT departments today is how to handle file layouts that change from one month to the next, and how the changes affect their historical data. For example, in one month there are five data fields, the next month there are six, the following month the same six, but the size of one has changed from three to four bytes. Fortunately, the engineers at Corworks have designed dynamic data schema handling into the Corworks data object architecture. The flexible and independent design allows Corworks to accommodate these changing schemas. Each data object is encapsulated with the data layout for that specific point in time and may be treated as an independent entity that can be queried on its own or combined with other data objects creating a time series perspective.

All the Data Close at Hand

Studies show that in most cases 80% of the end user's effort is spent on gathering, reformatting, and transforming data to be used for building predictive models. This leaves only the remaining 20% of the time to actually use the data to generate business intelligence. If you could give the data modeler easy access to all the relevant data (depth and breadth) and minimize the data collection effort to a simple, painless task, you could reverse the 80/20 norm and make the data modeler's efforts significantly more productive. In addition, by giving data modelers access to every piece of information, not just a subset, they are

empowered to build better, more predictive models.

Model Building Made Easy

The ubiquitous tool of choice for many data modelers is SAS from the SAS Institute. Using the Corworks SAS interface module, modelers can read Corworks Data Objects as if they are generic flat files, only without the huge space overhead. The Corworks SAS interface establishes a data pipe between the executing SAS application and the Corworks compressed Data Object. This allows the modeler to sequentially read through the data and build their models in a familiar environment with a tool they are comfortable using. In some instances, the SAS code developed during the analysis phase is converted into a 'C' callable model for high-performance processing directly within the Corworks processing environment. The Corworks SAS interface module is currently in use at several prominent customer sites to access massive amounts of compressed data and perform high-end analytics.

Conclusion

Business Benefits

Corworks leverages and maximizes your hardware investment (cost savings) with hardware neutral technology. Corworks open architecture runs seamlessly across multiple hardware platforms including IBM, Sun, HP, and Compaq.

PREDICTIVE MODELING

Corworks provides the capability to capture and store entire data sets of a customer's complete record layout (width) as well as a customer's entire historical data (depth) over multiple years of data (not just months). Because Corworks stores the historical data in detailed form, you can better predict the future by using the built-in time series processing capability.

STRONGER CUSTOMER RELATIONSHIPS

Corworks enables effective 1-to-1 marketing because you have a more complete historical and external perspective on each customer.

DISK STORAGE SAVINGS

Corworks' average 80% compression ratio will significantly reduce companies' disk storage costs and protect technological investments by providing a scalable framework that grows with business needs. Scalability is as simple as compressing more data. You don't need to eliminate data to accommodate size. As the amount of data under management increases, the price per terabyte of Corworks declines making Corworks increasingly attractive as data volumes rise. In most high volume settings, Corworks will more than pay for itself in disk storage savings alone.

AVOIDS TERAFLATION

Corworks further reduces expenditures on disk space because it does not require extensive work space, sort space or overhead in order to function effectively. A traditional relational database storing two terabytes of data will require a total of between six and ten terabytes of disk space in order to hold the data and manage it effectively. Thus, a large implementation can "teraflate" to many times the raw data size. In contrast, a two terabyte Corworks implementation generally requires only 400 GB of disk space to hold raw data and maintain efficient amounts of staging and output results space.

LOW-COST IMPLEMENTATION

Corworks graphical user interface, Knowledge Center, simplifies data preparation and analysis resulting in an implementation typically requiring fewer resources. Corworks takes the guesswork out of data warehousing. You can save all your data, every row, and every column. Because Corworks can be rapidly installed and flexibly configured for the customer's data needs, project implementation times are generally between two and five months for a fairly large warehousing project. This is compared with between twelve and eighteen months of consulting time for a large relational database implementation.

PERSONNEL SAVINGS

Corworks eliminates the need to manually integrate incompatible files and manage incompatible platforms – freeing your valuable personnel to work on more productive tasks.