December 2006

Maximizing the Economic-Performance of Data-Repositories: Configuring the Optimal Time-Span

Adir Even
Boston University School of Management

G. Shankaranarayanan
Boston University School of Management

Maximizing the Economic-Performance of Data-Repositories: Configuring the Optimal Time-Span

Adir Even
Boston University School of Management
Boston, MA, USA
adir@bu.edu

G. Shankaranarayanan
Boston University School of Management
Boston, MA, USA
Gshankar@bu.edu

ABSTRACT

Corporate data-repositories within data-warehousing environments are extensively used for decision making and business-intelligence applications. Managing the growing repositories involves high implementation and maintenance costs that are largely quantifiable. Conversely, the value-contribution of such repositories has rarely been quantified and organizations often debate whether the value-gained justifies the costs incurred. Building upon an optimization model proposed earlier, this study furthers our understanding of the value of data-repositories. Based on the premise that design choices associated with a data-repository affect its value-contribution and costs, this study explores formulations that model these effects. Such models are powerful tools for cost-benefit assessment of data-repositories and can direct economically-optimal design in data management environments. The model is demonstrated for configuring the optimal time-span for a data-repository.

Keywords

Data Management, Data-Warehousing, Business Intelligence, Information Value, Optimization, Design

INTRODUCTION

In the last decade there has been a significant growth in the use of organizational data-repositories for managerial decision-making and business-intelligence applications. As these repositories are typically supported by a data-warehouse (DW) infrastructure, investments in DW environments are increasing and organizations debate whether the value-added by these investments justify the cost. Quantitative assessment of DW investments poses challenges - while technological issues and the related costs are quantified, the value-contribution of data-repositories is rarely explored and quantified. We suggest quantifying the contribution of data-repositories and including economic-performance considerations in their design process is important to the success of data management environments and merits further exploration.

The economic-performance of DW environments (and information systems in general) reflects both value and costs. Arguably, these are affected by design and maintenance choices (e.g., hardware configuration, database design, software programming, and administration procedures). This study is built on the premise that modeling and quantifying the effect of design choices on economic-performance can inform the design of data-repositories and the systems that manage it to maximize profits. Extending the preliminary framework in Even et al. (2006), this study identifies design characteristics that influence economic-performance, and attempts to model their possible effect. Such cost-benefit optimization models can help determine optimal design decisions within given constraints. This approach is demonstrated in this study for configuring the time-span - the time range of data that a repository should cover. A larger time-span can be advantageous for business analysis, but managing the larger data volume might increase implementation and storage costs. The analytical model developed here highlights the possible cost-benefit trade-offs and offers a quantitative tool for optimizing time-span.

This research offers a few significant enhancements to the preliminary framework. The preliminary framework examines only a limited set of dataset design characteristics (time-span, record density, field structure, and level of data quality), while this study broadens the discussion to other important DW design characteristics at different system-architecture levels. This study also examines multiple alternate formulations to understand the effect of these characteristics on the cost and value of the data-warehouse and its repository. Finally, to illustrate the differences in design outcome, this study models the effect of the system-characteristics in addition to time-span that was previously examined, to identify optimal designs.

The remainder of this paper is organized as follows. We first provide relevant background on data management in DW environments, highlighting both the technology and the business/economic perspectives. We then discuss possible approaches for modeling the effect of design choices on economic-performance factors such as value and cost. An application of this analytical approach is then demonstrated for the time-span configuration, highlighting possible dependencies on other design factors. We conclude by outlining the contributions of this study and proposing directions for further research.

RELEVANT BACKGROUND

Two data management perspectives underlie this study: business and technical. From the business viewpoint, data-repositories and the systems that manage them are an essential resource for operating business processes and supporting managerial decision-making. From the technical viewpoint, data management environments are complex, introduce implementation and maintenance challenges, involve many design factors, and require significant resources. This study explores possible relationships between these two perspectives – can design choices affect economic-performance? Can the maximization of economic benefits direct better design?

This study focuses on the data-warehouse (DW) – a complex IS/IT environment that supports decision making. A DW typically manages large datasets that are integrated from multiple sources, stores them in repositories, and delivers them to data-consumers. The collection of DW processes and sub-systems forms a multi-stage architecture, viewed as a Data Manufacturing Process (DMP) (Ballou et al., 1998, Shankaranarayanan et al., 2003). The DMP-output is a collection of Information Products (IP) that can be used internally or sold to others. The DMP/IP approach is adopted here for conceptualizing the economic aspects of data management. The design of the DW is complex and challenging due to the large number of design characteristics, the interdependencies among them, and the business and technical constraints imposed. The set of design characteristics defines the design space (Baldwin and Clark, 2000), and the design process is viewed as searching for optimality within this space. The challenges of implementing and maintaining DW environments have been broadly discussed (e.g., Kimball et al., 2000, Shankaranarayanan and Even, 2004). Many of these challenges have been successfully addressed by adopting dedicated architectural concepts (e.g., a centralized repository and data marts), database design methodologies (e.g., the "Star Schema" model), and commercial-off-the-shelf (COTS) DW products.

Unlike the technical aspects of the DW design, the related economic factors have not been significantly explored. The DW is assumed to contribute business-value by supporting managerial decisions and providing the infrastructure for managing IP's that are sold externally (e.g., financial quotes and consumption behavior). However, the contribution of data resources and the effects of design are rarely linked explicitly and quantitatively to economic-performance (although performance targets are commonly reflected implicitly in the business requirement that direct the design). This study explores a more explicit modeling of design decisions that may affect on economic-performance. Concepts of information economy and the information value offer important insights towards the development of such models. While the economic literature commonly views business firms as entities that maximize economic-performance, the economic-contribution of IS/IT is often not apparent and difficult to quantify. Investments in IS/IT resources alone do not necessarily guarantee competitive advantage, but rather materialize through contextual use and successful integration into business processes, together with complementary resources (Davern and Kauffman, 2000). The value of information is often viewed as the payoff-margin between perfect information versus imperfections that result in inferior outcomes and a lower willingness-to-pay (Banker and Kauffman, 2004). The tangible value can be linked to technical IS characteristics more explicitly by using the concept of utility functions (Ahituv, 1981). Quantitative mapping of design choices to utility/cost has been used for assessing quality tradeoffs and DMP-optimization (Ballou et al., 1998). This concept of utility (and cost) mapping directs the development of our model.

MODELING THE EFFECT OF DATA-REPOSITORY DESIGN ON VALUE AND COST

This section describes different formulations for modeling the effect of design decisions on the contribution of data resources to economic-performance. Even et al. (2006) suggest a microeconomic framework that informs design decisions in data management by mapping design choices explicitly to value and cost. This preliminary framing of value/cost assessment suggests the following formulation of profit as the objective-function for optimal design. Value, cost, and profit factors in this formulation are all measured in monetary units, what reflects their relative importance from the business perspective.

(1) $$P(X) = U(X) - C(X) = \sum_{i=1..I} U_i(X) - \sum_{j=1..J} C_j(X)$$, where

- X – a vector of design characteristics
- $\{U^i(X)\}$ – Value attributed to I contextual usages, indexed by *[i]*

- *{Cj(X)}* – Cost attributed to *J* cost factors, indexed by *[j]*
- *P(X)* – The overall profit contribution

A set of characteristics were shown to significantly affect value/cost trade-offs and the preliminary study models the effects of these characteristics for defining an optimal design of a tabular dataset:

- Time-span – the maximal age of the data records in the dataset (age can be typically determined from a time-stamp field)
- Record density – the number of records per time-unit (e.g., hour, day, or week, assuming possible aggregation)
- Field structure – the set of fields (attributes) that form a record (assuming identical fields for all records)
- Quality – the targeted level of a data-quality measurement (e.g., accuracy, completeness), represented as a ratio between [0, 1] (e.g., targeting 0.99 accuracy implies that in average the dataset will have 1% of inaccurate records at the most).

However, within the context of a DW environment these characteristics, although important, are a small subset of the many design and maintenance factors that can affect value and cost. These factors are reflected in the six high-level metadata-categories that have been identified (Shankaranarayanan and Even, 2004): (a) *Infrastructure*: the configuration of the underlying technical IS/IT components – e.g., hardware, operating systems, networking protocols and database servers. (b) *Model*: the database conceptual and logical design – e.g., entities, relationships, attributes, and value domains. (c) *Process:* configuration of back-end data processing – e.g., data sources, storage-targets, data transformations, quality monitoring procedures and the business-rules that drive them. (d) *Content:* metadata the configuration and the contents of the actually-stored datasets – e.g., the number of records, the granularity, and the quality level. (e) *Presentation:* the preparation of data for consumption – e.g., data formatting, report generation, and delivery, and (f) *Administration:* managing the data acquisition, processing and use – e.g., security configuration, access privileges, and process or usage tracking. Many of these factors can be shown to introduce significant value/cost trade-offs. In general, enhancing the DW capabilities along all categories can improve (directly or indirectly) the benefits derived from the information products that the DW supports. At the same time these enhancements involve additional costs. Table 1 provides some examples of DW design characteristics along the six functional categories and briefly explains their anticipated effect on value and cost.

Modeling the value/cost trade-offs introduces significant challenges. First, due to the large number of design characteristics, and the possible dependencies among them, the design-space might be large and complex. It is therefore important to reduce the complexity by identifying characteristics that have stronger impacts and/or encapsulating subsets of characteristics with mutual inter-dependencies (i.e.., following the concept of modular design (Baldwin and Clark, 2000)). Second, cost/value effects of the design characteristics are likely to change over time and/or be subject to uncertainty due to technology progress and/or changes in usage patterns. Therefore, modeling these effects has to consider random and dynamic behavior. Third, the analytical representation of the effect of design decisions on value and cost may take different "shapes and forms" – for example, a choice within a discrete set of design alternatives (e.g., selection of a database server) versus selecting a value along a continuum (e.g., setting quality-level targets), or monotonous input/output relationships (e.g., increasing value with broader time-coverage) versus non-monotonous (e.g., initially increasing and then decreasing value due to information overload effects). *This study addresses the latter challenge – examining the large variety of cause-effect relationships by exploring analytical formulation alternatives.*

The preliminary framework by Even et al. (2006) takes a deterministic approach and formulates dataset design as a mixed-integer optimization problem. Value is argued to have a diminishing return with time-span and record density, and diminishing marginal-return with the decrease in data quality. Cost is assumed to have both fixed and linear components; and both value and cost were modeled as growing incrementally with field structure. These preliminary formulations can be potentially used for other value/cost effects, and alternative formulations may turn to be useful as well. Table 2 provides examples of fundamental utility formulations and their possible use in design scenarios. Some of these formulations have not been examined in the preliminary study - S-shaped utility, for example, may reflect demand patterns of information products that are subject to positive network effects; and declining utility may turn to be useful for modeling information-overload effects or decrease in usability when data delivery is delayed.

Table 3 presents similar examples of cost formulations. Choosing the appropriate formulation from possible alternatives may depend on the information-system architecture and/or on the business environment. In certain contexts, for example, data-quality is easier to manage and therefore the cost of error-free datasets is possibly bounded (e.g., small datasets from a single internal source). In other contexts, maintaining high data-quality is challenging and reaching an error-free level (i.e., quality-ratio of 1) is practically impossible (e.g., large datasets, collected from multiple internal and external sources). In such cases, the unbounded formulation of quality effects appears to be more appropriate.

The formulations in Table 2 and 3 are presented in a parametric form, where the actual parameter values will be influenced by implementation factors such as technological infrastructure, human skills, organizational knowledge and managerial overhead. These values can be obtained from known information (e.g., performance measurement, or pricing), assessed empirically, or estimated using decision-calculus methodologies (Little, 1970). *Evaluating these alternative formulations is the second contribution of this work.* However, other formulations are possible and ought to be further explored in the future.

Design Category	**Design Characteristics**	**Possible Utility Effects**	**Possible Cost Effects**
Infrastructure	• Hardware • Software platforms • Database server • Network	• Higher capacity to store and process large datasets may increase utility potential	• Higher capacity and performance typically require higher investment in technology and programming efforts
Model	• Table structure – entities, attributes, relationships • Constraints and value-domains	• Richer data-structure has higher utility-contribution potential • Some fields may offer only small marginal contribution	• Richer data structure implies higher data acquisition cost • Complex dependencies increase the chance of data quality hazards and, hence, costs
Process	• Data sources • Process configuration • Business rules • Processing software	• Larger variety sources increases data richness • Advanced processing utilities (e.g., automated ETL – Extraction, Transformation, Loading) enhance the capability to manage larger volumes of data • A proper set of business rules can improve the data fitness-to-use	• Handling a large variety of sources may increase acquisition and processing costs • Advance ETL software might bye costly to purchase and program • Complex or vaguely-defined business rules increase the chance of data quality hazards
Content	• Data volume • Granularity • Data quality level	• Utility increases with volume and granularity, but possibly at a diminishing return • Utility diminishes as quality degrades • Overly-detailed data increases the risk of information overload and might lower utility	• Higher volume and granularity increase storage-space • Maintaining high data quality requires investment in monitoring and cleansing • Low data quality damages reputation and causes opportunity loss
Presentation	• Reporting tools • Business-Intelligence (B.I.) applications • Automated delivery mechanisms	• Timely and error-free reports increase utility • Sophisticated B.I. increases analysis capabilities • Efficient data delivery may increase external demand for data products	• Reporting and delivery comes at a cost (formatting, software, delivery technologies) • Sophisticated front-end tools introduce training costs
Administration	• Access authorization • Security • Usage monitoring	• Higher security increases trust and potentially the willingness-to-pay • Usage tracking provides important input to system improvement	• Better administration capabilities require establishing proper procedures and investment in utilities

Table 1. Examples of DW Design Characteristics and Their Possible Effect on Value and Cost

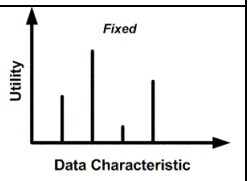
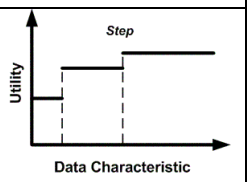
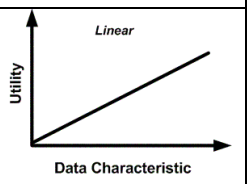
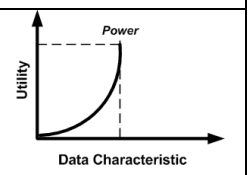
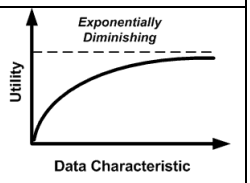
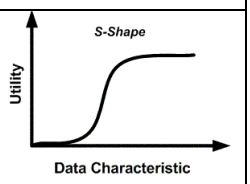
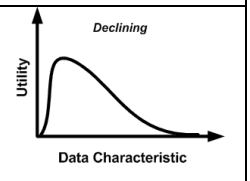
Utility	Response to Input Level	Possible Applications	Parametric Formulation	Illustration
Fixed	Discrete set of feasible options, each with a fixed-level utility effect	• Different subscription-levels to data or report services • Increasing the number of fields in a given dataset • Enhancing a set of reports	$U(x)=\begin{cases}K1, & x=x1\\ K2, & x=x2\\ K3, & x=x3\ldots\end{cases}$	
Step	Fixed utility within certain level-ranges	• Incremental charge for data services, based on data-volume ranges	$U(x)=\begin{cases}K1, & 0\le x<x1\\ K2, & x1\le x<x2\\ K3, & x2\le x\end{cases}$	
Linear	Utility is linearly proportional to input level	• Per-record charge for data services	$U(x)=ax$	
Power	Bounded utility, where marginal value increases with level	• Utility increase with higher quality level	$U(x)=Kx^{g}$	
Exponentially Diminishing	Bounded utility, where marginal value decreases with level	• Utility increase with higher data volume (Intra-firm use)	$U(x)=K(1-e^{-ax})$	
S-Shape	Bounded utility, where marginal value first increases and then decreases with level	• Utility increase with higher data volume (Inter-firm use)	$U(x)=\frac{K}{1+e^{-a(x-b)}}$	
Declining	Utility declines beyond certain level	• Information overload – interpretability decreases with too-detailed data • Timeliness – usability decreases with latency	$U(x)=Kxe^{-ax}$	

Table 2. Utility-Formulation Examples

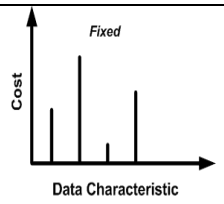
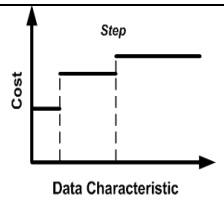
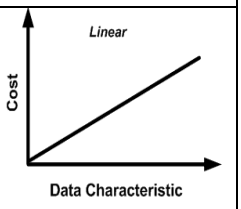
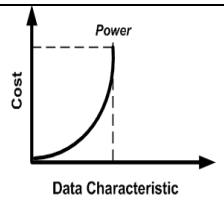
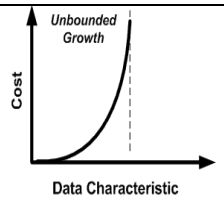
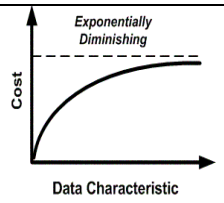
Cost	Response to Input Level	Possible Applications	Parametric Formulation	Illustration
Fixed	Discrete set of feasible options, each with a fixed-level cost effect	• Purchase-price of technology platforms • Programming efforts related to possible enhancements of a data model	$C(x)=\begin{cases}K1, & x=x1\\ K2, & x=x2\\ K3, & x=x3\ldots\end{cases}$	
Step	Fixed cost within certain level-ranges	• Bulk end-user software licensing, per ranges of user-base • Investment in hardware/software platform as a function of the data volume	$C(x)=\begin{cases}K1, & 0\le x<x1\\ K2, & x1\le x<x2\\ K3, & x2\le x\end{cases}$	
Linear	Cost is linearly proportional to input level	• Data acquisition cost (with per-record pricing policy) • Data storage cost • Manual error detection and correction	$C(x)=ax$	
Power	Bounded cost, where marginal cost increases with level	• Quality maintenance cost, assuming that perfection can be obtained at a finite cost	$C(x)=Kx^{g}$	
Unbounded	Unbounded cost, where marginal cost increases with level	• Quality maintenance cost, assuming that perfection cannot be obtained at a finite cost	$C(x)=\left(\frac{X}{b-X}\right)^{a}$	
Exponentially Diminishing	Bounded cost, where marginal cost decreases with level	• Programming and administration efforts with growing data volume	$C(x)=K(1-e^{-ax})$	

Table 3. Cost-Formulation Examples

APPLICATION: OPTIMIZING THE TIME-SPAN COVERAGE

This section demonstrates the use of economically-driven modeling for optimizing design decisions, where the design objective is an important configuration parameter in DW environments – the time-span. The preliminary model considers the time-span as one among a set of independent variables. Here we focus the attention on this variable, but take into account related system-configuration decisions that may also affect a profit-maximizing design. The result is a different and

significantly enhanced value/cost formulation. *This new formulation, and the optimization-search algorithm that we develop for obtaining design optimality within it, is an additional important contribution of this research.*

As in the preliminary framework, we assume here that recent data is more valuable and that the time span coverage parameter *(T)* acts as "cut-off"– only data items more recent than *T* will be kept in the data-repository. The model developed here attempts to answer the design question of choosing *T* such that the benefit within the value/cost tradeoffs is maximized. DW datasets typically contain a large number of records *(N)* with an identical field-structure (columns or attributes). We assume that record-age *(t)* can be determined (e.g., a timestamp that reflects the data record's most recent update) and that *t* can be reasonably treated as a continuous variable. We also assume that the records are uniformly distributed over time (i.e., *N(T) = R*T*, where *R* reflects a given record density). A record can contribute value within multiple business scenarios and the record value (aggregated along all usages) will depend on its recency (i.e., *v(t)*). Realistically, record-value may be affected by other factors such as its quality level or the actual data-content (Even and Shankaranarayanan, 2005). However, for our analytical development, we assume dependency on *t* only. The record-value is assumed to decline as *t* grows, with a value-decline slope that is negatively-proportional to the value-level:

$$\frac{dv(t)}{dt} = -\alpha v(t) \tag{1}$$

Solving (1) with $v(t = 0) = v_0$ (representing the value of most recent data records) yields:

$$v(t) = v_0 e^{-\alpha t} \tag{2}$$

The overall dataset utility *U(T)* for records with recency $t \le T$, is the aggregated sum of record value (Figure 1):

$$U(T) = \int_0^T Rv(t)dt = \frac{v_0 R}{\alpha}\left(1 - e^{-\alpha T}\right) = U_{\max}\left(1 - e^{-\alpha T}\right), \text{ where } U_{\max} = \frac{v_0 R}{\alpha} \tag{3}$$

This suggests that the dataset utility *U(T)* is monotonously increasing with the time span *T*, but the increase is exponentially-diminishing and the value is capped by U_{max}. The value cap is linearly-proportional to v_0 and *R*, but inversely-proportional to α. A high α also implies faster convergence towards the value cap and less dependency on outdated data. The actual values of α and v_0 and would be estimated empirically, or solicited from the appropriate manager using decision-calculus (Little, 1970).

The cost of maintaining the data-repository stems from the DW configuration – hardware and software components (e.g. computers, networks, database servers, and data-processing engines) and their integration. The design space, based of the selection of DW components, is a finite set of configuration-choices, indexed by *m*. We assume a maximal time-span capacity per configuration T'_m and a fixed setup cost of C^S_m. Additional cost can be attributed to the dataset time-span configuration (e.g., related to data acquisition, storage-space allocation, and data quality maintenance). Within a specific system configuration, we assume an average cost-per-record of c^D_m. The overall data cost $C^D_m(T)$ is therefore proportionally linear to the time-span *T*, to the record density, *R*, and to the per-record data-cost, c^D_m. Combining the setup cost C^S_m and the data cost C^D_m for each design configuration, we get:

$$C_m(T) = C^S_m + C^D_m = C^S_m + c^D_m RT \tag{4}$$

We assume that technical DW system-configuration does not affect the value but the cost. For a given time span T, the design configuration with the lowest cost will be preferred. The overall result (Figure 2) is a stepwise-linear, monotonously-increasing, and possibly discontinuous curve that represents a *cost efficiency-frontier* (highlighted in Figure 2).

Within these value/cost assumptions, what would be the *optimal time-span choice*? Optimality depends, of course, to the optimization objective. Here we optimize the profit defined as the difference between value and cost*: P(T) = U(T) – C(T).* For each segment of the cost efficiency-frontier (bounded by a time-span range $[T^-, T^+]$), the profit is given by (Figure 3):

$$P_m(T) = U(T) - C_m(T) = U_{\max}\left(1 - e^{-\alpha T}\right) - C^S_m - c^D_m RT = \frac{v_0 R}{\alpha}\left(1 - e^{-\alpha T}\right) - C^S_m - c^D_m RT, T^- \le T \le T^+ \tag{5}$$

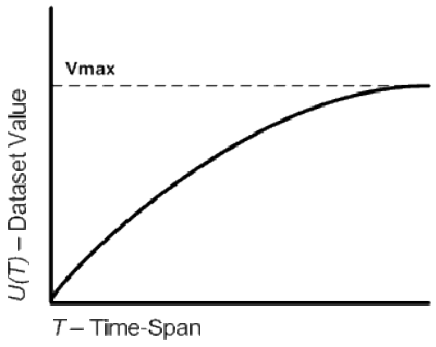


Figure 1: Marginal Added-Value Decease vs. Time-Span

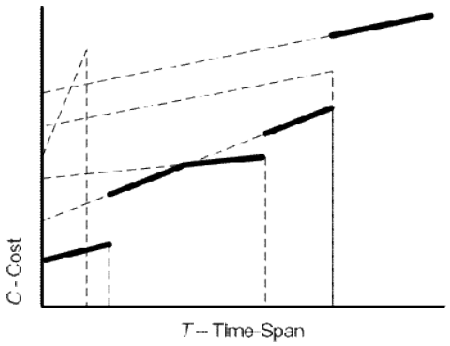


Figure 2: Cost vs. Time-Span (Efficiency-Frontier Highlighted)

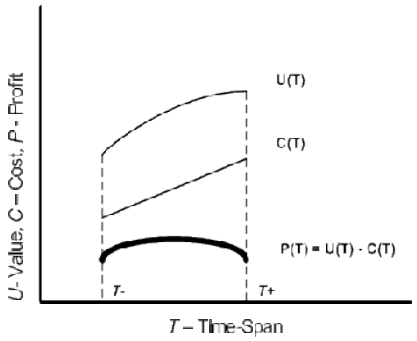


Figure 3: Value, Cost, and Profit vs. Time-Span (per Design Configuration)

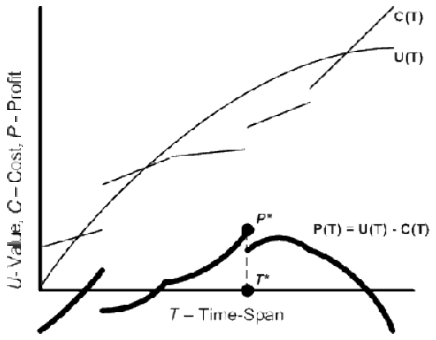


Figure 4: Value, Cost, and Profit vs. Time-Span

The optimal span within the segment can be obtained from $\partial P(T)/\partial T = 0$:

$$T_m^{OPT} = Ln\left(\left(\frac{v_0}{c_m^D}\right)^{\frac{1}{\alpha}}\right) \quad (6)$$

The profit at this optimal point can be shown to be:

$$P_m\left(T_m^{OPT}\right) = \frac{R}{\alpha}\left(v_0 - c_m^D\left(1 - Ln\frac{c_m^D}{v_0}\right)\right) - C_m^S \quad (7)$$

The second derivative of (5) is:

$$\frac{\partial^2 P}{\partial T^2} = -\alpha v_0 \,\mathrm{Re}^{-\alpha T} \quad (8)$$

This second derivative is always negative; hence, *P* is a concave function. If T^{OPT} is within the range of $[T^-, T^+]$, it has a maximum-profit point for this segment. Otherwise, *P(T)* has to be monotonous within the segment and the maximum profit is either at T^- or at T^+. These possible options can be calculated using (5) and than compared. Assuming a finite set of segments, the optimal time-span can now be obtained by repeating this evaluation procedure for each segment (Figure 4).

The process for obtaining the optimal time-span configuration can be algorithmically summarized as

1. Estimate the value curve parameters: R, v_0 and α, and calculate V_{max}
2. For each feasible system configuration, obtain the cost curve parameters: T'_m, C^S_m and c^D_m
3. Comparing the cost of all feasible configurations, obtain the segments of the cost efficiency-frontier curve and their boundaries ($[T^-, T^+]$).
4. Initiate $T^* = 0$, $P^* = 0$, and $m^* = 0$
5. For each segment of the cost efficiency-frontier, repeat the following steps:
    a. Obtain T_m^{OPT} from (6)
    b. If T_m^{OPT} is within the segment boundaries:
        i. Obtain P_m^{OPT} from (7)
        ii. If $P_m^{OPT} > P^{OPT}$, set $T^* = T_m^{OPT}$, $P^* = P_m^{OPT}$, and $m^* = m$, and proceed to the next segment
        iii. Otherwise retain T^*, P^*, and m^* and proceed to the next segment
    c. Otherwise, if T_m^{OPT} is not within the segment boundaries:
        i. Obtain $P_m(T^-)$ and $P_m(T^+)$ from (5)
        ii. If $P_m(T^-) > P_m(T^+)$ and $P_m(T^-) > P^*$, set $T^* = T^-$, $P^* = P_m(T^-)$, $m^* = m$ and proceed to the next segment
        iii. If $P_m(T^+) > P_m(T^-)$ and $P_m(T^+) > P^*$, set $T^* = T^-$, $P^* = P_m(T^-)$, $m^* = m$ and proceed to the next segment
        iv. Otherwise retain T^*, P^*, and m^* and proceed to the next segment
6. The variable m^* now reflect the design configuration that yields the optimal profit. T^{OPT} reflects the optimal time-span setup within this design configuration, and P^{OPT} reflects the anticipated profit.
    a. Notably, if finally $m^* = 0$ *(and obviously* $T^* = 0$ and $P^* = 0$*)*, the anticipated value does not exceeds the cost for any design configuration and, hence, the entire implementation can not be profitable and should be reconsidered

Illustration: A chain of retail stores plans to establish a DW for the firm's sales transactions, to support the analysis of consumption behavior. The designers of the DW debate the time-span to be covered. For assessing the potential value, the average number of data records per month is estimated as $R=100{,}000$; the maximal value contribution per record as $v_0 = \$5$, and the value decline slope as $\alpha = 0.1$. The estimated value cap is therefore $U_{max} = v_0 R/\alpha = \$5{,}000{,}000$. Four system design alternatives are considered, derived from infrastructural choices (e.g., hardware, operating system and database server) and their cost parameters are summarized in table 4. These alternatives significantly differ in their cost and capacity. For example, the more advanced alternative (#4) has a significantly higher fixed setup cost, but supports a higher time-span with a lower variable cost compared to all other alternatives.

Design Choice (m)	Time-Span Capacity (T'_m, in months)	Fixed Setup Cost (C^S_m)	Variable Data Cost Per Record (c^D_m)	Variable Data Cost Per Month ($R*c^D_m$)
1	6	$500,000	$2	$200,000
2	24	$1,500,000	$1.8	$180,000
3	48	$2,000,000	$0.8	$80,000
4	60	$3,500,000	$0.3	$30,000

Table 4. Cost Parameters of Design Alternatives

Given these estimates, *what would be the optimal time-span coverage, and which design system configuration should be chosen?* A first step is obtaining the cost efficiency-frontier (Figure 5). It can be shown analytically that within the range of 0-6 months, design alternative 1 would have the lowest cost, within the range of 6-30 months it would be alternative 3, and within 30-60 months – alternative 4. Figure 6 illustrates the value, cost and profitability at all segments and the evaluation steps are summarized in table 5. From this the optimal profit can be obtained by setting the time-span coverage to

approximately 18.32 months and choosing design alternative 2. This time-span choice implies storing the most recent 1,832,000 data records and the anticipated profit is $733, 934.

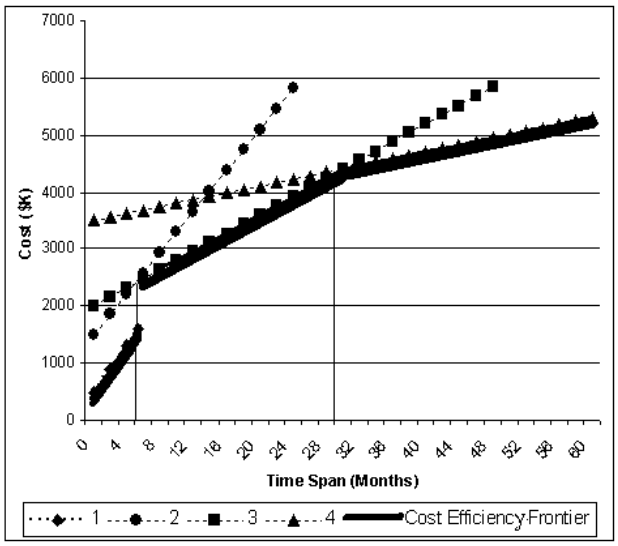


Figure 5: Cost vs. Time-Span (Efficiency-Frontier Highlighted)

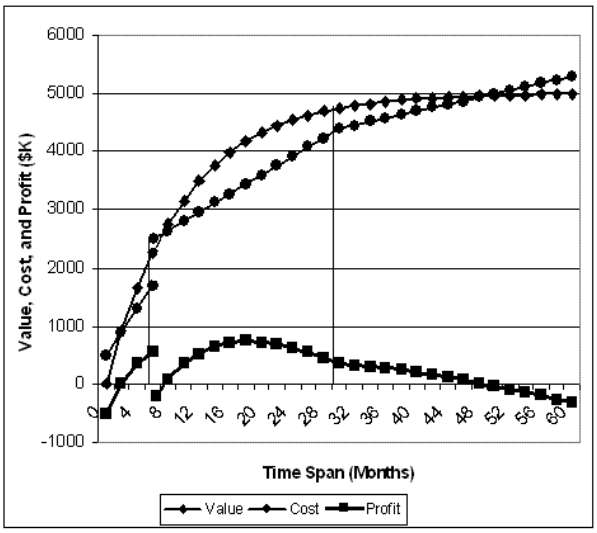


Figure 6: Value, Cost, and Profit vs. Time-Span

Time-Span Range (m)	Design Choice	T^{OPT}	Within Boundaries?	$P(T^{OPT})$	$P(T^{-})$	$P(T^{+})$	Optimal T	Optimal P
0-6	1	9.16	No	-	(-$500,000)	$555,941	6	$555,941
6-30	***2***	18.32	Yes	$733,934	-	-	***18.32***	***$733,934***
30-60	4	28.13	No	-	$351, 064	(-$312,394)	30	$351,064

Table 5. Profit Evaluation for Different Design Alternatives

This illustrative example highlights the benefits of economically-driven design. First, technology considerations alone (e.g.., maximizing the volume within a given capacity limit) might turn to be economically sub-optimal. Here, maximizing the time-span to 60 (or beyond) results in a significant loss. Second, quantitative value/cost models provide a powerful approximation tool for determining the optimal configuration. Third, such modeling highlights dependencies on other design factors (here, the system configuration) and can help configuring them as well. Notably, the time-span coverage is only one among many design parameters and including these in the optimization schema is likely to significantly change the results.

CONCLUSION

Observing the rapid growth of data-warehouses, this study explores possible economic drivers behind DW implementation. Assuming that the main driver behind this implementation is the enhancement of the firm's economic-performance, we suggest that the design of DW repositories should be driven not only by technological and functional considerations, but also by economic factors such as value-contribution and costs. By enhancing a preliminary framework for value-driven optimization, this study identifies DW design characteristics that may affect value and cost and explores modeling alternatives for understanding the value/cost effects. To demonstrate these enhancements, this study develops further the decision-model for time-span coverage. As illustrated by the model, the time-span decision may affect not only dataset design, but also the entire system configuration, as the time-span coverage of certain configurations might be limited. The suggested evaluation procedure examines the value-cost tradeoffs along the cost efficiency-frontier and helps detecting the optimal time-span coverage as well as the system configuration that supports it in the most cost-efficient manner.

As illustrated, considering value-cost tradeoffs may significantly affect design decisions. However time-span coverage, although important, is one design decision among many others and considering them all is likely to lead to a significantly more complex decision-scenario. Therefore, this research would benefit from identifying other economically-influential design factors, quantifying their possible effects, and developing design decision models accordingly. Future enhancements need to reflect some realistic considerations, such as usage uncertainty, mutual dependencies among design characteristics, and dynamically-changing behavior over time. The concept of design for optimal economic-performance can be extended to the design of information systems in general. Although the economic aspects of IS have drawn significant attention, their impact on IS design methodologies is not apparent. Examining economic considerations, integrating them into the IS/IT design process, and quantifying the possible effects can improve IS design from the business perspective.

REFERENCES

1. Ahituv, N. (1980) A Systematic Approach Towards Assessing the Value of Information Systems, *MISQ*, 4, 4, 61-75
2. Baldwin, C. Y., and Clark, K. B. (2000) Design Rules, Vol. 1: The Power of Modularity", MIT Press, Cambridge, MA,
3. Ballou, D. P., Wang, R., Pazer, H., and Tayi, G. K. (1998) Modeling Information Manufacturing Systems to Determine Information Quality, *Management Science*, 44, 4, 462-484
4. Banker, R. D., and Kauffman, R. J. (2004) The Evolution of Research on Information Systems: A Fiftieth-Year Survey of the Literature in Management Science, *Management Science*, 50, 3, 281-298
5. Davern, M. J., and Kauffman, R. J. (2000) Discovering Potential and Realizing Value from Information Technology Investments, *Journal of Management Information Systems*, 16, 4, 121-143
6. Even, A., and Shankaranarayanan, G. (2005) Value-Driven Data Quality Assessment, *Proceedings of the Tenth MIT International Conference on Information Quality*, Boston, MA, 221-236.
7. Even A., Shankaranarayanan, G., and Berger, P. D. (2006) Economic-Based Design of Data Management Systems, *Proceedings of the 1st International Conference on System Design*, Claremont, CA
8. Kimball, R., Reeves, L., Ross, M., and Thornthwaite, W. (2000) The Data Warehouse Lifecycle Toolkit, Wiley Computer Publishing, New York, NY
9. Little, J. D. C. (1970) Models and Managers: The Concept of a Decision Calculus, *Management Science*, 18, 466-484
10. Shankaranarayanan, G., Ziad, M., and Wang, R. Y. (2003) Managing Data Quality in Dynamic Decision Making Environments: An Information Product Approach, *Journal of Database Management*, 14, 4, 14-32
11. Shankaranarayanan, G., and Even, A. (2004), Managing Metadata in Data Warehouses: Pitfalls and Possibilities, *Communications of the AIS*, 2004, 14, 247-274