Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2006 Proceedings

International Conference on Information Systems (ICIS)

December 2006

Data-Warehouse as a Dynamic Capability: Utility/Cost Foundations and Implications for Economically-Driven Design

Adir Even
Boston University

G. Shankaranarayanan
Boston University

DATA WAREHOUSE AS A DYNAMIC CAPABILITY: UTILITY/COST FOUNDATIONS AND IMPLICATIONS FOR ECONOMICS-DRIVEN DESIGN

Design Science

Adir Even
Boston University School of Management
Information Systems Department
Boston, MA
adir@bu.edu

G. Shankaranarayanan
Boston University School of Management
Information Systems Department
Boston, MA
gshankar@bu.edu

Paul D. Berger
Principal and Chief Consultant
PDBerger Consulting
Southboro, MA
pdberger@alum.mit.edu

Abstract

IS design today is driven primarily by technical and functional requirements, and the economic implications for design are not yet well understood. This study argues that system design and architecture must reflect assessments of economic trade-offs besides satisfying technical/functional requirements. Modeling the economic performance structure behind IS design can highlight these trade-offs and help economically assess design alternatives. This study examines economics-driven design in the context of the Data Warehouse (DW). The DW environment is treated as a dynamic capability, providing the capacity for managing data resources and turning them into useful information products. These products contribute value when used for exploitative and/or explorative business processes. Recognizing possible uncertainties in usage, DW capacities are evaluated as real-option investments toward the development of a framework for modeling cost-utility effects of DW design decisions. This framework is used to evaluate important design scenarios along the layers of a DW stack architecture and optimize design outcomes accordingly.

Keywords: Data management, data warehousing, design, real-options, dynamic capabilities

Introduction

To what extent does the design of an information system (IS) affect economic performance? Can maximizing performance direct design? These questions highlight an important gap – an insufficient economic underpinning of the IS design. This study suggests that IS design can be enhanced by rigorously and explicitly linking design decisions to their economic consequences, arguing that these decisions influence the utility of information products and their production costs. This is achieved by identifying design characteristics that drive economic trade-offs and constraints and modeling their effects toward assessing design alternatives and optimizing design outcomes. The

economics-driven design is examined in the context of a Data Warehouse (DW) – an IS-environment that manages large data resources that support data-driven business processes. DW environments involve high implementation and maintenance costs that have been investigated and quantified. Conversely, the business-value contribution of data repositories has rarely been assessed, and firms often debate whether the benefits justify the costs. The study links this cost-benefit trade-off to DW design. It argues that understanding the utility/cost structure of the DW design can improve design outcomes and better assess investment decisions.

To further this, we explore factors that drive utility and costs in DW environments. We conceptualize the DW as a dynamic capability for managing critical data resources. These resources support established business processes, referred to as exploitative usage, and can simultaneously promote explorative use. This ambidextrous usage of the DW permits examining design alternatives as real-option investments – "lean" design that targets exploitative usage with relatively certain value contributions, versus investments in slack capacity to accommodate exploration with potentially high but often uncertain gains. We examine the possible utility/cost implications of this interplay between investing in excess capacity versus exploitative optimization along the layers of a typical DW stack architecture: data usage, data delivery, data resources, and system infrastructure. This layered-model provides the foundation for quantifying economic effects – utility gained from using information products versus their manufacturing costs. It allows modeling the effects of design characteristics and highlights possible vertical dependencies across the architectural layers. These models are first described in a high-level analytical form and then developed for some illustrative design scenarios to demonstrate their use as tools for DW design.

A key contribution of this study is exposing the economic structure behind the DW architecture and the related design decisions. This perspective complements existing DW design methodologies; while satisfying the technical/functional requirements is clearly necessary, enhancing it with a rigorous economic analysis can benefit the design from a business perspective. In the rest of this paper, following a literature review, we lay the theoretical foundations for the economics-driven design of a DW environment by examining DW design as an optimization problem, proposing a stack architecture for it, and framing capacity decisions in such environments as real-options investments. We then model the utility/cost effect of key DW design characteristics and demonstrate related decision scenarios. Finally, we highlight limitations and propose directions for further research.

Relevant Background

We examine the link between economic performance and design decisions in the context of a DW. Conceptualizing the DW as a dynamic capability, we explore relevant modeling methods. These help define the underlying utility/cost structure for DW design and develop an economics-driven design framework.

Data-Warehousing Environments

A DW is an IS/IT environment that manages large-scale data repositories and supports decision-making (Kimball et al., 2000). DW environments support a large variety of usages, such as transforming corporate strategies (Cooper et al., 2000), segmenting customers (Lee et. al., 2004), optimizing supply chains (Shin, 2003), improving operational efficiency (Srivastava and Chen, 1999), and delivering on-line data products (West, 2000). Their popularity can be attributed to benefits such as covering a broad range of business perspectives by integrating multiple data sources, allowing reuse and leveraging investments in data-collection, and shortening the implementation cycles for new information products (Srivastava et al., 1999; Counihan et al., 2002). These benefits are supported by the rapid advances in technologies for data storage, processing (e.g., ETL [Extraction, Transformation and Loading] engines), and delivery (e.g., reporting and business-intelligence tools). However, implementing the DW is challenging both technically, due to the many components and configuration characteristics involved (Shankaranarayanan and Even, 2004), and organizationally, due to the substantial managerial support, user participation, and financial resources needed (Wixom and Watson, 2001). This study links the potential costs and benefits of the DW to design decisions.

Data Warehouse as a Dynamic Capability

DW environments help transform data resources into information products. The value of information products materializes through usage (Shapiro and Varian, 1999) and integration into business processes (Sambamurthy et al., 2003; Melville et al., 2004). Important for understanding DW usage is the differentiation between *exploitative* versus *explorative* business processes (e.g., Tushman and Anderson, 1986; Henderson and Clark, 1990). March

(1991) suggests that exploration includes search, variation, risk-taking, experimentation, play, flexibility, discovery, and innovation, while exploitation includes refinement, choice, production, efficiency, implementation, and execution. Exploitation promotes continuous learning and knowledge accumulation, optimizes cost-efficiency through repetitiveness and refinement, and provides systematic returns. However, an over-investment in exploitation might cause stagnation, inertia, and low adaptability to changes. Exploration helps address external changes such as new technologies, demand shifts, and emerging competition. While important to the firm's continuous growth, exploration is highly uncertain and might yield insignificant returns. This interplay between managing both exploitative and explorative activities and allocating organizational resources accordingly may introduce organizational challenges. Stemming from the Resource-Based View of the firm is the notion of *dynamic capabilities* (e.g., Conner and Prahalad, 1996, Teece at al., 1997) – organizational processes that shift, alter, and recombine resources to match organizational changes. Resource reallocation is influenced by environmental dynamics: relative stability typically promotes exploitation, while high volatility increases exploration. Organizational *ambidexterity* (e.g., Benner and Tushman, 2003) describes the organizational ability to simultaneously maintain strategies, structures, and processes to support both exploration and exploitation.

We view the DW environment as a dynamic capability. The DW supports collecting, processing, and integrating multiple data resources. By enabling flexible access to these resources, it allows data consumers to efficiently and effectively recombine them into new utility-contributing information products. The DW is ambidextrous as it supports exploitative business processes (e.g., status reporting, corporate accounting), and simultaneously allows ad hoc explorative usage (e.g., ad hoc inquires, data-mining, and business analysis). We suggest that conceptualizing the DW as a dynamic capability has important implications for its design, as the design must account for the need to use its information products in both explorative and exploitative contexts. We examine the value contribution of both usage types, and evaluate the effect of the differences in uncertainty and risks on DW architecture, data contents, performance, capacity, and technology infrastructure choices.

Modeling the Effect of Design on Economic-Performance

Data management literature has discussed design largely from technical/functional perspectives such as architecturally separating data and applications, translating requirements into data models (e.g., ER diagrams), metadata configuration, and DW architectures (e.g., Kimball et al., 2000; Elmasri and Navathe, 2006). Although these perspectives clearly reflect business benefits, they do not explicitly link design to economic contributions. Such a link is addressed to some extent by utility functions (Ahituv, 1980), which can map design choices to economic outcomes. The utility concept has been used in IS/IT research to model and optimize data processes and products (e.g., Ballou et al., 1998; Even et al., 2005), and direct data-mining (e.g., Kleinberg et al., 1998). Utility-driven models have been proposed for the optimal design of products (e.g., Kohli and Sukumar, 1990), services (e.g., Eriksen and Berger, 1987), and production lines (e.g., Cooper and Slagmulder, 2004).

DW design decisions and related investments must deal with uncertainties stemming from non-systematic explorative usage and/or cost volatility. Such decision scenarios, which deal with irrecoverable investments under uncertain outcomes, can benefit from real-options modeling. Kogut and Kulatilaka (2001) frame the establishment of organizational capabilities as a real-option investment problem, and Sambamurthy et al. (2003) view investments in IT resources as a process of building "digital options". Some studies have applied real-options for modeling investments in IS-resources (e.g., Benaroch and Kauffman, 1999, Schwartz and Zozaya-Gorostiza, 2003), and others have applied it to conceptualize design (e.g., Sullivan et al., 1999, Baldwin and Clark, 2000). We next describe a framework that applies a real-option perspective for utility/cost mapping and assessing alternatives for determining an optimal DW design.

Toward an Economics-Driven Design of the Data Warehouse Environment

We view design as a goal-driven activity targeting the creation of new artifacts (Simon, 1996; Hevner et al., 2004). This section develops the theoretical foundations for an economics-driven design framework and sets the stage for understanding the economic motivations behind design decisions in a DW environment. Organizations may have multiple DW environments. In this study we address the design of a single DW environment, assuming that the others may serve as data sources or data consumers. The theoretical model (Figure 1) views the goal of design as

maximizing the economic performance. The maximization criterion is the net benefit – the difference between overall utility (attributed to data usage) and overall cost (attributed to implementation and maintenance). The decision variables are the design characteristics. These are the subject of design decisions, which are assumed to affect both utility and cost. The utility/cost effect of design decisions is assumed to be moderated by implementation and usage factors such as the existing IT environment, complementary resources (e.g., human skills, business-process setup), data consumption patterns, or other information resources, possibly competing with or complementing the DW data. The design characteristics, their value domains, and the possible constraints and dependencies among them define a design space (Baldwin and Clark, 2000). The design process is interpreted as evaluating alternatives within this space and determining optimality with respect to the design objective (here, maximize the net benefit). We assume a multi-phase design process, where each stage may have a different design scope (i.e., a subset of characteristics to be evaluated). Following evaluation, the designer may choose to change the configuration of some characteristics at the current stage, while deferring the decision on others to a later stage. The term "designer" refers to a managerial entity with the authority to approve/reject design decisions. The DW design process, however, may involve a few contributors with different managerial and/or technical skills.

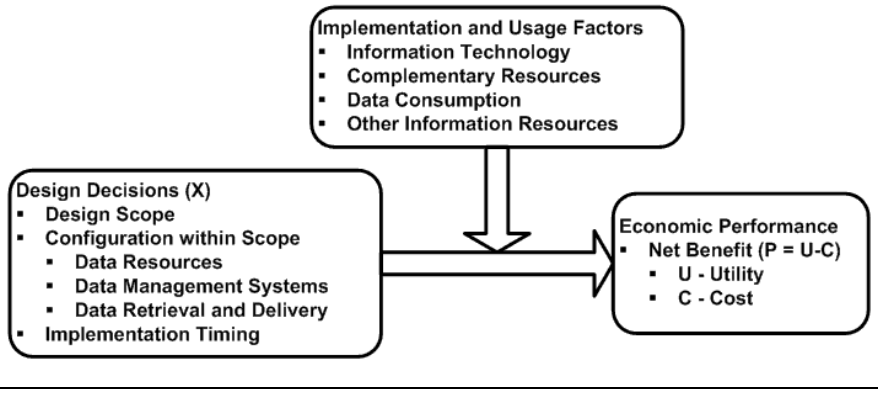

Figure 1. The Effect of Design Decisions on Economic-Performance

Following the decision-calculus methodology (Little, 1970), we parameterize the utility/cost effects, highlighting trade-offs for an optimal design. A high-level analytical formulation of this model is:

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

where

X –	A vector of design characteristics
$P(X)$ –	Net benefit, the difference between overall utility and overall cost
$\{U_i(X)\}$, $U(x)$ –	Utility attributed to I data usages indexed by $[i]$, and the overall utility
$\{C_j(X)\}$, $C(X)$ –	Cost attributed to J cost factors indexed by $[j]$, and the overall cost

This formulation can be viewed as the objective function for optimization – setting the design characteristics X to maximize the expected net benefit $E[P]$. Developing this formulation into a useful design tool requires further analysis of utility/cost contributions and design decisions that influence them. Mutual dependencies and moderation effects are possible and can affect the utility/cost mappings $\{U_i(X)\}$ and $\{C_j(X)\}$. Before expanding these factors in the context of DW design, we present some concepts that guide our model development: the stack architecture view of the DW capacity versus utilization considerations, and real-options framing of the related investment decisions.

Stack architecture, in which system components (e.g., hardware, software, and networking) are organized into layers, is common in IS design (Messerschmitt and Szyperski, 2003). Each layer provides services to the layer "above", and relies on services from the layer "below", such that inter-dependencies between components are reduced and implementation flexibility is improved. At a high level, data management systems employ a typical layered architecture: *system-infrastructure, data resources,* and *data delivery* (Figure 2.) The design of data resources is largely independent of the infrastructural platforms that manage them (i.e., computational resources and database servers), allowing better data transferability between platforms. Similarly, the design of data-delivery

components (e.g., data retrieval, formatting, and presentation) is largely independent of the design of data resources, and delivery mechanisms can be altered while minimally affecting the underlying data resources and vice-versa.

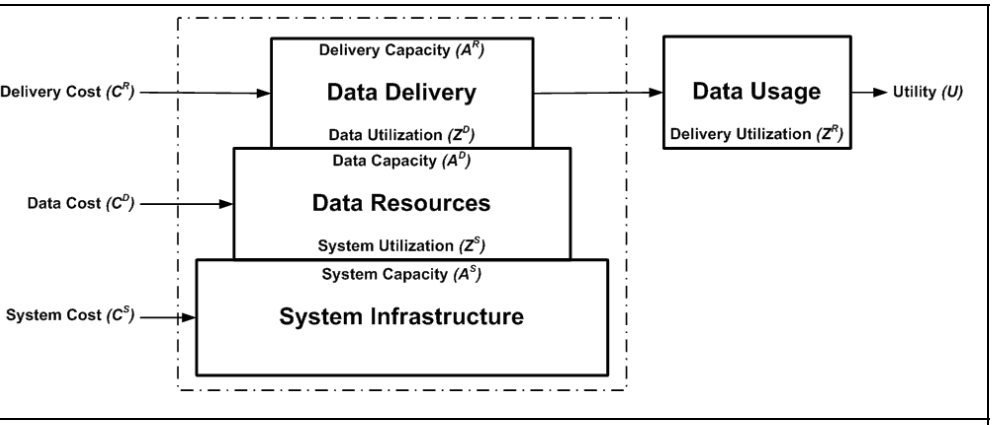

Figure 2. Stack Architecture of Data Management Systems

Figure 2 illustrates the utilities and costs in the stack architecture. The architectural layers provide a logical foundation for modeling the cost structure along *(a) System cost* (C^S), attributed to investments in system infrastructure, *(b) Data cost* (C^D), attributed to investments in data resources, and *(c) Delivery cost* (C^R), reflecting investments in data retrieval and delivery mechanisms. The system outputs are information products, and their *Utility (U)* is attributed to their actual or potential use. Data resources and the systems that manage them do not have stand-alone value but are attributed with value when information products are used. Hence, system layers are not associated with utility but only with costs. However, as later discussed, the design characteristics of the system layers do influence the creation of information products and, consequently, their utility.

As layering implies a high degree of design independence, we assume that the high-level cost categories (C^S, C^D, C^R) and the utility *(U)* can be reasonably treated as independent. However, some dependency exists across layers, conceptualized as an interaction between *capacity (A)* – the set of resources, services and capabilities available for use, versus *utilization (Z)* – their actual use. This interaction exists along all inter-layer interfaces and imposes design constraints – the utilization in one layer cannot exceed the capacity provided by the layer below ($Z^S \le A^S$, $Z^D \le A^D$, *and* $Z^R \le A^R$.) The following section further develops both capacity and utilization in DW environments and links them explicitly to design characteristics, cost, and utility. Capacity decisions are typically affected by planning horizon and usage uncertainty – a designer may provide slack capacity (or flexibility to increase capacity) to accommodate future growth. Capacity decisions involve trade-offs – minimally designing for the specific needs at a minimal cost, versus investment in slack capacity to support future utility-gain opportunities.

We posit that capacity configuration choices are real-option investment decisions. To demonstrate this, we first apply a simplified two-stage design-process model: decisions are made at stage 1 to support future usage, which is yet uncertain. As more information about actual usage is revealed, these decisions can be altered at stage 2 accordingly, possibly with some delay penalty. This model represents utility as a binomial variable: U with a probability of p, or 0 with a probability of $1\text{-}p$. Costs in this model are assumed deterministic and unchanging between stages. Utility and cost may have fixed and variable components, which are aggregated into a single net present value (NPV) measurement. Two possible usage modes are considered: *(a) Exploitative:* data consumption within relatively-predictable business processes. Exploitation is assumed to contribute U^a utility with high certainty (i.e., probability of $p^a \approx 1$). Generating the information products to support exploitative usage (IPR^a) requires system capacity $A^{a/S}$ costing $C^{a/S}$, data capacity $A^{a/D}$ costing $C^{a/D}$, and delivery capacity $A^{a/R}$ costing $C^{a/R}$. Assuming $U^a \ge C^{a/S}+C^{a/D}+C^{a/R}$, exploitative usage yields a positive net benefit and, hence, the system that supports it will be implemented. *(b) Explorative:* the utility-contribution U^b of explorative business processes is uncertain (i.e., $p^b < 1$). Explorative business processes rely on a set of information products IPR^b (in addition to IPR^a); hence, slack capacities are necessary ($A^{b/S}$, $A^{b/D}$, and $A^{b/R}$) at additional costs ($C^{b/S}$, $C^{b/D}$, and $C^{b/R}$). Since $U^b = 0$ is practically possible, explorative usage might not be beneficial and requires further evaluation.

Explorative usages are typically ad hoc, and timely provision of data is critical. Delays may damage utility due to opportunity loss or failure to address immediate needs. Utility, therefore, is expected to decline with time delay *t.*

We assume an exponential decline $U^b e^{-\tau t}$ with sensitivity parameter $\tau \geq 0$ (higher sensitivity implies steeper decline). The time to implement delivery mechanisms for explorative usage (once the required data resources are available) is $t^{b/R}$. An additional time $t^{b/D}$ is needed to prepare these data resources and $t^{b/S}$ to implement the system for managing them. Exploitative usages, though sensitive to time delays, can be planned for, as the system for supporting exploitative-usage is (certainly) implemented in stage 1. Hence, we assume a negligible time delay effect for exploitative use *($t^{a/S}$, $t^{a/D}$, and $t^{a/R}$)*. In real life, however, DW implementations (for both usages) might be sensitive to timeliness due to operational failures, but these effects are not addressed in this study.

Table 1. Parameters Involved in the Real-Option Framing of DW Design

	Exploitative Usage *(IPR^a)*	**Explorative Usage *(IPR^b)***
Utility	U^a with probability $p^a \approx 1$ Negligible implementation-time sensitivity	$U^b e^{-\tau t}$ with probability p^b, *0* otherwise t – implementation-time τ – implementation-time sensitivity-parameter
System Infrastructure	$C^{a/S}$, $t^{a/S}$ – System cost and implementation time	$C^{b/S}$, $t^{b/S}$ – System cost and implementation time
Data Resources	$C^{a/D}$, $t^{a/D}$ – Data cost and implementation time	$C^{b/D}$, $t^{b/D}$ – Data cost and implementation time
Data Delivery	$C^{a/R}$, $t^{a/R}$ – Delivery cost and implementation time	$C^{b/R}$, $t^{b/R}$ – Delivery cost and implementation time
Capacity and Utilization	*No incentive to implement slack capacity:* $A^{a/S}=Z^{a/S}$, $A^{a/D}=Z^{a/D}$, $A^{a/R}=Z^{a/R}$	*Possible incentive to implement slack capacity:* $A^{b/S} \geq Z^{b/S}$, $A^{b/D} \geq Z^{b/D}$, $A^{b/R} \geq Z^{b/R}$

Given the parameters (Table 1), the following decision alternatives can be considered:

(a) Minimal implementation (M): supporting exploitative usage only *(IPR^a)* at current stage and deferring the decision to implement additional data resources and capacity for explorative usage *(IPR^b)* to a later stage. Assuming that the uncertainty with explorative usage will be resolved at a later stage, the additional resources and capacity will be implemented only if positive utility is expected. However, deferring implementation will cause time delays and will cause utility reduction. Applying *(1)*, the anticipated net benefit P^M is:

$$P^M = U^a - \left(C^{a/S} + C^{a/D} + C^{a/R}\right) + p^b\left(U^b e^{-\tau\left(t^{b/S}+t^{b/D}+t^{b/R}\right)} - \left(C^{b/S} + C^{b/D} + C^{b/R}\right)\right) \quad (2)$$

(b) Full implementation (F): supporting both explorative and exploitative usage with data (IPR^a and IPR^b) and implementing system capacity accordingly. Applying *(1)*, the anticipated net benefit P^F is:

$$P^F = U^a - \left(C^{a/S} + C^{a/D} + C^{a/R}\right) + p^b U^b - \left(C^{b/S} + C^{b/D} + C^{b/R}\right) \quad (3)$$

This alternative can be interpreted as a real-option investment – the availability of additional resources allows faster implementation and avoids utility loss. The real-option gross-benefit (i.e., prior to subtracting cost-differentials) is $ROGB^F = p^b U^b\left(1 - e^{-\tau\left(t^{b/S}+t^{b/D}+t^{b/R}\right)}\right)$. The cost differential (in favor of *M*) in this case is the portion of the initial cost that cannot be recovered if usage is low: $ROCD^F = \left(1-p^b\right)\left(C^{b/S} + C^{b/D} + C^{b/R}\right)$. Full implementation is superior to minimal implementation if $ROGB^F > ROCD^F$. This will depend, of course, on the actual utility and cost parameters: if the expected utility from explorative use is relatively high, the delay penalty is significant, and/or the costs associated with explorative use are relatively low, then the investment in full implementation may be beneficial.

(c) Partial implementation (P): the designer may consider fully supporting only the exploitative use *(IPR^a)* at the current stage but implement some slack capacity toward supporting additional usage in the future. For example, the designer may choose (a) investing in additional system capacity now, while deferring implementation of additional data resources and delivery capacity to a later stage, or (b) investing in additional system capacity and data resources now, while deferring the decision on additional delivery capacity. Considering the former case (the latter can be similarly evaluated) and applying *(1)*, the anticipated net benefit P^P is:

$$P^P = U^a - \left(C^{a/S} + C^{a/D} + C^{a/R}\right) + p^b\left(U^b e^{-\tau\left(t^{b/D}+t^{b/R}\right)} - \left(C^{b/D} + C^{b/R}\right)\right) - C^{b/S} \quad (4)$$

This alternative can also be interpreted as a real-option investment – the added capacity allows faster implementation and, hence, prevents utility loss (although to a lesser extent than full implementation). Relative to minimal implementation, the real-option gross benefit is $ROGB^P = p^b U^b e^{-\tau\left(t^{b/D}+t^{b/R}\right)}\left(1 - e^{-\alpha t^{b/S}}\right)$. The cost differential is attributed to costs that cannot be recovered with low usage: $ROCD^P = \left(1 - p^b\right)C^{b/S}$. This alternative is superior to minimal implementation if $ROGB^P > ROCD^P$, i.e., if the relative utility loss due to delays in implementation is anticipated to be higher than the cost penalty. Many DW environments adhere to this scenario – some slack capacity enhancements to an existing infrastructure are relatively cheap (e.g., adding disk-storage space, upgrading computation power), but the implementation time might still be significant. However, in some cases adding capacity may require infrastructure upgrades (e.g., replacing database-servers) with significant costs. In these cases, or when there is no significant utility loss due to delay, minimal implementation may be more beneficial.

When $ROGB^F - ROCD^F > ROGB^P - ROCD^P$, full implementation will outperform the partial since the relative utility loss due to delays in implementing data/delivery-capacities is higher than the penalty cost. In some DW scenarios, data cost may be relatively low (e.g., acquisition from intra-organizational sources) and, hence, early inclusion of data may be beneficial. In other cases, the data cost is high (e.g., purchased from external sources), but once purchased, if the needed system capacity is already available, the data can be added quickly. The time delay penalty may not be severe, and deferring the data-capacity implementation decision may be beneficial.

None of these real-option evaluations involve exploitative use. Since the probability p^a was assumed to be nearly *1*, there is no incentive to implement higher capacity than required and, hence, capacity will be utilized to the maximum: $A^{a/S}=Z^{a/S}$, $A^{a/D}=Z^{a/D}$, and $A^{a/R}=Z^{a/R}$. Conversely, with explorative use there may some incentive to invest in slack capacity that may not be fully utilized, hence $A^{b/S} \geq Z^{b/S}$, $A^{b/D} \geq Z^{b/D}$, and $A^{b/R} \geq Z^{b/R}$. In general, we can argue that higher utility/cost certainty lowers the need for slack capacity and capacity-utilization margin.

In the preceding discussion we modeled certain design aspects of a DW using real-options. Treating the capacities in each stack layer as design options allows us to examine DW design efficiency in the light of its explorative versus exploitative usages. If supporting only exploitative use, the cost of this design is minimal and predetermined strictly by known functional/technical requirements. As explorative use can yield larger benefits but with some uncertainty, the designers may choose to plan for extra capacity, knowing that it will certainly increase the cost and understanding that corresponding benefits are not guaranteed. Balancing the support for explorative and exploitative usages during the design process is critical. Understanding the economics and the real-option considerations involved can create a radically different design of the DW that can be economically justified.

Design Decisions in a Data-Warehousing Environment and Their Utility/cost Effect

This section models the utility gained from DW usage and evaluates sample design scenarios in each architectural layer: selection of data-delivery platforms, acquisition of data resources, and data-storage infrastructure setup. Figure 3 presents a high-level DW architecture, mapping key components to the stack architecture. It highlights *(a) system boundaries:* data processes typically follow three high-level stages – collection, storage/maintenance, and usage (Lee and Strong, 2003). We treat the DW (within system boundaries) as the storage/maintenance stage, while data collection (by the sources from which data is acquired) and data usage by consumers are treated as exogenous. We assume that DW designers make design decisions within system boundaries only. However, they may choose the sources for acquiring data and the data components acquired from each, as well as the extent to which usages will be supported, thus impacting these exogenous stages. *(b) Organizational boundaries:* data collection and consumption can be internal and/or external to the organization. As later illustrated, this differentiation affects the design scope and factors that define utility/cost formulation – e.g., data uniqueness, supply/demand uncertainty, competition, pricing structure, potential for quality hazards, and the ease of system/data integration.

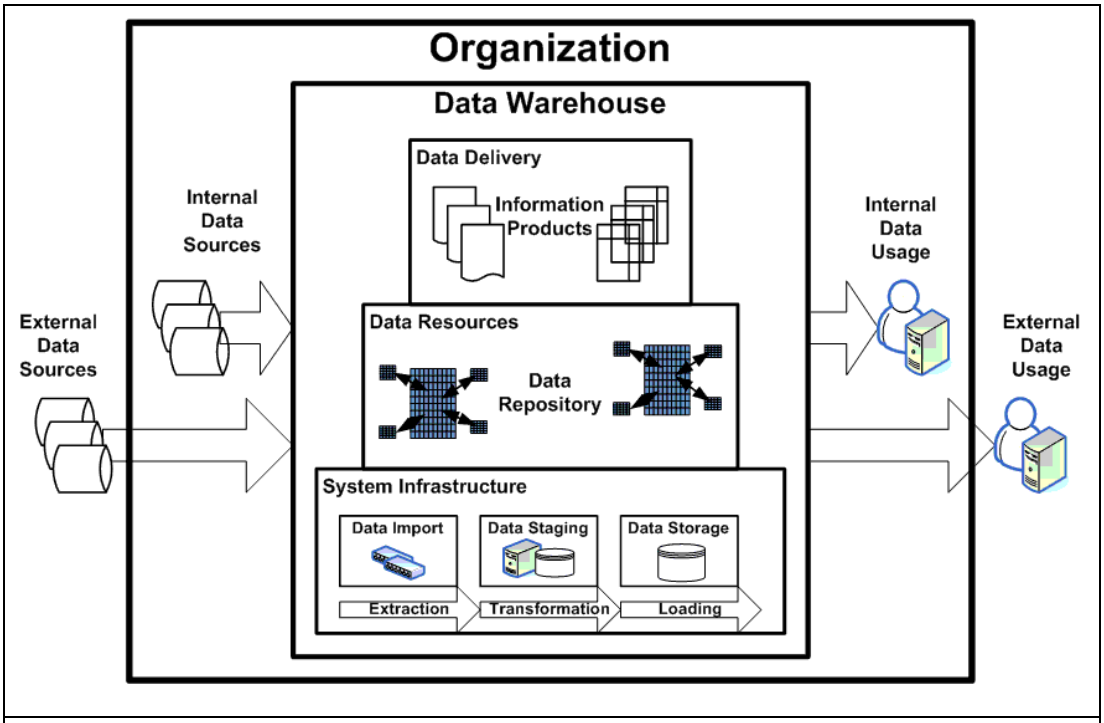

Figure 3. The Data Warehouse Architecture

We evaluate design decisions using parameterized models representing cost/utility effects, dependencies, and trade-offs. The models are presented along the DW layers in a "top-down" order, starting with the usage of information products. This order reflects the capacity/utilization interplay discussed – the utilization requirements of a layer are bounded by the capacity of the layer below and, hence, affect its design requirements. For example, the choice of certain data-delivery mechanisms will require careful integration and design of data resources, and a decision to acquire certain data resources will depend on available infrastructure (e.g., storage space and data-processing capacity). The illustrative design scenarios evaluated address a limited set of design characteristics within a layer, often using simplified analytical formulations to allow a parsimonious presentation. Additional design characteristics, layer interdependencies, and enhanced analytical formulations should be further explored.

Data Usage

The DW output is a set of information products used by internal and/or external consumers. A typical information product combines underlying dataset(s) and presentation format(s) that fit the consumer's data-gauging capabilities (e.g., GUI-based reports and data-visualization tools for humans, structured text-files for computer systems). Each information product (total of I) is represented by a vector X_i of design characteristics that affects its utility $U_i(X_i)$. The characteristic vector X_i is assumed to have G components $\{X_{i,g}\}_{g=1..G}$. In this study, we initially assume a one-to-one mapping between utility and information product (can be later generalized – a single information product may serve multiple usages and a usage may require multiple information products). The information product utility U_i is a random variable representing the consumer's willingness to pay (Ahituv, 1980.) It is shaped by possible interactions between information product characteristics (subject to design), and exogenous characteristics such as the usage-task (e.g., explorative versus exploitative), the business environment (e.g., stable versus turbulent), and competing and/or complementary resources (e.g., other information products, human knowledge). Exogenous characteristics, which we refer to as *usage context*, are treated as moderators and affect the formulation of U_i. We adopt the *production function* for modeling U_i – mapping both the relevant decision variables and possible controls to an assumed parameterized form, such that the parameters can be later estimated or assessed empirically. We adopt the Cobb-Douglas input-output mapping, as it is commonly used for empirical validation. This formulation represents multiple inputs, assuming possibly different output sensitivity to each:

(5) $$U_i = U_i^* \prod_{g=1}^{G} X_{i,g}^{\chi_{i,g}}$$

where U_i – Utility contribution of information product *[i]*

U^*_i – A constant utility component

$X_{i,g}$ – Component *g* (out of *G*) of the characteristic vector X_i

$\chi_{i,g}$ – Utility-sensitivity parameter to $X_{i,g}$

We further assume that the characteristics $\{X_{i,g}\}$ are represented either as a binary decision (1-inclusion, 0-exclusion) or continuous variables within a finite range that can be normalized to [0,1]. The utility is assumed to be non decreasing with inputs, hence, $\chi_{i,g} \geq 0$. The utility effect (Figure 4) is convex with $\chi_{i,g} > 1$, linear with $\chi_{i,g} = 1$, concave with $\chi_{i,g} < 1$, and fixed with $\chi_{i,g} = 0$ (here, for a binary variable $0^0 = 1$). Under these assumptions, U^*_i represents the *maximal utility* when all the input characteristics equal *1*.

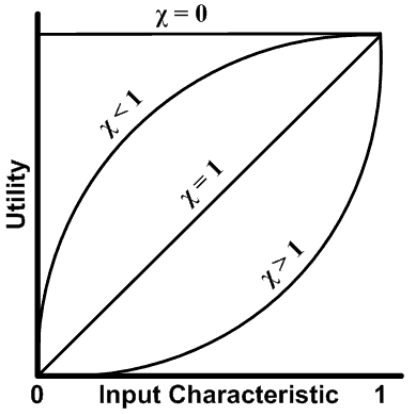

Figure 4. Possible Utility Effects of Design Characteristics

We consider some information product design characteristics that typically affect utility in a DW environment and classify them into three categories – *dataset, presentation, and time*.

Dataset – information products are generated by retrieving data from DW datasets. DW repositories commonly use a "Star Schema" model – a large fact table containing numeric measurements of business activities (e.g., quantity, sale- mount), linked to smaller "dimension" tables containing business entities that describe the transactions (e.g., customers, products, locations). We consider the design of a large fact table that supports multiple information products and hence, significantly affects utility/cost trade-offs (note: the model applies to the design of smaller datasets, but the related utility/cost trade-offs may be less significant). Fact tables consolidate data from multiple sources and in this study we initially assume a single fact table per information product. An information product is assumed to offer a higher utility contribution with richer content and better data quality. Its richness and quality are determined by the design of the dataset (here, the fact table) from which the information product is derived (Even et al., 2005). Content richness is enhanced by covering a broader business scope, including more historical data, and/or increasing granularity. We represent historical depth by TS_i – the time span coverage of information product *[i]*, rescaled to *[0,1]* by taking the ratio between the included time span TS^p_i (actual decision variable) and the maximal time span TS^m available $(TS_i = TS^p_i / TS^m)$. The granularity is reflected by record density, the number of records per time unit RD_i. It is rescaled to [0,1] by taking the ratio between the record density implemented RD^p_i (actual decision variable) and the maximal record density RD^m possible $(RD_i = RD^p_i / RD^m)$. The business scope breadth is reflected in the set of attributes that the information product includes. Assuming *M* possible attributes, the binary decision variable $AT_{i,m}$ indicates inclusion *(=1)* or exclusion *(=0)* of attribute *[m]* in information product *[i]*. In some cases, the utility is sensitive to the inclusion or exclusion of a certain attribute (i.e., if an attribute is needed but excluded, the utility is *0,*) while in others the utility is indifferent to inclusion/exclusion. Data quality is typically viewed as being multi-dimensional, where some dimensions are objective and impartial to the data content, while others are subjective and context dependent (Wang and Strong, 1996). The suggested framework considers the former type,

and in this study we model the *accuracy* (AC_i) – a commonly discussed impartial data quality dimension (e.g., Ballou et al., 1998), typically measured as a *[0,1]* ratio. Other quality dimensions (such as completeness or timeliness) can be added to the model in future extensions, using a similar utility formulation approach (or alternately, multiple dimensions can be combined into an "overall" quality measure on the *[0,1]* scale). With these assumptions, we propose the following model for the utility effect of dataset characteristics:

$$U_i \propto TS_i^{\alpha_i} RD_i^{\beta_i} AC_i^{\delta_i} \left(\prod_{m=1}^{M} AT_{i,m}^{\eta_{i,m}} \right) \tag{6}$$

where
- TS_i – Time span coverage measure
- RD_i - Record density measure
- AC_i – Accuracy level
- $\{AT_{i,m}\}$ - Inclusion *(=1)* of attribute *[m],* versus exclusion *(=0)*
- $\alpha_i, \beta_i, \delta_i, \{\eta_{i,m}\}$ – Utility-sensitivity parameters to TS_i, RD_i, AC_i, and $\{AT_{i,m}\}$, respectively

Presentation – the utility contribution of the information product depends on presentation capabilities (Nelson et al., 2005), e.g., fixed presentation (e.g., a text-file structure) versus flexible (e.g., interactive visualization), or aggregated (e.g., average and sum statistics) versus detailed. We represent presentation by a set of binary variables – the delivery layer offers a variety of presentation capabilities and the designer may choose to include certain capabilities in the information product and exclude others. Depending on usage context, utility may be sensitive to inclusion/exclusion of certain presentation capabilities (i.e., *0* utility with exclusion), and indifferent to others:

$$U_i \propto \prod_{n=1}^{N} PC_{i,n}^{\gamma_{i,n}} \tag{7}$$

where
- $\{PC_{i,n}\}$ - Inclusion *(=1)* of presentation capability *[n],* versus exclusion *(=0)*
- $\{\gamma_{i,n}\}$ - Utility-sensitivity parameters to $\{PC_{i,n}\}$

Time – Delay in providing information products can damage utility. We differentiate between one-time delays attributed to designing the information product and establishing the data and system resources for it, versus recurring delays due to production failures. In this study we address the former, assuming high-quality implementation that eliminates recurring time delays (in reality, production delays can be significant and must be examined). The time delay *TM* can be "infinite", and we assume decline in utility as it grows. For consistency (non-decreasing effect, *[0,1]* range), we apply an exponential transformation:

$$U_i \propto \left(e^{-TM_i} \right)^{\tau_i} \tag{8}$$

where
- TM_i - Time delay
- τ_i - Utility-sensitivity parameter to TM_i

Combining (6), (7), and (8), the overall utility can be represented by the Cobb-Douglas formulation:

$$U_i = U_i^* TS_i^{\alpha_i} RD_i^{\beta_i} AC_i^{\delta_i} \left(\prod_{m=1}^{M} AT_{i,m}^{\eta_{i,m}} \right) \left(\prod_{n=1}^{N} PC_{i,n}^{\gamma_{i,n}} \right) \left(e^{-TM_i} \right)^{\tau_i} \tag{9}$$

This model maps a set of information product design characteristics to utility. The sensitivity parameters are assumed to be fixed, although in reality they may change over time and, hence, need on-going re-estimation. Assuming independent data usages, the overall utility U is modeled as a sum:

$$U = \sum_{i=1}^{I} U_i^* TS_i^{\alpha_i} RD_i^{\beta_i} AC_i^{\delta_i} \left(\prod_{m=1}^{M} AT_{i,m}^{\eta_{i,m}} \right) \left(\prod_{n=1}^{N} PC_{i,n}^{\gamma_{i,n}} \right) \left(e^{-TM_i} \right)^{\tau_i} \tag{10}$$

This utility model is one side of the objective function in *(1)* for the design-optimization problem, the other being the cost. We further demonstrate the use of this model by analyzing the cost-structure and optimizing the net benefit for different design scenarios.

Design Scenario 1 – Delivery Cost and the Selection of Delivery Platform

The data-delivery layer provides the capacity for producing information products and delivering them to consumers. Information product implementation typically includes specification of content (the underlying data), presentation, and delivery configuration (e.g., recipients, delivery mechanisms, and schedule). The design scope evaluated is the choice of delivery platform for designing and distributing information products – a reporting or business-intelligence tool, which can be developed in-house or purchased (e.g., Business-Objects and MicroStrategy). The platform choice affects the delivery cost *(C^R)*. We choose one among *R* possible platforms (indexed by *[r]*), each associated with an overall purchase and setup cost of C^R_r. Two delivery-factors (among others) influence utility/cost trade-offs: (a) the implementation time *(t^R_r)*: an advanced platform will require a longer implementation time, higher learning efforts, and consequently a longer delay in information product delivery. This implementation time imposes an upper bound on the timeliness of information products and defines a set of design constraints per *[r]*: $\left\{TM_i \geq t^R_r\right\}$. (b) The interplay between *presentation-capacity* and *presentation-utilization*: *N* possible presentation capabilities are considered and $A^R_{r,n}$ indicates whether platform *[r]* supports capability *[n] (=1)* or not *(=0)*. A presentation capability *[n]* is needed if demanded by at least one information product (i.e., $PC_{i,n}=1$). We define the utilization of presentation capability *[n]* as $1-\prod_{i=1}^{I}\left(1-PC_{i,n}\right)$ - i.e., *1* if demanded at least by one information product *($PC_{i,n}=1$,)* and *0* otherwise. The result is a set of presentation-capacity constraints $A^R_{r,n} \geq 1-\prod_{i=1}^{I}\left(1-PC_{i,n}\right)$.

Using *(1)*, the selection of a delivery platform is framed as a design-optimization problem - select a delivery platform (from a set of *R* possible platforms, indexed *[r]*), to maximize net benefit *(P^R)*:

(11) $$P^R_r = U^R_r - C^R_r = \sum_{i=1}^{I} U_i^{Max}\left(\prod_{n=1}^{N} PC_{i,n}^{\gamma_{i,n}}\right)\left(e^{-TM_i}\right)^{\tau_i} - C_r$$

s.t. $TM_i \geq t^R_r$ per *[r]*

$\{A^R_{r,n}\}$, $\{PC_{i,n}\}$ are *0/1* integers, per *[i]*, *[r]*, and *[n]*

$A^R_{r,n} \geq 1-\prod_{i=1}^{I}\left(1-PC_{i,n}\right)$, per *[r]* and *[n]*

where

P^R_r, U^R_r, C^R_r –	Overall net benefit, utility and delivery cost, respectively
U_i^{Max} -	The maximal utility of information product *[i]*, considering the effect of other characteristics in *(9)*, i.e., $U_i^{Max} = U_i^* TS_i^{\alpha_i} RD_i^{\beta_i} AC_i^{\delta_i}\left(\prod_{m=1}^{M} AT_{i,m}^{\eta_{i,m}}\right)$
C_r ,$\{T^D_r\}$–	Implementation cost and time, respectively
$\{A^R_{r,n}\}$ -	Support for presentation capability *[n] (=1)*, vs. no support *(=0)*
t^R_r -	The implementation time of the delivery platform
TM_i, τ_i, $\{PC_{j,n}\}$, $\{\gamma_{i,n}\}$ -	see *(7)* and *(8)*

Illustrative Example 1: A bank evaluates implementing a DW of historical transactions, considering two possible usage types: *a) Financial reporting* – recurring reports, used internally for accounting and performance tracking. These reports introduce simple presentation requirements – displaying data in a tabular format, with some charts (e.g., bar, pie), and summary statistics (e.g., sum, average). Their estimated utility *(U_a)* has a timeliness-sensitivity of τ_a, and this usage is treated as exploitative. *b) Business-analysis tools* for advanced analysis of banking activities, such as customer/product segmentation, trend detection, and development of new credit and loan policies. These tools require fancy reporting capabilities, such as interactive tables and charts and advanced statistics (e.g., regression and time-series analysis), and sophisticated delivery configurations to support possible external use (e.g., allowing customers to analyze their own activities). Such tools have never been offered before, and hence are considered explorative. Their maximal utility contribution is estimated as U_b with timeliness sensitivity of τ_b, and a success-probability of $P_b<1$. A comprehensive usage analysis is suggested to assess whether the explorative usage will be successful. This analysis will require a time-period of t^*, and will cost C^*.

The DW designer considers two delivery platforms. The first (indexed by *"1"*) is an inexpensive reporting tool that can be implemented immediately at a cost of C_1 and can only support exploitative usage. The second (indexed by *"2"*) is an advanced reporting system that offers a full range of capabilities. This is pricier $(C_2>>C_1)$ and involves longer implementation $(t_2>>t_1)$. The designer may postpone the decision until after the usage analysis is performed. Given these parameters (Table 2), what should the designer choose? Should part of the implementation be deferred?

Table 2. Example 1 – Parameters

	Exploitative	**Explorative**	**Basic**	**Advanced**
Basic capabilities	1	1	1	1
Advanced capabilities	0	1	0	1
Maximal Utility	U_a	U_b		
Probability	~1	P_b		
Timeliness Sensitivity Parameter	τ_a	τ_b		
Usage-analysis Time	*N/A*	t^*		
Usage-analysis Cost	*N/A*	C^*		
Cost			C_1	C_2
Implementation Time			*~0*	t_2

Using *(11)*, the designer evaluates the following alternatives:

1) Supporting exploitative usage only **–** the cheapest alternative, with an expected net benefit of $P = U_a - C_1$

2) Supporting both usages now **–** the expected net benefit is $P = U_a + p_b U_b e^{-\tau_b t_2} - C_2$. Compared to the exploitative-use-only alternative, the real-option gross-benefit is $ROGB_2 = p_b U_b e^{-\tau_b t_2}$ at a cost-differential of $ROCD_2 = C_2 - C_1$. This alternative will be preferred over the first if $ROGB_2 > ROCD_2$.

3) Deferring the decision on explorative usage and support only exploitative now, at a net benefit of $P = U_a - C_1 + p_b\left(U_b e^{-\tau_b(t^*+t_2)} - C_2\right) - C^*$. Compared to the exploitative-use-only alternative, the real-option gross-benefit in this case is $ROGB_3 = p_b U_b e^{-\tau_b(t^*+t_2)}$. This benefit is attributed to the potential of adding utility from explorative-usage at a cost-differential of $ROCD_3 = p_b C_2 + C^*$. This alternative is preferred over the first if $ROGB_3 > ROCD_3$. Due to the added time delay (t^*), $ROGB_3$ is lower than $ROGB_2$, but $ROCD_3$ could either be lower or higher than $ROCD_2$. Overall, this alternative is preferred over the second if $ROGB_3 - ROCD_3 > ROGB_2 - ROCD_2$.

4) Deferring the decision on both usages **–** The advanced delivery platform will be chosen if the additional analysis reveals successful explorative usage, and the basic platform will be chosen otherwise. The net benefit in this case is $P = U_a e^{-\tau_b t^*} - (1 - p_b) C_1 + p_b\left(U_b e^{-\tau_b(t^*+t_2)} - C_2\right) - C^*$. Compared to the exploitative-use-only alternative, the real-option gross-benefit is $ROGB_4 = p_b U_b e^{-\tau_b(t^*+t_2)} - \left(1 - e^{-\tau_b t^*}\right) U_a$ - while some utility is added by potential explorative-usage, some exploitative-utility is damaged by the time delay (t^*) effect. The cost differential is $ROCD_4 = p_b C_2 + C^* - p_b C_1$, and this alternative is preferred over the first if $ROGB_4 > ROCD_4$, over the second if $ROGB_4 - ROCD_4 > ROGB_2 - ROCD_2$, and over the third if $ROGB_4 - ROCD_4 > ROGB_3 - ROCD_3$.

These alternatives highlight important trade-offs in selecting delivery platforms and explain common design errors – designers often choose quick-and-simple reporting solutions that later fail to support new usages. Further, expensive and advanced reporting capabilities may be unnecessary for most utility-generating usages. The evaluation also highlights the possible benefit of postponing investments in reporting tools until after assessing actual usage, and avoiding "rushed" decisions at the current stage. Delivery costs and utility damages due to poor design decisions are very expensive in a DW environment. Being aware of alternatives and using quantitative assessment are critical.

Design Scenario 2 – Data Cost and the Selection of Data Sources

The data resources layer contains datasets that generate information products. In a typical DW architecture (Figure 3), this layer contains a centralized repository of historical data, collected from (possibly) multiple data sources, internal and/or external to the organization. The scope of this design scenario is the selection of data sources for data acquisition. Given *K* available sources (indexed by *[k]*), the binary decision variables $\{S_k\}$ represent the decision to acquire data from a source $(S_k=1)$, or exclude it $(S_k=0)$. A factor that affects the inclusion and, hence, the data cost (C^D), is the accuracy level of a source $(0\leq\varepsilon_k\leq1)$, which can be higher or lower than the targeted accuracy level of the dataset *(AC)*. We assume that the accuracy level of information products is influenced primarily by the accuracy of the dataset, hence, for all information products, $AC_i=AC$.

Data cost is also affected by the attributes available in each source. The set of parameters $\{\omega_{k,m}\}$ indicate whether attribute *[m]* is available in source *[k]* $(\omega_{k,m}=1)$ or not $(\omega_{k,m}=0)$. Some attributes may be unique to a source, while others are present in multiple sources. The set of decision variables $\{W_{k,m}\}$ indicates whether attribute *[m]* from source *[k]* is included in the dataset $(W_{k,m}=1)$ or not $(W_{k,m}=0)$. This is possible only if the source *[k]* is selected $(S_k=1)$ and contains attribute *[m]* $(\omega_{k,m}=1)$ and, hence, the set of constraints $\{W_{k,m}\leq S_k\omega_{k,m}\}$ per *[k]* and *[m]*. A source will be used for data acquisition $(S_k=1)$ only if it contributes at least one attribute $(W_{k,m}=1)$, hence, the set of constraints $\left\{S_k = 1-\prod_{m=1}^{M}\left(1-W_{k,m}\right)\right\}$ per *[k]*. However, if multiple sources offer attribute *[m]*, we assume that this attribute will be acquired from at most one source, hence, the set of constraints $\left\{\sum_{k=1}^{K} W_{k,m} \leq 1\right\}$ per *[m]*.

An information product cannot be implemented if an attribute it needs is not in the dataset. For the *M* possible attributes we specify an attribute-capacity $\{A_m\}$ that indicates whether the attribute is included in the dataset $(A_m=1)$, or not $(A_m=0)$. The utilization of attribute *[m]* is $1-\prod_{i=1}^{I}\left(1-AT_{i,m}\right)$, i.e., *1* if attribute *[m]* is used at least by one information product $(AT_{i,m}=1)$, and *0* otherwise. Hence, the set of data-attribute constraints: $\left\{A_m \geq 1-\prod_{i=1}^{I}\left(1-AT_{i,m}\right)\right\}$ per *[m]*. The attribute *[m]* can be included in the dataset $(A_m=1)$ only if it is acquired from at least one source $(W_{k,m}=1)$ resulting in a set of constraints $\left\{A_m = 1-\prod_{k=1}^{K}\left(1-W_{k,m}\right)\right\}$ per *[m]*. Combining the constraints for A_m, we get: $\left\{1-\prod_{k=1}^{K}\left(1-W_{k,m}\right)\geq 1-\prod_{i=1}^{I}\left(1-AT_{i,m}\right)\right\}$per *[m]*.

We assume a fixed setup cost per utilized data source $C^{D/S}{}_k=c^{D/S}{}_kS_k$ (e.g., for design efforts and establishing connectivity), and a variable acquisition cost ($C^{D/A}{}_k$), attributed to on-going data imports (e.g., payment to data vendor).We assume that this variable cost increases with each attribute imported. If the source accuracy ε_k is below the target accuracy *AC*, quality improvement efforts are needed and these will further increase the acquisition cost. We assume that the cost effect of accuracy is proportional to $(max\{AC/\varepsilon_k,\ 1\})^{\lambda}$, *where* λ is the cost-sensitivity parameter for accuracy improvement. We further assume no cost effect if the provided accuracy (ε_k) is higher than the targeted *(AC)*. The suggested cost model for source *[k]* is:

$$C_k^D = C_k^{D/S} + C_k^{D/A} = S_k\left(c_k^{D/S} + \left(\max\left\{AC/\varepsilon_k, 1\right\}\right)^{\lambda}\left(\sum_{m=1}^{M} c_{k,m}^{D/A} W_{k,m}\right)\right) \tag{12}$$

where

- $C^D{}_k$, – Data cost, summing setup cost ($C^{D/S}{}_k$) and acquisition cost ($C^{D/A}{}_k$)
- $\{S_k\}$- Source inclusion *(=1)* versus exclusion *(=0)*
- $\{W_{k,m}\}$- Inclusion *(=1)* of attribute *[m]*, versus exclusion *(=0)*
- $c^{D/S}{}_k$ - Setup cost if source included $(S_k=1)$
- *AC* - Targeted dataset accuracy
- $\{\varepsilon_k\}$ - Given source-accuracy
- λ - Cost-sensitivity parameter to the accuracy ratio (AC/ε_k)
- $c^{D/A}{}_{k,m}$ - Incremental acquisition cost of attribute *[m]*, if included $(W_{k,m}=1)$

Following *(1)*, we can formulate the data source selection as a design-optimization problem – select the target accuracy level *(AC)*, the data sources $\{S_k\}$, and the attributes to import from each data source $\{W_{k,m}\}$, to maximize net benefit (P^D). The result is a mixed-integer, non-linear optimization model:

$$(13) \qquad P^D = U^D - C^D = \sum_{i=1}^{I} U_i^{Max} \left(\prod_{m=1}^{M} AT_{i,m}^{\eta_{i,n}} \right) AC^{\delta_i} - \sum_{k=1}^{K} S_k \left[\begin{array}{l} c_k^{D/S} + \\ \left(\max\left\{ AC/\varepsilon_k , 1 \right\} \right)^{\lambda} \left(\sum_{m=1}^{M} c_{k,m}^{D/A} W_{k,m} \right) \end{array} \right]$$

s.t. $\{AT_{i,m}\}$, $\{W_{k,m}\}$, *and* $\{S_k\}$ are *0/1* integers, per *[i], [k],* and *[m]*

$\{W_{k,m} \le S_k \, \omega_{k,m}\}$ per *[k]* and *[m]*

$\left\{ S_k = 1 - \prod_{m=1}^{M} \left(1 - W_{k,m} \right) \right\}$ per *[k]*

$\left\{ \sum_{k=1}^{K} W_{k,m} \le 1 \right\}$ per *[m]*

$\left\{ 1 - \prod_{k=1}^{K} \left(1 - W_{k,m} \right) \ge 1 - \prod_{i=1}^{I} \left(1 - AT_{i,m} \right) \right\}$ per *[m]*

where P^D, U^D, C^D– Overall net benefit, utility and data cost, respectively

U_i^{Max} - The maximal utility of information product *[i]*, considering the effect of other design characteristics in *(9)*, i.e., $U_i^{Max} = U_i^* TS_i^{\alpha_i} RD_i^{\beta_i} \left(\prod_{n=1}^{N} PC_{i,n}^{\gamma_{i,n}} \right) \left(e^{-TM_i} \right)^{\tau_i}$

$\{\omega_{k,m}\}$- Availability *(=1)* of attribute *[m]*, versus no availability *(=0)*

Other parameters – see *(6)* and *(12)*

Illustrative Example 2: Using "loyalty cards", a retail-chain can link sale transactions with individual customers. It considers a DW that integrates sale-transaction data with customer data. Elementary customer details (e.g., name, address, and phone) are currently maintained by the firm's customer-relationship-management (CRM) system. These details are insufficient for the two possible DW usages that the firm evaluates: *a) Promotion management* – customizing promotions based on segmentation and consumption-behavior analysis. This exploitative usage requires additional demographics (e.g., marital status, number of children, education). Its anticipated utility (U_a) has an accuracy-sensitivity parameter δ_a. *b) Advanced consumption analysis* for strategic decisions – e.g., analyzing revenue potential for new locations, developing new promotion policies, or identifying changes in consumption patterns. The corresponding data resource has revenue-generating potential if sold to manufacturers or to firms that analyze market behavior. These explorative usages need additional demographics, beyond what exploitative usage requires (e.g., neighborhood ranking, credit status, and value of property owned). The anticipated utility is U_b, with accuracy sensitivity parameter δ_b. This utility is uncertain with a success probability of $P_b<1$, and failure-probability (*0* utility) of $1\text{-}P_b$.

The DW designer evaluates two possible sources for the additional demographics. The first is enhancing the current CRM system requiring a setup investment of C_1 (e.g., redesign to support requested changes) and additional data acquisition costs: $C_{1,B}$ for collecting basic demographics (assumed relatively low), and $C_{1,A}$ for advanced demographics (relatively high). This solution is managed entirely in-house; hence, data-integration failures are less likely and the anticipated accuracy level ε_1 is relatively high. The second possible source is a data vendor specializing in collecting and selling customer demographics. The cost of setting up the connectivity to this data source (C_2) is assumed to be lower than setting up the internal source $(C_1>C_2)$. The acquisition cost of basic demographics $(C_{2,B})$ is assumed to be higher than in-house collection $(C_{1,B}<C_{2,B})$, but the acquisition cost of advanced demographics $(C_{2,A})$ is assumed to be lower $(C_{1,A}>C_{2,A})$. Since this data source is not managed in-house, the anticipated accuracy is lower $(\varepsilon_2<\varepsilon_1)$.

Table 3. Example 2 – Parameters				
Customer Information	**Exploitative**	**Explorative**	**Internal Source**	**External Source**
Basic demographics	1	1	$C_{1,B}$	$C_{2,B}$
Advanced demographics	0	1	$C_{1,A}$	$C_{2,A}$
Maximal Utility	U_a	U_b		
Probability	~1	P_b		
Accuracy Sensitivity Parameter	δ_a	δ_b		
Setup Cost			C_1	C_2
Accuracy Level			ε_1	ε_2
Cost Sensitivity Parameter to Accuracy Ratio			λ	

Adapting *(13)* to the parameters for this scenario (Table 3), the design-optimization model is to select the target accuracy level *(AC)*, the data sources $\{S_1, S_2\}$, and the demographic information to be imported from each data source $\{W_{1,B}, W_{1,A}, W_{2,B}, W_{2,A}\}$ so as to maximize the net benefit (P^D):

$$(14)\ P^D = U_a AT_B AC^{\delta_a} + p_b U_b AT_B AT_A AC^{\delta_b} - \left[\begin{array}{l} S_1\left(C_1 + \left(\max\{AC/\varepsilon_1, 1\}\right)^{\lambda}\left(C_{1,B}W_{1,B} + C_{1,A}W_{1,A}\right)\right) + \\ S_2\left(C_2 + \left(\max\{AC/\varepsilon_2, 1\}\right)^{\lambda}\left(C_{2,B}W_{2,B} + C_{2,A}W_{2,A}\right)\right) \end{array}\right]$$

s.t. $\{AT\}, \{S\}, \{W\}$ are *0/1* integers

$W_{1,B} \le S_1$, $W_{1,A} \le S_1$, $W_{2,B} \le S_2$, and $W_{2,A} \le S_2$

$S_1 = 1-(1-W_{1,B})(1-W_{1,A})$, and $S_2 = 1-(1-W_{2,B})(1-W_{2,A})$

$W_{1,B} + W_{2,B} \le 1$, and $W_{1,A} + W_{2,A} \le 1$

$(1-AT_{a,B})(1-AT_{b,B}) \ge (1-W_{1,B})(1-W_{2,B})$, and $(1-AT_{a,A})(1-AT_{b,A}) \ge (1-W_{1,A})(1-W_{2,A})$

Solving this yields different outcomes, depending on actual parameter values:

Supporting exploitative usage only **–** this occurs when the costs of acquiring advanced demographics (from either source) and/or quality improvement exceeds any benefit from explorative usage. The basic demographics will be obtained either from the internal or external source, depending on the actual cost and accuracy level (the internal source was assumed to have higher accuracy and lower data acquisition cost, but higher setup cost.) The corresponding net benefit is:

$$(15)\qquad P^{D/a} = U_a AC^{\delta_a} - \min\left\{\left(C_1 + \left(\max\{AC/\varepsilon_1, 1\}\right)^{\lambda} C_{1,B}\right), \left(C_2 + \left(\max\{AC/\varepsilon_2, 1\}\right)^{\lambda} C_{2,B}\right)\right\}$$

The optimal accuracy-level can be obtained for each data source. For instance, considering the internal source, assuming initially that the given accuracy (ε_1) is lower than the targeted *(AC)*, the net benefit is:

$$(16)\qquad P^{D/a} = U_a AC^{\delta_a} - \left(C_1 + \left(AC/\varepsilon_1\right)^{\lambda} C_{1,B}\right)$$

Comparing the first derivative to 0:

$$(17)\qquad \partial P^{D/a} / \partial AC = \delta_a U_a AC^{\delta_a - 1} - \lambda \varepsilon_1^{-\lambda} C_{1,B} AC^{\lambda - 1} = 0$$

The candidate optimal solution is:

$$(18)\qquad AC = \left(\frac{\lambda \varepsilon_1^{-\lambda} C_{1,B}}{\delta_a U_a}\right)^{\frac{\lambda}{\delta_a}}$$

If this solution is not within the $[\varepsilon_1,1]$ range, and/or if the second derivative of *(16)* is positive (i.e., indicating a minimum, not maximum), the optimal solution is either (a) above *1*, suggesting that the maximum possible accuracy should be approached (i.e., $AC=1$), or (b) below ε_1, in which case the accuracy level of $AC=\varepsilon_1$ is already guaranteed by the data source at no added cost. After obtaining the optimal accuracy per data source, *(16)* can be used to select the source that maximizes net benefit.

Supporting both exploitative and explorative usage when under at least one configuration the expected benefit from explorative usage exceeds the advanced demographics cost. Depending on parameter-values, the firm may acquire all demographics from the same source (internal or external), or acquire the basic from the internal source and advanced from the external. We can evaluate real-option gross benefit *(ROGB)* and cost differential *(ROCD)* accordingly, e.g., assuming (a) internal source for exploitative-usage alone, at accuracy level of $AC_a>\varepsilon_1$ $\left(P^{D/a}=U_a AC_a^{\delta_a}-\left(C_1+\left(AC_a/\varepsilon_1\right)^{\lambda}C_{1,B}\right)\right)$, versus (b) internal source for basics and external for advanced, at accuracy $AC_b>\varepsilon_1,\varepsilon_2$ $\left(P^{D/b}=U_a AC_b^{\delta_a}+p_b U_b AC_b^{\delta_b}-\left(C_1+\left(AC_b/\varepsilon_1\right)^{\lambda}C_{1,B}+C_2+\left(AC_b/\varepsilon_2\right)^{\lambda}C_{2,A}\right)\right)$. Here, $ROGB=U_a\left(AC_b^{\delta_a}+AC_b^{\delta_a}\right)+p_b U_b AC_b^{\delta_b}$, $ROCD=\left(AC_b{}^{\lambda}-AC_a{}^{\lambda}\right)\varepsilon_1^{-\lambda}C_{1,B}+C_2+AC_b{}^{\lambda}\varepsilon_2^{-\lambda}C_{2,A}$, and the latter option is preferred if $ROGB>ROCD$.

Avoiding implementation altogether - if acquisition and accuracy-improvement costs exceed all possible benefits.

Supporting explorative usage only is unlikely in the current model – the data resources that support explorative usage can also support exploitative usage at no additional cost. However, a slightly different model may yield such an outcome – for example, if exploitative usage requires a very high accuracy (lower-bound constraint on accuracy level), while explorative usage has substantially lower accuracy requirements.

The data-resource model and design scenarios highlight the importance of data source evaluation. DW designers typically prefer internal, well-structured, and easily accessible sources. While this may guarantee faster implementation, the limited scope can damage usability and, hence, utility. Alternately, too many sources can increase complexity and cost, resulting in sub-optimal benefits. This model encourages adopting an economic perspective for assessing data quality in a DW. There is a shift toward viewing DW quality from usage and business-value perspectives (e.g., Wixom and Watson, 1998; Nelson et al., 2005). When both utility and cost are considered, maximizing data quality can be economically sub-optimal if quality-improvement costs offset utility. These trade-offs emphasize modeling the effects of source selection and quality improvement decisions more explicitly. This model is a step in this direction and can be extended to address complex scenarios.

Design Scenario 3 – System Cost and the Selection of Database Server

The system cost (C^S) includes purchasing, setting up, and maintaining the DW system infrastructure (Figure 3), which provides the hardware, software, and network resources needed for establishing data repositories. Two key capacity aspects affect C^S: *a) Storage-capacity:* DW environments manage large datasets and require large storage capacity. Implementing storage capacity involves the purchase of database management systems server software (e.g., Oracle, MS-SQL), the hardware to support it (e.g., server, disk space,) and labor. *b) Processing-capacity:* the DW-repository is populated and refreshed with new data using ETL processes – *Extracting* data from different sources into the DW environment, *Transforming* data into the required format, and *Loading* it into the DW repository. Processing capacity is typically established by incorporating ETL-engines – software, either developed in-house or purchased (e.g., Informatica, MS-DTS), for configuring, scheduling and executing the ETL processes.

We model the storage capacity and the related DBMS choices. We consider a single large tabular dataset (e.g., a fact table) that affects storage capacity. The design scope is the time span coverage *(TS)* – a [0,1] ratio, which determines the number of records and thus storage. A tabular dataset implies a linearly increasing volume with the number of records. Assuming an approximately fixed number of records per time-unit, the dataset volume will be linearly proportional to *TS*. The time span coverage sets a capacity constraint for each information product: $TS_i \le TS$. The utility of each information product *[i]* is assumed to increase with TS_i. For maximal net benefit, each information product will utilize the maximum-available time span, hence, at the optimal point, $TS_i=TS$, for all *[i]*.

We now consider *L* possible DBMS-configurations (indexed by *l*). Each has a fixed cost $C^{S/F}{}_l$ (e.g., DBMS-software, hardware) and variable cost that increases (assumed linearly) with data volume (e.g., owing to storage space). Hence with the time span $C^{S/V}{}_l TS$ ($C^{S/V}{}_l$ represents the variable cost required for covering the entire time span available). We

also assume an upper limit on storage capacity per configuration A^S_l, as certain DBMS-configurations cannot manage high data volumes. Hence, if θ is the data volume for covering the entire time span, the storage constraints are $\{\theta(TS) \leq A^S_l\}$.

Following *(1)*, the DBMS configuration can be framed as a design-optimization model – select a configuration (indexed by l) and time span coverage TS such that the net benefit (P^S) is maximized:

(19) $$P^S_l = U^S_l - C^S_l = \sum_{i=1}^{I} U_i^{Max} TS^{\alpha_i} - C^{S/F}_l - C^{S/V}_l TS$$

s.t. $0 \leq TS \leq min\ (1, A^S_l/\theta)$ per l

where

P^S_l, U^S_l, C^S_l - Overall net benefit, utility and system cost, respectively

U_i^{Max} - Maximal utility of information product *[i]*, considering the other characteristics in *(9)*, i.e., $U_i^{Max} = U_i^* RD_i^{\beta_i} AC_i^{\delta_i} \left(\prod_{m=1}^{M} AT_{i,m}^{\eta_{i,m}}\right)\left(\prod_{n=1}^{N} PC_{i,n}^{\gamma_{i,n}}\right)\left(e^{-TM_i}\right)^{\tau_i}$

TS - Time span coverage

α_i - Utility-sensitivity parameter to the time span coverage

$C^{S/F}_l$ - Fixed setup cost

$C^{S/V}_l$ - Variable cost of covering the entire time span available

$\{A^S_l\}$ - Data-volume capacity

θ - Data-volume required for entire time span coverage

Illustrative Example 3: A large hospital evaluates a DW for analyzing treatment history (e.g., treatments, medications, and labs.) Possible DW usages are: *(a) exploitative* – the DW can support on-going reporting and monitoring needs such as inventory tracking, resource utilization, and treatment-history for specific patients. The anticipated utility from these usages (U_a) has a relatively low time span sensitivity parameter (α_a). *(b) Explorative* – the DW can potentially support advanced analyses such as detecting shifts in resource utilization, identifying patterns of reactions to drugs, and segmenting treatment history along demographic and socio-economic attributes. The anticipated utility is U_b with success probability of $(P_b<1)$ and failure probability (*0* utility) of $1\text{-}P_b$, and a relatively high time span sensitivity parameter (α_b).

Covering the entire time span *(TS=1)* requires a large data volume (θ). Hence, the designer considers including only a portion of the available time span. L possible database configurations (indexed by l) are considered for implementing the DW, each having a different fixed cost (C^F_l), variable cost for covering the entire time span (C^V_l), and data-volume capacity (A_l). Based on these parameters (Table 4), which system configuration *(l)* should be chosen, and what should be the dataset time span coverage *(TS)*?

Table 4. Example 3 – Parameters

	Exploitative	**Explorative**	**DBMS-configuration *l***
Maximal Utility	U_a	U_b	
Probability	~1	P_b	
Time Span Sensitivity Parameter	α_a	α_b	
Maximal Data-Volume Requirement			Θ
Data-Volume Capacity			A_l
Fixed Cost			C^F_l
Maximal Variable Cost			C^V_l

For optimality, we evaluate *(19)* along all *L* configurations:

(20) $P_l = U_a TS^{\alpha_a} + p_b U_b TS^{\alpha_b} - C_l^F - C_l^V TS$, s.t., $0 \le TS \le min\ (1, A^S_l/\theta)$

Setting the first derivative to 0, we can obtain a candidate optimal-solution:

(21) $\partial P_l / \partial TS = \alpha_a U_a TS^{\alpha_a - 1} + \alpha_b p_b U_b TS^{\alpha_b - 1} - C_l^V = 0$

If this solution is outside the *[0, min (1, A^S_l/θ)]* range, and/or if the second derivative of *(20)* is positive (indicating a minimum, not maximum), the optimal solution is either (a) above the range, hence, we select the lower of the entire time span *(TS=1)* and the capacity/size ratio *(TS= A^S_l/θ,)* or (b) below the range, hence, the constraint enforces *TS=0*. The latter option implies that the evaluated configuration cannot yield a positive net benefit and hence, should not be considered. After obtaining the optimal time span and the net benefit as per configuration *(20)*, we choose the one that maximizes net benefit.

Unlike the two previous examples, where under certain circumstances explorative usage was not implemented, here it will always be supported to some extent, assuming that at least one configuration yields positive net benefit. However, real-option considerations may arise with a change to the model – e.g., a lower-bound constraint on the time span coverage for explorative use (i.e., $TS_b \ge TS^*$). Under this constraint, we may have one configuration (indexed *l/a*) that optimizes exploitative usage only with time span coverage of $TS_a < TS^*$ and offers a net benefit of $P_{l/a} = U_a TS_a^{\alpha_a} - C^F_{l/a} - C^V_{l/a} TS_a$, and another configuration (indexed *l/b)* that optimizes both usages, with time span coverage of $TS_b \ge TS^*$, and a net benefit of $P_{l/b} = U_a TS_b^{\alpha_a} + p_b U_b TS_b^{\alpha_b} - C^F_{l/b} - C^V_{l/b} TS_b$. Supporting all usages has a real-option gross-benefit of $ROGB = U_a \left(TS_b^{\alpha_a} - TS_a^{\alpha_a}\right) + p_b U_b TS_b^{\alpha_b}$ and a corresponding cost differential of $ROCD = C^F_{l/b} + C^V_{l/b} TS_b - \left(C^F_{l/a} + C^V_{l/a} TS_a\right)$. This is chosen if *ROGB≥ROCD*.

Storage capacity is a key design-decision in DW environments (as is processing capacity, not evaluated here). Implementing a large capacity is expensive – DBMS software that supports high-volume storage is expensive to license and requires extensive labor. Alternately, a cheaper DBMS might limit storage capacity and, hence, the ability to enhance the DW to support new information products and usages. Similarly, time span coverage can significantly impact utility, storage capacity, and investments in DBMS. Extending the model for quantifying the effects of time span choices will offer the DW designer a useful design tool. Record density and field-structure also affect the dataset volume and consequently the storage capacity and should be examined.

Conclusions and Directions for Future Research

This study contributes to design research by developing an economics-driven framework and evaluating related constructs for a critical task, the design of a DW. It explores the link between design decisions and economic benefits and suggests that modeling the effects of design decisions on economic performance can enhance design processes. The innovative approach proposed here is founded on this notion. The need to support both exploitative and explorative usages in a DW environment has important design implications – usages may substantially differ in data-utilization patterns and levels of uncertainty, and these differences can influence and direct design decisions. Similarly, design decisions influence and can be influenced by costs – higher capacity, faster performance, and sophisticated capabilities require larger investments. Modeling the effect of design decisions on utility and costs can assess trade-offs and identify economically optimal designs.

The DW environment is modeled as a layered stack. We treat capacity in each layer as a real-option investment decision. The utility gained by data usage (topmost layer) is often uncertain. Adding capacity to the layers below allows timely and often cheaper support to new types of usages and, hence, enhances opportunities for utility contribution. Our real-option model addresses common DW design trade-offs – investing in expensive slack capacity for explorative usage with a potential for utility gains at a substantial risk, versus optimizing capacity for exploitative usages with more certain returns. The design scenarios demonstrate this, highlighting that real-option considerations exist along all DW-layers.

Although the economic-aspects of IS have been recently discussed, their impact on IS design is not apparent. Examining IS/IT design from an economic perspective offers substantial benefits. We view this framework as a

significant first step in this direction. The evaluated decisions models address specific design scopes and limit the number of design characteristics within each. In practice, DW environments involve a complex set of design decisions with many inter-dependencies, constraints, and economic effects. Our preliminary models simplify the formulation of utility/cost effects and must be enhanced. Turning these into useful design tools will require not only analytical enhancements but also quantitative parameter assessments. A common approach for parameter assessment is decision-calculus (Little, 1970), with a long history of successful applications. Using this, parameters can be solicited from managers or estimated empirically. A plethora of successful modeling and parameter-estimation methodologies have been described in literature (e.g., Little, 1970; Eriksen and Berger, 1987; Hanna et al., 2005). These can help extend the models that we suggest here into useful design tools.

References

Ahituv, N. "A Systematic Approach towards Assessing the Value of Information System," MIS Quarterly (4:4), 1980, pp. 61-75

Baldwin, C.Y., and Clark, K.B. "Design Rules, Vol. 1: The Power of Modularity", MIT Press, Cambridge, MA, 2000

Ballou, D.P., Wang, R., Pazer, H., and Tayi, G.K. "Modeling Information Manufacturing Systems to Determine Information Quality," Management Science (44:4), 1998, pp. 462-484

Benaroch, M., and Kauffman, R.J. "A Case for Using Real-options Pricing Analysis to Evaluate Information Technology Project Investments," Information Systems Research (10:1), March 1999, pp. 70-86

Benner, M.J., and Tushman, M.L. "Exploitation, Exploration, and Process Management: The Productivity Dilemma, Revisited," Academy of Management Review (28:2), 2003, pp. 238-256

Counihan, A., Finnegan, P., and Sammon, D. "Towards a Framework for Evaluating Investments in Data Warehousing," Information Systems Journal (12), 2002, pp. 321-338

Conner, K.R., and Prahalad, C.K. "A resource-based theory of the firm: Knowledge versus opportunism," Organization Science (7:5), 1996, pp. 477-501

Cooper, B.L., Watson, H.J., Wixom, B.H., and Goodhue D.L. "Data Warehousing Supports Corporate Strategy at First American Corporation," MIS Quarterly (24:4), 2000, pp. 547-567

Cooper, R., and Slugmulder, R. "Achieving Full-Cycle Cost Management," MIT Sloan Management Review (46:1), 2004, pp. 45-52

Elmasri, R., and Navathe, S.B. "Fundamentals of Database Systems (5th Ed.)," Addison Wesley, Redding, MA, 2006

Eriksen, S.E., and Berger, P.D. "A Quadratic Programming Model for Product Configuration Optimization," Zeitschrift fűr Operations Research (31:2), 1987, pp. 143-159

Even, A., Shankaranarayanan, G., and Berger, P.D. "Profit-Maximization with Data Management Systems," Proceedings of the Intl. Conference of Information Systems (ICIS), Dec. 2005, Las Vegas, NV

Hanna, R., Berger, P.D., and Abendroth, L. "Optimizing Time Limits in Retail Promotions: an Email Application," Journal of the Operational Research Society (56), 2005, pp. 15-24

Henderson, R., and Clark, K. "Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms," Administrative Science Quarterly (35), 1990, pp. 9-30

Hevner, A.R., March, S.T., Park, J., and Ram, S. "Design Science in Information Systems Research," MIS Quarterly (28:1), 2004, pp. 75-105

Kimball, R., Reeves, L., Ross, M., and Thornthwaite, W. "The Data Warehouse Lifecycle Toolkit," Wiley Computer Publishing, New York, NY, 2000

Kleinberg, J., Papadimitriou, C., and Raghavan, P. "A Micro-Economic View of Data Mining," Data Mining and Knowledge Discovery (2:4), 1998, pp. 311-324

Kogut, B., and Kulatilika, N. "Capabilities as Real-options," Organization Science (12:6), 2001, pp. 744-758

Kohli, R., and Sukumar, R. "Heuristics for Product-Line Design Using Conjoint Analysis," Management Science (36:13), 1990, pp. 1464-1478

Lee, Y.W., and Strong, D.M. "Knowing-Why about Data Processes and Data Quality," Journal of Management Information Systems (20:3), 2003, pp. 13-39

Lee, S.M., Hong, S., and Katerattanakul, P. "Impact of Data Warehousing on Organizational Performance of Retailing Firms," International Journal of Information Technology and Decision Making (3:1), 2004, pp. 61-79

Little, J.D.C. "Models and Managers: The Concept of a Decision Calculus," Management Science (18), 1970, pp. 466-484.

March, J.G. "Exploration and Exploitation in Organizational Learning," Organization Science (2:1), February 1991, pp. 71-87

Melville, N., Kraemer, K., and Gurbaxani, V. "Review: Information Technology and Organizational Performance: an Integrative Model of IT Business Value", MIS Quarterly (28:2), June 2004, pp. 283-322

Messerschmitt, D.G. and Szyperski, C. "Software ecosystem: understanding an indispensable technology and industry," MIT Press, 2003, Cambridge, MA

Nelson, R.P., Todd, P.A., and Wixom, B.H. "Antecedents of Information and System Quality: An Empirical Examination within the Context of Data Warehousing," Journal of Management Information Systems (21:4), Spring 2005, pp. 199-235

Sambamurthy, V., Bharadwaj, A., and Grover, V., "Shaping Agility through Digital Options: Reconceptualizing the Role of Information Technology in Contemporary Firms," MIS Quarterly (27:2), 2003, pp. 237-263

Schwartz, E.S., and Zozaya-Gorostiza, C. "Investment under Uncertainty in Information Technology: Acquisition and Development Projects," Management Science (49:1), 2003, pp. 57-70

Shankaranarayanan, G., and Even, A. "Managing Metadata in Data Warehouses: Pitfalls and Possibilities," Communications of the AIS (2004:14), 2004, pp. 247-274

Shapiro, C. and Varian H.R. "Information Rules," Harvard Business School Press, 1999, Cambridge, MA

Shin, B. "An Exploratory Investigation of System Success Factors in Data Warehousing," Journal of the AIS (4), 2003, pp. 141-170

Simon, H.A. "The Science of the Artificial (3rd edition)," The MIT Press, MA, 1996

Srivastava, J., and Chen, P.Y. "Warehouse Creation - A Potential Roadblock to Data Warehousing," IEEE Transactions on Knowledge and Data Engineering (11:1), 1999, pp. 118-126

Sullivan, K., Chalasani, P., Jha, S., and Sazawal, V. "Software Design as an Investment Activity: A Real-options Perspective," In Real-options and Business Strategy: Applications to Decision Making. L. Trigeorgis, ed. Risk Books, 1999

Teece, D.J., Pisano, G., and Shuen, A. "Dynamic Capabilities and strategic management," Strategic Management Journal (18), 1997, pp. 509-534

Tushman, M.L., and Anderson, P., "Technological Discontinuities and Organizational Environments," Administrative Science Quarterly (31), 1986, pp. 439-465

Wang, R.Y. and Strong, D.M. "Beyond Accuracy: What Data Quality Means to Data Consumers," Journal of Management Information Systems (12:4), 1996, pp. 5-34

West, L.A. Jr. "Private Markets for Public Goods: Pricing Strategies of Online Database Vendors," Journal of Management Information Systems (17:1), 2000, pp. 59-84

Wixom, B.H., and Watson, H.J. "An Empirical Investigation of the Factors Affecting Data Warehousing Success," MIS Quarterly (25:1), 2001, pp. 17-41