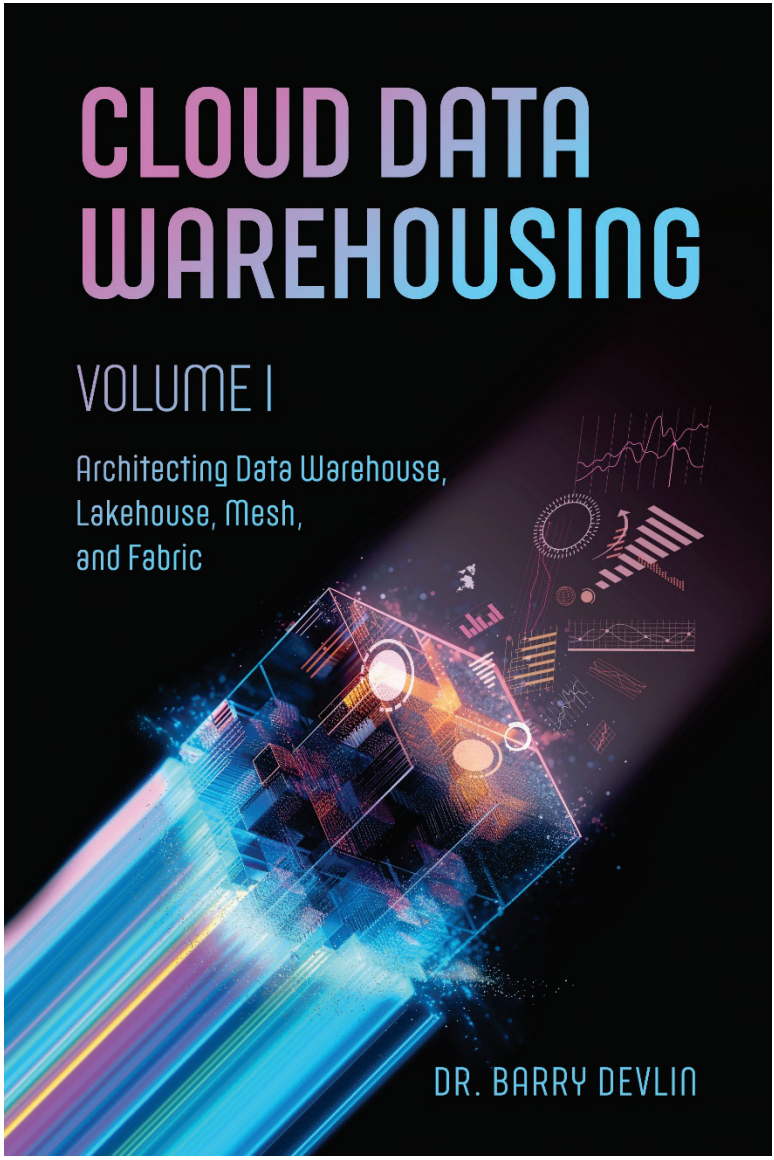

https://technicspub.com/cloud-dw-series/

Praise for Cloud Data Warehousing

"Barry Devlin has taken on the task of examining how the cloud changes data warehousing. No, it's not just about being "cloud native." Barry outlines what's different, why it matters, and what to do about it. A must read." — Merv Adrian, Founder and Principal, IT Market Strategy

"Once again, Barry Devlin has sharpened the focus on new trends in data architecture, explaining what's new, what's not, and what matters. For more than 30 years, Barry has placed enterprise data architectures on a firm footing with his insights and frameworks. For data veterans and newbies alike, this new volume will edify, instruct, and keep readers on the right path."— Wayne Eckerson, President, Eckerson Group

"The cloud database and analytics industry is evolving rapidly, with new approaches and radical enhancements. As one of the original thinkers in data warehousing, Barry Devlin's forty years of experience and knowledge of current thinking brings welcome perspective. He explains trends and navigates the landscape in a clear, practical manner, to help design, implement, and benefit from emerging systems. As always, his writing is thought provoking and enjoyable, and will be of interest to anyone working in this area."— Henry Cook, Gartner

"In 1988, Devlin and Murphy published a vision of analytic architecture. Millions of jobs and billions of dollars later, we're almost finished building it. Barry Devlin's new cloud books are the design pattern for the next era."— Dan Graham, PM director for IBM SP2 and Teradata 6700.

"Barry Devlin sets out a soup to nuts overview of how all that is new is old in data management. He makes a compelling case for the importance of fundamental concepts when navigating the choppy waters of the data ocean." — Daragh O Brien, CEO, Castlebridge

"In the data and analytics space today, there is so much chasing of the next big thing. Barry Devlin has a unique ability in building good scaffolds that help us to think clearly and focus on the jobs to be done. 'Context-setting Information' is a first-class, concise, and intuitive framing term for metadata. And the 'Three Thinking Spaces for Cloud Data Warehousing' will aid streamlining today's architectural discussions. Highly recommended for both modern old-timers and newbies!" — Thomas Frisendal, Data Architect and Modeler, ISO database language standards committee National Expert

"The data warehouse is not dead! It is an essential part of information management. Barry Devlin brings fresh ideas and new thinking at a time when we really need to rethink data warehousing. This book is filled with thought-provoking wisdom from the original architect of the business data warehouse."— Dave Wells, Information Management Architect, Consultant, and Educator

Cloud Data Warehousing

Volume I: Architecting Data Warehouse, Lakehouse, Mesh, and Fabric

Dr. Barry Devlin

Technics Publications

Published by:

115 Linda Vista, Sedona, Arizona USA
https://www.TechnicsPub.com

Edited by Jamie Hoberman
Cover design by Lorena Molinari

First Printing 2023

ISBN, print ed.	9781634623360
ISBN, Kindle ed.	9781634623377
ISBN, ePub ed.	9781634623384
ISBN, PDF ed.	9781634623407

Library of Congress Control Number: 2023938546

Data Warehousing—Purpose and Principles

The intelligent have plans; the wise have principles.

Raheel Farooq

Our wander down the lanes of data warehousing history was purposeful. Of the principles that we began with in the 1980s, some have remained solid in the interim decades. Some have evolved as we've better understood the needs of business and the nature of data. A few of our principles have not withstood the test of time. This chapter[1] delineates the purpose and principles of data warehousing as currently understood and how they support current business requirements.

Spoiler alert: The purpose and principles of cloud data warehousing do not differ from those of data warehousing as it has existed for multiple decades.

---

[1] Parts of chapters 3 and 4 are reproduced by permission of the Insurance Data Management Association, from their Course Book, *Approaches to Data Design, Engineering, and Development* (IDMA and Devlin, 2023).

The Purpose of Cloud Data Warehousing

I'm using the term *cloud data warehousing* here to emphasize the forward-looking emphasis of all that follows. However, the only difference between cloud data warehousing and previous incarnations of data warehousing is that the data is being stored and processed largely or entirely off premises "in the cloud." The data storage locations and processing approaches or, indeed, any legal constraints or contractual arrangements around them can have no relevance to the business purpose intended.

The purpose of cloud data warehousing emerges from all the same reasons that drove data warehousing four decades ago. It needs repeating only because modern thinking tends to focus solely on the newest, sexiest business support technologies—analytics, predictive and prescriptive algorithms, AI, and ML—and what they offer their users. Therefore, we must keep in mind that the primary purpose of data warehousing, cloud or otherwise, is to provide the information to support businesspeople's insights in decision making and action taking. Both span from the smallest and fastest activities to the largest and slowest processes.

This purpose clearly includes the provision and preparation of the necessary data/information at the appropriate levels of timeliness, consistency, usefulness, usability, and quality,

commensurate with the value offered and rigor demanded by the business activity in hand. To declare the purpose of cloud data warehousing to be solely to support "analytics"—as is often done today—is to completely underestimate the true scope of what is needed by the business and offered of the technology.

First, however, we must begin with that old conundrum—data or information—and try, once more, to put a stake in the ground. Or through data's heart 😊.

ONE LAST TIME: DATA VS. INFORMATION

Most discussions about the difference between data and information start with data and then describe why information is different. For example, data may be equated to "facts" and information defined as facts with useful context added. But what is a fact, especially in a world full of discordant opinions and fake news? Which context should be added: the context of creation, first use, or reuse? Starting the discussion with information, however, leads to far more clarity.

Information is what people create—directly (by typing or speaking into a smartphone, by taking a photograph or making a video, for example) or indirectly (such as by friending someone on Facebook, searching on Google, or researching products on Amazon). It is their means of expressing

themselves and for communicating with and relating to others. Information can be anything from a poem to a purchase order, from a click trail to a call home, from an address label to an advertising video, as detailed in *Table 3.1*.

Type / Class of Information	Examples
Text	E-mail, Memos, Word-processing documents, Stories, Novels, Poems, SMS or Text Messages, Facebook posts, Tweets, Audio transcriptions, Forms (blank or filled), Field names, Descriptions and help in computer applications, Handwritten messages via optical character recognition
Audio Media	Audio recordings, Phone and VOIP calls, Music, Audio conferences, Interactions with Siri, Alexa, etc.
Video Media	Video conferences, recordings, and calls, TV programs, Movies, Virtual reality and holographic media
Images	Photographs, Scans, Maps, Digitized drawings, Holograms

Type / Class of Information	Examples
Computer Interactions	Clicks, Gestures, Selections in computer applications, Likes, Friends, and Links in social media
Future types	Emerging examples, beginning with information generated by AI, chatbots, etc.

Table 3.1: Types/Classes of Information

The examples are deliberately drawn as widely as possible from all walks of life. Not all are necessarily applicable in a business environment. However, the expansion so far in this millennium of information types used for different business purposes is instructive. It indicates that we may anticipate a growing range that piques business interest as technology enables new possibilities and society evolves.

An important characteristic of information is its creation context. A person who creates information does so in a certain time, place, and context, using specific tools. All of this may be relevant to the meaning or use of the information thus created. A letter written with a pencil is less durable than one written with a pen, and its content therefore deemed less reliable. A Facebook photo, stamped with date and place, may offer context of importance in a criminal case. A contract

created in South Africa conforms to Roman Dutch law even though the content may not explicitly reference it. Traditional metadata represents some—mostly technical—aspects of the creation context of information.

Today, all information, almost without exception, is digitally encoded, processed, and stored either transiently or permanently in some sort of computing or communications device. The expectation is that it will be permanently available, readable, and searchable while also secure and private as required. Experience suggests that these assumptions are far less reliable than may be imagined, a fact that influences architectural decisions in many ways.

Business, information is, of course, what people (and, increasingly, AI) use to support management, analysis, decision making, and action taking throughout the organization. It should be noted that these usage contexts of information—and there are generally many for every set of information—most often differ from its creation context. A creation context and a set of usage contexts associated with an information set is vital to understanding, interpreting, and effectively benefiting from it.

Some context may be explicit, such as the date/timestamp and GPS location included in the metadata of a photograph. Most or all of such metadata's value and use is obvious. In

contrast, implicit context may well exist in the content of the information. The use of certain slang words in a text message may offer clues to the community of the writer. The shadows in a photograph may indicate the time of day.

The next chapter will have some more to say on context, but there is a reason why it is emphasized here. Context, it turns out, is the distinguishing feature between information and data. Naked data is a subset of information from which context has been stripped to the maximum extent. In the simplest terms:

`Information = Naked Data + Context`

The qualifier *naked* is added to make clear the difference between the word *data* on its own, often used to mean different things, sometimes even to stand for *information*.

Who stripped the context away and why? It was the developers of software tools and IT applications who dropped context from information, transmuting it into naked data. In the early days of computing, they did this mainly to reduce storage and processing needs. In the 1960s, computing was generally known as *data processing*, a moniker that was literally true then and remains so in many cases today.

The approach has persisted to the present day, even though storage is now immense and processing power enormous.

The reason remains largely technical. Mathematical calculations are easier when performed only on data (numbers) rather than information (contextualized numbers). A junior accountant enters sales in a spreadsheet with the currency symbol included in the cell: £30.50. As a result, monthly autosummations fail, much to his manager's annoyance. Spreadsheets, like most applications, only deal well with naked data. When a bill is paid in Euro, however, the value of entering €50.35 becomes apparent. The summation still doesn't work, but that may be a good thing because it would be wrong. Excel would need to be a lot more sophisticated, as well as knowledgeable about exchange rates and which to apply when in order to use information (naked data qualified by currency context) rather than naked data in its calculations.

From this perspective, starting with a definition of information and deriving an understanding of data, we see that information is fundamental, while data is just a subset of it. In fact, naked data is a degenerate subset of information, degraded by the stripping away of the context explicitly or implicitly embedded in information. We must be aware that when we focus on data rather than information, we are at risk of misunderstanding or misinterpreting the meaning and relevance of the information from which that naked data was

derived[2]. Furthermore—and this becomes a central pillar of all data warehousing architectures, including cloud data warehousing—we need to understand where that context goes or resides when it is sucked out of information.

The answer is surprisingly complex and varied, both logically and physically. And the cloud complicates it further. We return to this topic when we define context-setting information (CSI) in chapter 4 and explore the manifest meaning model (m^3) in depth in Volume II.

SEVEN DEADLY SINS OF DATA WAREHOUSING

Psychology tells us that we all carry unexamined assumptions from early learning and experience that color current behaviors without our conscious awareness of their power or even existence. Data warehousing experts, vendors, and implementers are no different. In my experience, there exists seven deadly sins of data warehousing—dating back many decades. They point to thinking that we unconsciously apply to design decisions, even though many of the problems they purport to address have long since been solved or mitigated.

[2] Throughout the rest of these books, I'll mostly use data as a shorthand for naked data for ease of reading.

1. Never mind the information, feel the data

We dealt with this in the previous section. Dating from the earliest days of computing in the 1950s, this category error has proven impossible to dislodge: the confusion between data and information. Its consequences persist still. Why do Chief *Information* Officers deal with technology and data management and Chief *Data* Officers with the governance of what is actually information?

2. Operational and informational systems are separate

Dating from the 1970s and incorporated into thinking about data warehousing from its earliest days, this remains a hidden and uncontested assumption behind many approaches to data warehousing even today.

Its origin is both business-driven and technological. Historically, decision makers operated on weekly or monthly cycles, often deliberately ignoring the daily fluctuations of business activity. On the technology front, applications were hand-crafted and run on mainframes operating at the limits of their computing power and storage. It made sense to have separate operational and informational systems: business users liked it, and anyway, the technology could not support decision-

making systems running on operational databases (or files).

3. *Data integration is possible only with a warehouse*

In the 1980s, data warehousing was driven by the need to integrate data from disparate operational systems. These sources were often complex, poorly designed by today's standards, enormous black-box applications, built separately over many years, and never designed to work together. They could not provide a consistent view of the full business. Their data was often incomplete, inaccurate, and inconsistent across different sources. The only viable place where these problems could be fixed and data combined was the data warehouse.

4. *Delivering quality data requires central control*

In a direct follow-on to the previous point, such integration was only possible in a centralized location; there was no other suitably large and powerful environment at the time. No other approach was possible, and once accepted, none other was necessary to consider.

5. *Layering is obligatory for speedy, easy querying*

Except possibly for Teradata's massively parallel processing (MPP) warehouse, the first data warehouses of

the late 1980s were exceedingly slow to deliver even simple query results, never mind the more complex queries often needed by business. Layering and subsetting of the data warehouse into an EDW and dependent data marts became almost mandatory. Although database performance improved, data volumes increased in tandem, and layering and subsetting has remained the go-to solution, despite its obvious drawbacks in development complexity and timely data delivery.

6. *An enterprise data model can exist only in the EDW*

Also in the late 1980s and into the 1990s, data modeling was becoming a popular way to translate business requirements into database structures, and the first forays into enterprise architecture and enterprise/industry data modeling were being made (Zachman, 1987), (Evernden, 1996). Given its complexity, an enterprise data model (EDM) was usually considered impossible to implement in operational systems. Performance would plunge. And the risks to daily operations were too high. The EDW (as well as master data management, MDM) was seen as an ideal environment to build out the EDM as a basis for data governance, integration, and quality.

7. *Innovative business usage of data threatens quality*

Spreadsheets became ubiquitous in the 1990s and remain businesspeople's primary data tool. IT has long regarded this as a threat to data quality. There is truth in this concern, but the innovative possibilities and ease-of-use that spreadsheets offer cannot be ignored.

This thinking is symptomatic of the central control-focused approach to data warehousing that prevailed throughout its early history. It doesn't sit well with the empowerment of businesspeople to which we aspire in digital business and cloud computing.

Each of the above postulates was valid at some stage of the evolution of data warehousing. Each and every one of them is untrue to some extent today. Some were invalidated as long as twenty years ago. However, they still cast extensive shadows in the thinking of some long-time practitioners. A brief examination of each will show where it limits current thinking and how it has led to specific design decisions in all the data warehousing approaches previously described. This rather straightforward exercise is left to the reader.

Today's practitioners also bring their own deadly sins of assumption. Theirs is a background of almost limitless computing power and storage distributed from smartphones to cloud data centers. Their software of choice is open-source, and their methodologies are Agile and DevOps. This leads to

beliefs such as "decentralized everything," "Agile is the only way," or "move fast and break things." In a decade or so, these deadly sins may look as outdated as those listed above.

FIVE FOUNDATIONAL PRINCIPLES OF CLOUD DATA WAREHOUSING

Now, let's forgive ourselves of our past and present deadly sins and jump to a set of valid principles for cloud data warehousing. Principles we can be proud of today and into the future. Let's begin to define a vision of how cloud data warehousing should be. However, do keep in mind that these principles may be no more immutable than our prior sins!

One over-arching design drive is evident in digital business today. That is the desire to eliminate the functional silos and the gaps between them that emerge in every business of more than a few people. The aim is to eliminate these gaps that confuse businesspeople, slow business processes, and sow confusion and inconsistencies in data and information.

In data warehousing, our focus is on analysis, decision making, and (the oft-forgotten) subsequent action taking. Eliminating the blindingly obvious gaps between these activities in today's implementations leads to the five modern principles of data warehousing—on-premises and in the cloud.

1. **A modern digital business seamlessly combines analysis, decision making, and action taking, requiring a logically integrated, coherent continuum of *all* valid information used by the business: the information space.**

   All—literally all—of the information that may potentially be used by businesspeople, irrespective of its source, is included in this information space. This incorporates all the process-mediated data produced by every operational application, whether on-premises or in the cloud. It includes all information and data received from the internet, other businesses, and regulatory authorities, as well as all information created during the processes of analysis, decision making, and action taking. This principle is emergent or evident in all modern approaches to cloud data warehousing.

   But how can we know what any piece of information or data actually means and how it can be used in any specific situation? There must exist some description, which may be called *metadata*, *context-setting information*, or *ontology* of the contents of the information space.

2. **The information space must be based on a comprehensive, supra-enterprise information model, spanning *all* information types used by the business.**

The complexity and development time/cost of an enterprise data model (EDM) are familiar to traditional data warehouse implementers. Many try to minimize its role or even avoid it altogether. Nonetheless, a model spanning from semantics to logical structure is necessary to define information meaning. It must relate all information across the enterprise and address the challenges of velocity and veracity of externally sourced data.

The further implication is that the model must span enterprise boundaries—be supra-enterprise—to enable viable inter-enterprise collaboration. It must also include all types of information shared between enterprises. The need for such broad coordination of context and meaning drives industry-wide and other ontologies, but their implementation is still in its infancy.

This principle is accepted—at least theoretically—in all approaches to cloud data warehousing, except data mesh, where a more localized formulation is preferred. Implementation therefore differs by approach and will be considered on a case-by-case basis.

3. This information space is best maintained as the minimum number of copies of each item of information, resorting only to transient layers or duplicates of specific subsets for specialized needs.

App developers are compulsive and unapologetic copiers of existing data. It's a control thing! It removes development dependencies. It allows small (allegedly unimportant) tweaks in the data structure or content. And anyway, storage has become incredibly cheap.

But the costs of and time required for data management and information governance increase dramatically—possibly exponentially—with the number of copies and near-copies of data stored. AI may reduce this cost, although that remains unproven. The cloud's distributed nature will increase the pressure to make more copies. This is a habitual development behavior that drives technical debt. It needs to be reined in, on principle.

Thus far, the focus has been firmly on information and its storage and management. However, how is it created, modified, enhanced, delivered, and deprecated? In brief, computer processes must be developed and operated to do all these activities of managing data. This is the software needed to deliver cloud data warehousing.

4. **An integrated, model-based, and closed-loop process space is needed to create, maintain, and use information and to support human activities.**

Traditionally, data warehousing was the preserve of "data geeks" and application development the concern of "coding jockeys". In the world of data warehousing, the processes needed to manage data were often relegated to a secondary role in support of the database designers and administrators. In cloud data warehousing, this situation is no longer viable as the activities of managing data in a completely distributed environment become increasingly complex and widespread.

The processes needed to create, maintain, and use information are as important as the information itself, and modern, comprehensive software engineering practices must apply. Just as the data must be integrated, so too must be the processes around it. These processes require knowledge of the structure and meaning of the information space and thus also depend on a model-based foundation. A closed-loop approach is required to ensure that tasks that are begun will terminate properly, that their outcome is confirmed to be as expected, and if not, that appropriate steps are taken.

This leads directly to the final principle. Processes act on behalf of someone and are designed to do that person's bidding. They are what support human activity in the computing environment and that enable human access to and use of information.

5. A supra-rational, flexible, and role-aware people space provides access to all process and through the process space to information.

   The real role of people and how and why they actually behave personally or in organizational roles is often largely ignored or underestimated in data warehousing and, indeed, generally in computing. This is significant because people—businesspeople, decision makers, data scientists, line workers, customers, regulators, and more—all interact with information in different roles with disparate objectives. Some of these objectives are explicit, some may be unconscious, some indeed, may be counter-productive to the business' goals, or perhaps may be illegal. People make decisions and take action; their motivations are important.

   The people space addresses these vital considerations. It is often largely absent from IT considerations and cloud data warehousing thinking. Indeed, business itself hardly considers the psychological and sociological underpinnings of how people behave within a business milieu, what motivates them, and the consequences of that—unless and until some crime is committed or some program radically fails. This principle declares that we must consider people- and organization-related aspects in the design and delivery of cloud data warehousing.

The above principles apply to all forms of data warehousing, including cloud data warehousing.

THREE THINKING SPACES FOR CLOUD DATA WAREHOUSING

The principles outlined above lead straight to a conceptual architecture for cloud data warehousing. The purpose of a *conceptual* architecture is to provide a shared structure and vocabulary to allow business and IT to discuss and settle on business drivers and information technology enablers. It is thus kept deliberately simple as an introductory image that both business and IT can keep in mind as they begin to discuss what cloud data warehousing means for the business and how technology drives and enables it.

This conceptual architecture was first published a decade ago (Devlin, 2013). It is uniquely based on a reevaluation from first principles of how data/information has been used in business since the beginning of data processing in the 1960s up to and including the advent in the early 2010s of data-driven business. As digital business has evolved since, this architecture—both the conceptual level described here and the logical level described in the following chapter—has so far withstood the test of time. It is thus equally applicable as a basis for designing and delivering cloud data

warehousing. More recently, I have begun calling this approach the Digital Information Systems Architecture (DISA). The aim is to emphasize that this architecture can underpin all aspects of the design of and transformation to a digital business, of which cloud data warehousing is a key part.

Figure 3.1 shows the basic conceptual architecture of cloud data warehousing. The three thinking spaces arise directly from the spaces identified in the last section. The addition of the word *thinking* emphasizes that these are conceptual in nature; they do not show physical implementations nor even logical components. However, the spatial arrangement is meaningful. Information is placed as the foundational block; the others build upon it. Cloud data warehousing begins and ends with information. Not data, but information. Data on its own lacks the meaning and context needed as a basis for decisions or actions.

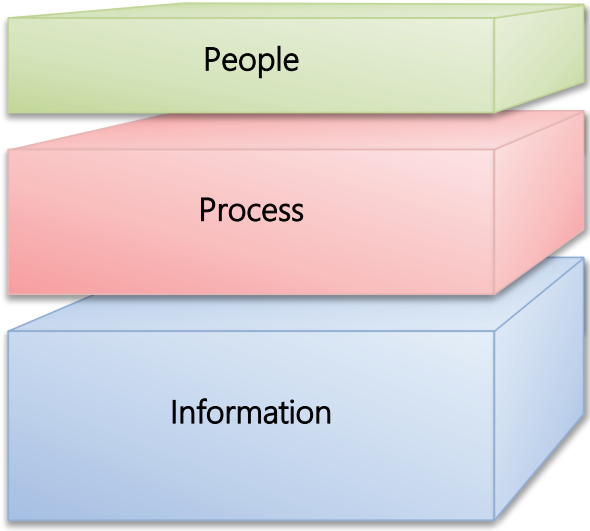

Figure 3.1: The cloud data warehousing conceptual architecture

Process is placed between people and information because people always use—create, access, or manipulate—information via some process. Processes may be formally defined and implemented or may be informal and *ad hoc* in nature. However, no matter how loosely defined, we recognize that using information is a process that can be modeled, managed, and improved as needed.

Finally, people—whether individually or collectively in organizational units—are the primary drivers of everything in the digital business. The people thinking space therefore sits above the other two blocks.

Also called the IDEAL architecture, the acronym lists its five key aspects: Integrated/Inclusive, Distributed, Emergent, Adaptive, and Latent. These characteristics are largely self-explanatory, except for *latent*, confirming that this conceptual view remains hidden in the logical architecture and implementation as we see later. This name emphasizes that a conceptual architecture is, in fact, an ideal that can only be approached but never perfected. In business and IT, trade-offs are always necessary. Furthermore, in a constantly changing environment, business needs and technical possibilities are always evolving, and an ideal must accommodate that. The conceptual architecture is therefore more an image of what we aim to achieve rather than what can be fully delivered. Further details can be found in *Business unIntelligence* (Devlin,2013).

You might argue that technology should also appear here. Indeed, there are long-standing methodologies for strategic planning that include technology, either instead of or in addition to data/information (Simon, 2021), (Shevlin, 2021). Technology does play a role in conceptual architecture, and could certainly be depicted as a fourth thinking space beneath information. However, in my experience, architects are often too focused on technology and may overemphasize it in discussions with the business. I therefore made a conscious decision to exclude it at this level of architecture,

knowing that IT will bring their knowledge of it to the conversation anyway. I generally reserve consideration of technology to the physical implementation stage of design.

Figure 3.1 is kept deliberately simple as the initial starting point of the business-IT conversation. When we need to dive to the next level of conceptual thought, we add to each thinking space a set of three axes as shown in *Figure 3.2*. Each axis represents a topic of significant importance in the discussion between business and IT, as experienced over the years of designing data warehousing solutions.

The choice of three axes is, again, driven by the goal of keeping the picture clear and simple, both in its representation and in its discussion. However, in a specific implementation, additional or other topics may rise in importance and need to be added to the picture or to replace/extend the existing axes. The individual axes, particularly those of the information space, will be considered in more depth as we explore each of the three thinking spaces.

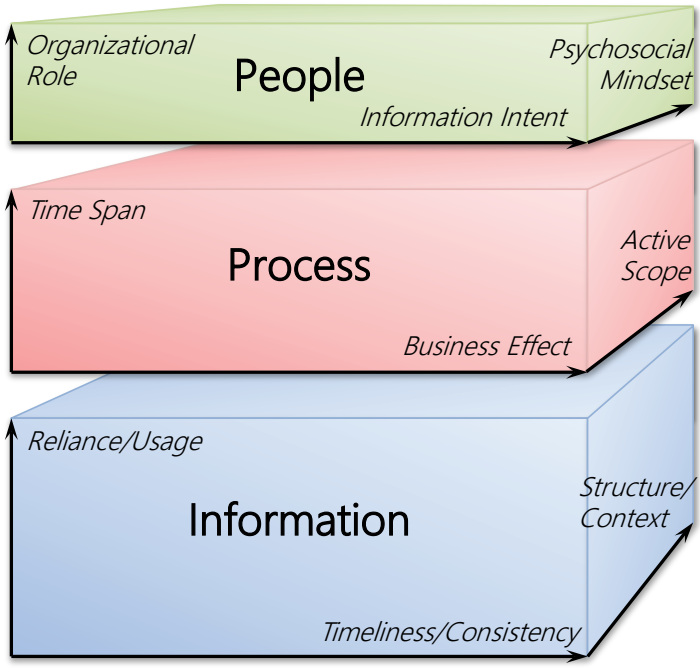

Figure 3.2: The axes of the conceptual thinking spaces

The information thinking space

The information thinking space is a conceptual view of the *entire* information resource of the enterprise—whatever its source, structure, or storage—that the business uses or may use to achieve its goals. Such information may reside in the data center, the cloud, on personal computers, smartphones, or any network-attachable device with computing function—in short, anywhere. The information may be formally the property of the enterprise or belong to its customers or partner enterprises, such as social media firms, governments, and

so on. This inclusion of *all* information—whatever its location or ownership—is necessary because a modern business uses a much broader scope of data/information than traditional businesses and implicitly depends on the availability and quality of all that information.

The three axes of the information space represent key conversations between business and IT, as well as vital trade-offs that must be made in delivering cloud data warehousing. Although it might be argued that there should be more than three topics, experience shows that the chosen axes and their composite character are generally as far as business will want to go and as deep as IT needs to delve to determine the most important design considerations.

As shown in *Figure 3.3*, each axis is labeled with a series of classes, indicating the state of the information at those positions. For example, on the timeliness/consistency axis, we have a series of classes that would be familiar to a data warehouse architect: live information is found in operational systems; reconciled in data warehouses; etc.

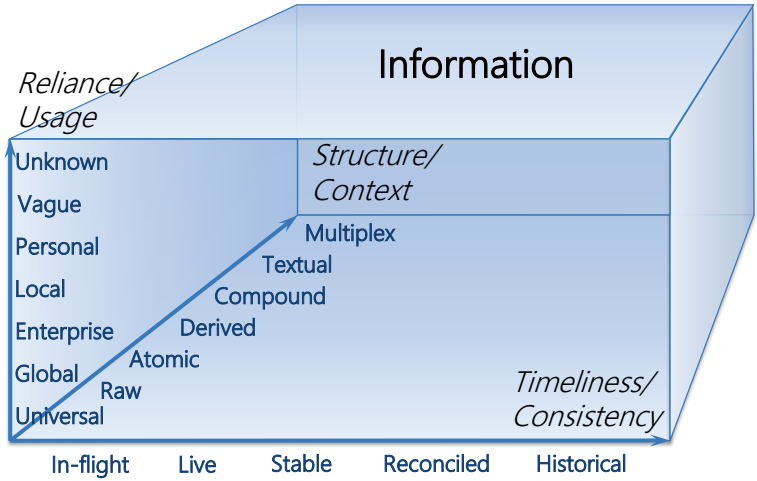

Figure 3.3: The three axes of the information thinking space

The idea is that a piece or set of information is positioned in the three-dimensional space based on the characteristics that can be read off the axes. However, all three axes are continua and the classes depicted are not discrete steps. When we identify information along some axis, we may find that it has some characteristics of two or more classes.

The Timeliness/Consistency Axis

In a digital business, with data coming at speed from multiple sources, understanding the trade-off between timeliness and consistency for any resulting information set is vital. If all the data needed comes from one source, its timeliness and thus the speed at which it can be processed is set only by that source, which also determines how internally consistent it is.

With multiple sources, however, consistency must be ensured after all the source data is available. This determines the timeliness of the information seen by the business. In addition, the faster the required reconciliation, the more expensive it is. The classes on this axis are largely self-explanatory, ranging from the timeliest and most transient on the left to the most consistent and permanent on the right.

The Structure/Context Axis

Big data is often described as *unstructured* or *semi-structured* in comparison to the *structured* data found in traditional systems. The terminology is unfortunate. Unstructured data cannot exist; that is just noise. However, the extent of structuring of data is an important characteristic that determines how it can be stored and processed. For example, the highly structured atomic data in an order entry system or that used to manage the business in a data mart is very different from the loosely structured information in a textual contract or in the multiplex videos or images on Facebook.

The related and, arguably, more important characteristic included on this axis is the context that is embedded with the data and makes it usable and useful for the business. Raw data, coming from sensors, for example, is just numbers. It requires extensive metadata and descriptions before it can be understood and safely used. The multiplex information of a video recorded at the scene of an accident, on the other

hand, contains a wealth of context that a human (or increasingly an AI system) may be able to extract without explicit metadata. Raw data is encoded and has a meaning and minimal structure that is defined by its designer, while multiplex information is more loosely structured with implicit/tacit meaning and context.

The balance of these two aspects—structure and context—of any information required or used by the business provides specific storage and processing requirements to IT and drives technology choices and implementation approaches.

The Reliance/Usage Axis

The third axis shows the level of trust that the business can place in information and how safely it can be used. The classes on this axis range from completely trustworthy, universally usable information to that which has unknown provenance and can only be used with great care. Enterprise information, for example, in an EDW, can be used across the entire organization. Global information can be shared with other organizations or regulators. At the other end of the spectrum, personal information (in spreadsheets, for example) and vague information, such as that arriving from YouTube, requires careful identification, management, and possibly restricted access within the enterprise.

Cloud implementation of data warehousing may lead to new considerations on this axis compared to on-premises implementations. This may occur when locating specific types of data/information in the cloud gives rise to varying reliance/usage considerations. For example, in the case of personally identifiable information (PII), the location or jurisdiction where data is stored gives rise to various legal and sovereignty issues that would appear on this axis.

The process thinking space

The process thinking space for cloud data warehousing, shown in *Figure 3.3*, positions ***tasks*** or types of activities to better understand their key characteristics. As in the case of the information thinking space, the classes are a continuum, merging from one to the next, rather than discrete divisions. The process space is somewhat simpler than its information counterpart; each axis here represents a single characteristic.

The first process dimension is **active scope**, which is based on microservices—and originally, Service Oriented Architecture (SOA)—thinking. The lowest level is an ***event*** in the real world. This might be a simple machine sensor measurement or a basic value calculation. It might be a communications event, such as a tweet or an email. Legally significant events or combinations thereof take on special significance and become ***transactions*** through cleansing and validation.

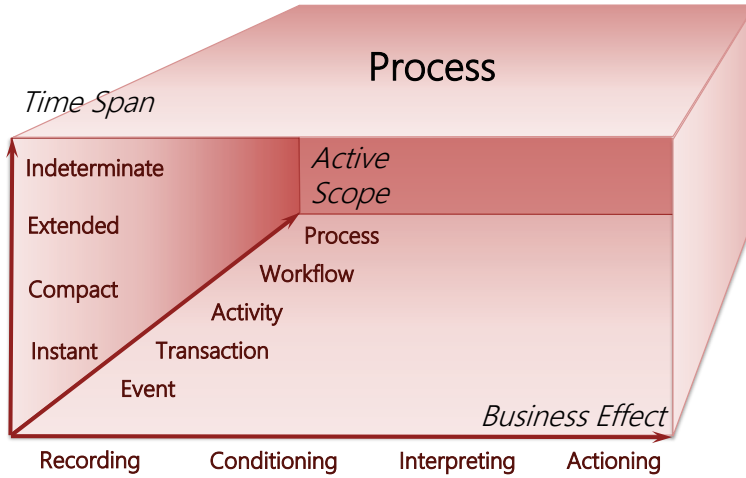

Figure 3.4: The three axes of the process thinking space

An ***activity,*** or service in SOA, is the smallest unit of function that makes sense to a business user, usually consisting of a number of related transactions and events. Services are defined by the business but implemented by IT to maintain internal consistency. A ***workflow*** links activities into flexible sequences as workflows. These are defined by businesspeople and may be implemented or changed by them, at least in principle. A ***process*** is the highest level of workflow delivering a value-generating set of interdependent activities.

The **business effect** dimension describes how any action taken relates to the underlying business information. ***Recording*** is the act of capturing any class of data/information from any source. This ranges from storing it in a database—where a permanent record is necessary—to simply "noting"

in-flight data or accessing it remotely. Metadata is also recorded.

Conditioning covers all subsequent changes to, calculations about, and derivations and deletions of the recorded information. *Interpreting* is the next step of applying intelligence to information to understand its business implications. It is thus a function found in both operational and informational systems. It ranges from validating and understanding the data entered to analytics. Finally, *actioning* reaches a decision on what to do and taking the appropriate action.

Time span represents the period of time over which a process element is active or open. The *instant* time span, associated only with events and the simplest of transactions, refers to the shortest measurable timing. Each such event or transaction deals with a single, integral piece of information. Operational transactions occur in a *compact* time span, a period of time within which a person might place an order for an item, for example. The *extended* time span recognizes that some activities can spread over hours, days, or more because of temporal dependencies either in the real world—closing a contract—or in the technical environment—running a batch job. The *indeterminate* time span applies to processes that could possibly go on forever and where no end condition can be defined. Business processes at the highest level have this characteristic.

The people thinking space

The people thinking space provides the foundation for understanding why and how businesspeople use and process information to drive business needs, to get their job done, and to succeed in their roles. In defining cloud data warehousing today, it is—unfortunately—given the least attention, simply because current technologies seldom consider these matters too deeply.

Figure 3.5 shows the three axes and their ranges of characteristics that must be considered.

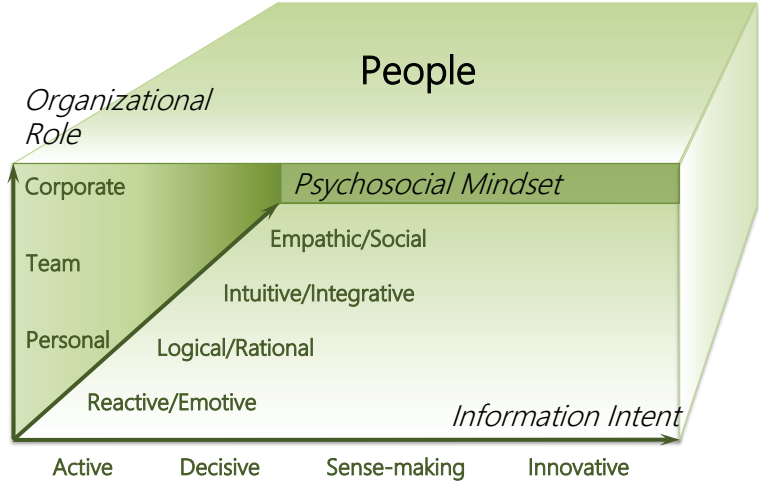

Figure 3.5: The three axes of the people thinking space

Organizational role is the most familiar and simplest axis. People use information and make decisions aligned to their

role in the organization, which is shown here in a typical but generic hierarchical structure.

The information intent axis provides the primary linkage from the people space to the spaces below it. *Active* intent—getting something done—is the simplest and most common motivation in business, especially in day-to-day operations and in time-constrained activities. It is supported through operational applications and operational BI. *Decisive* intent—reaching a decision—is seen most obviously in people the organization empowers to make decisions of significance: managers and executives. This is the most common area for application of traditional BI and dashboarding tools.

Sense-making intent is the behavior of seeking a story to explain some phenomenon and what to do about it. BI tools, as well as analytics—and spreadsheets—are most often applied here. Intuition and team working, with its social and empathic skills, are required. Significant expansion in IT support is needed here, driven by the recognition that human behavior is far more complex and nuanced than business typically imagines. *Innovative* intent drives the "ah-ha moments", the creation of novelty, the spark for new products or processes. It is the least understood, most poorly supported, and least managed of all four types of intent. It is the source of the most important behaviors and requires future business and IT focus.

Perhaps the most interesting and, sadly, least often considered aspects of business use of information and decision making lie on the **psychosocial mindset** axis. Much of academic and business school consideration of business decision making starts—and ends—with **rational decision theory** (March, 1994), which is a poor basis for understanding how real people behave, especially under stress.

Modern psychology and neurobiology show that—in real life—much more is involved. This leads to the concept of **insightful decision making** that is well-informed but cognizant of the decision maker's mental landscape. Combining left and right brain, it takes account of intuitive and emotional concerns to reach integrated, well-rounded decisions (Siegel, 2010). Human behavior operates at all psychosocial levels and applies also to decision making in business settings.

At the lowest level, ***reactive/emotive*** impulses may be unconscious, driving actions that may not be in the best interests of the business or the person. Such impulses cannot be eliminated; rather, the goal is to integrate the negative and positive aspects to create more holistic functioning. ***Logical/rational*** thinking is so deeply embedded in Western culture that we imagine it to be the most desirable and prevalent mode of thinking. Both are myths. While vital for many tasks, logic alone can miss the bigger picture and come to distinctly inhuman conclusions, a danger becoming

apparent in emerging horror stories from the application of AI to social problems (O'Neill, 2016).

In the right brain, *intuitive/integrative* thought sees the forest for the trees and finds those "ah-ha insights". This is the source of creativity and the mother of invention. Working together with the structured and rational left brain, this drives the creativity that modern business seeks. Finally, *empathic/social* thought is the foundation of relating at the personal, group, and societal levels. Collaboration springs from here, and with it, our best opportunities for innovation.

Further academic research and development are needed to elaborate this thinking space and allow it to take its proper and vital role in future cloud data warehousing design.

TAKEAWAYS

- The purpose of cloud data warehousing is the same as that of data warehousing in general. This purpose is to enable and support businesspeople in their analysis, decision making, and action taking required to run and manage the business. It provides them with the data and information they need at suitable levels of timeliness, consistency, usefulness, usability, and quality, commensurate with the value offered and the rigor demanded by the business activity in hand.

- Information is what the business needs from cloud data warehousing. Data, stripped of context to varying degrees, is not sufficient.

- Examining the history of data warehousing shows seven mistaken or outdated assumptions that may underpin poorly considered decisions about the scope of cloud data warehousing, what it can achieve, and its fundamental limitations.

- The five foundational principles of cloud data warehousing set its scope as *all* the information that may be used by the business and the processes needed to deliver and manage that information. People are positioned as the ultimate arbiters of what is to be delivered.

- The three thinking spaces of the conceptual architecture of cloud data warehousing are information, process, and people, with information as the foundation and people at the apex. Each thinking space provides the basic design topics that must be discussed and agreed between the business and IT.

- These same three spaces allow us to tell the simplest and most elegant story of the purpose of cloud data warehousing: *People process information.*