

## The Seven Faces of Data

### *Rethinking data's basic characteristics*

November 2011

A White Paper by

Dr. Barry Devlin, 9sight Consulting  
[barry@9sight.com](mailto:barry@9sight.com)

*We live in a time when data volumes are growing faster than Moore's Law and the variety of structures and sources has expanded far beyond those that IT has experience of managing. It is simultaneously an era when our businesses and our daily lives have become intimately dependent on such data being trustworthy, consistent, timely and correct. And yet, our thinking about and tools for managing data quality in the broadest sense of the word remain rooted in a traditional understanding of what data is and how it works.*

*This paper proposes seven fundamental traits of data structure, composition and use that enable IT professionals to examine existing and new data sources and respond to the opportunities and challenges posed by new business demands and novel technological advances. These traits can help answer fundamental questions about how and where data should be stored and how it should be protected. And they suggest how it can be securely made available to business users in a timely manner.*

*Using the Data Equalizer, a tool that graphically portrays the overall tone and character of a dataset, IT professionals can quickly evaluate the data management needs of a specific set of data. More generally, it clarifies how technologies such as relational databases and Hadoop, for example, can be positioned relative to one another and how the data warehouse is likely to evolve as the central integrating hub in a heterogeneous, distributed and expanding data environment.*

Sponsored by:  
Teradata Corporation  
[www.teradata.com](http://www.teradata.com)

**TERADATA.**

## Contents

**Data—time to restructure  
our thinking?**

**So—what do you need  
of your data?**

**Seven data traits:**

- Horizon
- Composition
- Anatomy
- Temporality
- Temperature
- Access
- Trust

**Using the Data Equalizer**

**Conclusions**

## Data—time to restructure our thinking?

We are living through the greatest explosion of data ever seen on this planet. An explosion that is set to continue at an ever increasing pace for the foreseeable future. According to International Data Corporation (IDC)<sup>1</sup>, the volume of data that will be generated in the digital world in 2011 is 1,800 Exabytes (EB), or 1,800 million Terabytes and set to grow almost 40% in the next year to 2,500 EB. By 2020, IDC predicts the number will have reached 35,000 EB, or 35 Zettabytes (ZB), and apparently not even enough disk space to store it all!

Such figures are beyond comprehension. Of course, much of this data is comprised of video, audio and image data generated by a general public waving smart phone cameras wherever they go and perhaps we can argue sensibly that IT managers don't need to worry so much. But even in the relatively staid world of enterprise IT, the numbers get a bit scary. Figure 1 narrows the focus to enterprise data showing traditional, structured and unstructured<sup>2</sup> volumes. In 2005, we stored 4 EB of structured data; by 2015 it will grow to 29 EB, a compound annual growth rate (CAGR) of over 20%. The figures for unstructured business data

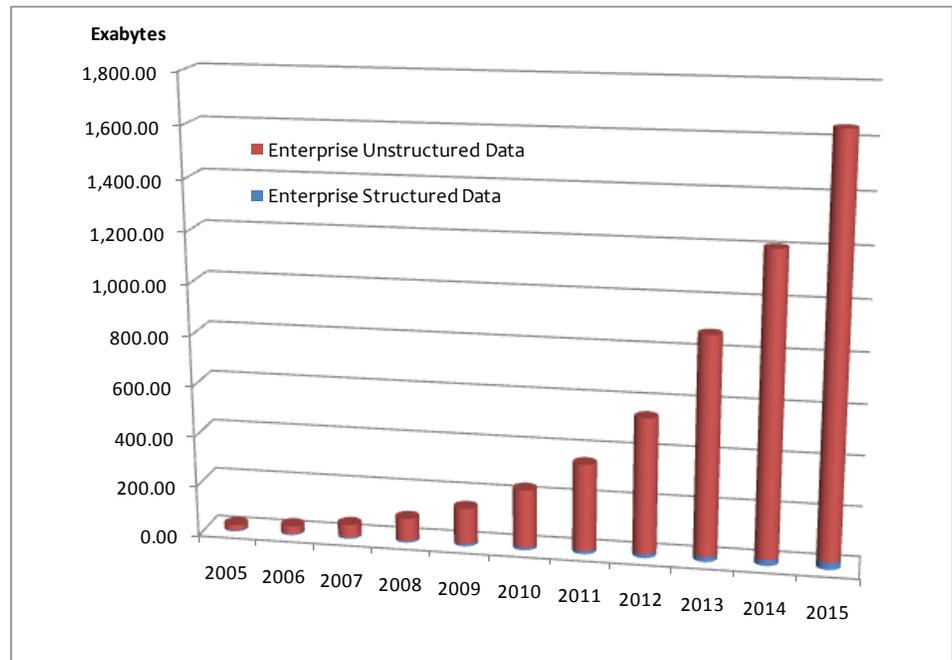


Figure 1:  
Total  
enterprise  
data growth,  
2005-2015

confirm what all the experts say: such data now far exceeds structured data in volumes and is growing even faster. In 2005, it amounted to 22 EB; and reaches 1,600 EB by 2015. That's a staggering CAGR of approximately 60%. In 2005, the type of data we're comfortable with as data warehouse and operational systems experts comprised 15% of the total; this year it's down to 4%.

But while the figures may be scary, they tell only part of the story. The enormous growth in unstructured data volumes and proportion shows up explicitly. What doesn't appear is equally important for data management. The main sources of data are migrating dramatically from internal and operational to external and user-generated, data quality from known to unknowable, from pre-defined conditions of use to conditions that have to be inferred at use time. More importantly, the figures say very little about the intrinsic value of the data. That 4% of structured data above still represents probably 80% or more of the value of data, simply because it describes the fundamentals of any business: customers, products, sales, profits and every other core measure of interest.

These dramatic changes cannot be ignored. The term *big data* has become common recently and hints at the ongoing sea change. But it is vague and unsatisfactory, focusing only on data volumes. A number of analysts and vendors have begun to discuss the topic in terms of volume, variety and velocity and more<sup>3</sup>. However, there has been no serious attempt so far to define a set of characteristics of today's enormous and highly heterogeneous data resource that would help IT to make sensible decisions on how to manage, store, process and make it available. That is the goal of this paper.

### What is this thing called data?

Before computer professionals corrupted the word, data was the plural of datum. Datum, from the Latin "something given" is defined as (1) a single piece of information, such as a fact or statistic and (2) in philosophy, any fact assumed to be a matter of direct observation or any proposition assumed or given, from which conclusions may be drawn<sup>4</sup>. This definition emphasizes that data and reality are

distinctly different. Computer data is at least one and usually two steps away from what exists in the real world. First there is the object or event in the real world, such as the smart phone you want to buy or the click on the website that found it. At the second level, is the specific and very personal human interpretation of the object or event. Is the smart phone an object of desire or simply a communications device? Is that website click an indication of buying interest or a random slip of the mouse? Depending on the person, their interests, attention and the context, we attribute different meanings and interpretations to everything we see. Finally, at the third level is data: the formalization of the information we choose or chance to see about the real world. At this level, the computerized level, IT must determine the appropriate data and structure required to give a useful representation of the object or event coupled with the needs of the business users.

### ***And why is it important?***

All very well, I hear you say, but why should I care? Taken together, the explosive growth of data volumes and the technological advances over the past decade is changing the ground rules for every aspect of data management in its broadest sense. Data professionals are faced with difficult decisions about how best to gather, model, store, manage, protect, make available, decommission data and more. Instead of hammering new data into existing platforms and tools, the model shown in this paper provides a framework for assessing new data sources early on and separating requirements from products.

Most IT professionals recognize data management life cycle shown below:

- **Sourcing:** What must we do to ensure data quality or mitigate quality concerns during creation or ingestion of data?
- **Processing:** How must data be (pre-)processed to ensure optimum balance between timeliness and consistency, reliability and innovative use, cost and value and a variety of other trade-offs?
- **Storage:** Where is data best stored in terms of physical devices, logical domains (e.g. cloud, enterprise, personal storage) and geographical locations (local, distributed, etc.)?
- **Protection:** What are the appropriate levels of protection that must be applied, from recovery in the event of disaster to defense of confidentiality and privacy?
- **Usage:** How can innovative and integrated, controlled yet creative use of data throughout the organization be facilitated?

What we need now is a set of characteristics of data as it exists in our systems and applications that enable practical and pragmatic data management decisions at every phase of the data life cycle.

---

## **So—what do you need from your data?**

A survey of data-centric sources of information reveal almost thirty data characteristics considered interesting by different experts. Such a list is too cumbersome to use. Narrowing down the almost thirty characteristics to a more manageable number, can be based on two criteria. First is the practical usefulness of the characteristic: how does the trait help IT make decisions on how to store, manage and use such data? What can users expect of this data based on its traits? Second, can we measure the trait in a way that helps us manage the data better?

The following seven characteristics reflect key data management concerns and are linked to the most relevant stages in the data life cycle on which they impact:

1. **Horizon [sourcing]:** How reliable is the data and how broad its possible usage based on its original data source type and subsequent manipulation
2. **Composition [processing]:** The overall composition of the data set; how the records relate to facts or events in the real world and how they relate to one another

3. **Anatomy [processing, storage]:** How complex and predefined is the data structure within each record (in the broadest sense of the word “record”)
4. **Temporality [storage]:** The degree to which the latency of data delivery and validity over time matches the needs of the consuming individuals or processes
5. **Temperature [storage, usage]:** The level of demand for and use of the data, which may vary at different times
6. **Access [usage]:** Whether data is public or has various levels of restriction in access and use, and the extent to which access to data is restricted appropriately to maintain its security and privacy
7. **Trust [usage]:** The extent to which data is regarded as true and credible by users

## Example

*Fire and Life Insurance Protection (FLIP) Ltd. finally joined the Internet age acquiring a web-based company with a substantial on-line automotive insurance business among under-30s, a strategic market segment for FLIP.*

*Penny Wise, the CIO, has been tasked with bringing FLIP into the 21st century using clickstream data from the newly combined websites and sentiment information from social networking sites such as Facebook and Twitter, in addition to the non-trivial task of consolidating the open-source motor insurance BI environment with FLIP’s extensive, successful data warehouse. The first business requirement is to integrate web information into FLIP’s campaign management system to enable the sales force to offer existing and reliable vehicle customers life and property insurance at a safe but attractive discount.*

Envisioning and dealing with more than three characteristics of anything is tough! Seven is beyond the scope of most people. We need a visual metaphor to enable us to deal with such complexity.

### The Data Equalizer

The data equalizer analogy is based on the audio sound mixers (equalizers) used to set the tone, levels, and mood of a musical recording. There is a subtle difference, however. The audio equalizer is a control device: sliders set the characteristics of the music output, amplifying or reducing different parts of the audio spectrum to produce, for example, a sound like a concert hall from a recording made in a small studio.

As shown in figure 2, the seven data traits are depicted as sliders with between four and six positions that represent the measures for each trait. The sliders are positioned left to right relative to the stage of the data life cycle that most influences their setting. Any data set of interest can thus be mapped against the seven traits and the resulting pattern of the slider positions observed.

This pattern of slider positions on the data equalizer is broadly predictive: it identifies the characteristic “tone” of the each type of data. When data is abstracted this way, both the IT professional and business user can foresee the processing needed, the best location to store the data, its usage and value, and extrapolate to the appropriate budget required. Consequently, the data equalizer will help identify platforms, software tools, requirements and costs.

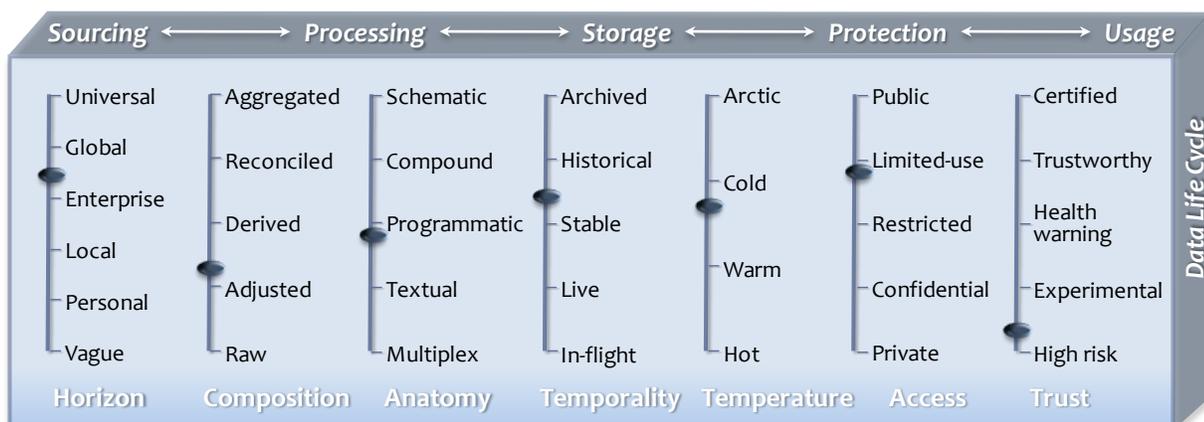


Figure 2: The Data Equalizer

**B**ack in the “good old days”, you could be sure of one thing about your business data—it all came from the one place, your mainframe. And it stayed there, too. And because mainframes were not cheap, you could be relatively certain that a lot of effort had gone into making that data as clean, accurate and consistent as possible. As a result, and by design, the horizon from which it came was the enterprise as a whole; and you could apply it with confidence enterprise-wide.

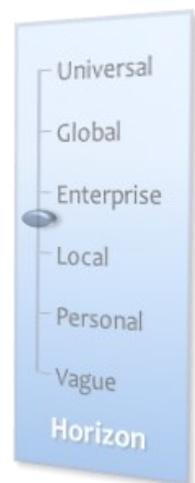
Today, data comes from every source imaginable, and a few we haven’t yet imagined. And with that variety of sourcing comes great uncertainty about the reliability of the data and, as a consequence, what users can validly do with it. It’s not just about the original source of the data; it’s also about any manipulation it may have undergone before it became available for use. We must therefore consider the likelihood that the horizon can be narrower or broader than the enterprise level, spanning from vague and personal all the way to universal acceptance.

### Measures

Horizon ranges from extremely limited to the broadest and most reliable possible. Some Internet data is such that its provenance ranges from largely unknowable to poorly defined or volatile. Such **vague** data carries many risks, but may be the best that can be obtained in some circumstances; competitor prices or blogs may fall into this category, but may still be deemed useful in the absence of more reliable data. **Personal** data comes from a known, trusted personal source. Known and trusted is, of course, a value judgment; in practice, it generally means an employee or contractor of the enterprise. Spreadsheet data originates on this horizon, but can be moved higher through peer review or more formal expert evaluation processes. Much personal data originates at a wider horizon (see below), but is “degraded” by unmanaged personal manipulation.

**Local** data is at the next wider horizon; it originates from an IT system formally designed and constructed to perform a particular task or capture a specific measurement—in traditional terms, an application. It thus includes data generated by operational systems, human data entry and machine-generated data. Such data has a well-defined scope and application, within which it can be relied upon. Users, and IT, must be aware that data originating at a local horizon and used at a different or wider horizon requires great care. For example, consider RFID sensor data that has been gathered for the purpose of tracking the speed of goods through the supply chain from manufacturing to store. Its local horizon is goods tracking. Reusing this data for detecting “shrinkage” or theft (a different horizon) may give invalid results when delivery routing is changed, even though the original application results remain true.

Most of us are familiar with moving local data to the wider **enterprise** horizon—it’s the basis of an enterprise data warehouse (EDW). It is, in many cases, an expensive process involving widespread modeling, cleansing and reconciliation of data from different local scope sources. This illustrates the general principle that moving data from a narrower to a wider horizon involves substantial human expertise, technology and investment. The **global** horizon takes this one step further, indicating data that can be used across enterprise boundaries, either between trading partners or with regulatory bodies. Some parts of an EDW may be designed for use at the global horizon; but it’s important to document those which aren’t—some measures are meant for internal use only! The final horizon is **universal**, indicating data that can be relied upon irrespective of time or place. It includes universal physical constants like the gravitational constant, as well as data that is declared by a relevant authority as final and fixed, such as specific taxes, time, or corporation names and addresses.



### Example

*Penny is faced with data from a variety of horizons, some new to FLIP. Social networking data is vague; a spike in negative sentiment, for example, may be real or may be due to a subversive campaign by a competitor. Clickstream data is local; it is under IT control but using it to understand customer needs was not its original design point. Combining social data or clickstreams with enterprise horizon data in the data warehouse does not raise the reliability of such data. Decisions based on these combinations must be treated with care. Such data combinations might need to be quarantined to avoid contamination of data used at the enterprise or global horizon for auditing or financial reporting.*

## Why it matters

In general, the use of data within a narrower horizon than that at which it was created is perfectly fine. The problem arises going the opposite direction; creating data with a broad horizon is more expensive, and often considerably more expensive, than at a narrower level. The financial consequences of misplaced trust in narrower horizon data for use in a broader context can be catastrophic.

Data has a natural horizon determined by its source and previous history of manipulation. Use of data beyond this horizon carries significant risk of misinterpretation and errors of judgment. This situation is well-known today in the case of spreadsheets created in whole or in part from personally generated data which are subsequently used in enterprise level decision-making processes and have led to serious financial and legal exposures<sup>5</sup>. A good example is the recent derivatives financial meltdown in the USA where brokers, banks, and investors were betting billions of dollars on fewer and fewer facts about the underlying mortgage loans. Data sourced from the Internet carries high levels of risk, given the uncertainty around its provenance.

Risk is the negative side of this equation, of course. But, in a more positive view, one may ask how we can deal with data that comes increasingly from diverse sources. The answer is twofold. First, we must define these different horizons and tag data accordingly, providing business users with an understanding of the limits of reliability of particular data and the risks of exceeding them. Second, we can undertake the necessary steps to certify the data at a higher usage horizon. This will certainly involve costs, some of them significant, to investigate data provenance, cleanse the incoming data if possible, or find alternative and more reliable sources. Data integration vendors do a fine job of handling data lineage (source mapping) but mostly miss the broader implications of the data horizon.

---

## Composition

[PROCESSING]

Each time we design a system, we make decisions about which data we collect and store and which we discard. We also decide at what level of detail we capture the data. We decide how long to keep it. The outcome of all these decisions defines the composition of the data set as a whole, describing the level of processing to which the data is subjected prior to its storage and use.

In some cases, particularly for traditional business activities, the composition of the data needed is relatively well known and defined. ATM transactions, for example, generate data containing the machine identifier, date and time, details about the person making the withdrawal, their account, and the amount of money dispensed. We need to keep the raw data indefinitely. On the other hand, the ideal composition of data describing driver behavior gathered from electronic sensors in automobiles and used for determining insurance premiums is far less clear. Partially, this is because the application is still evolving, but also because the data is only indirectly indicative of the characteristic being measured. The raw data consists of engineering measures, such as GPS location, acceleration or deceleration forces, tire pressures and more. This data may need to be refined or augmented to take account of poor road or weather conditions which the sensors do not capture. The actual data we're interested in is derived by sophisticated algorithms from the adjusted raw data and probably aggregated over time or distance driven. Some of this data is ephemeral; others must be retained.

## Measures

Composition runs from raw, unprocessed to highly processed data. **Raw** data is exactly as recorded, either by an electronic sensor / device or as input by a human operator. While we often assume that the data we use daily is raw, the truth is that most data is processed before it is stored. **Adjusted** data has undergone basic processing to take account of known errors or simple conversions. In business data, adjustments involve code conversions, format changes and similar procedures. In scientific / engineering data it might be to remove outliers, interpolate missing measurements or even change measurement systems. Note that adjustment may thus involve data loss or creation; a characteristic that must be recorded and communicated. In simple terms, this is the data cleansing step for most data warehouses. Some other types of adjustment also involve data loss, such as conversion of raw image data to the jpg format.



**Derived** data has undergone more significant processing, usually involving a combination of two or more fields in raw or adjusted data. Examples include calculating profit from cost and selling prices, determining net salary after tax and deductions, and so on. **Reconciled** data is the result of combining across multiple sources and ensuring it is consistent in meaning and timing—a process that often demands significant processing. The EDW is the prime example of this category. In all of the above categories, the data remains congruent with the real world; that is, for every event or item in the real world there exists one or multiple records in the data. **Aggregated** data, on the other hand rolls up multiple identical records to create an overview record that is often more relevant to human perception, for example, profit per month by geographical region. Aggregation is, by definition, a lossy process—information details are lost as data is aggregated.

### Why it matters

Application needs drive required composition characteristics of data. As new applications are defined or old ones evolve, data composition may change. In some cases, prior decisions about data composition limit what we can do with the data in future. Future flexibility of use may be traded off against other factors, such as storage or processing costs. A common decision in business intelligence (BI) over the years has been whether to store detailed data or aggregated data, knowing that aggregation involves a loss of information, but detailed data requires substantially larger storage and more processing. In some cases, multiple composite forms of the raw data may be in use, for example a clickstream, its reconciled form, and an aggregate.

### Example

*FLIP's data warehousing team is very familiar with this classification of data, having used it extensively for internal data. Internal data is adjusted and new data derived before being reconciled across different sources as it is fed into the EDW. Many data marts contain aggregated data. Clickstream data can be adjusted and partially reconciled in a similar manner; the lack of a userid on much of the data limits the amount of reconciliation possible. Social networking data is stored and used in raw format, with text analytics providing some derived and aggregated information.*

## Anatomy

[PROCESSING, STORAGE]

**W**e hear a lot about unstructured vs. structured data today. It's a very misleading debate. First, there is no such thing as "unstructured data"; that would be just noise. Second, it's not a binary choice; data exists in a range of levels of structure. We call that data anatomy. In very basic terms, the anatomy of data is determined by its creators and depends entirely on how they expect to process it in a computer-friendly way. Anatomy can range from very simple to rather complex, from pre-defined to highly flexible and from logical (independent of physical storage structure) to physical.

### Measures

Anatomy ranges from a structure that is simpler and looser to one that is more complex and pre-defined. **Multiplex** data is the simplest and most loosely defined class, consisting of image, video, and audio data, followed by **textual** data, containing all forms of documents, e-mail and so on. Commonly mislabeled as unstructured, these classes do exhibit minimal structuring such as "To", "From", and "Date" fields in e-mail or various parameters embedded in image files. However, they also contain large blocks of data whose content must be discovered by inspection and at the moment of use, rather than defined by field names and types. Such discovery is most easily done by humans; but text, image and other forms of analytics are gradually expanding the scope of computers' ability to parse such data and extract meaning from it on the fly.

**Programmatic** data consists of a variety of simply structured data that, as the name implies, has typically been designed to support preplanned access to and use of the data. Typical structures include key-value pairs; comma separated variable (CSV) and similar structures, as well as stream data, triples and graph databases. A key characteristic of these structures is that they are amenable to easy extension as the need to store additional data fields arises. **Compound** data adds an additional level of complexity by allowing the inclusion of multiplex or textual data within the constructs. XML is the most widely used form of compound data.



**Schematic** data is the most structured and predefined class of data, and is characteristic of general purpose databases. Today, this is mainly in the relational model, but hierarchical and network models have also been successfully used. In this class of data, meaning is isolated from the data values by the prior definition of a schema (a specific form of metadata) that defines data types and relationships between them. Such data is ideal for use by a wide range of more generic applications that can read the metadata and infer data meaning and usage. However, the need for a previously defined schema limits future flexibility for expansion. Slow changes to the schema has its advantages (plan carefully, do it right) and disadvantages (fast moving opportunities missed).

### **Why it matters**

Life would be so much simpler if a conceptually single set of data, let's say a customer information file, was to be used only for a single business task—we could thus structure it in a way that best supported that task. Unfortunately, that customer information file fulfils a wide range of purposes: finding a customer delivery address, updating a customer name, analyzing the geographical customer distribution, combining it with order information to define shipping routes, to name but a few. Each may require a unique optimum structure for most efficient processing, which might imply storing and managing multiple copies of the same data, with obvious storage and management costs.

While we can, of course, create data with any structure we like, much of the data we use is created beyond our control and arrives in a particular form, or is more suited to a certain form because of its predominant usage. In the light of current rapid advances in software, storage and processing technology, understanding data anatomy is vital to deciding how to trade-off processing impacts of having a single copy of the data against storage and management impacts of creating multiple copies.

## **Example**

*Penny's past experience in BI has been focused entirely on the schematic relational data structure. Clickstream data requires new thinking—the volumes are too large and the need for analysis too urgent to load and store such data permanently in the warehouse. The programmatic anatomy of this data in its log files (adjusted to improve clarity and ease of use) lends itself to a more procedural approach to processing it in a Hadoop environment. Social networking data is largely textual today, but the need to handle multiplex data from images and video is growing.*

---

## **Temporality**

[STORAGE]

**L**ike foodstuffs, data has a shelf-life: a period of time when it is useful. Some data is like shellfish, usable beyond a very short period only at a significant risk to health; other data is like preserved food, usable indefinitely. And often the value of data changes over time. Data temporality must be declared by its owner or steward because it is not an inherent property of the data itself. In the past, the temporal nature of data was often understood only by the programs that created and used it. For example, an order entry system may assign different levels of temporality to different records. An order and its status may be recorded only in its current state, and all previous states discarded. A customer record, on the other hand, is likely to be considered to be longer-lived, and past addresses may be maintained as well as the current value. However, to enable other people and programs to safely and appropriately use the data, dates and times must be recorded with the data.

In fact, almost everything changes—perhaps very slowly—with time, so measurements taken at different times give different results. Data designers thus must decide whether to keep an historical record of past values or to simply overwrite and replace old records. *Historicized* data contains time-stamps, usually indicating the beginning and end date (and time) of the validity of the record and can be manipulated as a time series. *Overwritten* data contains only one record for each real world event or item, removing the previous state, as opposed to one record for each event. Historicized data always grows in volume over time and requires more storage than overwritten data. Typically, temporality is recorded on a record by record basis, although data of the longer temporality can be gathered together and that class assigned to the data set as a whole.

## Measures

Temporality ranges from shorter to ever longer validity of the data involved. *In-flight* refers to data that was just now created and is on the network being processed “as it passes by”; its period of validity is limited to the present moment and no more. Complex event processing (CEP) uses in-flight data as its primary data source.

When in-flight data is persisted it becomes *live*. Most traditional computer applications create and manage such data, which is stored on a persistent medium (often in a database on a disk today) but can be overwritten by the managing application at any time. This behavior means you may not be able to retrieve the original facts at some future time, which makes the data temporally less reliable. This is the characteristic of *stable* data, consisting of time-stamped records, with periods of validity typically defined by start- and end-dates and times. Stability is a characteristic of most data warehouse data and many enterprise content stores. Live and stable data may be either *valid* (the current date/time lies between the start and end timestamps) or *invalid* (the current date/time lies before or after the period of validity). Any historical or temporal processing of data is heavily dependent on such (or similar) timestamps, versions of which have been incorporated in many relational databases.

*Historical* and *archived* data are simply special classes of stable data which are declared to be a never-changing, permanent record of some aspect of the business, and thus always valid. They differ only in that historical data is used regularly, while archived data is seldom used and is usually stored in a safe, protected environment for some unforeseen eventuality. Much of the data in the EDW and data marts is historical.

Note also that some data has no indication of its temporality; any use of such data where timeframe is important requires extreme caution.

## Why it matters

While the fundamental meaning of temporality is universally incorporated into applications, its importance increased with the rise of business intelligence. Decision making requires the ability to track changes in data values over time and to be able to compare values from the same time period. Both of these needs drove the understanding and implementation of ways of showing the temporal nature of data within the data warehousing environment. These include timestamps on data records and storing information on temporality in metadata. Operational BI extends the demand to understand and take account of temporality as data is used ever closer to real time.

As the importance and volumes of data from non-traditional sources grows, temporality also becomes a major concern in these areas. As text, voice and other unstructured information is increasingly combined with traditional structured information, users need to understand clearly what time periods the entire information set describes, if they are congruent and how they can be compared. This is always an area of particular complexity when populating data warehouses; the diversity of sources for non-traditional data up the ante considerably.



## Example

*Both clickstream and social networking data fall into the live category in terms of temporality. As a consequence, Penny must ensure that they are captured and stored in stable form regularly, so that they can be used with stable and historical data in the BI environment. Given the volumes and use of both raw clickstream and social networking data, neither is a likely candidate for historical storage.*

## Temperature

[STORAGE, USAGE]

**T**emperature indicates the level of demand for and use of data by business users and applications, and is a characteristic that has been of interest for some considerable time, typically being used to determine with data is best kept in memory (in addition to disk) as a way of reducing expensive disk I/O. In business terms, temperature is defined by data popularity—the number of people issuing the most requests for a specific piece of data—and can be directly tied to service level agreements or business user expectations of system performance.

Data temperature typically declines, like most things, with age. Old data is generally less frequently used than newer data. However, the relationship between temperature and age is more complicated than that simple statement suggests. The temperature pattern also depends on the use being made of the data. Operational uses of data have a temperature peak much earlier and more briefly in time than informational uses. Certain time periods, such as quarter- or year-end, show predictable increases in data temperature for certain classes and subsets of data as financial reports are being prepared. Other temperature peaks are less predictable, being driven by media hype, world events or fashion trends. Consider a CFO challenged with a fraud investigation where numerous analysts and accountants must research data going back five years. For a few weeks, a dataset of information going back five years is dramatically raised in temperature, then cools off once the investigation completes.

Some systems automatically and intelligently manage data placement based on temperature, placing the most frequently used or *hot* data on the fastest storage units and the least used or *cold* data on the slower storage units. This supports high-performance access to hot data for real time decision making, and an automated lifecycle management process as data ages to migrate it to less expensive drives.

### Measures

Temperature is a measure broadly running from hot to cold. The temperature of data starts off unknown, until sufficient information has been gathered by the system about its usage pattern. Most new business data exhibits a high temperature during the early stages of its life. **Hot** data is in constant use and is usually placed in the fastest, most resilient storage possible as well as in memory. **Warm** data is distinguished by frequent usage. Fast storage is also preferred here, either SSD or fast hard disk drives. **Cold** data is accessed less regularly; it typically resides on slower and less expensive HDD. **Arctic** data is seldom accessed and maintained largely for regulatory and audit purposes. It is usually stored offline or archived (see temporality) and must be brought online for use.



### Why it matters

Understanding data temperature enables data to be stored on the medium that provides the optimal balance between cost of storage and access speed, given the added cost of higher performing storage. Today, in the light of rapidly increasing data volumes, demands for faster access and new storage options, data temperature is of growing interest. Data temperature affects the choice of which platform or device is best suited to hold the data. For example, arctic data might be moved to a specialized system for archival or large quantities of cold data might move to an experimental platform.

The development of most interest is the rapid reduction in the cost and improvement in technology of solid state memory. Solid state drives (SSD) have attracted most attention recently, with drives now available sporting 100s of GB of memory and access speeds some 20 times faster than hard disk drives (HDD) but is much more expensive. Designed with identical interfaces and form factors to those of hard drives, they are easily incorporated into existing hardware. SSDs have provided an entirely new price/performance niche on the tape / removable HDD / slow HDD / fast HDD hierarchy. The next development, already unfolding, is the rapid growth in main memory sizes which are increasing the opportunity to create ever-larger memory caching and driving research into entirely novel database architectures.

### Example

*Raw/adjusted clickstream and social networking data are warm for a short period (a few days) after capture and are then discarded. Summaries derived from the original data are also warm for a short time, but cool rapidly to cold and eventually arctic after that. During some phases of campaign management, cold summaries may become warm again as users review results and plan new campaigns.*

## Access

[USAGE]

The level of access granted to data is constrained by a number of factors such as relevance (is this data required to do your job?), security (are there risks to the business in accessing the data?) and privacy (are there risks to individuals in accessing the data?). While these factors, and

others, need to be evaluated independently in relation to specific actions in these areas, from the point of view of the data itself, all of these factors lead to a cumulative access rating of the data that results in some limit to what data can be accessed by whom.

To further complicate matters, data access may differ for an individual record and for a collection of records, where a collection may be part of one data set or span multiple data sets. Furthermore, when data is aggregated, as described under *composition*, different access conditions may apply to the aggregate than that found for the individual records. For example, access to a single record from the business' customer file may be harmless; access to the entire file or a summary is a security exposure. Combine that with access by the wrong people to open orders or complaints, and you have a potential catastrophe in waiting. Access is thus combinatorial as well as singular in nature. And if this weren't enough, some aspects of access—privacy in particular—are subject to different regulations in different countries, leading to different ratings depending on where the data is stored or used. And in many countries, data movement across country boundaries is highly regulated.

### Measures

Data access is a single slider running from totally private to fully public. As with many of these traits, data may have unknown access characteristics, especially when obtained externally; the aim, of course, is to classify it as soon as possible. **Private** data is accessible only to its owner—the person(s) or application(s) that created it—and must be controlled by the most stringent access rules. **Confidential** data extends access by the owner to an identified and controlled set of people or applications for a limited period of time. **Restricted** data is limited to those people or applications required to get a particular task or process performed as a common part of running the business. Most data used by the business falls into this class. **Limited-use** data covers all remaining data that can be used within the enterprise by contracted business partners or relevant regulatory bodies without any restrictions. **Public** data, as the name implies, can be made available to anybody.



### Why it matters

Defining the extent of access to data granted to users and applications has long been considered vital for reasons of security and need-to-know in all businesses. Privacy has become a significant factor more recently and continues to grow in importance. The increasing number of non-traditional and external data sources poses considerable challenges in determining access rules that ensure safe and legal use of such data. In addition, as data is shared and combined in ever more numerous ways, the possible security and privacy exposures become more widespread and complex. A simple way of defining and describing access limits for data records and collections is therefore vital and will drive decisions about where data can be stored, distributed and processed.

### Example

*Based on privacy concerns, Penny has decided to declare both click-stream and raw social networking data private. The adjusted data has all personally identifiable data removed or encrypted, and is thus available for restricted use. Summary data is limited use to preserve campaign management business advantage.*

## Trust

[USAGE]

The level of trust that business users can place in the data available to them depends on a wide variety of factors, not least of which is the reputation of the IT department that delivers it. Related to the data itself, these factors include judgments about its accuracy and completeness, objectivity and validity, its suitability for the intended use and, importantly, how well it is defined. An enterprise-wide data quality project is the foundation for defining an initial trust level for data. However, trust is a highly subjective characteristic for business users that can easily change over time and by department, so ongoing monitoring and recording by collaborative tools, for example, is required.

### Measures

Trust ranges from data that should be avoided in most circumstances to data that is highly regulated. Prior to some formal evaluation through a data quality review, the trust level of a set of data is unknown. Poorly defined or described data from unknown or untrusted external sources is classified as

**high risk**, and its use should be limited to highly-bounded, low impact decisions. **Experimental** data may combine data from trusted and untrusted sources or may be trusted data used for a variety of experimental purposes in controlled conditions. For example, data coming direct from the data warehouse, a highly trusted source, becomes experimental data when extended or analyzed in novel ways. Experimental data gives analytic results that have a low confidence factor. Data that carries a **health warning** (in the associated metadata) is poorly defined or has timing or consistency concerns but comes from largely trusted sources; it can be used with care by qualified business users.

There are two classes of more trusted data. **Trustworthy** signifies the majority of information used for standard business intelligence purposes, being well defined and coming from known and well-defined sources through metadata-driven ETL or Virtualization technology. To be trusted, there must be a data cleansing process coupled with full sourcing lineage of each field and business user validation. **Certified** data has similar characteristics as trustworthy data but is, in addition, sourced from highly controlled data sets, such as the EDW or master data management (MDM) stores.

### Why it matters

Data quality has been a long-standing concern in most well-managed businesses; it is the basis for understanding how trustworthy data is, so that it can be used reliably in decision making and action taking at all levels of business. Until recently, most data used by businesses was generated internally or by business partners. Even under these limited circumstances, ensuring and maintaining data quality has often proven a challenge.

However, there is an increasing dependence on data sourced from outside the enterprise and its partners, both in direct use and incorporated with internally-sourced data into other data sets. This has led to a much greater level of uncertainty about how much trust can be placed in the data made available to business users today, and that situation is only going to become more challenging as further data sources are added.

### Example

*Based on its highly unverifiable sourcing, social networking data is clearly high risk. When combined with trustworthy or certified data from the EDW or campaign management data marts, the resulting data is deemed suitable only for experimental use. Penny has further agreed with the business that clickstream data should carry a health warning about certain aspects of its content, limiting its use to marketing.*

## Using the Data Equalizer

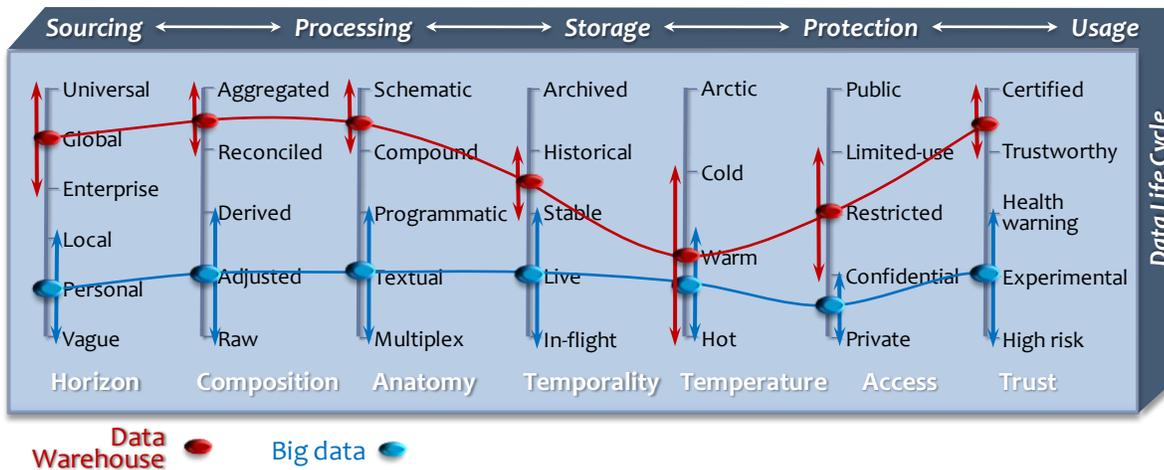
Having seen the individual characteristics and their use of in data management, let's briefly look at how the entire set of seven characteristics work together to characterize data and support decision making on storage and distribution of data. In each case, we examine the alternatives and show the mean and range of approximate slider positions for each generic workload. When using this approach for your specific data set, the range of slider positions will be much narrower and, in many cases, narrow down to a single value.

### Data warehouse or "big data"?

The choice between using the data warehouse infrastructure or a largely file-based, parallel-processing environment loosely termed "big data" (such as Hadoop, for example) for emerging data types has recently become somewhat contentious. It is usually presented solely in terms of the volume of data involved, often with the suggestion that big data will make traditional data warehousing redundant. The data equalizer allows us to take a more nuanced view and see that both approaches have specific strengths.

Figure 3 shows that for data in the EDW, sliders related to sourcing and processing on the left and usage on the far right of the equalizer are near the top of the scale, reflecting the highly managed and structured needs of such data and the significant level of trust users place in it. The trough in the middle shows the ongoing use and need for temporality of such data. Data suitable for use in big data platforms, on the other hand, shows a much more flexible and loose profile in terms of sourcing, processing and storage. This leads to limitations in access and a lower level of trust by its users.

Figure 3: EDW and “big data” on the Equalizer



The distinctly different data profiles for the two cases make it clear that there is an ongoing role for both data warehousing and big data approaches. If you look at your specific data sets, there is likely to be a distinct divergence between the characteristics of the data sets that indicates suitability for data warehousing or big data. You may further observe that some types of data and processing that have traditionally been performed in the data warehouse may be more appropriately undertaken in the less regulated big data environment. Such data may have been routed through the data warehouse even though it had a more local horizon and was used at an experimental level of trust, simply because there was no other more suitable environment. The emergence of big data approaches may provide a more cost-effective solution for such data.

### Core or optional data?

We can take this thought further into the concept of **Core Business Information**, which can be thought of as that information which defines the heart of the business and which, if lost or erroneous, would likely cause an irretrievable breakdown of the business—either through an inability to perform daily operations or to manage and track the business internally or in the context of regulatory reporting. The data equalizer in figure 4 shows that a subset of EDW data, focusing on that which is reconciled and certified, forms a key part of the core. However, in addition, we can see that the ranges of temporality and access are wider than in the case of the EDW, indicating a need for certified real-time data found, for example, in a master data management (MDM) implementation.

Optional data, in contrast, has strikingly different characteristics. Like big data above, the sliders on the far left are low, indicating more local and specific interest. On the right, however, the sliders rise, due in part to the wider range of storage and usage options. Of more interest, perhaps, is the high position of the access slider, which shows how wide a use of such data is permitted. Optional data is thus suitable for distribution over multiple and various platforms, even personal storage, while core

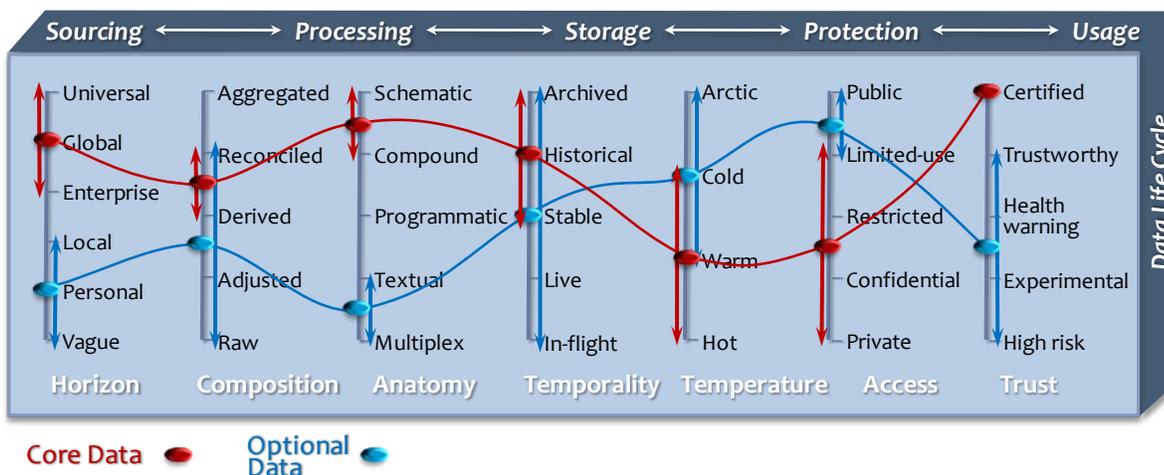


Figure 4: Core and optional data on the Equalizer

business information is likely to be highly centralized and managed. These considerations clearly apply to BI, but the range indicator on the temporality slider shows that we can also apply this thinking to operational data, especially to transactional apps on mobile devices as we discuss next.

### Cloud, personal or internal IT-managed storage?

In the previous two examples, we used the equalizer to distinguish between some generic categories of business data. In this last example, we'll use the tool to decide where a specific data set should reside—in the traditional IT-managed environment, in the cloud or distributed on personal devices.

We are reaching a significant three-way tipping point in where we store data and how we manage it. The options have been around for some time, but the circumstances haven't been right to force broad impact decisions—until now. For a variety of reasons—technological, organizational, and more—IT has traditionally stored and managed all data of value to the business on centralized servers. Over past decades, the volume and value of data stored on personal devices such as PCs and laptops has grown substantially. Today, we see an explosion in these volumes as smart phones and tablets redefine personal computing. In addition, in recent years we have seen rapid growth in cloud computing—on-demand storage and processing of data in a utility model. The infrastructure for this model is maturing, and it is becoming increasingly attractive for reasons of limited capital investment and growing data volumes.

The question posed by these developments is: given a specific set of data, how can we make a clear and conscious choice among the three options? Figure 5 shows the data profile (in red) for a data mart used to track sales of country-specific variants of a range of cell phones across Europe. Obviously, internal IT-managed storage can be configured with any required characteristics—its range of possibilities (not shown) spans the entire equalizer and can meet the data needs. The green bars representing personal storage characteristics hardly overlap the data traits at all—basing this application on such a platform is shown as not realistic, as might be expected.

The more interesting aspect is to look at the cloud environment characteristics shown in blue. While, in theory, the cloud offers as wide a range of characteristics as internal IT-managed storage, practical limitations exist in most implementations, as we can see in figure 5. On the left of the equalizer, focusing on sourcing- through to storage-driven considerations, the cloud offers a good match to the traits of this particular data mart. However, as we move towards the right, protection and usage considerations show a much poorer correspondence. In this case, the equalizer analogy clearly illustrates that temperature, access, and trust requirements aren't met by some public cloud offerings.

Clearly, the choice of cloud provider narrows dramatically or the business can decide to sacrifice these traits if the application allows it.

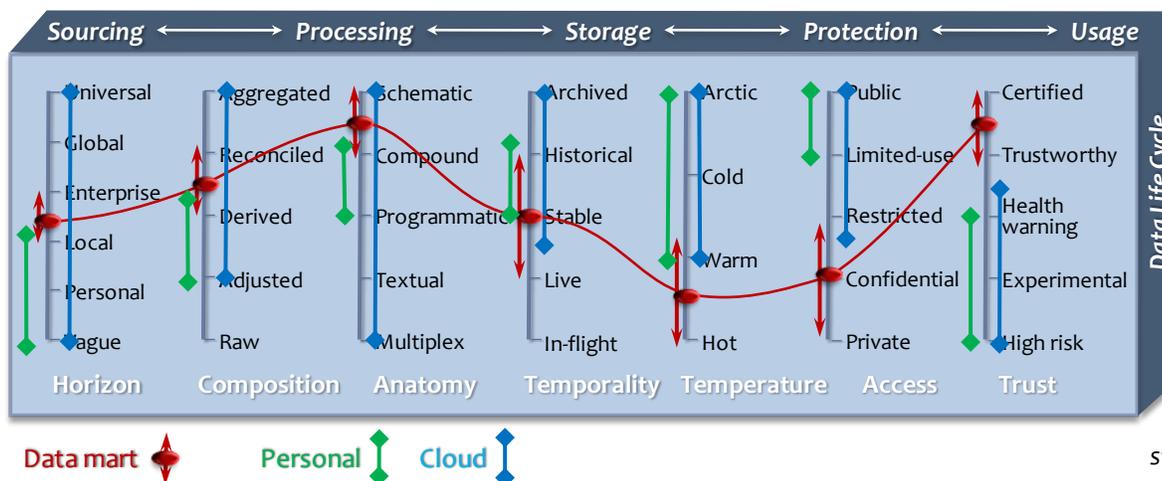


Figure 5: Evaluating data storage needs on the Equalizer

---

## Conclusions

This paper reexamines the fundamental characteristics of data as found in computing today and proposes seven important characteristics that must be considered when making decisions about how it should be sourced, processed, stored, protected, and used. These fundamental characteristics are: (1) horizon, (2) composition, (3) anatomy, (4) temporality, (5) temperature, (6) access and (7) trust.

Individually, each of these seven data traits provides IT professionals with specific insights into different aspects of data management, such as the most appropriate technology to store the data, how it should be maintained and how it must be protected. Taken together, these characteristics can be plotted on the data equalizer to gain an instant overview of the overall tone and character of a data set, enabling early judgments about likely storage approaches, such as whether it should reside in the data warehouse or in a distributed file store, whether it can be centralized or distributed, if it can safely reside on personal devices or in the cloud, to name but a few examples.

Understanding the fundamental characteristics of data today is becoming an essential first step in defining a data architecture and building an appropriate data store. The emerging architecture for data is almost certainly heterogeneous and distributed. There is simply too large a volume and too wide a variety to insist that it all must be copied into a single format or store.

The long-standing default decision—a relational database—may not always be appropriate for every application or decision-support need in the face of these surging data volumes and growing variety of data sources. The challenge for the evolving data warehouse will be to ensure that we retain a core set of information to ensure homogeneous and integrated business usage. For this core business information, the relational model will remain central and likely mandatory; it is the only approach that has the theoretical and practical schema needed to link such core data to other stores.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing. He is a widely respected consultant, lecturer and author of the seminal book, “Data Warehouse—from Architecture to Implementation.” Barry’s current interest extends to a fully-integrated business, covering informational, operational and collaborative environments to offer an holistic experience of the business through IT. He is founder and principal of 9sight Consulting, specializing in the human, organizational, and IT implications, and design of deep business insight solutions.



### About Teradata Corporation

Teradata is the world's largest company solely focused on data warehousing and integrated marketing management through database software, enterprise data warehousing, data warehouse appliances, and analytics. Teradata provides the best database for analytics with the architectural flexibility to address any technology and business need for companies of all sizes. Supported by active technology for unmatched performance and scalability, Teradata’s experienced professionals and analytic solutions empower leaders and innovators to create visibility, cutting through the complexities of business to make smarter, faster decisions.

Simply put, Teradata solutions give companies the agility to outperform and outmaneuver for the competitive edge. Visit [www.teradata.com](http://www.teradata.com)

Teradata Corporation  
10000 Innovation Drive  
Miamisburg, OH 45342

Teradata and the Teradata logo are trademarks or registered trademarks of Teradata Corporation and/or its affiliates in the U.S. or worldwide. Brand and product names mentioned in this paper may be the trademarks or registered trademarks of their respective owners.

<sup>1</sup> Based on IDC “Expanding Digital Universe” 2007-2011 sponsored by EMC, [http://bit.ly/IDC\\_Digital\\_Universe](http://bit.ly/IDC_Digital_Universe) (Some figures are estimates and extrapolations from the published work and are believed to be within the correct order of magnitude.)

<sup>2</sup> The terms *unstructured* and *structured* are widely used to categorize data. As we shall see in this paper, these terms are both misleading and inadequate, but are used in this section in the broadly accepted meaning.

<sup>3</sup> See, for example, “‘Big Data’ Is Only the Beginning of Extreme Information Management”, Gartner Inc, April 2011

<sup>4</sup> Abstracted from <http://dictionary.reference.com>

<sup>5</sup> European Spreadsheet Risks Interest Group, <http://www.eusprig.org>