CortexDB Reinvents the Database*

JUNE 2019

ThoughtPoint 1 of a 5-part Series by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

CortexDB is the first example of a new class of information management tool that allows data and context to be managed with both independence and integration, a feature that is central to reuse of information for different business purposes.

Over more than thirty years as a thought leader in information management, my observation is that most advances are incremental and built on existing ideas. I have encountered truly novel thinking only rarely. I have thus been delighted to meet a small German software company that has incorporated some innovative and potentially revolutionary thinking in its product.

Cortex AG designates the core or its product, CortexDB, as a database, and indeed as nothing less than a "Multi-Model NoSQL DBMS". I believe they undersell their work. This focus on data and database models, and the internal structure of the database distracts from the fundamental difference between what CortexDB does and what all previous databases do.

I believe that CortexDB may be the first of a new type of product: an adaptive information context management system (ICMS)

To understand why and what this means, we need to look at the history of databases and go back to an old debate: the difference between data and information—the fundamentals of which I described in "*Business unIntelligence*"—as a prelude to redefining what CortexDB is and can do.

---

* *A German version of this material is available at Informatik Aktuell.*

In the mid-1960s, researchers began defining *data bases* (yes, two words) as sets of data that could be used by more than one application, instead of having a specific dataset for every instance. When data is owned and used by only one application, it can be stored simply and efficiently as a sequential string of values (like a CSV file without a header). If you want to share data, you need at least to name the individual fields. Ideally, you also need some idea of the logical relationships between individual fields. Data stores thus became databases, as these "naked" data values were named, described and dressed up in hierarchical and later, relational structures.

These names, descriptions and relationships are ***context-setting information (CSI)***, also called metadata. CSI forms the bridge between data and information. I use the following definitions:

Information: the recorded and stored symbols and signs we use to describe the world and our thoughts about it, and to communicate with each other. Information consists of data and CSI. It's information that we really need and use as people and businesses.

Data: "facts"—measurements, statistics, the output of physical sensors, etc.—in the form of numeric or simple textual values. I put *"facts"* in quotes to signify that facts are seldom as hard or fixed as we assume; they are deeply dependent on the frame of reference of those who create and/or store them. As a result, CSI—as a representation of that frame of reference—is also the responsibility of these data creators.

Data base vs. Information base

The discussion above suggests that every database lies somewhere on the spectrum from data to information management. A key-value store is much closer to data. A relational database with a fully populated set of description tables in much closer to information. And because information is what we as people need, relational databases have come to dominate computing.

However, there is an additional consideration. The context—and by extension, meaning—of any piece of information in a database is precisely and only that intended by its ***creator(s)***, including engineers who understand the data and business experts who know the intended usage (CSI) of that data. Using the resulting information for another purpose may lead to errors if the meaning assumed by the second application aligns poorly with that defined for the first.

This is the exact challenge found when we take information from the operational environment to the world of BI, even though both are built on relational databases. The basic problem is that a relational database can only represent one context at a time: that of its creator. As a result, in the world of BI, we must build and populate a second database to hold the context needed for analytical tasks. In fact, we find we must build multiple databases for many different analytical contexts.

A second—and more pressing problem for modern digital business—is that of onboarding externally sourced data. Here, information from the external environment is often largely decontextualized (for example, in a CSV format file) as it is passed over the Internet to the receiving enterprise. Data wrangling is, in effect, the rebuilding of context needed to interpret and use the incoming data.

The underlying problem is the same in both cases: We have failed to recognise that there are multiple usage contexts for the same base data set. The database creator embeds the CSI for their original usage scenario in the database design. Subsequent usage scenarios pose varying levels of mismatch to that original design. As usage scenarios become ever more varied and complex, we need something beyond a traditional database to manage the mix and match of uses.

An Information Context Management System

This is where CortexDB comes in. It provides a system that almost completely separates the management of "naked data"[†] from management of CSI, and further ensures that the two areas are completely synchronised in the event of changes to one or the other. I call this an adaptive Information Context Management System (ICMS).

CortexDB stores the naked data in a document store where the only CSI is the names of the fields and record IDs for the documents. (These are, of course, the minimum requirement for joining to the extended CSI.) Meanwhile, the CSI resides in a 6th normal form (6NF) relational structure. I'll describe the structure in deeper detail in my next article.

In the case of differing application needs that demand conflicting ways of envisaging the data, multiple sets of CSI that structure data access and use in incompatible ways can easily be created and maintained in parallel. Of course, where applications use information the same way, they all use the same set of CSI.

By providing a single store of naked data, supported by multiple sets of CSI, CortexDB can support both operational/transactional applications and informational/analytical applications on the same data (provided, of course, that the naked data is stored at the appropriate atomic level of detail).

In the case of onboarding externally sourced information, the initial data store can be designed and created with only minimal knowledge of the detail of how the incoming data is internally interrelated or how it relates to existing internal data. Those elements of context and structure can be added later and incrementally as the understanding and use on this onboarded data evolves.

This latter case is an important example of how the context / meaning asserted by the data creator (e.g. in the Internet of Things) may be in large part unknown to the eventual users of the data (in this case, within the receiving enterprise). Here, the ICMS provides the means to define and explore multiple ways of interpreting and using poorly described information. As digital businesses create and exchange ever more data, this issue is set to become ever more common.

[†] I've chosen to use "naked data" rather than "raw data" because the latter phrase has other uses, especially in onboarding external data. Naked data is information that is "stripped" of as much CSI as possible, consisting mostly of "pure" value data.

These two important areas of information processing—and others—are united in requiring far more and far deeper agility in the creation, structuring and use of data than traditional data processing. As the programming world has adopted the precepts of agile development, now too must data management.

In the second ThoughtPoint of this series, I dive into some more detail on how CortexDB is structured internally and show that it supports data agility.

Links to all ThoughtPoints in the series originally published on LinkedIn:

ThoughtPoint 1: CortexDB Reinvents the Database – June 2019
ThoughtPoint 2: Making Data Agile for Digital Business – June 2019
ThoughtPoint 3: Managing Data on Behalf of Different Actors – June 2019
ThoughtPoint 4: Distilling Deeper Truths from Dirty Data – July 2019
ThoughtPoint 5: CortexDB Drives Agile Digital Transformation – July 2019

*Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI Upside, and more, Barry is based in Bristol, UK, and operates worldwide.

Making Data Agile for Digital Business*

JUNE 2019

ThoughtPoint 2 of a 5-part Series by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

CoxtexDB combines SQL and NoSQL technology to offer business the freedom of highly flexible operations together with the structure needed to manage its fundamentals. In this second of a series of ThoughtPoints, we see how this is the foundation of true data agility and digital business transformation.

In the first ThoughtPoint of this series, "CortexDB Reinvents the Database", I suggested that CortexDB was rather more than a database. Rather, I see it as a first example of an adaptive Information Context Management System (ICMS).

From a technology viewpoint, CortexDB uses a combination of SQL and NoSQL to support both the structure and flexibility, the planning and agility that characterise an ICMS. In doing so, it offers business the freedom of completely flexible operations together with the ability to manage its fundamentals.

Like the brain, from which the product derives its name, CortexDB is conceptually simple. The semantic structure of all types of data is preserved in a document store, which is automatically described and modelled in an integrated, relational sixth normal form (6NF), multi-dimensional index that points to every piece of data in every document in the store.

In contrast to a traditional relational database, a document store offers an infinite range of structuring. A classical, simple text document consists of characters arranged in words and sentences. A video or audio file is a binary large object with some metadata fields. An

---

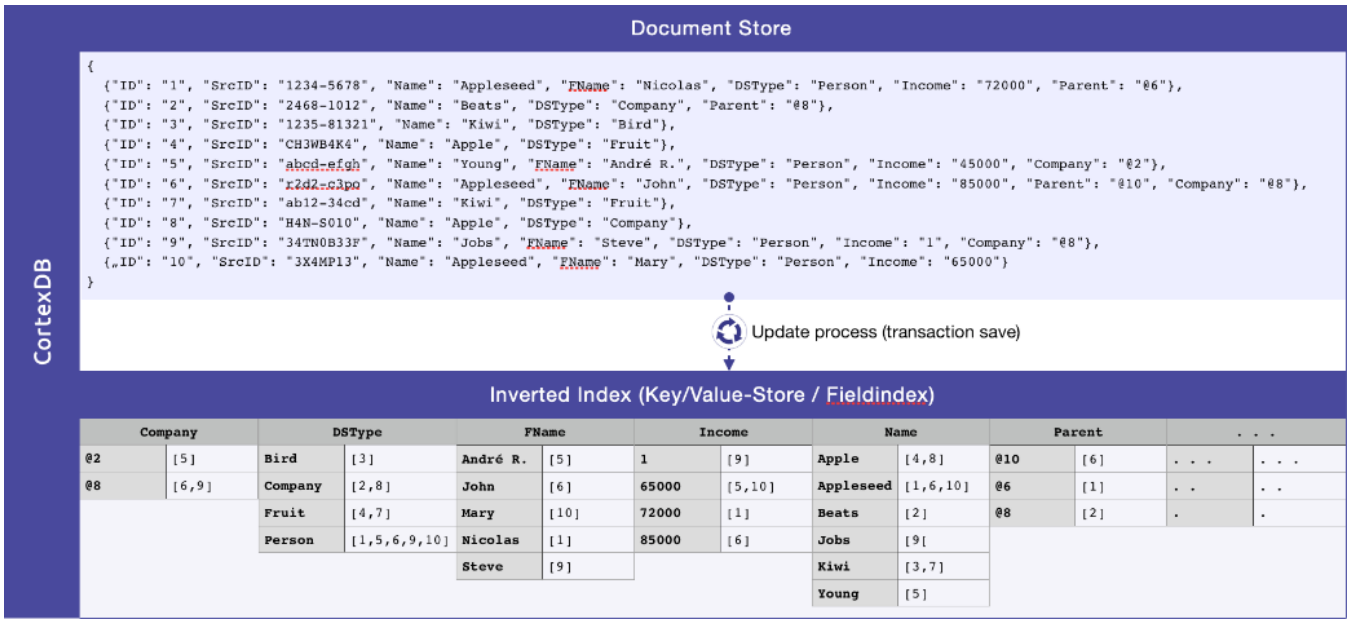
* *A German version of this material is available at Informatik Aktuell.*

e-mail document consists of a few defined and named fields and their values (key/value pairs), such as Sender and Subject, as well as the body of the e-mail, which is simple text. A field-based document, such as XML or JSON, consists of an arbitrary set of key/value pairs such as FirstName: "John", DoB: "19451123", and so on. A document store can contain any of these data types and more.

Obviously, the level of structuring in a document store determines the processing possible. In a simple text document store, we can find and count all occurrences of a word such as "Macintosh". However, it is only when the document contains defined and named fields that we can distinguish between "Macintosh" as a model of computer, a type of coat, or Scottish last name, and find and count them separately.

Finding and counting (and, indeed, all types of more complex data processing) lead to the need for indexing to speed up access. A simple text document store is processed via an inverted index, which is simply an ordered list of all non-trivial words in the store together with pointers to the documents where they appear and their positions within those documents. In a document store that contains more structured documents, the contents of every individual field can be indexed in a similar manner.

CortexDB provides an inverted index of every value of every field in the data store that allows every instance of "Macintosh" as a last name, for example, to be easily found in the document store. As a result, these indexes can be used to process—count, average, max and min, etc.—data in each field. More interestingly, the complete set of indexes for a document store represents the most discrete level of model (6NF) possible for the fields and contents of the document store. Simple set operations, such as union, intersection and complement, can thus be easily performed. At 6NF level, such set operations map directly to the more familiar language of "select" and "join" in SQL terminology, allowing all the Macintoshes (of any type) in London to be located. Furthermore, this is the correct level of normalisation at which to apply temporal concepts, graph structures, and other fundamental data management approaches. Further details can be found in Thomas Kalippke's article on Medium.



The consequence of having such a multi-dimensional, 6NF set of indexes may be surprising to data designers who are trained in traditional modelling methods. Because 6NF is the most fundamental and simple representation of data, all other normalization levels can be subsequently built from it. However, in a traditional relational database, the use of 6NF is complex and cumbersome due to the number of joins of very narrow tables required and performs poorly.

CortexDB's use of 6NF inverted indexes overcomes these issues and supports the full power of the 6NF approach. At design and initial load time of the data store, modelling is reduced to the identification, naming and typing of the individual fields. Any other more complex modelling needed—such as defining dimensional or 3NF structures, creating temporal databases, etc.—can be addressed at a subsequent design stage.

This separation of design-time concerns from run-time concerns is what makes CortexDB a context-aware system and elevates its view from data to information, an adaptive Information Context Management System.

This is also the meaning of true *data agility*. Any required business use of data can be defined and implemented after the data has been stored, using the initial data set. Given the performance characteristics of the 6NF structure, it is seldom necessary to create further copies of the data in different formats.

For business insight applications, such as business intelligence and analytics, that are at the heart of digital transformation, eliminating—or even significantly reducing—the need for multiple copies of the same data in different structures drives new levels of agility and speed of delivery in IT. Similar considerations apply to the new operational applications required by digital business.

This novel approach to data structure enables new ways of thinking about data, allowing a focus on business needs rather than database constraints, and improving IT productivity as well as time to market for new applications. Such data agility is at the core of digital business.

From Data Agility to Digital Business

Digital business has become the buzzword for what business needs to do as this decade comes to a close. Despite its popularity, its definition is often unclear. In its technical essence, it centres around a change in sourcing of business data, from mainly internally from business processes to largely externally from the real world. Allow me to explain…

Way back in the last millennium, the data needed by business to operate was well-structured, (reasonably) well-governed, and almost entirely sourced from within the walls of the business itself. This is *process-mediated data*: the legally binding foundation of business, as described in my book *Business unIntelligence*. Relational databases — highly structured, carefully managed, but resistant to change — ruled the data world. For decision making, data warehousing was king.

However, within the last decade, the data landscape has changed completely. Complexity and data volumes are increasing not only as a result of the automation of processes and the requirement for more agility. Social media, click streams, and the Internet of Things have brought huge volumes of rough and raw data: externally sourced, poorly structured,

loosely governed, and varying in structure over time. The focus of today's digital businesses has turned to insights from this *human-sourced information* and *machine-generated data*.

Relational databases fell out of favour, to be replaced by NoSQL data stores of varying forms. Data flooded into lakes, threatening to wash away the dusty, old warehouses. Except it didn't. Because managing the legal foundation of business, using process-mediated data, is still mandatory.

These conflicting requirements, together with the above explosion of data volumes, demand that IT must now become as agile in data design and management as it has in development. Data created in one environment must be readily used and reused in multiple contexts. This leads us back to the Information Context Management System and its implementation in CortexDB.

Subsequent ThoughtPoints in this series will look at some of the applications of this technology and how CortexDB has been used in support of digital business and in all its various forms.

Links to all ThoughtPoints in the series originally published on LinkedIn:

ThoughtPoint 1: CortexDB Reinvents the Database – June 2019
ThoughtPoint 2: Making Data Agile for Digital Business – June 2019
ThoughtPoint 3: Managing Data on Behalf of Different Actors – June 2019
ThoughtPoint 4: Distilling Deeper Truths from Dirty Data – July 2019
ThoughtPoint 5: CortexDB Drives Agile Digital Transformation – July 2019

*Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI Upside, and more, Barry is based in Bristol, UK, and operates worldwide.

Managing Data on Behalf of Different Actors*

JUNE 2019

ThoughtPoint 3 of a 5-part Series by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

In a digital world, every piece of data serves multiple purposes for different people and organisations. In the third ThoughtPoint of this series, we see how CoxtexDB addresses the new and difficult issues that arise when the same data has multiple owners.

In the first ThoughtPoint of this series, "CortexDB Reinvents the Database", I mentioned one of the key initial drivers for the development of databases: the need to support multiple applications from a single data set. At that time—the Flower Power era of the 1960s—the multiple applications under consideration and, indeed, all the data they used belonged to the specific enterprises that built them. Data seldom if ever crossed enterprise boundaries. Data ownership, if considered at all, related to business departments within the enterprise.

Today, as digital transformation proceeds apace, these simplifying constraints are falling away. As data crosses and re-crosses enterprise boundaries, is gathered and reconstituted in shared-use systems, questions of data ownership and usage authorisation between different actors beyond the enterprise become more complex. Furthermore, with the enforcement of the European Union's General Data Protection Regulation (GDPR), the privacy of personally identifiable information (PII) must be protected in all circumstances and specific rules determine how access to and use of such data must be controlled.

These issues pose specific challenges to the traditional approaches to storing and managing data in old-style database systems. And, just as a database solved the problem of

* *A German version of this material is available at Informatik Aktuell.*

sharing data between applications, an adaptive Information Context Management System (ICMS) such as CortexDB offers the solution to managing data and sharing information between different actors. The IT industry is only starting to understand and address the complexity that arises, but Germany's CarPass system provides an excellent example of what needs to be and, indeed, can be done.

CarPass—a Digital Twin of your Vehicle

CarPass is a system currently rolling out in Germany and eventually across the EU, the goal of which is to create and store a complete digital record of any vehicle's service history that is the property of the (current) owner but can be updated by repair workshops, dealers, and insurance companies, and eventually passed on the new car owner when the vehicle is sold. While the longer-term goal is to include all service and component information about the vehicle—thus creating the vehicle's digital twin—the initial data being stored is the original information from the vendor, service history, mileage, number of owners and each changed/repaired part in the car since production. One initial business focus is to address mileage fraud that currently costs billions of Euro across the continent. In the longer term, the aim is to provide the current car owner with the right to view a complete history of the vehicle without compromising the privacy of previous owners.

At first glance, CarPass seems like a simple application: create a relational database and give car owners and relevant organisations appropriate access to it. The reality is far from simple.

Although the current data is restricted, the future scope envisages a wide range of data types, many of which will vary from vehicle to vehicle. The data is thus semi-structured, in a variety of types, and variable over time. Such characteristics suggest a document store rather than a relational database as the base technology to avoid problems such as schema change and null value proliferation. However, a traditional document store cannot address data privacy issues when all vehicle data associated with a known owner/vehicle combination resides in a single document/record. Some or all of this data has PII characteristics under the GDPR, as has been stated by federal data protection authority for Lower Saxony in Germany. In addition, authorisation to access or update different parts of the vehicle record must be assigned to and revoked from different parties over time. Even if this were technically feasible in a traditional document store, the administrative overhead of managing access in accordance with the GDPR would be enormous.

As an Information Context Management System (ICMS), CortexDB offers solutions to all these problems and delivers a highly performant system on which to deliver the CarPass application. As described in the second ThoughtPoint of this series, "Making Data Agile for Digital Business", all the data fields in the CortexDB document store are indexed in a 6NF structure that is created and maintained in parallel with the document store. Authorisation and access to the naked data records in the document store are managed through this index, allowing access to individual fields within documents/records to be granted and revoked. The vehicle owner, as current owner of the CarPass record for his/her vehicle, can, for example, grant write access for a new mileage field in the record to a workshop that performs the regular service on the vehicle, revoke it and grant it to another workshop as s/he sees fit. And when the vehicle is sold on to a new owner, the access rights to the CarPass record can be easily passed to the new owner.

It's not just CarPass alone…

CarPass is but one example of a new class of information management applications where data ownership and responsibility for data creation and access is distributed among different actors. As different parties—both human and machine—create and access data across the entire Internet, issues of authorisation and protection of personal privacy are arising with increasing frequency. Addressing these issues requires the adoption of ICMSs such as CortexDB.

In the next ThoughtPoint, I'll look at another example of how the unique structure of CortexDB offers a new way to address another common problem for digital businesses: dealing with dirty data.

Links to all ThoughtPoints in the series originally published on LinkedIn:

ThoughtPoint 1: CortexDB Reinvents the Database – June 2019
ThoughtPoint 2: Making Data Agile for Digital Business – June 2019
ThoughtPoint 3: Managing Data on Behalf of Different Actors – June 2019
ThoughtPoint 4: Distilling Deeper Truths from Dirty Data – July 2019
ThoughtPoint 5: CortexDB Drives Agile Digital Transformation – July 2019

*Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** was published in October 2013.*

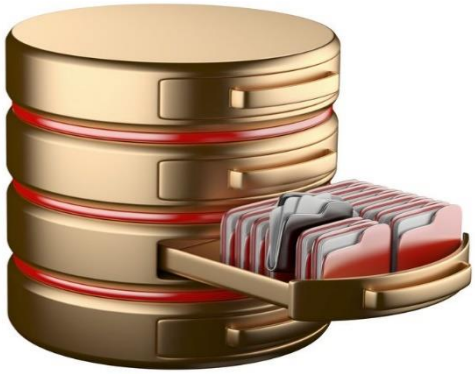
Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI Upside, and more, Barry is based in Bristol, UK, and operates worldwide.

Distilling Deeper Truths from Dirty Data*

JULY 2019

ThoughtPoint 4 of a 5-part Series by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Dirty data is not the real problem. CortexDB reframes the challenge: figuring out the true meanings hidden in the grime created when data sources—from people to sensors—deliver low quality data. In the fourth ThoughtPoint of this series, we explore data distillation.



Data scientists' jobs are the most data-driven—as well as, allegedly, the sexiest—jobs in digital business. The goal is to seek out business insights from the extensive data available in a digitised world. Before it became glamourous, data science was called data mining, arguably a better name. Like physical mining, the initial challenge for data scientists is to refine relatively tiny nuggets of data gold from enormous quantities of raw data ore and, equally important, to generate clean and pure gold-standard data.

Much of the dross in raw data comes from how we misspell names, abbreviate words, misplace values in manual data entry or as speech-to-text applications misunderstand our speech. Further problems arise when we fail to fully specify the context of the data generated or captured by the billions of sensors deployed on the Internet of Things. The incoming data is dirty in many different ways, but cleansing each instance is only a part of the problem. The real challenge arises when, from millions of arriving data records, we

* *A German version of this material is available at Informatik Aktuell.*

SPONSORED BY
www.cortex-ag.com

need to distil those that refer to the same real-world entity and uniquely specify the truth meanings of each. Traditionally, name and address data is the most commonly cited example of this challenge, but it applies to any type of loosely structured information.

Solutions to this type of problem—often called *record linkage*—date back many decades and aim to create a "golden record" of cleansed and reconciled data about each entity. Traditionally, the process consists of multiple, sequential, partially overlapping cleansing and comparison steps. The choice of which steps and in which order is often manually determined, depending on the types of data involved, the categories of problems seen, and the researcher's skills and experience.

Often called *data wrangling* today, this process often falls to data scientists, wasting up to 80% of their time. Their skills are better employed analysing data and producing insights rather than cleaning up dirty data. Data wrangling tools can certainly ease data scientists' pain, but mostly simplify and visually augment existing manual cleansing steps, rather than addressing the underlying issues beneath the "dirty data" problem.

Context is the Answer… But What's the Question?

To solve this problem in a digital business—large numbers of individually dirty data records differing inconsistently across enormous data sets from multiple, often conflicting sources—we must reframe our thinking. The real business need is not to cleanse dirty data; rather the need is to distil from it uniquely identifiable identities for unique real-world entities, even while individual records may still contain unreconcilable differences within their data. This shifts our focus from errors in the data values (naked data) to the context of data creation and use (context-setting information, CSI), allowing us to work around those errors.

By separately storing and continuously aligning naked data and CSI, as described in "CortexDB Reinvents the Database", an Information Context Management System (ICMS), such as CortexDB, enables the creation of an integrated and highly automated system for reconciling and cleansing data from multiple disjoint sources. A lifelike scenario demonstrates what this means.

360° Customer Data Management

A fictitious Large European Automobile Producer (let's call them LEAP for short) sells and services its vehicles through a network of dozens of dealers in the Far East. Each dealer runs its own IT systems, which vary in size and complexity depending on dealership size. Each manages its own customer data, sales and services in its own language and according to local laws. Furthermore, LEAP also has multiple, partially inconsistent IT systems for different business functions and/or regions, due to acquisitions and IT legacy development.

How can LEAP become an integrated digital business, consolidating data from all dealers and internal departments to optimise its operations, define and track KPIs at local and international levels, and deliver sales and service excellence to all its customers and partners? How can it even hope to achieve even a small part of that aim when there is no complete, consistent master list of its customers?

Attempts to cleanse and consolidate the customer master set using data wrangling tools fail early. Each dealer's data must be imported and cleansed individually, but this approach cannot account for customers who buy from multiple dealers. Furthermore, such manual systems are exceedingly difficult to apply to continuously changing data that must support real-time operational and reporting systems.

As an ICMS, CortexDB enables the creation of an integrated set of all customer data that can be distilled—cleansed and reconciled—continuously and automatically.

Every different customer record from every system, both internal and external, is stored individually, time-stamped, and in its raw format in the CortexDB document store. No standard structure, naming nor ordering of data fields needs to be defined in advance. New or changed schemata can be handled with equal ease. Every record, with full historical sequencing, consists of an unordered and unconstrained set of key:value pairs, stored forever. Storing a complete set of customer records in original form as naked data is a prerequisite to its ongoing distillation, as well as any required auditing.

As each record is loaded into the document store, CortexDB adds its CSI into the associated sixth normal form, inverted index database. Using a combination of techniques, including deterministic methods based on likely-unique fields such as social media IDs, probabilistic phonetic and linguistic methods, and semantic graph analysis, high probability matching records are identified and tagged. As new records are added daily, they are automatically integrated into the database and assigned system-wide IDs with high probability of uniqueness with reference to real-world entities. In a highly automated approach, human involvement is limited to oversight and validation of unusual cases.

The Meaning of Context

The scenario above shows how an ICMS can support the automated distillation of dirty customer identity data from multiple sources. It can be easily extended to a range of other types of data, such as product names and descriptions, contracts, medical diagnoses, and call transcripts.

By taking a step back from the common, simplistic view of "cleaning dirty data" and focusing on the context-setting information, we can see that the real need is to distil uniquely identifiable entities from poor-quality data. CortexDB's unique indexed schemaless structure allows us to step beyond the data to the meaning of the information being stored and processed.

The fifth and final ThoughtPoint in this series positions Information Context Management Systems in the larger picture of digital transformation and shows how CortexDB offers broader possibilities than traditional database systems.

Links to all ThoughtPoints in the series originally published on LinkedIn:

ThoughtPoint 1: CortexDB Reinvents the Database – June 2019
ThoughtPoint 2: Making Data Agile for Digital Business – June 2019
ThoughtPoint 3: Managing Data on Behalf of Different Actors – June 2019
ThoughtPoint 4: Distilling Deeper Truths from Dirty Data – July 2019
ThoughtPoint 5: CortexDB Drives Agile Digital Transformation – July 2019

*Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI Upside, and more, Barry is based in Bristol, UK, and operates worldwide.

SPONSORED BY

cortex

www.cortex-ag.com

CortexDB Drives Agile Digital Transformation*

JULY 2019

ThoughtPoint 5 of a 5-part Series by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

As CortexDB transitions from customer-driven projects to a roadmap-based product development approach, it is well-positioned to occupy a leading position in the emerging Information Context Management System (ICMS) market and deliver key functionality in support of digital transformation.

In the first of the preceding four ThoughtPoints of this series, I've positioned CortexDB as the forerunner of a new and important class of information management product—an adaptive Information Context Management System. I've also illustrated and described its uses in solving some key issues facing organisations that are embarking on a digital transformation journey. That is a lot of responsibility to put on the shoulders of one small software company! Allow me to explain why I believe they can carry it.

A strong conceptual foundation

Traditional databases are based on a belief that the context of the data is known in advance of storing it. They therefore begin with a model of the meanings and relationships of the data as defined by its creators. Schemaless data stores shun such structuring, preferring to push the identification of context and structure to users of the

* *A German version of this material is available at Informatik Aktuell.*

SPONSORED BY
www.cortex-ag.com

naked data[†]. Both approaches have well-known pros and cons. And neither works the way our brains do. Our brains have neither predefined schemas nor big data dumps. We actually build associations between facts on the fly and retrieve and process information via those associations.

CortexDB, as the name suggests, replicates this cognitive / associative approach, but with traditional software—a document store combined with a 6NF (6th normal form) inverted index array—that allows every data "fact" and its context to be directly and rapidly located based on its value. I call this an Information Context Management System (ICMS). Beyond the advantage of using mature software, CortexDB also avoids the black box problems of modern artificial intelligence (AI) neural networks that cannot explain how they reached their conclusions.

This novel approach confers several important advantages. Data can be stored at first with little or no structure / context, allowing multiple, useful schemas to be discovered or developed over time depending on usage. This "schema-by-usage-pattern" leads to agile data modelling and application development, bridging the gap between schema-on-write and schema-on-read. Supporting multiple schemas on the same naked data allows a reduction in the number of copies of data that needs to be stored and maintained. A prime example is the possibility to move from separate operational and informational systems to a single instance, sometimes called Hybrid Transactional / Analytical Processing (HTAP). As discussed previously, specific, high-impact issues for digital transformation can be addressed, such as distributed data ownership and distillation of dirty data.

A clear organisational mindset

Database management system (DBMS) development is rather different from that of other software components. Historically, DBMSs can take a decade or more to mature to the stage where businesses are comfortable to trust them with their precious data. Successful databases have often matured in situations where the developer and a few key partner customers hone the system through specific and diverse projects in the early years, before the vendor moves to a more typical product development roadmap. The same small and dedicated team of engineers guide the design and development over the whole multiyear development cycle. Their work is shielded from annual or quarterly stock market targets, so that decisions are taken for the longer term good of the product.

The same considerations apply to ICMS development, and CortexAG match the above criteria. As a small, privately held company, it can focus on the longer-term success of its product. In its almost two decades of development, the same team of a dozen or so engineers and leaders have driven development according to a vision of mimicking the way the human brain works. The development has proceeded with a wide variety of projects with demanding customers, such as Volkswagen, Bundesdruckerei (a German

[†] Naked data is information that is "stripped" of as much context-setting information (CSI) as possible, consisting mostly of "pure" value data. See the first ThoughtPoint in this series for details.

manufacturer of banknotes, identity documents, driving licences and other federal documents), Airbus, BMW, and other small and mid-size companies.

With multiple successful customer-driven projects contributing to the funding and direction of CortexDB, the company is ready to embark on the next stage of its evolution: the transition to development driven by a more formal roadmap.

An emerging mainstream product

CortexAG have now reached the transition point from project-led development to product-roadmap adoption. This involves focusing product direction toward a set of predefined goals, in addition to responding directly to the specific needs of key partner customers. A key step in this transition is the identification of an overall strategic area of development focus. In the case of CortexDB, this is now defined as the Information Context Management System.

The database market—both structured and unstructured—is crowded with literally hundreds of competitors, from start-ups to mega-vendors. CortexDB's focus on the cognitive / associative model of creating, collecting, storing, and managing context-setting information in close connection with the underlying naked data allows it to grow and prosper in a relatively novel space, but one that is garnering widespread attention as data governance issues take centre stage in digital businesses and data agility becomes a mandatory requirement for digital transformation.

Links to all ThoughtPoints in the series originally published on LinkedIn:

ThoughtPoint 1: CortexDB Reinvents the Database – June 2019
ThoughtPoint 2: Making Data Agile for Digital Business – June 2019
ThoughtPoint 3: Managing Data on Behalf of Different Actors – June 2019
ThoughtPoint 4: Distilling Deeper Truths from Dirty Data – July 2019
ThoughtPoint 5: CortexDB Drives Agile Digital Transformation – July 2019

*Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI Upside, and more, Barry is based in Bristol, UK, and operates worldwide.

SPONSORED BY
www.cortex-ag.com