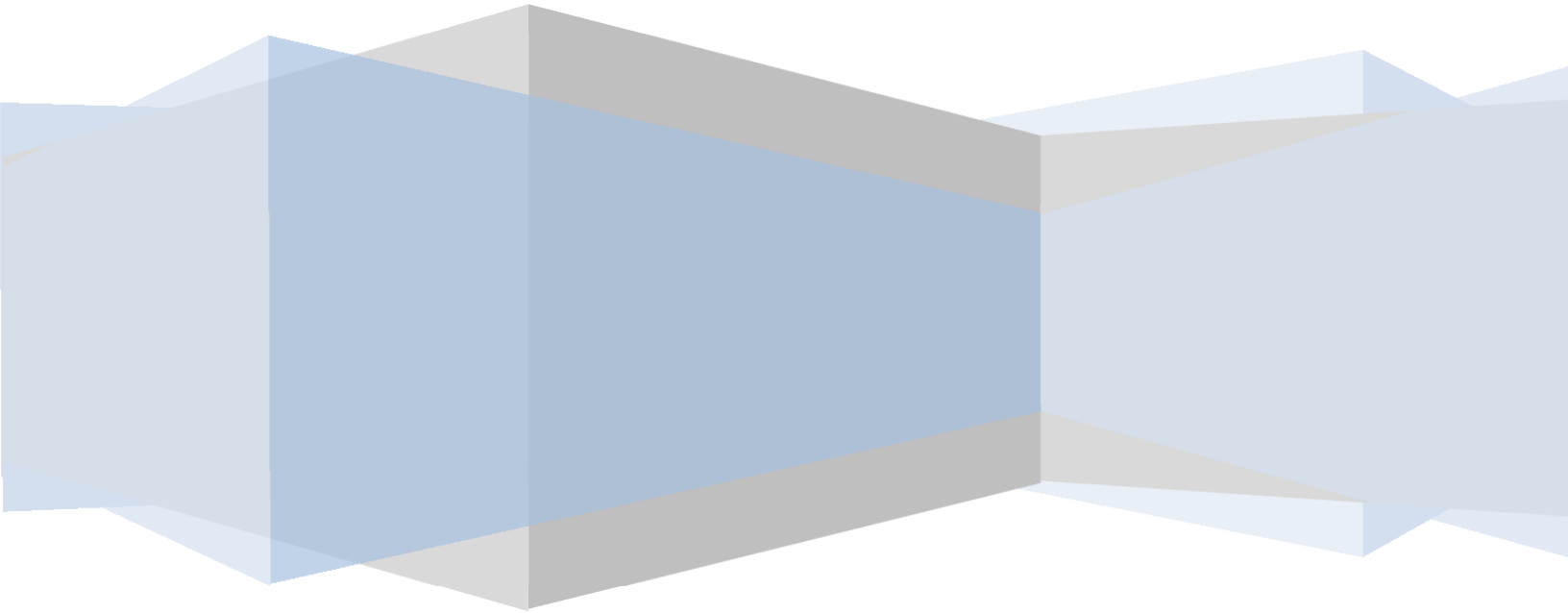
ACORD and MarkLogic

Making Sense of Big Data in Insurance

Puneet Bharal, ACORD and Amir Halfon, MarkLogic



Disclaimer and Reservation of Rights

This document is provided “as is” and the authors make no representations or warranties, express or implied, including, but not limited to, warranties of merchantability, fitness for a particular purpose, non-infringement, or title; that the contents of the document are suitable for any purpose; nor that the implementation of such contents will not infringe any third party patents, copyrights, trademarks or other rights.

This document is provided for general informational and educational purposes only and does not constitute legal advice. The information contained herein is intended, but not promised or guaranteed, to be current, complete, and up-to-date. You should not act or rely on any information contained in this document without first seeking the advice of an attorney.

The authors will not be liable for any direct, indirect, special or consequential damages arising out of any use of the document or performance or implementation of the contents thereof.

Permission to use, copy, and distribute the contents of this document in any medium for any purpose and without fee or royalty is hereby granted, provided that you, in all cases, clearly reference the document and the authors thereof.

Copyright

About the Document

The insurance industry is data-dependent. Today, carriers and intermediaries are engaged in improving data capture to help them to better manage their business, manage their risk and know their customers. Business and regulatory drivers are pushing the industry to manage its data better. For the past 40 years, ACORD has been defining data standards to help with this.

Recently, though, one of the most talked about trends in IT is **Big Data**. Some commentators and vendors have attached great promise to the capabilities of Big Data, even arguing that other data types are no longer necessary. According to Gartner, Big Data has reached the peak of its Hype Cycle this year. "Hype" connotes an over-selling of potential and leads to inflated expectations.

ACORD members are interested in understanding what Big Data can offer them in their businesses and how these new data management technologies interact with their existing database systems and ACORD standard data. In this paper, ACORD and MarkLogic demystify some of the hype around Big Data and provide a pragmatic review of the technologies available, their strengths and weaknesses and some examples of Big Data technologies in use in the Insurance industry.

About ACORD

ACORD (Association for Cooperative Operations Research and Development) is a global, nonprofit organization serving the insurance and related industries. ACORD facilitates the development of open consensus data standards and standard forms, and works with its members and partner organizations to drive implementation of those standards.

Implementing ACORD Standards has been shown to improve data quality and flow, increase efficiency, and realize billion-dollar savings to the global industry. ACORD members worldwide include hundreds of insurance and reinsurance companies, agents and brokers, software providers, financial services organizations and industry associations.

ACORD also presents events, videos, research papers and seminars on standards implementation as well as current technology and business topics. ACORD maintains offices in New York and London.

ACORD's Vision

ACORD envisions an insurance industry that embraces a global and enterprise view of information,in which relevant business solutions all include or provide for ACORD Standards. ACORD wants all trading partners to be able to easily exchange information.

ACORD's vision includes implementation of best practices for enterprise architecture, including systems made up of interchangeable components based on ACORD Standards that provide a 360-degree view of people, organizations, and risks. Products and services built upon these components will be highly configurable and will enable a wide range of consistent transactions and processes across the entire insurance value chain.

About MarkLogic

Since 2001, MarkLogic has focused on building a platform that enables our customers to capture more data -- and do more with it. We give our customers an unmatched competitive edge through a powerful and trusted Enterprise NoSQL (Not Only SQL) database that enables organizations to turn all data into valuable and actionable information. With search and application services embedded into the technology stack, MarkLogic allows companies to streamline operations and quickly develop advanced database applications that search through terabytes of information to store and retrieve any type of data. Designed from the onset to be enterprise-grade, MarkLogic has been building the enterprise-hardened performance, security, and reliability features that CIOs expect from their enterprise databases. Our more than 360 enterprise customers use MarkLogic to innovate, grow their businesses, and, even make the world a safer place.

Contents

Summary

The insurance industry has been struggling to get a good handle on its data for decades, both on the transactional and the risk management sides. And the recent emphasis on utilizing new sources of data that extend beyond traditional sources, often referred to as *Big Data*, has created renewed interest in data management across the industry. Data variety and diversity in particular are pushing the traditional, relational database management technologies to their limits, and are raising more and more interest in new approaches to data management.

Some commentators and vendors have attached great promise to the capabilities of Big Data, even arguing that other data types are no longer necessary. According to Gartner, Big Data has reached the peak of its Hype Cycle this year. "Hype" connotes an over-selling of potential and leads to inflated expectations.

So, are relational databases on their way to extinction? Not anytime soon. They are still the most efficient way to handle highly structured tabular data, especially in the context of Online Transaction Processing. Instead, companies should look to benefit from new technologies associated with Big Data where these can provide value beyond the core transaction processing associated with policy administration and claims management.

The recommended adoption strategy when it comes to Big Data is therefore a **hybrid** approach, utilizing both relational and non-relational enterprise technologies, considering the opportunities offered by the new technology paradigm, while continuing to utilize the old paradigm where appropriate. Especially in this era of limited, shrinking budgets, it is important to use technology appropriately, and find the use cases where new technology would provide specific strategic benefits.

Background

The insurance industry has been struggling to get a good handle on its data for decades, both on the transactional and the risk management sides. And the recent emphasis on utilizing new sources of data that extend beyond traditional sources, often referred to as *Big Data*[1], has created renewed interest in data management across the industry. Data variety and diversity in particular are pushing the traditional, RDBMS[2] technologies to their limits, and are raising more and more interest in new approaches to data management.

One can draw a parallel to the early days of relational databases, which freed data from the reins of the original application with which it was associated, by providing decoupled access via SQL[3]. Today, new technologies are freeing the data from the reins of pre-determined data schema[4], allowing it to remain in its original form, and thus enabling organizations to use any and all their data without the need for expensive, time consuming development cycles associated with data integration.

This paper will discuss the implications, characteristics, and benefits of the new data management era in Insurance, paying particular attention to specific use cases driving new technology trends. It will also explore the capabilities of both new and traditional data management technologies within this context. To that extent, the paper will also explore the relevance of industry data standards, such as ACORD, in this evolving data management paradigm.

Data Management Examples

The paper now outlines some of the industry use cases associated with Big Data technologies, which can provide analyses otherwise impossible through traditional data capture and store practices. Big Data technologies can enable companies to assess unstructured data in to an actionable degree.

Fraud Detection

Insurance providers are looking beyond algorithmic fraud detection techniques that are claim-centric, to ones that are person–centric. These techniques focus on analysing beneficiary behavior across claims, providers, and other sources of information (e.g. how many similar claims were submitted by the same individual, reported by the same individual), and extend to data sources beyond the firewall to analytics based on external information (e.g. cohort analysis - using a person's social graph to look for similar activities among connected individuals), and considering networks of people rather than just individuals.

This person-centric approach requires integrating information across all providers involved in a claim, including counter-parties as well as partners (e.g. auto repair shops) requiring the schema-agnostic

[1] Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. (Source: Gartner)
[2] RDBMS: Relational DataBase Management Systems
[3] SQL: Sequential Query Language
[4] Data schema: The formal description of a data model

approach to data management mentioned earlier. Even when all the data lives within the firm, the agility provided by this approach makes it much more feasible to turn that data into useable information.

Customer Experience and Insight

A similar shift to customer-centricity is also occurring from marketing and regulatory perspectives. Firms are looking to go beyond policy administration to increase customer retention and satisfaction, and offer tailored solutions based on a deep understanding of their customers' needs and behavior (sometimes referred to as personalized insurance). Similarly, Know Your Customer (KYC) and Anti-Money Laundering (AML) regulations are also pushing firms towards a deeper customer insight and understanding. Both translate into a need for a single customer view, aggregating analytics across multiple channels and lines of business, which both relies on, and stretches the limits of traditional RDBMS-based technical approaches. Even today, many carriers would greatly benefit from easier customer data capture, store and share technologies based on industry standards, such as ACORD. ACORD standards can provide the basis for capturing and storing reliable customer, straight from the customer or intermediary into the carrier's databases, without manual re-keying or delay. This data can be supplemented by the rich and varied data available from other sources, and indeed, other formats (i.e. Big Data).

The main challenge is data variety and diversity – different systems have different shapes and forms of data that need to be aggregated, which require long development cycles devoted to schema design and ETL[5]. This is especially true when it comes to data that doesn't neatly fit into tabular rows and columns. A few of examples are:

- Customer onboarding documentation, containing a wealth of information that can be unlocked if these documents can be easily searched and integrated into the enterprise data management platform. These can span various lines of business, from Life and Health for non-standard, unstructured medical reports, to Reinsurance and Large Commercial P&C for property-, employee- and fleet schedules.
- Customer Care call logs, containing free-form representative comments about incoming customer calls, which could be mined for *sentiment* in order to identify customers that had negative conversations, and tied to structured enrollment data to determine if perceived negative customer service affected customer dis-enrollment. This can help to develop strategies and best practices to improve customer retention and reduce customer churn.
- Clickstream data from the customer-facing website, which can be analyzed to find browsing patterns that indicate customer tendencies, especially when correlated to call center logs to find those instances when customers called immediately following a web-interaction.

[5] ETL: Extract, Transform and Load

- Auto telematics, which monitor driver behavior directly, enabling a carrier to provide usage-based auto insurance, with premiums based on very fine-grained risk assessment.
- Geo-spatial data, which is valuable in analyzing the location of assets such as ships and vehicle fleets (for marine and other commercial insurance), as well as analyzing the impact of weather events such as hurricanes to determine customer follow-up or pro-active advisory (e.g. severe weather precautions). The value-adding activities not only assist in reducing potential claim size, but also provide the customer with positive interactions with the carrier. It is a truism that most customers only ever speak with their insurer or agent when obtaining premium quotes or managing a claim, and so carriers and/or their agents get few opportunities to develop a strong affinity with the customer.

Claims Management

Non-relational data is also quite relevant to claims management, with carriers looking to maintain images, video and text notations right alongside the claims they support (e.g. notations from police inspector or tow truck operator for auto insurance claims). Combined with the push towards aggregating payer and beneficiary information across several entities (beneficiaries, payers, providers) to achieve the purposes mentioned previously, this also creates tremendous challenges for traditional relational technologies, which can't scale to meet these new demands.

Even without considering these new pressures, there's much to be gained from using non-relational data management technology in claims management: The widely used ACORD standard for message exchange is hierarchical in nature, and newer technologies such as document oriented databases can load these messages as-is from the bus, without the need for transforming them into normalized schemas before than can be processed. Mainframe computers, which are still used in claims management, often use hierarchical data structures that also lend themselves to new-generation document oriented databases that can aggregate and process this data as-is. This can also enable firms to reduce their mainframe consumption (measure in Million Instructions Per Second, aka *MIPS*) by offloading transaction processing to more cost-effective systems without the need for long, expensive data transformation efforts.

In short, ACORD standard data messages can provide structured data from first notification of loss onwards, along with workflow triggers to streamline and manage the claims process. This process can be supplemented with unstructured text and multimedia information which can speed up the decision-making process and save time and cost for the carrier. The single greatest opportunity to impress (or otherwise) the client is at the time of a claim, so better claims processing is a chance to not only reduce time and expense, but to differentiate on service rather than premium price alone. Consider the possibility of allowing customers to notify of their claims with a simple app-based form and attaching photos, videos and police reports using their smartphones, feeding straight into carriers' data systems.

Underwriting

In the Reinsurance and Large Commercial insurance sectors, large quantities of supporting information are provided as part of placing submissions (e.g. loss histories, property schedules, fleet vehicle schedules and Directors' details). They are often presented in widely diverse file formats such as Excel spreadsheets, Word files, PowerPoint presentations, Access databases, PDF documents etc., and rarely in a standard format. Even elements as simple as *Address* can be presented in any number of ways. In order to extract all the data and make it *query-able*, users would have to re-key it – a time-consuming and error-prone task, regarded as a high-cost, low-value exercise.

Entering this data into a relational database while conforming to its data definitions would be an onerous task. Instead, this extensive data is sifted through and assessed manually by the Underwriter at the quotation stage and then simply filed away, only to be re-examined by Claims experts if a claim is subsequently made against the policy. These processes are time-consuming and unreliable, and virtually impossible to audit. Moreover, the information remains locked in the documents, with little or no analytical use.

Non-relational technologies allow insurers to very quickly store and access any data so they can run analyses to highlight anomalies, patterns and red flags - things which are very difficult for people to do when manually reading a collection of documents. The ability to automate the data management, of underwriting support documentation allows insurers to create a risk and client profile that's used consistently across the firm, is auditable, and provides rich analytics on top.

Big Data from a Technical Perspective

From a technical perspective, new data management platforms are available, which are nothing short of game-changing. The term Big Data can sometimes be misleading, as new technology adoption is driven by increasing data diversity and variety, and not just increasing volumes.

One of the most substantial shifts occurring today is from structuring data before it could be managed (schema on write) to having self-described data, containing many forms and shapes of structure (schema on read). Two key technology categories comprising this shift are Hadoop[6] and NoSQL[7], both offering highly scalable, non-relational solutions that complement each other in handling different workloads.

Hadoop is most often used for long-running analytics, which typically require some software development cycles. It is essentially a distributed processing and storage framework that enables developers to take advantage of cheap, commodity compute and storage resources in a highly parallel

[6] Apache Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware. Hadoop was derived from Google's MapReduce and Google File System (GFS) papers. (Source: Wikipedia)

[7] NoSQL: Can refer to a database which does not use SQL. More recently, NoSQL is regarded as *Not Only* SQL.

fashion, without having to develop the intricate mechanisms associated with distributed computing. Several tools are also built on top of this framework to provide additional capabilities such as machine learning.

NoSQL databases are most often used for more interactive, operational and analytical use cases associated with traditional databases, and are similar to them from a usage perspective. Their distinguishing difference is that they alleviate the need for extensive schema design up front, instead supporting many structures contained within the data itself. Some also integrate full text search and SPARQL with structured querying (as is the case with MarkLogic), to tightly integrate structured and unstructured data management within a single platform.

The main benefit of these new technologies is their ability to handle any structure data – all your data – without requiring extensive data integration efforts. This allows for increased agility in responding to ever-changing business demands, and opens the door for innovation while cutting the costs associated with relational data management platforms.

Traditional Relational Database Management Systems (RDBMS) on the other hand, are focused on providing users with the ability to query sets of clearly defined data items, allowing them to define the universe of their data. There are inherent strengths to this approach, chiefly the knowledge that queries will compare *apples* with *apples* and not *oranges,* but the downside is the rigidity that this approach imposes, and the need to fit all the data into neatly defined schemas before it can be examined. Over the years, some databases grow to such size and complexity that schema additions and modifications require such effort as to become prohibitive, especially in support of what can be regarded as one-off queries.

New technologies also offer better scalability models than traditional databases, as they typically scale horizontally on commodity hardware, rather than require expensive infrastructure to run efficiently. It is important to note that while many new database entrants sacrifice certain qualities such as full consistency in order to scale, this is not an imperative. Horizontal scalability and full transaction support (providing Atomicity Consistency Isolation and Durability across records) are orthogonal, and in fact MarkLogic has been providing both for many years, in addition to other enterprise capabilities such as role-base security authorization and full backup and recovery.

Conclusion

So, are relational databases on their way to extinction? Not anytime soon. They are still the most efficient way to handle highly structured tabular data, especially in the context of Online Transaction Processing. Replacing them in this context just for the sake of using the latest "shiny object" would certainly be unwise, and few if any firms are looking to "rip and replace" all of their data management platforms. What they are looking to do instead is benefit from new technologies associated with Big Data where these can provide value beyond the core transaction processing associated with policy administration and claims management.

The recommended adoption strategy when it comes to Big Data is therefore a hybrid approach, utilizing both relational and non-relational enterprise technologies, considering the opportunities offered by the new technology paradigm, while continuing to utilize the old paradigm where appropriate. Especially in this era of limited, shrinking budgets, it is important to use technology appropriately, and find the use cases where new technology would provide specific strategic benefits. This paper provided some of these, and additional ones may be found at any given firm. The common thread is that they providing specific, tangible business value from the agility associated with schema-on-read, and the expanded operational and analytical capabilities associated with the ability to manage all any data, regardless of its form. And this value is what's driving the current shift in data management technology.

The Authors

Puneet Bharal
Director, Global Development
ACORD

pbharal@acord.org
+44 (0)207 617 6406

Amir Halfon
Chief Technologist
MarkLogic

amir.halfon@marklogic.com
+1 917 312 5040