CAPCO

THE CAPCO INSTITUTE JOURNAL OF FINANCIAL TRANSFORMATION

DATA MANAGEMENT

Unlocking value through data lineage

THADI MURALI | RISHI SANGHAVI | SANDEEP VISHNU

DATA ANALYTICS

50TH EDITION | NOVEMBER 2019

THE CAPCO INSTITUTE

JOURNAL OF FINANCIAL TRANSFORMATION

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

CONTENTS

DATA MANAGEMENT

10 **The big gap between strategic intent and actual, realized strategy**
Howard Yu, LEGO Professor of Management and Innovation, IMD Business School
Jialu Shan, Research Fellow, IMD Business School

24 **Data management: A foundation for effective data science**
Alvin Tan, Principal Consultant, Capco

32 **Synthetic financial data: An application to regulatory compliance for broker-dealers**
J. B. Heaton, One Hat Research LLC
Jan Hendrik Witte, Honorary Research Associate in Mathematics, University College London

38 **Unlocking value through data lineage**
Thadi Murali, Principal Consultant, Capco
Rishi Sanghavi, Senior Consultant, Capco
Sandeep Vishnu, Partner, Capco

44 **The CFO of the future**
Bash Govender, Managing Principal, Capco
Axel Monteiro, Principal Consultant, Capco

DATA ANALYTICS

54 Artificial intelligence and data analytics: Emerging opportunities and challenges in financial services
Crispin Coombs, Reader in Information Systems and Head of Information Management Group, Loughborough University
Raghav Chopra, Loughborough University

60 Machine learning for advanced data analytics: Challenges, use-cases and best practices to maximize business value
Nadir Basma, Associate Consultant, Capco
Maximillian Phipps, Associate Consultant, Capco
Paul Henry, Associate Consultant, Capco
Helen Webb, Associate Consultant, Capco

70 Using big data analytics and artificial intelligence: A central banking perspective
Okiriza Wibisono, Big Data Analyst, Bank Indonesia
Hidayah Dhini Ari, Head of Digital Data Statistics and Big Data Analytics Development Division, Bank Indonesia
Anggraini Widjanarti, Big Data Analyst, Bank Indonesia
Alvin Andhika Zulen, Big Data Analyst, Bank Indonesia
Bruno Tissot, Head of Statistics and Research Support, BIS, and Head of the IFC Secretariat

84 Unifying data silos: How analytics is paving the way
Luis del Pozo, Managing Principal, Capco
Pascal Baur, Associate Consultant, Capco

DATA INTELLIGENCE

94 Data entropy and the role of large program implementations in addressing data disorder
Sandeep Vishnu, Partner, Capco
Ameya Deolalkar, Senior Consultant, Capco
George Simotas, Managing Principal, Capco

104 Natural language understanding: Reshaping financial institutions' daily reality
Bertrand K. Hassani, Université Paris 1 Panthéon-Sorbonne, University College London, and Partner, AI and Analytics, Deloitte

110 Data technologies and Next Generation insurance operations
Ian Herbert, Senior Lecturer in Accounting and Financial Management, School of Business and Economics, Loughborough University
Alistair Milne, Professor of Financial Economics, School of Business and Economics, Loughborough University
Alex Zarifis, Research Associate, School of Business and Economics, Loughborough University

118 Data quality imperatives for data migration initiatives: A guide for data practitioners
Gerhard Längst, Partner, Capco
Jürgen Elsner, Executive Director, Capco
Anastasia Berzhanin, Senior Consultant, Capco

DEAR READER,

Welcome to the milestone 50th edition of the Capco Institute Journal of Financial Transformation.

Launched in 2001, the Journal has covered topics which have charted the evolution of the financial services sector and recorded the fundamental transformation of the industry. Its pages have been filled with invaluable insights covering everything from risk, wealth, and pricing, to digitization, design thinking, automation, and much more.

The Journal has also been privileged to include contributions from some of the world's foremost thinkers from academia and the industry, including 20 Nobel Laureates, and over 200 senior financial executives and regulators, and has been co-published with some of the most prestigious business schools from around the world.

I am proud to celebrate reaching 50 editions of the Journal, and today, the underlying principle of the Journal remains unchanged: to deliver thinking to advance the field of applied finance, looking forward to how we can meet the important challenges of the future.

Data is playing a crucial role in informing decision-making to drive financial institutions forward, and organizations are unlocking hidden value through harvesting, analyzing and managing their data. The papers in this edition demonstrate a growing emphasis on this field, examining such topics as machine learning and AI, regulatory compliance, program implementation, and strategy.

As ever, you can expect the highest caliber of research and practical guidance from our distinguished contributors, and I trust that this will prove useful to your own thinking and decision making. I look forward to sharing future editions of the Journal with you.

Lance Levy, **Capco CEO**

FOREWORD

Since the launch of the Journal of Financial Transformation nearly 20 years ago, we have witnessed a global financial crisis, the re-emergence of regulation as a dominant engine of change, a monumental increase in computer processing power, the emergence of the cloud and other disruptive technologies, and a significant shift in consumer habits and expectations.

Throughout, there has been one constant: the immense volume of data that financial services institutions accumulate through their interactions with their clients and risk management activities. Today, the scale, processing power and opportunities to gather, analyze and deploy that data has grown beyond all recognition.

That is why we are dedicating the 50th issue of the Journal of Financial Transformation to the topic of data, which has the power to change the financial industry just as profoundly over the coming 20 years and 50 issues. The articles gathered in this issue cover a broad spectrum of data-related topics, ranging from the opportunities presented by data analytics to enhance business performance to the challenges inherent in wrestling with legacy information architectures. In many cases, achieving the former is held back by shortcomings around the quality of, and access to, data arising from the latter.

It is these twin pillars of opportunity and challenge that inform the current inflection point at which the financial industry now stands. Whilst there is opportunity to improve user experiences through better customer segmentation or artificial intelligence, for example, there are also fundamental challenges around how organizations achieve this – and if they can, whether they should.

The expanding field of data ethics will consume a great deal of senior executive time as organizations find their feet as they slowly progress forward into this new territory. In my view, it is critical that organizations use this time wisely, and do not just focus on short-term opportunities but rather ground themselves in the practical challenges they face. Financial institutions must invest in the core building blocks of data architecture and management, so that as they innovate, they are not held back, but set up for long-term success.

I hope that you enjoy reading this edition of the Journal and that it helps you in your endeavours to tackle the challenges of today's data environment.

Guest Editor
Chris Probert, **Partner, Capco**

UNLOCKING VALUE THROUGH DATA LINEAGE

THADI MURALI | Principal Consultant, Capco
RISHI SANGHAVI | Senior Consultant, Capco
SANDEEP VISHNU | Partner, Capco[1]

ABSTRACT

Data and information lifecycle management challenges in a financial services organization (FSO) can be daunting, especially when they relate to data security, integrity, or availability. Large FSOs recognize this and are willing to make investments to address IT cyber risk, data management, and data governance, specifically when the payoff is clearly articulated. In a world of big data, current techniques for information risk and control assessment fall woefully short as they do not provide adequate visibility around data nor do they assist the business in decision making. Data lineage can fill this gap. Thus far, data lineage has largely been directed towards regulatory initiatives focused on risk and finance. However, the broader business use of data lineage is relatively unexplored, in part due to a lack of industry standards or methodologies to guide organizations to realize the full potential of data lineage. This article explores how data lineage standards and patterns can drive substantial value beyond regulatory compliance by holistically considering control optimization and cost reduction.

1. INTRODUCTION

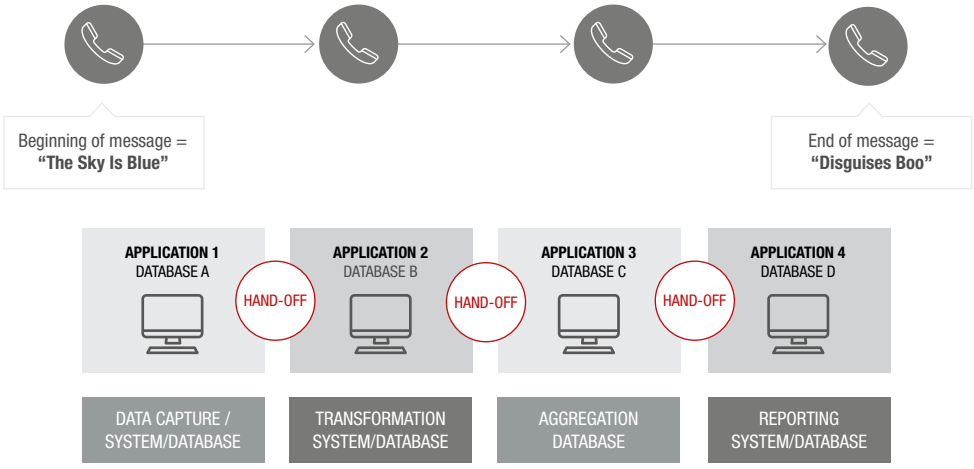
Banks have always held vast amounts of data inside of their organizations, however, with the recent exponential growth in data volume, velocity, variety, and veracity, efficient usage and governance of data has become a critical success factor and a source of competitive advantage for financial institutions. The ability to identify, monitor, interpret, and extract value from data is something that many organizations have historically struggled to achieve, mostly due to poor tracking of data across the enterprise. Data lineage serves as a tool to track data from origination, through transformation(s), and ultimately to consumption. Financial institutions have an opportunity to provide major value to their organizations by using data lineage to provide benefits along three dimensions: (1) regulatory compliance, (2) control optimization, and (3) cost reduction.

2. WHY DATA LINEAGE?

To understand how data lineage is useful, it is important to acknowledge the obvious: that data in a large organization is not held in a single repository but flows across many systems and databases. As this proliferation happens there is a risk to data quality, security, and availability. For example, those familiar with the "telephone game", know that despite best intentions, the original message from the first person (e.g., "The sky is blue.") changes to something completely different when it ends with the last person (e.g., "Disguises Boo") (Figure 1). This is the same with information – data can get compromised every time it goes to a new system or database or there is a hand-off between systems. This can occur even when there is no malicious intent on the part of the users of the systems. If there is malicious intent, the risk is exponentially higher. In such situations, data lineage has an even greater role as it provides the traceability needed to mitigate risks with appropriate controls.

[1] The authors would like to thank Mayssam Jahansoozan, Consultant, Capco, Tyler West, Associate Consultant, Capco, and Clara Steiner, Associate Consultant, Capco for their contributions to this article.

Figure 1: Analogy between "telephone game" and enterprise data movement



3. TRADITIONAL DATA LINEAGE AND CHALLENGES

Data lineage rose to prominence due to regulatory requirements after the financial crisis, when regulators required evidence to substantiate that the Comprehensive Capital Analysis and Review (CCAR) stress-test reporting for banks was accurate. Other regulations that followed, like BCBS 239 (the Basel Committee on Banking Supervision's principles of risk data aggregation), reinforced the need for sound data lineage. BCBS 239 guidelines are designed to improve data aggregation, accountability, and reporting across financial markets. Since then, regulations, such as Markets in Financial Instruments Directive II (MiFID II), General Data Protection Regulation (GDPR), Fundamental Review of the Trading Book (FRTB), FDIC 370, and others all require financial institutions to implement data lineage procedures to demonstrate the reliability of their reporting.

However, currently data lineage is underutilized – it largely focuses on the mechanical movement of data and less on its contextual flow. Additionally, it is often targeted towards an IT audience. Most financial institutions use data lineage to map technical data, which typically consists of tables and columns. Programmers then use the mapping to update their software code. In applying this traditional approach to data lineage, businesses are missing out on the full potential that can be realized through the insights provided by having clarity on data movement and transformation. For example, in a bank, team members decided to explore payment transactions to come up with rules to prevent fraud crimes. When project managers included data lineage details, such as who owns and accesses the data, why and how the data is transformed, it provided additional insight that helped mitigate payment fraud risk. By understanding the data lineage, the right preventative controls were developed to address vulnerable areas. Furthermore, the bank also reduced the time and effort on transaction monitoring, as they concluded that preventative controls from lineage were more effective.

For data lineage to be meaningful for business purposes beyond regulatory compliance, traditional lineage needs to expand from just the what and where to address the following dimensions: who, what, where, when, why, and how. Having industry standards or accepted practices would help provide a structure for capturing this list of dimensions and drive business value.

4. THREE STANDARDS TO MAKE LINEAGE USEFUL

Many FSOs have a vastly scattered data landscape. A big challenge for organizations that want to use data lineage is that there are no standards on how to depict data lineage. There exist large differences in representation of lineage, from spaghetti diagrams, which are overwhelming to a business audience, to process diagrams that often leave out data, and to technical representations of architecture and infrastructure that obfuscate nuances of transformation.

There are three critical guiding principles to make data lineage useful and standardized: (1) make it business friendly, (2) highlight context and ownership of data, and (3) show how data is transformed and used. The standardization of these three principles is explored in greater detail below.

Figure 2: Standard 1 – make it business friendly

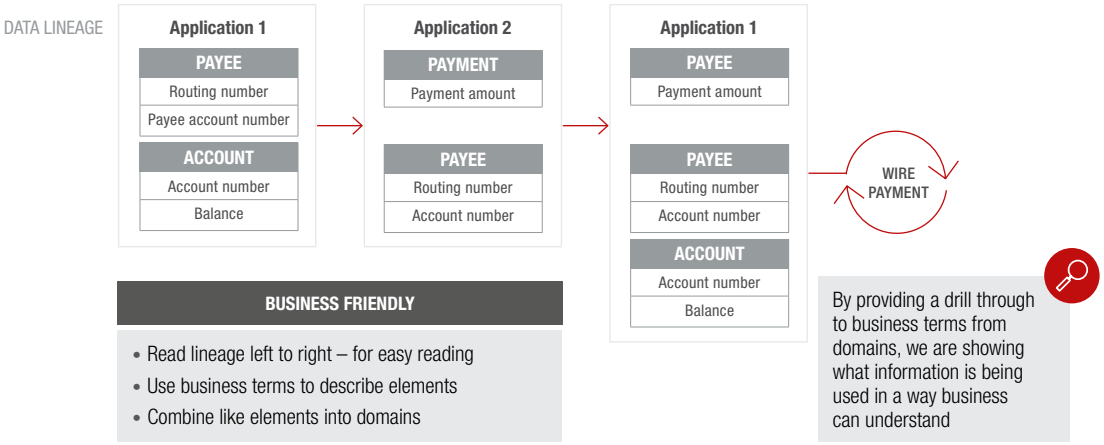


Figure 3: Standard 2 – show why (context) and who (ownership) of the data

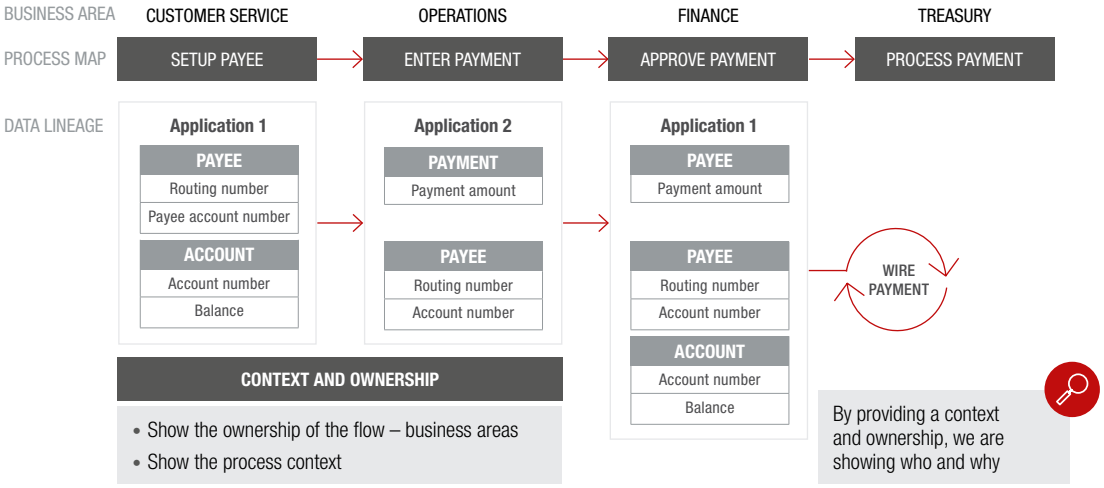
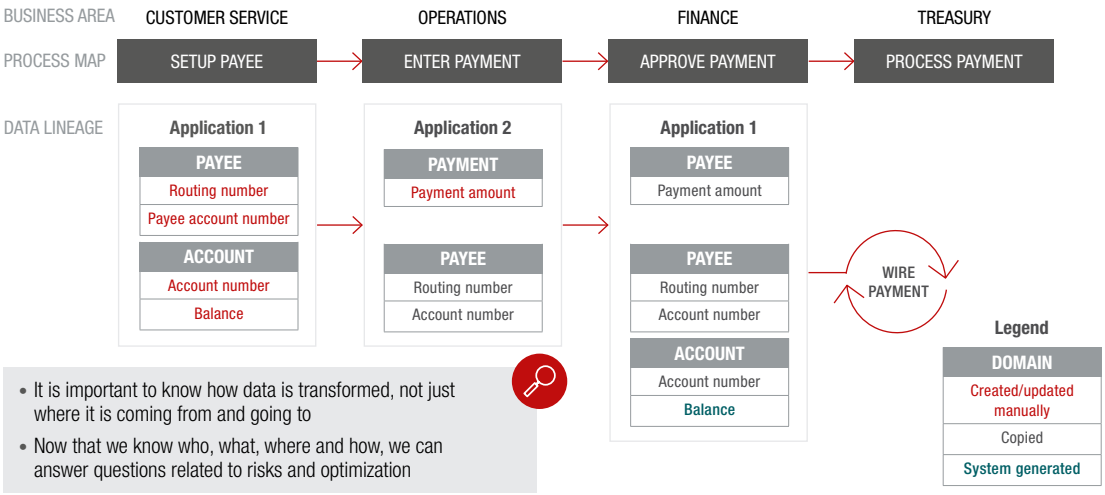


Figure 4: Standard 3 – show how data is used/transformed



The font color is used to show how data is used/transformed. For example, red font denotes that the data is manually entered or updated.

The first step is making it friendly to business. This can be done by ensuring readability and enhancing understandability. It includes using business-centric nomenclature and taxonomy as well as presenting data in easy to consume forms and applications. Figure 2 shows that a left to right read with business terms to define a data element and using domains that group related data elements can enable this. It is also important with data lineage to show only those elements that are critical to the output, or would impact the quality of the output, if compromised. These are typically referred to as 'critical data elements' (CDEs) or 'key data elements' (KDEs).

Understanding the context of data and who owns the data are vital to setting up standardized data lineage (Figure 3). Organizations need to start by determining the connections that show who owns the process, the application, and, at a finite level, the data element. Given the size and complexity of a financial institution, this can be an enormous undertaking. Taking a top-down approach by division, process, application, data element, and critical data element creates a route to drill through to business terms from the subject area; the goal is to present data in ways that organizations can understand and use. Once the data ownership structure has been established, the enterprise can begin to align the process and create a visual diagram of data lineage.

To fully grasp how a process is aligned, each step needs to be documented from beginning to end across all the divisions it touches. Visualizing the process by looking at the connections that show application ownership, process ownership, data quality, automated or manual processes, data usage, access request, and the number of outstanding data issues can help drive improvements, identify risk, and strengthen process governance. This visualization is created by defining the different types of ownership, including role of subject matter experts, managing the changes in, and versions of, data lineage diagrams, and incorporating commentary from appropriate stakeholders into the diagram. Once each step is documented and unique data elements are identified and validated by the subject matter experts, the organization will have clarity on who owns each step of the process and what data elements are associated with that step. Consequently, the enterprise will be able to leverage process diagrams to help better understand what happens to the data.

It is essential to know how data is moving, not just where it is coming from and going to — by determining the who, what, when, where and how, we can answer questions related to risks and if they can be improved (Figure 4). Financial institutions will better understand what type of controls are needed around the critical data element based on the type of action that is performed on the data element and how the data element is acted upon – manually or system updated/ created or copied.

5. PATTERNS FOR COST AND PROCESS OPTIMIZATION

Using data lineage, financial institutions can easily visualize the flow of data through systems and their applications. By doing so, it becomes possible to recognize distinct patterns, good and bad, that exist within an organization's data. When

Figure 5: Pattern 1 – using data lineage to remove redundancy

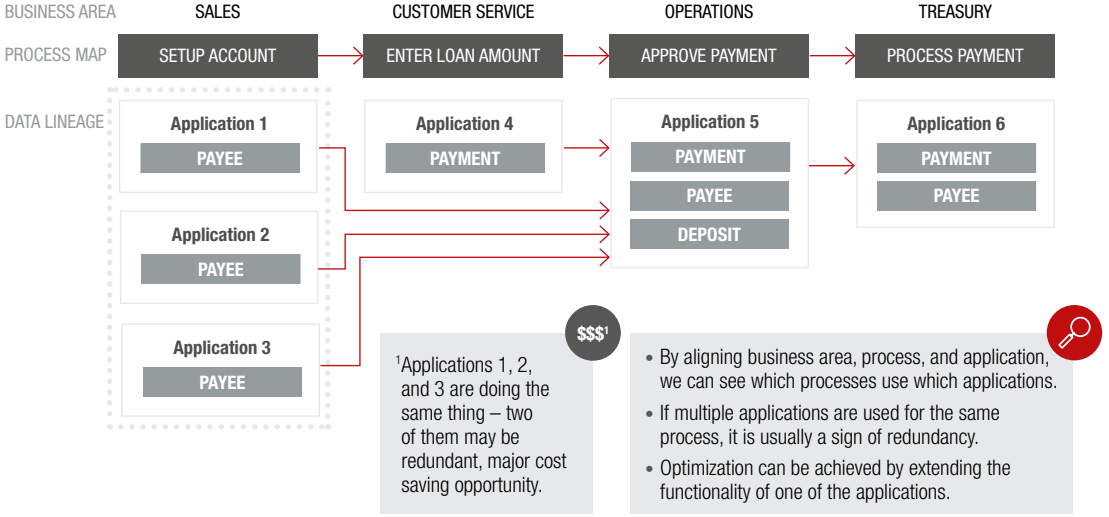
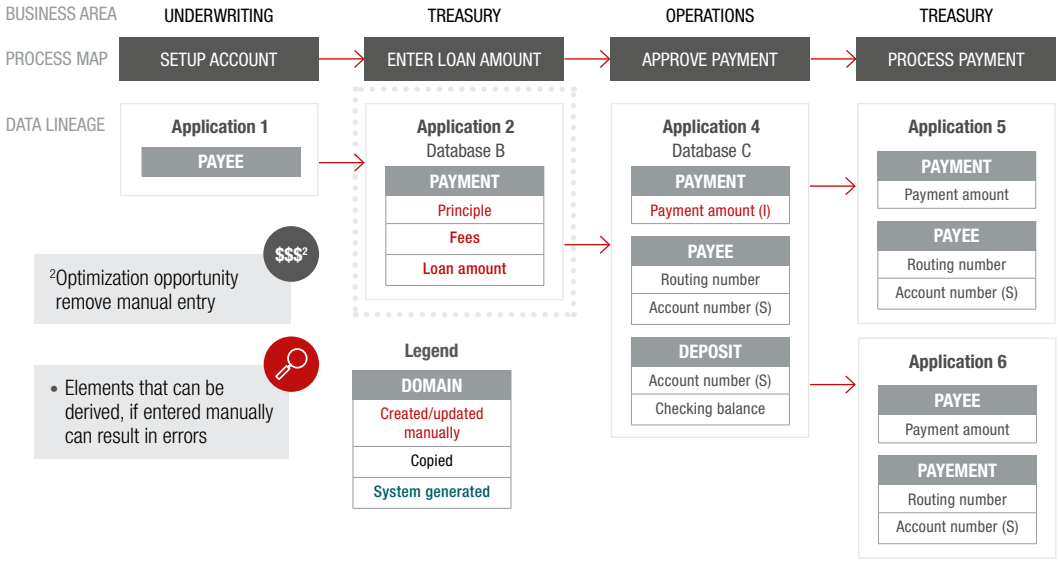


Figure 6: Pattern 2 – using data lineage to identify automation opportunities



examining the travel of data closely, organizations can use lineage to unearth insights into their data, which would otherwise be difficult to identify.

With data lineage, organizations have an opportunity to optimize cost and processes by minimizing repetition and redundancy within their systems.

In Figure 5, "account setup" represents a process, where an organization has three separate applications performing only one function. Whenever data is handled by multiple applications for the same task, it should serve as a red flag for the organization. This kind of pattern occurs often in organizations undergoing rapid growth through mergers and acquisitions or from new products and services.

Without data lineage, many organizations will miss out on identifying inefficiency, allowing waste to persist due to improper data visualization and lack of insight into applications. With data lineage, organizations can clearly align "business area", "process", and "applications" using visual diagrams that create a holistic picture of data management across the enterprise. Once a redundancy in data processes is discovered, organizations can rapidly eliminate the associated waste by enhancing one application to achieve all closely related functions and retiring the others. Data lineage provides the ability to visualize data usage, highlight inefficiencies of redundancy and repetition, and reduce cost.

In Figure 6, two data elements within application 2 – "fees and loan amount" – are bolded to show that they are derived from calculations based on the data element "Principal". In this example, fees are a percentage of the "principal and loan amount" is the sum of "principal and fees". However, the data lineage shows that these data elements are calculated elsewhere and entered into the system manually. When data is entered manually, organizations are more susceptible to inaccuracies resulting from human error in calculation or data entry. With data lineage, organizations can delineate which processes are handled manually and which through automation. Where processes are handled manually, organizations have the opportunity to automate, thereby reducing the risk of error and improving efficiency.

6. PATTERNS FOR RISK MANAGEMENT

Determining where to implement controls within a given data supply chain is crucial for maintaining data quality or integrity. Two significant types of controls are used for maintaining the quality of the data. The first type of control is an accuracy control, which is best implemented at the system of origin or the system where data is first created or entered. The second type of control is a consistency control, which supports and maintains the accuracy of the data throughout the entire data supply chain. The recommended implementation of the consistency or a reconciliation control is in downstream applications – i.e., downstream from the system of origin. These two controls, when effectively maintained, will reduce inefficiency and duplicative controls, thereby improving data quality across the lineage.

Figure 7 shows three examples, two of which have implemented controls correctly and a third where controls have been implemented incorrectly. However, organizations

Figure 7: Pattern 3 – using data lineage for designing better quality controls

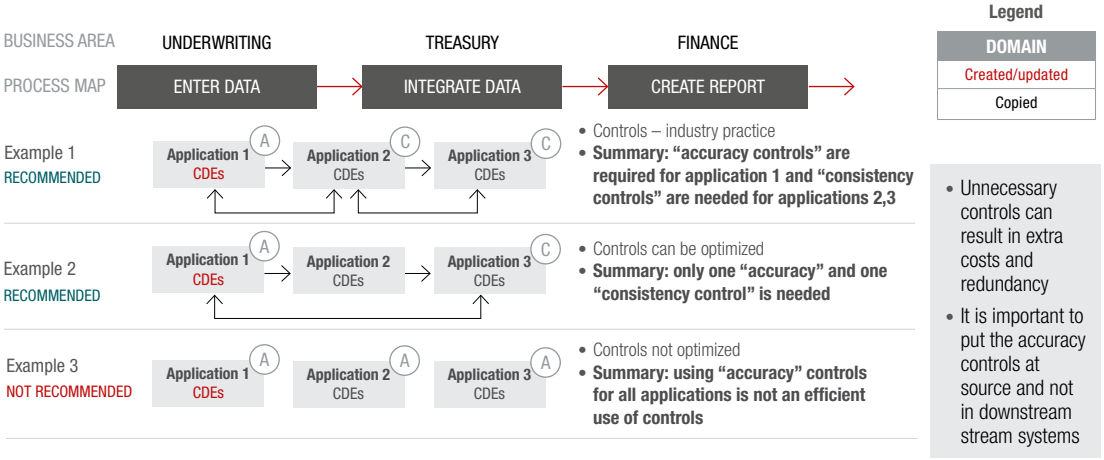
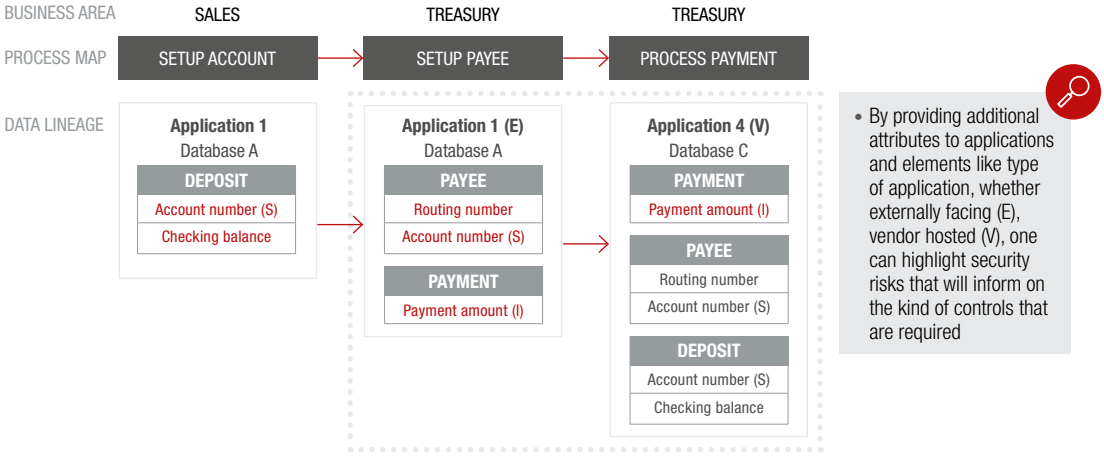


Figure 8: Pattern 4 – using data lineage for designing better security controls



often follow Example 3 and unnecessarily build multiple accuracy controls in downstream systems. A classic example is a downstream system like a data warehouse or a data lake where accuracy controls are built instead of consistency controls. This causes data in the data warehouse to be different from the system of origin, thereby creating a maintenance nightmare and leading to increases in cost. Data lineage will help point the right application for accuracy controls and the right application for consistency controls. If controls have been already implemented, it will highlight the redundancy allowing the organization to optimize on controls and save costs.

Within an organization, as the data flows from one application to the next, there is an inherent security risk. The applications that have the most vulnerability are the ones that are external facing or vendor hosted. As shown in Figure 8, data lineage can highlight applications that are externally facing, vendor-hosted, or end-user applications. In doing so, data lineage can be used to highlight security risks which in turn inform on controls. Organizations that have healthy security controls in place reduce cyber risk and protect data as it moves across the enterprise.

7. CONCLUSION

Data lineage will continue to evolve; however, its power in helping organizations think about their data, control frameworks, and process optimization, is still to be fully realized. Financial institutions that can successfully leverage data lineage will drive value through cost reduction by removing redundancy and unnecessary manual processing, while simultaneously mitigating risk related to data quality, integrity, and security through better controls. Adopting standards and infusing contextual content will facilitate business-friendly implementation of lineage and spur adoption. In today's world of "big data", lineage provides quantifiable business value for organizations in their journey towards harnessing data as a source of competitive advantage.

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward.

Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and asset management and insurance. We also have an energy consulting practice in the US. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Hong Kong
Kuala Lumpur
Pune
Singapore

EUROPE
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM

CAPCO