THE DEATH OF TRADITIONAL DATA INTEGRATION

HOW THE CHANGING NATURE OF IT MANDATES NEW APPROACHES AND TECHNOLOGIES

DAVID S. LINTHICUM

LINTHICUM
RESEARCH

> *“It’s not a matter of “if” we’re moving in new directions that will challenge your existing approaches to data integration, it’s “when.” Those who think they can sit on the sidelines and wait for their data integration technology provider to create the solutions they require will be very disappointed.”*

Table of Contents

Executive Summary

There is a data-related technology crisis that is almost upon us. Just look at the changing nature of IT, and how we share information from system to system, and from systems to humans. There is the exploding use of cloud-based resources, and database technology built for specific purposes, and the proliferation of devices that now produce and consume information at gigabytes per second. This only begins the conversation.

The patterns of integration are becoming more complex. Information externalized from existing and emerging systems ranges from complex behaviors bound to data that must be dealt with in very specific ways, to simple structured and unstructured data, and all points in between. The patterns of information change so quickly that existing data integration technologies will soon find an insurmountable divide between the emerging needs and the existing approaches to data integration.

Integration, in terms of problem and solution patterns, has not changed a great deal in the last 20 years. Many of today's approaches and technologies function much the same as they did in 1997. While the functionality has increased and the prices have dropped, the future gap between the requirements upon this traditional technology and what's actually on the market will be significant.

What you think you know and understand about integration is about to be tossed out the window. A new breed of technology providers will try to match the capabilities of newer platforms with newer integration strategies, approaches, and technologies. The good news? These more modern approaches and technologies will have the best chances of meeting your future integration needs. Indeed, they will have to meet your needs, because traditional data integration technology will soon be a thing of the past, given the changing nature of information technology.

So, what do you need to know? This report summarizes the changes that are occurring, new and emerging patterns of data integration, as well as data integration technology that you can buy today that lives up to these new expectations.

Conclusions reached in this paper include:

- New approaches to managing data, as well as the rapid growth of data, make traditional data integration technology unusable. Data no longer only comes in rows and columns, or semi-structured / unstructured hierarchical formats. It's no longer easy to predict. Thus, there is a need for late binding, declarative approaches, or the ability to determine the schema when reading the data. It often remains where it is, and can't be changed to accommodate emerging use cases. Modern data integration technology must be prepared to deal with these new changes.

- Cloud computing is turning enterprise IT into complex and distributed systems that span existing data centers, to public clouds. The use of the cloud changes the game, in terms of how data is be leveraged, including the mandate to leverage data where it exists, how it exists, and bring the data together into the right context for the business.

- The rise of services, and, now, microservices, changes the game, in terms of how we leverage and manage data. These services are the new dial tone for cloud computing, and are appearing within the enterprise as well. Data services are services married with data, and they will be the most common mechanism for accessing data as we move forward. Therefore, data integration technology must layer in service directories, service governance, and service identity-based security.

- Now is the time to reinvent your enterprise around these trends. Those who don't understand the strategic value that a new approach to data integration will have in the emerging world of computing within the next several years will end up caught without the technology they need to be successful. Those who foresee this event can learn to leverage their data assets for more strategic purposes, and thus provide a great deal more value to the business.

The Evolution of Data Integration

The evolution of integration began in the middle 1990s with the Enterprise Application Integration movement. This architectural pattern was a response to enterprise systems, such as SAP and PeopleSoft, which were beginning to appear in data centers and needed to synchronize information with other systems within the enterprise.

Older patterns of data and application integration are fairly simple to understand. Information is extracted from the source system(s), changed in structure and content (typically), and then placed in the target system(s). This usually occurs around events, such as adding a new customer in the accounting system, or updating the current status of inventory. This was a simple approach to data integration, a response to a very simple problem pattern.

In Figure 1, this is the "Initial Integration Technology / EAI" era, where the concept of EAI (Enterprise Application Integration) was introduced as both an approach and set of enabling technologies that could provide a simple solution to the lack of real-time data integration technology, which meant that data had to be re-keyed in order for it to be replicated from system-to-system. This older EAI technology still exists today, but newer generations of data integration technology have begun to appear since 2008/2009.

In the last 4 to 5 years (2009 – 2014), the focus has been on leveraging existing integration technology, traditional and not, to provide capabilities that included: Data replication, semantic mediation, data cleansing, and mass data migration (see Figure 1). These technologies are leveraged within enterprises, and even between enterprises, and have had reasonable success, considering the use cases and the state of enterprise technology.

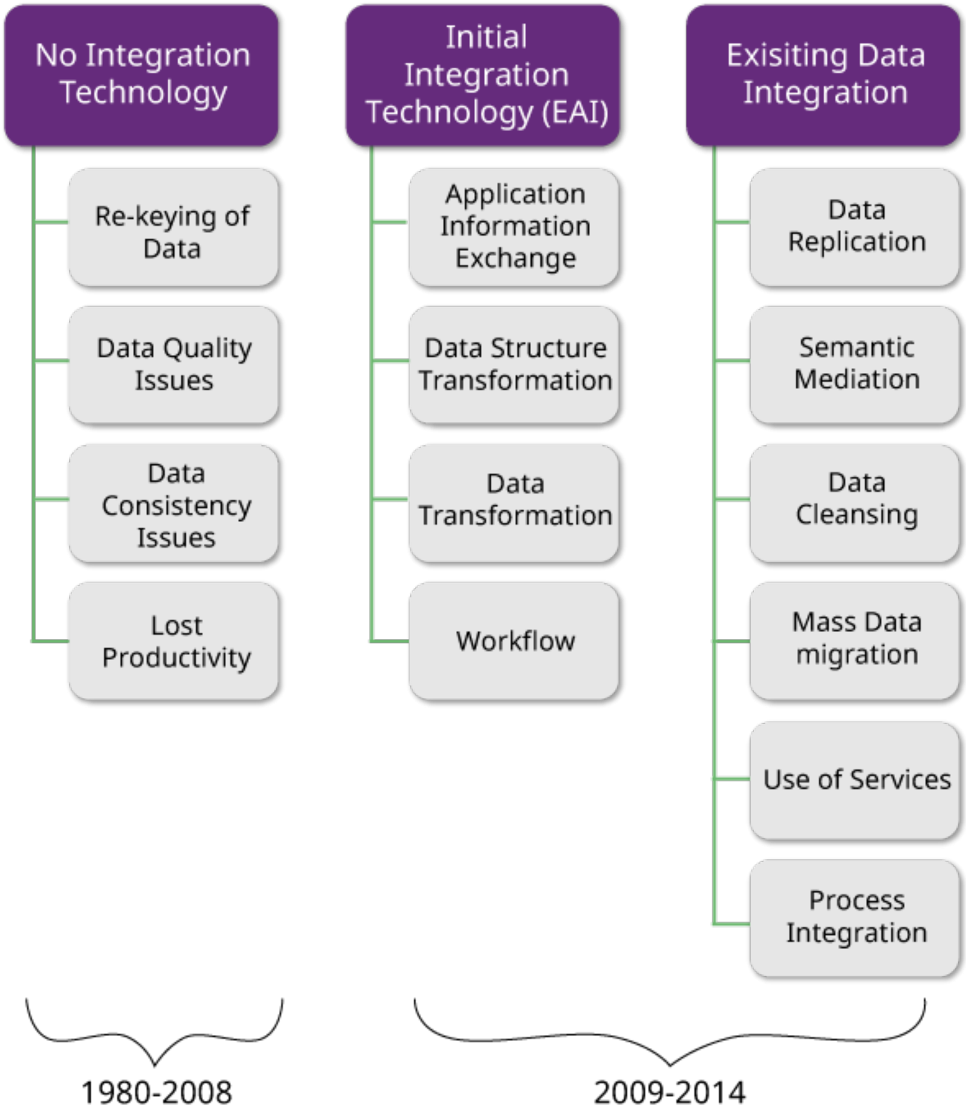


Figure 1: Timeline of major technology shifts from 2010 to 2020, and what integration technology needs to provide around those shifts.

The increase in existing approaches to data integration (labeled "Existing Data Integration" in this diagram) aligns with the use of public and private cloud-based systems, as well as the exploding sizes and number of data stores. The end result is the clear understanding that existing approaches to data integration won't meet future needs as the use of technology continues to change. Drastic measures must be taken now to prepare enterprises for the arrival of this technology, and to position enterprises to take full advantage.

> *"Older approaches to data integration, as well as older technology, can no longer provide the value that they once did."*

Additionally, there is a growing demand for users to do the integration work themselves. These self-service approaches are becoming the norm, now that users see the value in the data and thus desire more access to the data integration layer, as well as the analytical layer. The days of having just a few data integration specialists around are rapidly coming to a close. Indeed, modern integration platform as service (iPaaS) providers like SnapLogic report deployments with over 100 users. These types of deployments were not heard of just a few years ago.

As we consider all of this change, the larger question becomes, what is "Emerging Data Integration?" What's changing today that will drive the new requirements, and new concepts such as data services, microservices, data orchestration, and other technologies that should be understood and leveraged?

Changing Patterns of Information and Integration

Integration is changing around the evolving needs of information processing within the enterprise. For instance, what used to be well-structured data in applications, data warehouses, and other more traditional systems has given way to larger unstructured and structured data stores that may exist inside or outside of the enterprise's firewall.

These changes, or, evolutions, are apparent, in terms of their place in enterprise IT. There is no going back to simpler times when the data was structured, took up much less space, and was loaded on traditional servers within the enterprise data center. Those days are long over. Older approaches to data integration, as well as older technology, can no longer provide the value that they once did.

That said, it's helpful to understand the current path we're on, including the evolving use of technology, such as cloud computing, big data, and the rise of services and microservices. Understanding these evolutions provides insight into what things are changing, and why.

Evolution: The Rise of Cloud Computing

The rise of cloud computing provides a great deal of value within the enterprise. However, cloud-based systems are complex distributed computing platforms, when all is said and done. The integration solution patterns need to accommodate those types of architectures. As you can see in Figure 2, the 2014 survey of IT leaders showed that about 81 percent of those who responded to the survey have some migration efforts to the public cloud ongoing. Just two years ago, it was about a fourth of that number. By 2016, it's likely to have quadrupled.

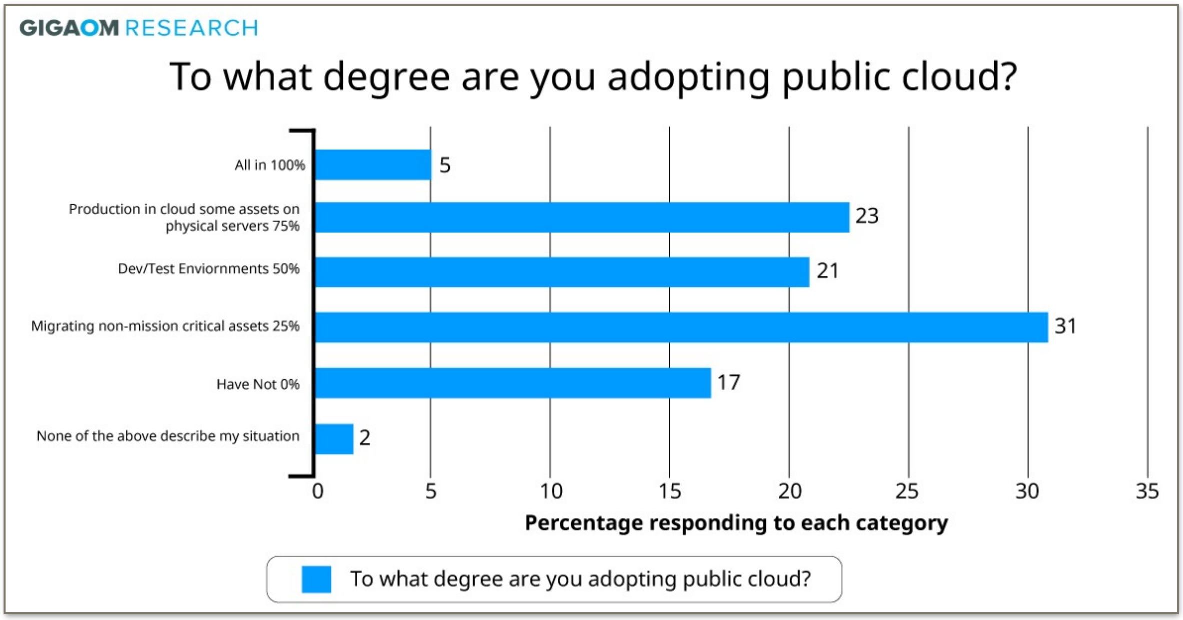


Figure 2: In a recent Gigaom survey, when asked about the degree of cloud adoption, the response showed that most in enterprise IT, about 81 percent, have ongoing migration programs to some public cloud(s). Survey researchers analyzed the inhibitors and drivers behind cloud adoption across a sample of 1,358 respondents.

While most consider the rise of cloud computing and big data as separate notions, they are linked by the changes they drive inside of modern enterprise IT, and will drive for the next 6 to 8 years. These changes include:

The move to manage massive amounts of structured and unstructured data that is physically distributed on traditional systems and new big data systems, as well as on public and private clouds.

- The continued rise of mobile computing, and thus the need to provide direct and reliable integration with these devices, as well as the ubiquitous back-end systems that support them.

- The move to service-orientation, including integration that's occurring at the service and microservices layers. SOA, as an architectural concept, was not as widely accepted a few years ago. The change to architectures that use APIs/services has exploded, which returns new approaches to SOA to favor.

- The move to security and governance systems that are systemic to the entire enterprise, including traditional systems, data, and the private and public cloud.

- The rise of the strategic use of data, including the ability to collect and see all data under management, as well as external data that may put enterprise data into a more understandable context.

- The focus on performance, including the expectation that any needed data will be delivered on demand to an application or user.

Prediction: The total volume of enterprise data is expected to grow at the rate of 50% annually, reaching around 40 Zettabytes by 2020.

IDC

In the next few sections we'll explore some of the general trends in data, the changing requirements around the integration of data, and how the technology needs to shift to accommodate these changing data integration requirements. We will also illustrate why traditional approaches to data integration, and data integration technology, are no longer effective.

Evolution: More Complexity, Less Structure

IDC predicts that the total volume of enterprise data is expected to grow at the rate of 50% each year. By 2020, IDC predicts that the volume of data will reach around 40 Zettabytes (1 billion terabytes equals 1 Zettabyte)[1]. Another important fact about this gigantic amount of data is that 90% of it will be unstructured data.

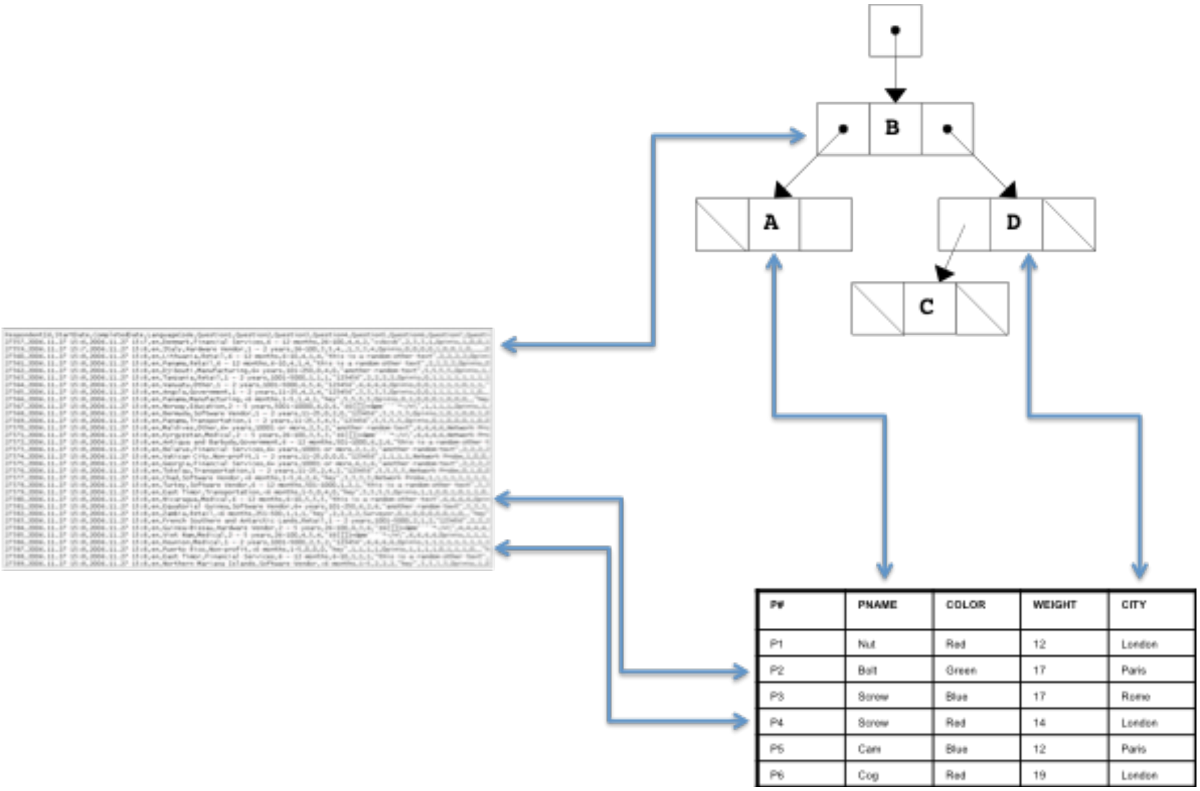
These days, unstructured data is not contained in the simple raw data storage systems from years ago, nor is it all binary data, such as videos or audio. The growth pattern is in unstructured data that

[1] http://www.emc.com/about/news/press/2012/20121211-01.htm

is also complex data. This means that we're dealing with massive amounts of data that's missing metadata. Moreover, that data is typically related to other structured or unstructured data, but those relationships are not tracked within the data storage systems.

Structure is typically applied at runtime, when the data is brought into a temporary structure as it is analyzed in context. It's the role of the integration technology to define a meta-structure for the complex and heterogeneous data, as well as substructures and restructures, for the instances of structured and unstructured data.

Figure 3 depicts the core issues around managing complex and unstructured data. In this case, unstructured data needs to be leveraged with semi-structured and highly structured data. While each data silo was typically created for a single purpose, now the data needs to be leveraged in a much larger context, and brought together to provide true value to the enterprises that own the data.



P#	PNAME	COLOR	WEIGHT	CITY
P1	Nut	Red	12	London
P2	Bolt	Green	17	Paris
P3	Screw	Blue	17	Rome
P4	Screw	Red	14	London
P5	Cam	Blue	12	Paris
P6	Cog	Red	19	London

Figure 3: As data grows, enterprises are finding it's a mix of structured and unstructured data, and there are relationships that need to be derived between data stores. This is where traditional data integration approaches and technology fall down.

Figure 3 is a relatively simple example. In many cases, there are hundreds, perhaps thousands of data stores that employ structured, semi-structured, and unstructured data. What's more, many of these data stores may exist outside of the enterprise on public clouds, or in distributed file systems, such as those employed by Hadoop, or other emerging big data technologies.

In many instances, new data integration approaches are the better choice, especially if they are built to take advantage of concepts such as late binding, and the ability to declare schema(s) when reading the source data vs. having schemas defined prior to the read. This works around the issues of dealing with a large amounts of data that, these days, does not support a native hierarchical structure, including most unstructured data. What's more, modern technologies that handle this type of data to support applications are accustomed to declaring a structure upon access, and not having databases that are dependent upon a structure.

Evolution: Mass Data Storage

Another clear trend is the growth of mass data storage. As outlined above, the data is growing at more than 50 percent per year, both in the cloud and within traditional data centers. This growth is largely spurred by the strategic use of data, including the growth of big data systems using noSQL databases, Hadoop, Spark and other technologies. Additionally, the growth of multi-media data such as video and audio files, as well as binary images of documents and documents, spreadsheets, etc., all of which bring value to the enterprise.

> *"The clear reality today is that mass data storage is something that enterprises must manage, along with integration services. This is something that most traditional data integration technologies are ill-prepared to do."*

The clear reality today is that mass data storage is something that enterprises must manage, along with integration services. This is something that most traditional data integration technologies are ill-prepared to do. At issue is the sheer volume of the data, and thus the ability to effectively process it. Traditional approaches to application and data integration focused on simple extraction of data, and then the transformation of data, in terms of structure, so the data appears to be native to the target. This process was not designed for the massive amount of data that is presently stored and managed. Indeed, this approach to integration won't be able to keep up with the volume of data that needs to be brought together to obtain the value of the information. Once again, the inability to support late binding slows things down considerably.

Another approach to integration – extract, transform and load (ETL), was designed to deal with larger volumes of data, typically in support of traditional data warehouses and data marts. ETL was designed around batch data processing, typically moving and transforming large amounts of data from one data store, such as a transactional database, to another database, such as a large relational database that is used as an enterprise data warehouse (EDW).

However, traditional ETL tools only focused on data that had to be copied and changed. Emerging data systems approach the use of large amounts of data by largely leaving data in place, and instead accessing and transforming data where it sits, no matter if it maintains a structure or not, and no matter where it's located, such as in private clouds, public clouds, or traditional systems. JavaScript Object Notation (JSON), a lightweight data interchange format, is emerging as the common approach that will allow data integration technology to handle tabular, unstructured, and hierarchical data at the same time. As we progress, the role of JSON will become even more strategic to emerging data integration approaches.

Evolution: Rise of Services

Considering that data is becoming both less structured and more complex, and that we're largely consuming data where it sits, many enterprises are looking to provide service-based access to core enterprise data, on premises or within public clouds.

Leveraging services, specifically data services, allows enterprises to provide well-defined access to structured and unstructured data. Data services are able to define structure within the services themselves, and thus read and write unstructured data without requiring that a structure exist within the source or target databases.

These services are typically built on top of existing data stores using any number of tools, and then the services are managed using a service management and governance layer. Figure 4 depicts a higher-level architecture, including the use of data services.These data services sit on top of existing data stores, providing interfaces into the data as we previously described.

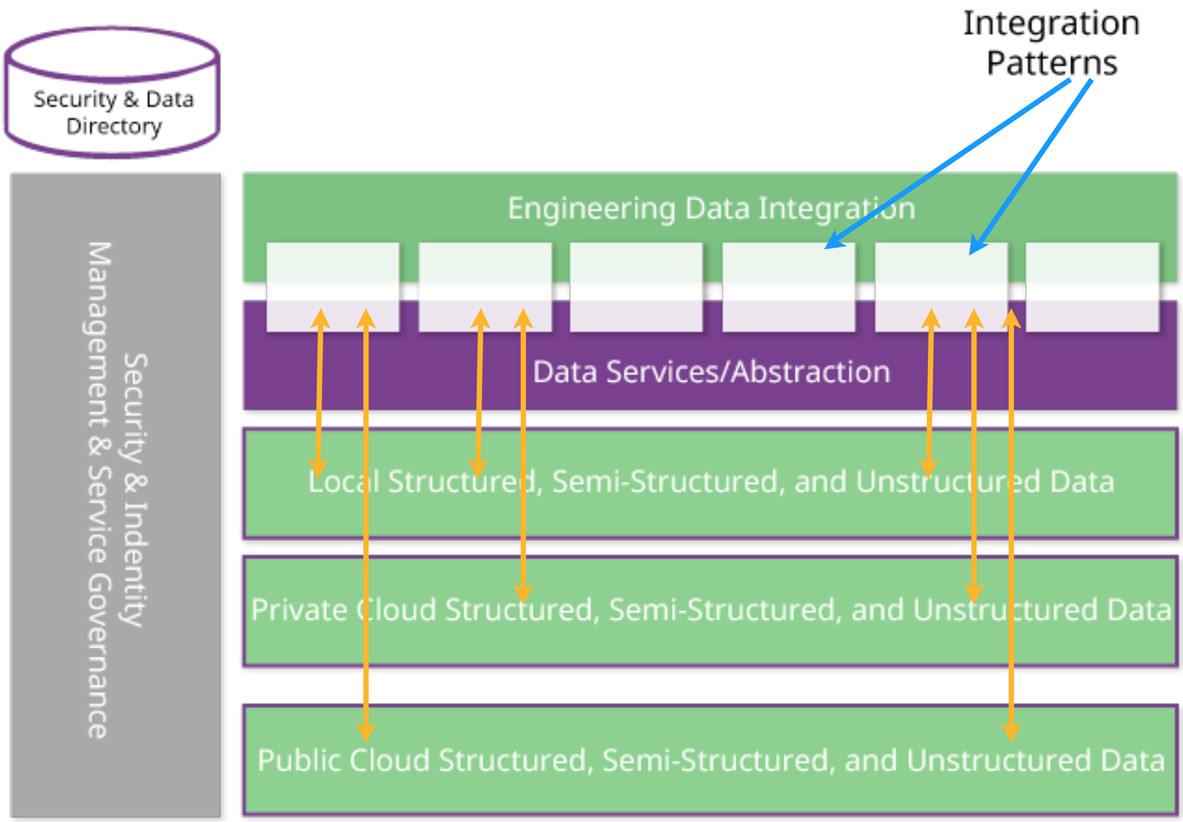

Integration Patterns
Security & Data Directory
Security & Indentity Management & Service Governance
Engineering Data Integration
Data Services/Abstraction
Local Structured, Semi-Structured, and Unstructured Data
Private Cloud Structured, Semi-Structured, and Unstructured Data
Public Cloud Structured, Semi-Structured, and Unstructured Data

Figure 4: Data services sit in front of structured and unstructured data, no matter where the data resides. Data services provide well-defined access to the underlying data, and, in most cases, define both the structure and the interactions with the data for the data consumer. Emerging data integration approaches and technologies need to consider these new usage patterns.

As data services and the use of data become a larger part of enterprise systems, data integration technology will need to adapt. While most data integration approaches and technologies available today can certainly process data services as a point of access to data, the emerging use cases will require a much wider degree of integration patterns that define access to data inside or outside of the enterprise.

For example, common patterns, such as data replication, data transformation, data quality processes, etc., will certainly continue to be in place. However, core to this service-oriented approach to integration is the use of services and microservices that are built and defined inside of a service directory (a.k.a., Service & Data Directory) that provides a centralized location to maintain information about these services, including location, invocation procedures, policies that define use, and other information that needs to be tracked.

Along with the information on services, is information on the data that the services are abstracting. This means metadata, including location, policies, relationships, and other information that is available and should be maintained about the data. As data becomes more distributed and heterogeneous around the rise of cloud computing and big data, these directories become even more important, and should be known and possibly even maintained by the data integration technology layer.

Evolution: Non-Persisted Data Streaming, Device Native Data, and Data Encryption

Another evolving aspect of enterprise IT is the increasing use of data streaming. **Data streaming** means that data is being sent to a data consumer from a data source, and the data is continuous and typically not persisted. This approach to data consumption has grown in popularity, as those charged with building and deploying business intelligence systems rely upon data streaming technology to gather data in real time, in support of operations, where processing a continuous data stream provides more immediate business value.

Device native data is around the mobile computing movement, as well as the rise of the Internet of Things (IoT). Devices have their own data storage structures that are native to the devices. The ability to deal effectively and directly with those data stores becomes a path to the efficient use of those devices. Thus, emerging data integration approaches must deal with the approaches and the mechanisms that devices or other machines utilize to store and manage data, including cell phones, tablets, MRI machines, industrial robots, surveillance drones, etc. This area is experiencing rapid growth, and will likely continue to grow in the future.

> *“Data integration is getting a reboot, and new players are likely to replace older players. Get a handle on that trend, now, and you'll be fine.”*

Finally, the need for **data security** has led many enterprises to take an “always encrypt” approach to data at rest, and in flight. Data integration technologies have not done a stellar job of supporting these approaches, and some of the problems include issues with performance. Modern data integration approaches and technology need to understand that data security, as well as data governance, should be systemic to most data integration activities, with information never being exposed to intermediaries, including cloud providers, unless they have the proper encryption keys.

Changing Data Integration Requirements

Considering the evolutions listed above, and what they mean to your enterprise, it's time to think about the next generation of data integration technology and what it looks like. The list of features and functions needed to solve the emerging data integration problems is massive. However, there are a few focus areas that will define the more innovative paths for this technology.

Referring to Figure 5, with cloud computing's use of services, big data, and other technology, clearly the world of integration will evolve as well. Let's call this "Emerging Data Integration" technologies, which will take us beyond the limitations of existing data integration approaches and technology. This includes adding or expanding integration capabilities such as:

- Intelligent Data Service Discovery
- Data Orchestration
- Data Identity
- Microservices
- Data Economy

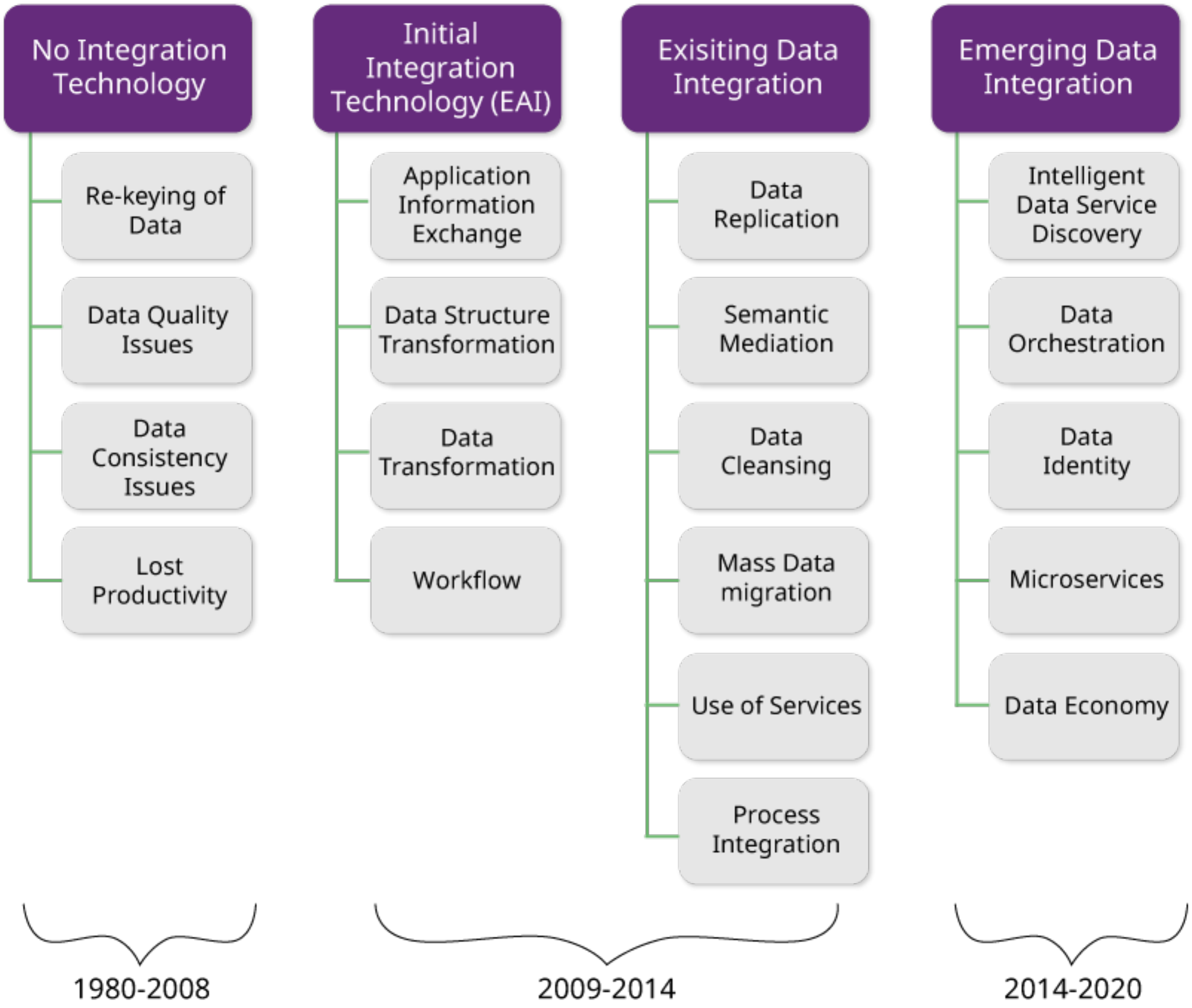


Figure 5: As data integration evolves to meet the changing needs of the enterprise, there are several new features that will become mandatory for modern data integration technologies over the next several years.

Intelligent Data Service Discovery refers to the ability of the data integration technology to automatically find and define data services that are becoming the primary mechanism for consuming and producing data from existing cloud and non-cloud systems. This means that we can discover and re-discover which data services exist within the enterprise, and, more importantly, which come from public clouds, noting where they are, what they do, and how to access them. Enterprises will leverage this tool as a way to understand all available data assets, which provides the ability to leverage the most meaningful data assets to support core business processes, owned or rented.

Data Orchestration refers to the ability to define how the data interacts together to form and reform solutions. Much like service orchestration, this defines composite data points, perhaps combining sales and customers, to form new data services that can be leveraged inside or outside of the enterprise. This allows those who leverage the data a greater degree of control over what the data means for each application view, keeping the physical structure and data content intact. This is important in the new world where data must largely remain in place, using whatever natural structure or lack of structure that exists.

Using data orchestration for data integration, those in enterprise IT can keep volatility, or, the ability to change things, within a domain. As data environments continue to become more heterogeneous and complex (see Figure 3), there will be little desire to bind these systems to tightly coupled integration flows, or even leverage them from composite applications. Instead, data orchestration layers are able to move data from place to place, structured or unstructured, without requiring that an application be written to bind the data together. Changes can occur within the configuration or orchestration layer, and typically not within the physical databases or applications that leverage them.

Data Identity refers to the ability to link data, both structured and data instances, to humans or machines. You can control who or what can consume the data, and see the contents. This makes living up to ever-changing and expanding regulations, and even internal data security policies, much easier to manage. The data containers control access to the data, which is set within the data. This becomes a common mechanism that spans enterprises, and public cloud providers.

Microservices is a software architecture design pattern in which complex applications are composed of small, independent services that provide well-defined APIs for integration and communications. Microservices are small, highly decoupled, and focus on doing a small task.

For instance, the ability to create an application that performs complex risk analytics. Microservices would provide data access, risk calculation, formation of results, and presentation of results. Derived from the world of service-orientation and service-oriented architecture, microservices are designed for a specific usage, and are typically well-defined, so they can be leveraged by many applications.

Data Economy refers to the use of data that is typically not owned by those using the data. Enterprises have access to more data in the second decade of the 21st century than anyone could have imagined just 10 or 20 years ago. From traditional data sources, like corporate databases and applications, to non-traditional sources, like social media, mobile devices and machines outfitted with data-generating sensors, data volumes are exploding and show no signs of abating. [2]

The data economy means that we'll be able to leverage massive amounts of new data, either through a subscription service, or even free-of-charge. This data provides access to patterns of information that may help an enterprise, government agency, or even a private citizens understand their data better, in context of data provided by the data economy, or even leverage that data for advanced analytics.

For instance, the ability to determine the likelihood that a product will succeed based upon trending keywords on Twitter. Or, perhaps the ability to determine sales for the next few years by looking for dependencies in economic data provided by the government, for example, how the rise of new home sales can be determined by key economic indicators, which also determine demand for gardening equipment. These predictive analytics require a great deal of data, mostly data which exists within the data economy, as well as data which resides in owned systems.

Call to Action

It's not a matter of "if" we're moving in new directions that will challenge your existing approaches to data integration, it's "when." Those who think they can sit on the sidelines and wait for their data integration technology provider to create the solutions they require will be very disappointed. Indeed, the likely case is that your legacy data integrating provider may not have viable technology to take them into the next generation, and thus they may join the world of dead technologies, as enterprise IT progresses too fast for them to keep up.

[2] http://wikibon.org/wiki/v/The_Data_Economy_Manifesto

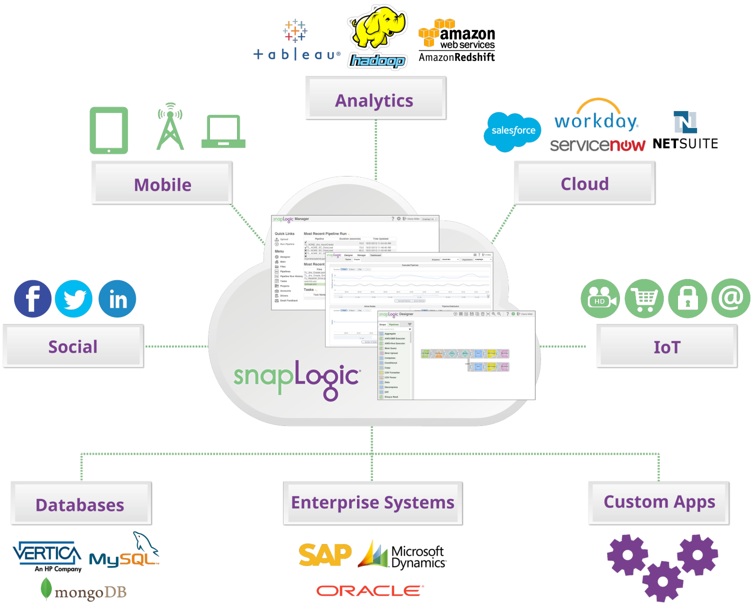
Everything you currently understand about data integration is changing. Assume that your traditional technology provider will soon be dead. Assume that the strategic use of technology and data will begin to provide even more value to most enterprises, and hopefully create a sense of urgency that things need to quickly change. Data integration is getting a reboot, and new players are likely to replace older players. Get a handle on that trend, now, and you'll be fine.

About the Author

David S. Linthicum is an internationally recognized industry expert and thought leader in the world of cloud computing and the author or co-author of 15 books on computing, including the best-selling Enterprise Application Integration, and his latest book, Cloud Computing and SOA Convergence. He is a blogger for InfoWorld, Intelligent Enterprise, eBizq.net, and Forbes, and he conducts his own podcast, the Cloud Computing Podcast. His industry experience includes tenure as the CTO and CEO of several successful software companies, and upper-level management positions in Fortune 100 companies. In addition, Linthicum was an associate professor of computer science for eight years and continues to lecture at major technical colleges and universities.

About SnapLogic

SnapLogic connects enterprise data and applications in the cloud and on-premises for faster decision-making and improved business agility. With the SnapLogic Elastic Integration Platform, organizations can more quickly and affordably accelerate the "cloudification" of enterprise IT with fast, multi-point and modern connectivity of big data, applications and things. Funded by leading venture investors, including Andreessen Horowitz and Ignition Partners, and co-founded by Gaurav Dhillon, co-founder and former CEO of Informatica, SnapLogic is utilized by prominent companies in the Global 2000. For more information about SnapLogic, visit www.SnapLogic.com.



+ a b l e a u
hadoop
amazon web services
Amazon Redshift
Analytics
Mobile
salesforce
workday.
servicenow
NETSUITE
Cloud
Social
HD
IoT
snapLogic
Databases
Enterprise Systems
Custom Apps
VERTICA An HP Company
MySQL
mongoDB
SAP
Microsoft Dynamics
ORACLE