CHAPTER

ARCHITECTING TO DELIVER VALUE FROM A BIG DATA AND HYBRID CLOUD ARCHITECTURE

3

Mandy Chessell[*], **Dan Wolfson**[†], **Tim Vincent**[‡]
[*]*IBM, Winchester, Hampshire, United Kingdom* [†]*IBM, Austin, TX, USA* [‡]*IBM, Toronto, Ontario, Canada*

3.1 INTRODUCTION

This chapter describes the enterprise architecture implications of making extensive use of big data and analytics. It is based on our experiences over the last four years of working with a variety of organizations, both large and small, from multiple industries that have all wanted to derive value from analyzing big data. It summarizes our observations of the different architectures they have employed, extracting what has been successful and why.

Ultimately, these solutions must deliver value to the organization. This could be in the form of better customer service, new products and services, better use of resources, and reduced risk. We have seen examples of ingenious analytics on big data that never moved beyond a proof-of-concept because it was impractical for one reason or another. So we will focus on the practicalities of a solution's operation and evolution in our evaluations. In particular:

- How easy is it to acquire the necessary data and build the analytics?
- How easy is it to deploy the analytics into the appropriate processes that then deliver value to the organization?
- How easy is it to gather data on the effectiveness of the analytics so that it can continuously be improved?

Most big data solutions interface with the real world and therefore need to adapt to their changing environment. They are also still pretty experimental, as their developers explore the different possibilities with the data they have. So a part of any big data architecture includes an aspect of agility in the way analytics and data are used. We call this agile process the "Analytics Lifecycle."

3.2 SUPPORTING THE ANALYTICS LIFECYCLE

The analytics lifecycle is the process that a person or team undertakes to develop new analytics. Fig. 3.1 is a simplified version of the CRISP-DM method [1], which is an industry standard analytics development method that supports the analytics lifecycle. There are three high-level phases.

Software Architecture for Big Data and the Cloud. DOI: 10.1016/B978-0-12-805467-3.00003-X

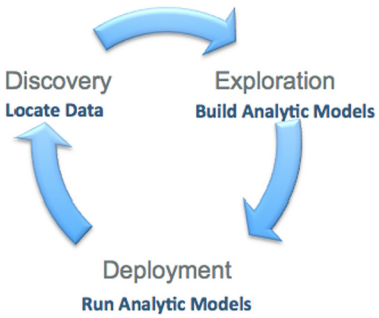

FIGURE 3.1

The analytics lifecycle

Discovery is the process of identifying the data that is potentially useful to feed the new analytics model. Often it involves searching for potentially useful data sets and acquiring them in a form that the analytics tools can operate on in the exploration phase. Typically, this means it can be transformed by the analytics tools without impacting the original source and includes a history of the values as they have changed over time.

Exploration is the iterative process of understanding the patterns in the data and building the analytics implementation to produce the new insight. This may involve further transformation and integration of new data followed by repeated execution of data queries and candidate analytic algorithms until the desired results are achieved.

Deployment is the process of taking the analytics implementation and integrating it with the data in a system that will bring new value to the organization.

All three phases of the analytics lifecycle have their challenges. Discovery often involves battling with the challenge of locating potentially useful information, getting permission to use it, and then getting reliable, on-going access to this data. Exploration is focused on understanding, correlating and identifying where the useful patterns in the data are located and how they can deliver value to the business. However, for many organizations, the major stumbling point in their big data projects is deployment. In simple terms, the analytics has to be deployed where its data is available and there is an opportunity to take action and record the result for future refinement of the analytics. The desire for real-time execution of analytics is expanding the scope of the deployment step beyond the analytics development environment to include the integration of the analytics into operational systems.

- The target operational system may require a different data structure to the one used to develop the analytic model.
- The analytic processing may need to be broken into a number of pieces that are deployed into different places and run at different times in order to access the appropriate data. The results must then be reunited to achieve the overall analytic result.
- Then there is the logic to take action on the resulting insight. This may be an automated action, an alert to a person, or new data displayed on a screen.

- The target operational system needs to collect data around the use of the model to provide the analytics teams with data for on-going validation and training of the model to absorb new data elements and adjust the models as the world evolves.[1]
- Finally, operational systems typically have stringent service level agreements around availability and integrity, which means that extensive quality assurance must take place before new function can go live.

This means the initial deployment of analytics on big data might be a fully-fledged software IT project with all of the checks and steps that implies.

Once the analytics is live, it will need to be constantly verified and improved to maintain its effectiveness as the business environment evolves.

Part of the original deployment of the analytics should have also included the collection of instrumentation data for the analytics. This comprises the data passed to the analytics, the results of the analytics execution, and the outcome of acting on this recommendation. The data scientists will use this instrumentation data to assess the effectiveness of the analytics and improve it as necessary. If the initial deployment is done well, refreshing the analytics implementation should be a routine process once the new analytics have been tested.[2]

Due to the workload generated by the development of new analytics, both the discovery and the exploration phases are typically supported by a specialized analytics system that stores copies of data from many sources. New data is constantly fed into its data stores, as it is generated by the organization, enabling the analytics developers to discover, review, select, and explore data from across the organization as they build their analytics. The specialized analytics system may also support the deployment of analytics.

For big data, where the volume and variety of data needs cheap storage and a flexible processing environment, the specialized analytics system is called a data lake and it is typically implemented using Apache Hadoop technology.

3.3 THE ROLE OF DATA LAKES

A data lake is an analytics system that supports the storing and processing of all types of data [2]. Typically, data from multiple systems, licensed data sets from external partners, and open public data are stored in their original form in the data lake. In the discovery phase, the team building analytic models select data sets from the data lake and potentially supplement this data with new sources.

In the exploration phase, they work with the data in their analytics tools to build new analytic models that can be used to improve operations in the organization. The analytic models may be used once to answer a specific question, deployed into the data lake itself to execute on incoming data as it arrives to derive new data (insight) or deployed in another system. Fig. 3.2 illustrates the data lake in action.

[1] While this is sometimes called model drift, it is more accurate to call this world drift, because it is the world that is changing, not the model.

[2] For many organizations – A/B or Champion/Challenger testing is an ongoing, never-ending process. They continually look for new and better approaches.

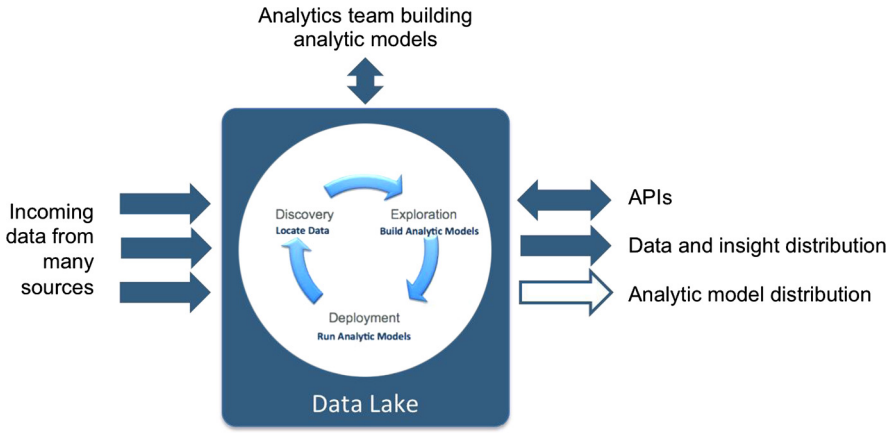

FIGURE 3.2

Interactions with a data lake

In Fig. 3.2, data is flowing in from the left side. Along the top, the analytics teams are working on building new analytic models through discovery and exploration. Some of these models may be deployed into the data lake.

On the right shows the output of the data lake. Systems may access the original data and new insight derived from running the analytic models through APIs. Alternatively, data and insight from the data lake may be distributed to other systems as events or in batch. Finally, the data lake may have distributed analytic models for deployment into other systems.

3.4 KEY DESIGN FEATURES THAT MAKE A DATA LAKE SUCCESSFUL

The data lake approach has received some criticism in recent years, and has even been characterized as a "data swamp" [3]. The root cause of this criticism is that without proper governance and cataloguing of data in the data lake, people are frequently not able to locate the data they need, and even when they do find some potentially useful data, they are not sure where it came from and do not trust it. On the supply side, there is often resistance to add valuable data sources to a data lake because there are no controls on how it will be used.

In our work with clients around data lakes, particularly in regulated industries, we have extending the notion of a simple data lake with metadata management and governance. This architecture is published under the title of a "data reservoir" to highlight that it is a managed data lake [4–6]. The aim is to create an ecosystem where there is trust both to share and consume data.

One of the key differences in the managed data lake is that it potentially includes multiple data platforms, such as Apache Hadoop, streaming technology, relational databases, and No SQL databases. The aim is to site workload and data on the most appropriate platform whilst governing all of the platforms consistently. The managed data lake is surrounded by services that create a consistent interface

to the data and analytics irrespective of which platform they are deployed on. The result is an environment where business users and data scientist can innovate with data and analytics whilst the IT team is able to take advantage of the latest innovations in technology.

The following components deliver the managed data lake:

- Data repositories that provide organized data sets from many sources.
- A catalog of the data repositories with details information about their content, lineage and ownership. This enables the discovery phase of the analytics lifecycle by helping the analytics team identify and locate the right data to use in their work.
- Support for self-service population and management of sandboxes, data preparation tools and tools for building new analytics. This supports the exploration phase.
- Production level support for the execution of analytics within the repositories. This supports the deployment phase.
- Ongoing exchange of data in and out of the data lake connecting it to the latest sources of data and distributing new insight.
- Operational information governance and data security services to protect and maintain the data within the care of the data lake.

The success of any data lake is largely due to the investment in the catalog and the governance around it because it must become an environment where there is trust and confidence both to the share data and to consume it. Data lakes that operate without this discipline become an ever-increasing collection of duplicated data where no one is sure what is available and so gets their own copy of data from the source systems for each project. The data left lying around in the data lake becomes a cost, security and privacy liability for the owning organization.

That withstanding, the centralized data lake is an increasing popular approach to creating a big data and analytics environment for an organization's general use. However, it is not always the most appropriate approach.

3.5 ARCHITECTURE EXAMPLE – CONTEXT MANAGEMENT IN THE IOT

Our next architecture relates to big data in an Internet of Things (IoT) solution. The example comes from the field of home monitoring for the elderly [7]. The idea is to monitor the activity in an elderly or vulnerable person's home to detect whether they are performing their usual activities, or there is a problem, such as they have fallen and are hurt, or they did not get out of bed, so may be ill. The aim is to provide simple monitoring without a major invasion of their privacy – such as through using cameras.

The solution involves adding sensors to chairs, kettle, bed, front door, bathroom, and other areas that can detect normal activity. The readings from these sensors can be used to determine if the individual living in the home is ok.

Early architectures for IoT big data solutions had all of the data from the sensors being pumped into a central data lake that was responsible for parsing the raw data, making decisions on actions and then sending the commands back to the devices if needed. However, this has proved impractical for a number of reasons:

- The data transfer times between the monitored environment and the data lake makes the control feedback sluggish.
- The monitored environment has no autonomous action if there are network communications issues.
- The data lake processing is complex since it has to understand all of the complexity of the monitored environments. There is a lot of variation in the types of sensors, how they work and the types of data they generate.
- The central data lake processing is fragile since it is impacted by changes in the monitored environments – such as a broken sensor being replaced by a new sensor from a different manufacturer.

Each of these reasons impacts the ability to scale the solution. The central processing also creates concerns over privacy of the individual given the volume of data about their lives that is being transmitted [8].

A more scalable design pushes processing close to sensors. So a small processing box in each home that manages the data from the sensors and outputs status and alerts as required. Any local changes to the sensors in the home are handled by the local system. The local system does not transmit details of every activity in the home – just that there is activity going on – or that something is potentially wrong.

The value of this approach is that details of the physical deployment of the sensors are replaced by meaningful messages such as "no activity detected since time *t*" are shared with the central data lake and the individual has an increased level of privacy because only relevant activity is shared beyond the home.

Processing IoT data close to it origin is becoming the best practice approach for IoT solution design. However, there is still a need to transmit a portion of the data back to a central processing point (such as a data lake) in order to enhance the analytic models, or perform historical analysis.

We have seen this pattern used in smart electrical power grids where the physical deployment of components in the power distribution equipment is so complex and volatile, whilst action must be taken very quickly when problems arise. A similar approach is being adopted for automobile automation.

3.6 BIG DATA ORIGINS AND CHARACTERISTICS

The IoT monitoring case study reminds us that big data is not born in the data lake. The areas of growth in data are broadly grouped into:

- Data coming from sensors that are reporting the state of the environment and the activity around them;
- Unstructured data such as text, audio and video media. This information is generated through social media and other collaboration technology and well as the wealth of document publishing channels that we have today.

However, structured data from operational systems (enterprise data) is still significant in big data analytics because it provides the context where the analytics will need to operate if they are to impact the way that the organization operates.

As we examine different big data architectures it is helpful to group related systems into categories that define the types of big data they are processing.

3.7 THE SYSTEMS THAT CAPTURE AND PROCESS BIG DATA

An organization that is focused on becoming a digital enterprise is typically investing in either or both of the following types of systems [9].

- Systems of Engagement (SoE) – these systems interact with people. They include mobile apps and social media services. They are systems that are dedicated to supporting people in many aspects of their daily life. As a result they generate a lot of data about the activities of each individual and have the potential to understand the interests and needs of these individuals. Analytics are often used to make personalized recommendations to individuals.
- Systems of Automation (SoA) – these systems interact with the environment, using sensors and other physical devices to capture data about an asset or a particular location. They are also called Internet of Things (IoT) systems. The big data from these systems is typically streams of events related to the activity around the location or asset. The system of automation uses this input to understand the situation and to make changes to various controls to correct an issue, or make use of an opportunity. Analytics are typically used to predict the likelihood of particular future events based on recent activity.

In addition, the category called System of Record (SoR) is part of this classification scheme. Systems of Record cover the traditional systems that drive an organization's operation. They are considered to represent a reliable, all but localized view of a particular part of the organization's operation. They may also be the target deployment for analytics. So although they do not produce big data from their core operation, they are relevant to this discussion because:

- Their data provides organization context to big data processing. For example, these systems record the transactions of the business. This is the ultimate gauge on the organization's success.
- The transactions also link to the people, products, assets and requests that drive the business. Often the identifiers of these objects are used in correlation in the big data environment.
- They may be a target deployment system for analytics or the insight generated from the big data processing.
- They may be instrumented or monitored by a process that produces log data that requires big data processing to parse, interpret and act on.

An organization's data lake can then be thought of as a System of Insight (SoI) [9] in this classification scheme. They gather data from a wide variety of data sources and aim to blend them together to create a broader picture of the activity across the organization. These systems generate what we refer to as "data gravity." This mean the presence of such a wide variety of data draws people, such as data scientists, and new applications towards it. It therefore also becomes an environment supporting the entire analytics lifecycle plus a distribution point for data and insights to a wide range of applications.

These system categories are illustrated in Fig. 3.3, along with the types of data likely to be generated by these systems.

Each of these types of systems is both collecting and processing key data that is needed by the organization to operate coherently across all of the channels it uses to connect with its customers, business partners and employees. Thus, there is considerable data flow between these systems.

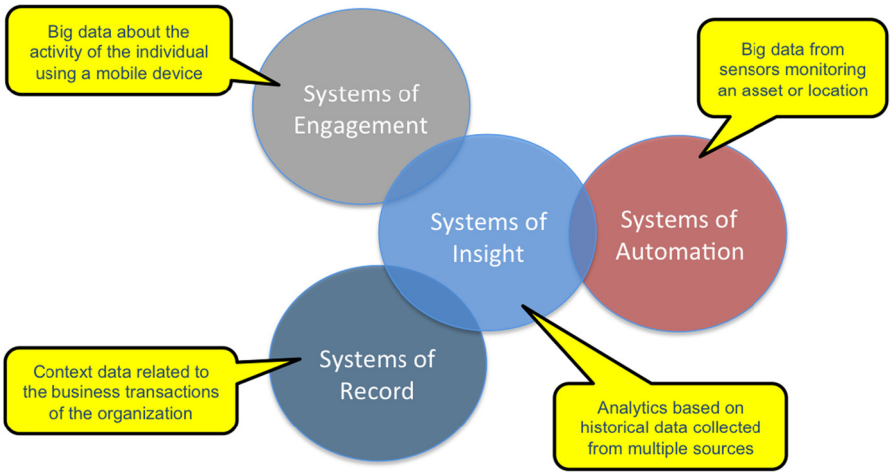

FIGURE 3.3

System categories and the data they produce

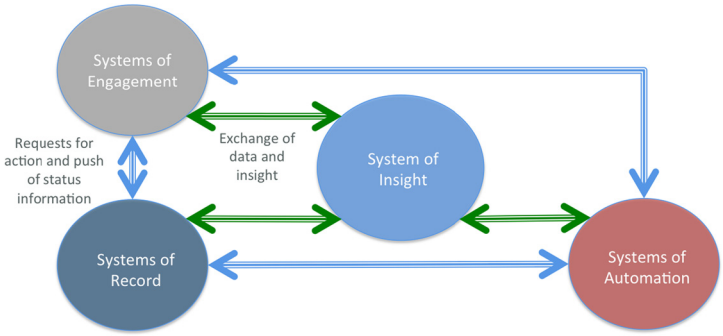

FIGURE 3.4

Interaction between the system categories

In addition, these systems may be hosted either on premise or in the cloud.

In Fig. 3.4, we have added the data integration flows. The blue arrows represent API calls requesting requests to perform an action or service and the green arrows represent bulk movement of the data itself for further processing and analysis. Notice how the data lake (System of Insight) typically acts as a hub for data movement and analysis.

If organizations were flat, egalitarian constructs, then Fig. 3.4 describes an interesting technical challenge to designing data formats and structures that allow each type of system to acquire the data it needs in an efficient form for its processing and run analytic models generated from the data lake. Data would be synchronized between systems as and when it is needed.

The reality is that many organizations are deeply siloed. These silos are designed to divide up the work of the organization into functional units that can be effectively managed. Data is generated by systems owned and operated within these silos. Some big data solutions can be localized within a single silo, combining only its data with potentially data from outside the organization. However, it is more common that a big data solution is aiming to create a coordinated decision making capability for the organization that therefore needs data to flow laterally between silos.

The increasing use of cloud-based services can add to this complexity, creating new technical silos that must be bridged. Cloud-based services extend the technological capabilities of an organization, potentially supporting innovative platforms and functions. Cloud-based services often retain and maintain data that is useful for analytics. The challenge is often in gaining access to that data and combining it with data from other places.

3.8 OPERATING ACROSS ORGANIZATIONAL SILOS

The broader the scope of the big data solution, the more complex it becomes. Not only does it involve crossing the organization's political silos, requiring negotiation and collaboration between people, but also crossing the organization's process silos, requiring the integration of data that has been created with very different assumptions and context.

Blending data from multiple processes as we have already discussed can be complex. Often raw data is full of local, tribal knowledge that makes it unintelligible to external systems. Examples of tribal knowledge include:

- Use of code enumerations to represent valid values for fields. For example, using 0 for "Mr", 1 for "Mrs", 2 for "Miss", 3 for "Dr", 4 for "Sir", 5 for Professor, and so on for courtesy title.
- Inventive use of data fields – for example, using the fifth line of the address for messages to the postman such as "Leave parcel by gate" because the application will then print it on the envelope.
- Local interpretations of common terms.
- Local knowledge on how much to trust the data in the systems and where it came from.
- Local identifiers for people, organizations and shared assets and services.

This tribal knowledge has to be encoded and associated with the data to make it relevant and understandable to an external team using the data. This can be done either by changing the data so it is self-describing, or having metadata that augments the data.

Thus more thought must be given to the way data is acquired and managed to ensure it is processed properly. A similar thought process is required for the insight generated from the data and any resulting action and outcome if it is going to be used outside of the system producing the insight.

Experience also tells us that no organization beyond the control of a single person finds it easy to have a single unifying system for innovation. The nature of innovation is such that it is often opportunistic, slightly maverick, and unplanned. So expecting it to all happen in a single system of insight is unrealistic. As data becomes more important to an organization, multiple silos are likely to develop big data solutions. However, there is clearly value in enabling an organization to act coherently and take advantage of the big data generated from its full range of activities.

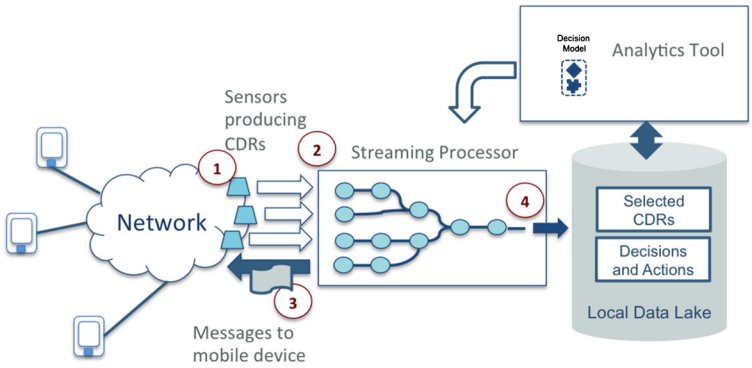

FIGURE 3.5

Mobile device use analysis for a CSP

Combining the need innovate with the need to act coherently – two seeming contradicting requirements – suggests that enterprise needs to accept that there will be multiple solutions using big data for innovation and the enterprise architecture must to account for this.

We will use two examples of next best action solutions that come from the telecommunications industry. Each aims to offer personalized customer service to people as they interact with a service.

3.9 ARCHITECTURE EXAMPLE – LOCAL PROCESSING OF BIG DATA

The first example of a next best action solution only processes the big data generated from a system of engagement [10].

In this example, a real-time streaming engine in a communication service provider (CSP) [11] is receiving Call Detail Records (CDRs) [12] and related data feeds from mobile phones where the subscriber has a pay-as-you-go service. With a pay-as-you-go service, the CSP has no information of their subscribers. The aim of the big data solution is to analyze the activity on each device to build a profile of the person's behavior that is then used to present offers to the individual. For example, the analytics may notice that the subscriber spends a significant time on Facebook, and so an offer could be made for an enhanced package that gives them unlimited, or faster response time on Facebook. The aim is to create a deeper relationship with the subscriber to reduce the chance of them moving to a different CSP.

Fig. 3.5 shows this architecture.

The interaction of the components shown in Fig. 3.5 is as follows:

1. Call detail records and related data that logs the communication activity from the mobile devices are generated by the network.
2. The streaming processor receives this data and parses, categorizes and analyzes the communication from each mobile device.

3. When appropriate, it sends messages back to an individual mobile device with an offer.
4. The streaming processor also logs the results of its analysis and the outcome of any offer, enabling the data scientists to tune the analytics from the local data lake.

The beauty of this type of big data application is that it is self-contained. All of the analysis is on the data from the mobile device, and the resulting action is executed with messages to the device. The project team can focus on understanding the data they are receiving, gaining proficiency with the streaming technology, and developing the analytics that determines the offers. They are likely to only have one stakeholder and the project can be rolled out incrementally as more advanced analytics are developed.

Make no mistake; this is still a challenging project, since the team is handling both volume and velocity of data. However, the variety of data is missing and this is an important simplification.

Our next architecture expands out from this first architecture, to analyze subscriber interaction from many different types of channels and products and take action on the combined results.

3.10 ARCHITECTURE EXAMPLE – CREATING A MULTICHANNEL VIEW

Mature CSPs tend to provide a broad range of services that reflect the development of the telecommunications industry. For example, they may offer landlines for homes and offices, broadband services, as well as mobile phone services. These services typically have one of more contracts associated with them, potentially grouped around a household. The challenge for the CSP is to offer the best package for the contract holder lest they lose all or part of the business to a competitor.

Fig. 3.6 shows a simplified version of the architecture for a multichannel big data solution [13]. The location of the big data solution shown in Fig. 3.5 is highlighted in red and labeled (1) – although some of its plumbing changes, as insight from the CDRs must now be passed to the centralized decision service, and actions that need to be delivered to a particular device are passed back.

The interaction of the components shown in Fig. 3.6 is as follows:

1. Specialized processes receive, parse and extract unstructured data from different sources to understand the behavior of people using each of the different channels of activity. When certain events occur, they trigger an event that is sent to the real-time decisions subsystem (see step 7).
2. The results of the specialized processes are gathered into the Data Lake and where appropriate, onto the Customer Activity system (this maintains a list of recent activity for each customer that is organized for real-time access).
3. The analytics team use data from the system of insight to build analytics.
4. The analytics is deployed to the specialized processes, into the system of insight and/or the real-time decisions subsystem.
5. The activity of the customer generates more data for the specialized processes, and where they interact directly with the CSP's systems, requests for recommendations on offers to make to the customer.
6. The request for a recommendation, either directly from the channel systems, or from an event detected by the specialized processes, results in a call to the real-time decisions subsystem.
7. Inside the real-time decisions subsystem is a Complex Event Processor (CEP) monitoring the events around a customer over different time windows. This may generate additional requests for recom-

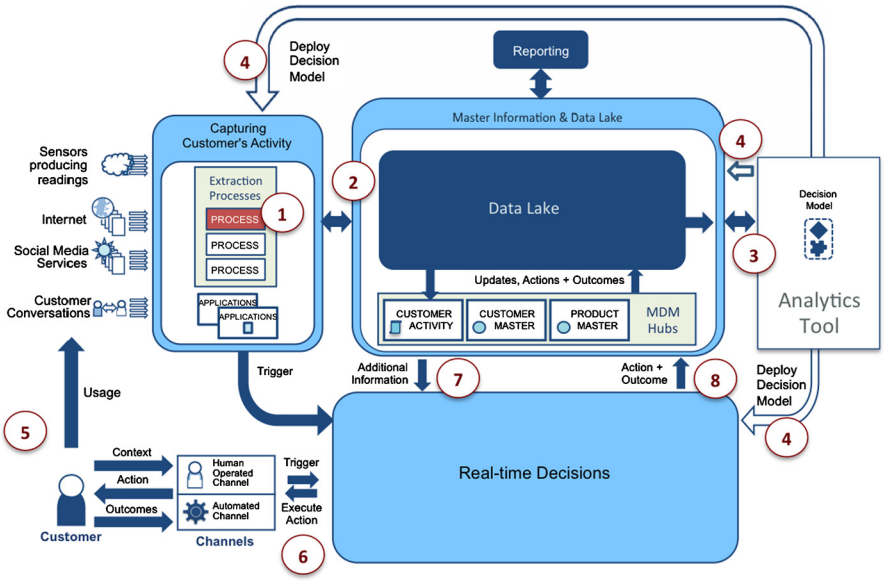

FIGURE 3.6

Multichannel next best action solution

mendations. There are one or more decision engines processing the requests for recommendations. They use the context from the request for a recommendation plus the data from the MDM hubs to make a decision. The resulting recommendation is either returned to the requestor, or sent asynchronously on a different channel.

8. The decision and any outcome are fed back to the data lake.

The striking difference between the architectures shown in Figs. 3.5 and 3.6 is the amount of data movement and data integration logic is required in addition to the analytics. This represents the delta for any big data solution that is processing a variety of data from many different sources. Many of the techniques found in traditional data warehouse solutions become necessary in the big data solution to capture, parse, enrich, correlate, and combine data. In addition, there is infrastructure required to communicate actions back out to the channels and gather results.

Projects of this size often experience delays as teams negotiate access to data and coordinate the rollout of enhancements to systems maintained by different teams. This adds politics to the complexity of the architecture in terms of access to the data, funding of changes required to existing systems and who realizes the benefit of the ultimate value from the actions taken by the big data solution.

In some organizations, the politics are so complex that they make the technical part of the solution seem simple. They are certainly a key factor in whether this type of big data solution is practical for a particular organization.

3.11 APPLICATION INDEPENDENT DATA

There is one type of system in Fig. 3.6 architecture that seems to sit a little uncomfortably within the system categories shown in Figs. 3.3 and 3.4. It is the Master Data Management (MDM) hubs.

An MDM hub is one of a number of reference data systems that greatly simplify the integration of all types of data across an ecosystem and as such is used by all of the different system categories. The MDM hub specifically manages the identifiers and core attributes of key objects that are the focus of the ecosystem. These objects could be about people and organizations (customers, employees, business partners), locations, products and offerings, or assets. These objects are often described in many of the systems producing the data for the big data solution and each system assigns a different identifier to their copies of the object. There are often differences in the attributes assigned to the object in each system as they gather and process the values under different conditions. The MDM hub maintains a registry of these objects, listing the identifiers from each system and the authoritative values for the core attributes. Ideally, the MDM hub is used to synchronize the core values in the other systems so the raw data entering the big data solution is reasonably consistent. Either way, the MDM hub is a key source of information for the big data solution when combining data from multiple systems. Organizations that want to operate complex big data solutions find that an investment in MDM pays dividends in facilitating the matching of data from many sources. They should be thought of as the oil that eases the friction of data exchange between the silos.

Other reference systems that aid the integration if data in addition to the MDM hub are the code hub for reconciling code table values [14] from different systems and the metadata catalogue of systems and data for the ecosystem.

3.12 METADATA AND GOVERNANCE

Metadata is descriptive data about data. In a data warehouse environment, the metadata is typically limited to the structural schemas used to organize the data in different zones in the warehouse. For the more advanced environments, metadata may also include data lineage and measured quality information of the systems supplying data to the warehouse.

A big data environment is more dynamic than a data warehouse environment and it is continuously pulling in data from a much greater pool of sources. It quickly becomes impossible for the individuals running the big data environment to remember the origin and content of all the data sets it contains. As a result, metadata capture and management becomes a key part of the big data environment. Given the volume, variety and velocity of the data, metadata management must be automated. Similarly fulfilling governance requirements for data must also be automated as much as possible.

Enabling this automation adds to the types of metadata that must be maintained since governance is driven from the business context, not from the technical implementation around the data. For example, the secrecy required for a company's financial reports is very high just before the results are reported. However, once they have been released, they are public information. The technology used to store the data has not changed. However, time has changed the business impact of an unauthorized disclosure of the information, and thus the governance program providing the data protection has to be aware of that context.

Similar examples from data quality management, lifecycle management and data protection illustrate that the requirements that drive information governance come from the business significance of the data and how it is to be used. This means the metadata must capture both the technical implementation of the data and the business context of its creation and use so that governance requirements and actions can be assigned appropriately.

Earlier on in this chapter, we introduced the concept of the managed data lake where metadata and governance were a key part of ensuring a data lake remains a useful resource rather than becoming a data swamp. This is a necessary first step in getting the most value out of big data. However, from the different big data solutions reviewed in this chapter, big data is not born in the data lake. It comes from other systems and contexts. Metadata and governance needs to extend to these systems, and be incorporated into the data flows and processing throughout the solution.

3.13 CONCLUSIONS

Deriving value from big data involves processing the right data in the right location and taking action on the results. Innovative data science is only the start of the journey. Big data projects that can process data and act on it close to its origin are more likely to be successful than projects that incorporate systems operated by multiple silos in the organization. This is because they are simpler technically; they meet fewer political hurdles and deliver value while the original stakeholder is still in place. However, it is this second type of project is often addressing the use cases that have the higher value.

As an industry we need to improve the time to value and success rate of big data projects. From a technical point of view, this is going to take:

- Better architecture method around identifying the appropriate systems that support the different types of big data processing needed within a solution.
- Tools and runtimes that automatically manage the metadata and context data necessary to pass data between processing systems.
- Standard structures for this metadata and context data to allow interoperability between cloud services and on premises systems. Hybrid cloud brokers and gateways could then support these standards.

Organizations will also need to rethink their attribute and relationship with data.

Organizations that wish to be data-driven and embark on these broad big data projects need to think deeply about the barriers created by their existing silos, and whether these silos are appropriate for their future digital business. For many, becoming a digital business is going to involve tearing up the current organization chart and organizing around data. This is likely to create new executive roles that bring data oriented skills to the boardroom.

Metadata management and information governance needs a greater focus at the business level. It must be deeply embedded in the systems that are involved in processing data – not just the data platforms associated with a data lake. Metadata must cover both the technical implementation of the data and its processing engines, as well as the business context of where the data was created, its use and the governance requirements associated with it. This metadata then must be an active part of the way data is managed, keeping it up-to-date and relevant to the needs of the organization.

3.14 OUTLOOK AND FUTURE DIRECTIONS

The big data space is still evolving. New types of data platforms and processing engines are appearing with a regular cadence and data-oriented roles, such as Chief Data Officer and data scientist, are in high demand. As cloud adoption grows, we see an increasing amount of data that is born on the cloud. This will increase the demand for big data processing systems to also reside in the cloud.

From a data perspective, these trends spell greater chaos, since the origin and consumption of data increasing occurs in systems operated by different organizations. If it was hard to get data management right within an organization then what chance of getting it right in a multiorganizational situation?

This suggests a fundamental change to IT technology in the way it manages data. A system should treat the data it holds as a sharable resource rather than as its own private asset. This means it need to be described both in a human and machine-readable way and accessible through open interfaces. We need a greater level of standardization in the way that data is described, at multiple levels:

- Cataloguing of the data presents and its structure,
- Ownership and custodian responsibility,
- Business meaning and the rules around its use,
- Levels of confidence in its quality and timeliness,
- Classifications, licensing and regulations around its use.

With systems consistently describing their data in this way, exchanging metadata with data as it is copied between systems becomes much easier. The consumer then receives data accompanied by a rich description of its origin, history and related characteristics.

Technically, this is not difficult to do. There are many metadata standards that we could adopt and the processing overload on a system is not that high. Today it does not happen because metadata is treated as an optional capability – used mainly for documentation. However, if metadata is used to give the business a greater visibility and control over the data stored in their many systems then its value rises and so does the investment in it.

There are vendors who sell metadata solutions to help capture and manage metadata. Tools to provide integration capabilities, reports or virtualization interfaces, sit on top of this metadata and use it to interact with the underlying data sources. So using metadata to drive software capability is not new. It is the lifecycle of this metadata that needs to change.

In today's tools, metadata is created retrospectively, an expensive undertaking – the cost coming from the time of subject matter experts to document the origin, meaning and use of metadata. Each vendor uses its own formats so metadata is only exchangeable with additional metadata bridges and brokers.

As data rises in importance to society, it is time to move metadata from an optional extra capability to an embedded capability that systems maintain by default. Open source is potentially offering us a solution in the new Apache Atlas project [15,16]. It aims to provide an open source implementation for an embeddable metadata management and governance capability.

The metadata standards that Apache Atlas adopts would become de facto standards. So where cloud platforms, big data platforms and tools vendors wish to use their proprietary implementation; they can implement these standards in their interchange code.

So what would be the benefit to a ubiquitous metadata and governance capability? Most importantly, data would become visible and consumable to big data and analytics solutions [17]. This would allow an organization to get greater value from their data. The increased value creates a greater interest from all parts of an organization, government and society in general and the human constraints on big data solutions begin to ease.

REFERENCES

[1] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth, CRISP-DM 1.0, Step-by-step data mining guide, ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.
[2] Data lake, https://en.wiktionary.org/wiki/data_lake.
[3] Gartner says beware of the data lake fallacy, Gartner Press Release, http://www.gartner.com/newsroom/id/2809117.
[4] Mandy Chessell, Ferd Scheepers, Nhan Nguyen, Ruud van Kessel, Ron van der Starre, REDP5120: governing and managing big data for analytics and decision makers, http://www.redbooks.ibm.com/redpieces/abstracts/redp5120.html?Open.
[5] Mandy Chessell, Nigel L. Jones, Jay Limburn, David Radley, Kevin Shank, SG24-8274-00, designing and operating a data reservoir, http://www.redbooks.ibm.com/Redbooks.nsf/RedpieceAbstracts/sg248274.html.
[6] Mandy Chessell, Building a data reservoir to use big data with confidence, http://www.ibmbigdatahub.com/blog/building-data-reservoir-use-big-data-confidence.
[7] Bolzano case study, http://www-03.ibm.com/press/us/en/pressrelease/28465.wss.
[8] A guide to the best practices in ensuring privacy in big data solutions is covered by the "Privacy by Design resolution", https://www.ipc.on.ca/site_documents/pbd-resolution.pdf.
[9] Brian Hopkins, Systems of insight will power digital business, http://blogs.forrester.com/brian_hopkins/15-04-27-systems_of_insight_will_power_digital_business.
[10] Arvind Sathi, et al., Advanced Analytics Platform (AAP), http://www.ibm.com/developerworks/library/ba-adv-analytics-platform1/index.html.
[11] Communications service provider, https://en.wikipedia.org/wiki/Communications_service_provider.
[12] Call detail record, https://en.wikipedia.org/wiki/Call_detail_record.
[13] Mandy Chessell, REDP4888, smarter analytics: driving customer interactions with the IBM next best action solution, http://www.redbooks.ibm.com/abstracts/redp4888.html?Open.
[14] Dan Wolfson, Going with the flow, http://www.ibmbigdatahub.com/blog/going-flow.
[15] Mandy Chessell, Insight Out: the case for open metadata and governance, http://www.ibmbigdatahub.com/blog/insightout-case-open-metadata-and-governance.
[16] Mandy Chessell, Insight Out: the role of apache atlas in the open metadata ecosystem, http://www.ibmbigdatahub.com/blog/insightout-role-apache-atlas-open-metadata-ecosystem.
[17] Tim Vincent, Bill O'Connell, Insight Ops: the road to a collaborative self-service model, http://www.ibmbigdatahub.com/blog/insight-ops-road-collaborative-self-service-model.