Big Data Fundamentals

Simple | Smart | Speed

CHOOLS

This page intentionally left blank

Big Data Fundamentals

Concepts, Drivers & Techniques

Thomas Erl,
Wajid Khattak,
and Paul
Buhler

ServiceTech
PRESS

BOSTON • COLUMBUS • INDIANAPOLIS • NEW YORK • SAN FRANCISCO
AMSTERDAM • CAPE TOWN • DUBAI • LONDON • MADRID • MILAN • MUNICH

PARIS • MONTREAL • TORONTO • DELHI • MEXICO CITY • SAO PAULO

SIDNEY • HONG KONG • SEOUL • SINGAPORE • TAIPEI • TOKYO

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, mar- keting focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact international@pearsoned.com.

Visit us on the Web: informit.com

Library of Congress Control Number: 2015953680

ISBN-13: 978-0-13-429107-9
ISBN-10: 0-13-429107-7

Text printed in the United States on recycled paper at RR Donnelley in Crawfordsville, Indiana.

First printing: December 2015

Editor-in-Chief
Mark Taub

Senior Acquisitions Editor
Trina MacDonald

Managing Editor
Kristy Hart

Senior Project Editor
Betsy Gratner

Copyeditors Natalie Gitt Alexandra Kropova

Senior Indexer
Cheryl Lenser

Proofreaders
Alexandra Kropova
Debbie Williams

Publishing Coordinator
Olivia Basegio

Cover Designer
Thomas Erl

Compositor
Bumpy Design

Graphics
Jasper Paladino

Photos
Thomas Erl

Educational Content Development
Arcitura Education Inc.

To my family and friends.

—Thomas Erl

I dedicate this book to my daughters Hadia and Areesha,
my wife Natasha, and my parents.

—Wajid Khattak

I thank my wife and family for their patience and for
putting up with my busyness over the years.

I appreciate all the students and colleagues I have had
the privilege of teaching and learning from.

John 3:16, 2 Peter 1:5-8

—Paul Buhler, PhD

This page intentionally left blank

Contents at a Glance

PART I: THE FUNDAMENTALS OF BIG DATA

PART II: STORING AND ANALYZING BIG DATA

This page intentionally left blank

Contents

PART I: THE FUNDAMENTALS OF BIG DATA

PART II: STORING AND ANALYZING BIG DATA

This page intentionally left blank

Acknowledgments

In alphabetical order by last name:

- Allen Afuah, Ross School of Business, University of Michigan
- Thomas Davenport, Babson College
- Hugh Dubberly, Dubberly Design Office
- Joe Gollner, Gnostyx Research Inc.
- Dominic Greenwood, Whitestein Technologies
- Gareth Morgan, The Schulich School of Business, York University
- Peter Morville, Semantic Studios
- Michael Porter, The Institute for Strategy and Competitiveness, Harvard Business School
- Mark von Rosing, LEADing Practice
- Jeanne Ross, Center for Information Systems Research, MIT Sloan School of Management
- Jim Sinur, Flueresque
- John Sterman, MIT System Dynamics Group, MIT Sloan School of Management

Special thanks to the Arcitura Education and Big Data Science School research and development teams that produced the Big Data Science Certified Professional (BDSCP) course modules upon which this book is based.

Reader Services

Register your copy of *Big Data Fundamentals* at informit.com for convenient access to downloads, updates, and corrections as they become available. To start the registration process, go to informit.com/register and log in or create an account.* Enter the product ISBN, 9780134291079, and click Submit. Once the process is complete, you will find any available bonus content under “Registered Products.”

*Be sure to check the box that you would like to hear from us in order to receive exclu- sive discounts on future editions of this product.

Part I

The Fundamentals of Big Data

Chapter 1 Understanding Big Data

Chapter 2 Business Motivations and Drivers for Big Data Adoption

Chapter 3 Big Data Adoption and Planning Considerations

Chapter 4 Enterprise Technologies and Big Data Business Intelligence

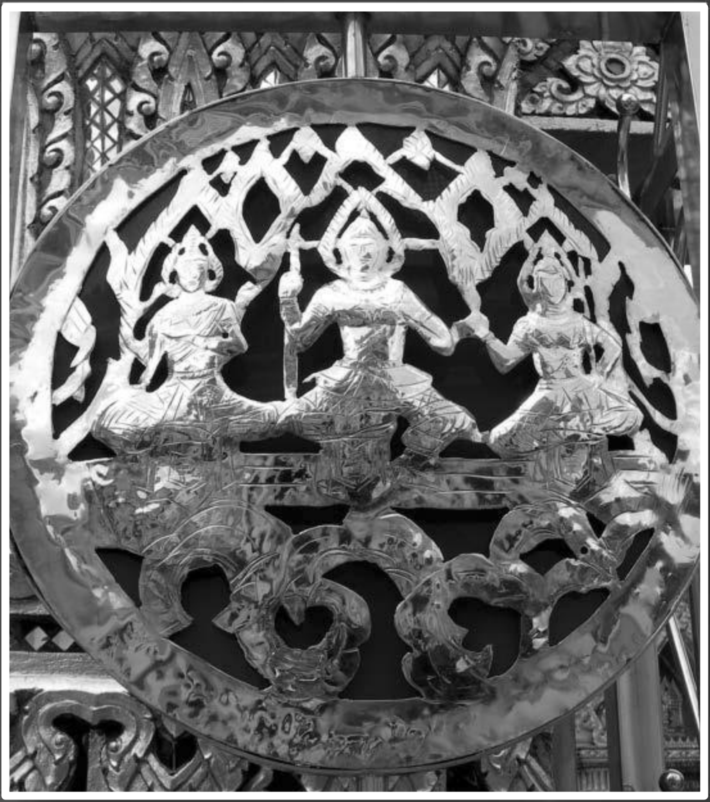
Big Data has the ability to change the nature of a business. In fact, there are many firms whose sole existence is based upon their capability to generate insights that only Big Data can deliver. This first set of chapters covers the essentials of Big Data, primarily from a business perspective. Businesses need to understand that Big Data is not just about technology—it is also about how these technologies can propel an orga- nization forward.

Part I has the following structure:

- Chapter 1 delivers insight into key concepts and terminology that define the very essence of Big Data and the promise it holds to deliver sophisticated busi- ness insights. The various characteristics that distinguish Big Data datasets are explained, as are definitions of the different types of data that can be subject to its analysis techniques.
- Chapter 2 seeks to answer the question of why businesses should be motivated to adopt Big Data as a consequence of underlying shifts in the marketplace and business world. Big Data is not a technology related to business transformation; instead, it enables innovation within an enterprise on the condition that the enter- prise acts upon its insights.
- Chapter 3 shows that Big Data is not simply "business as usual," and that the decision to adopt Big Data must take into account many business and technology considerations. This underscores the fact that Big Data opens an enterprise to external data influences that must be governed and managed. Likewise, the Big Data analytics lifecycle imposes distinct processing requirements.
- Chapter 4 examines current approaches to enterprise data warehousing and busi- ness intelligence. It then expands this notion to show that Big Data storage and analysis resources can be used in conjunction with corporate performance moni- toring tools to broaden the analytic capabilities of the enterprise and deepen the insights delivered by Business Intelligence.

Big Data used correctly is part of a strategic initiative built upon the premise that the internal data within a business does not hold all the answers. In other words, Big Data is not simply about data management problems that can be solved with technology. It is about business problems whose solutions are enabled by technology that can support the analysis of Big Data datasets. For this reason, the business-focused discussion in Part I sets the stage for the technology-focused topics covered in Part II.

This page intentionally left blank



Chapter 3

Big Data Adoption and Planning Considerations

Organization Prerequisites

Data Procurement

Privacy

Security

Provenance

Limited Realtime Support

Distinct Performance Challenges

Distinct Governance Requirements

Distinct Methodology

Clouds

Big Data Analytics Lifecycle

Big Data initiatives are strategic in nature and should be business-driven. The adop- tion of Big Data can be transformative but is more often innovative. Transformation activities are typically low-risk endeavors designed to deliver increased efficiency and effectiveness. Innovation requires a shift in mindset because it will fundamentally alter the structure of a business either in its products, services or organization. This is the power of Big Data adoption; it can enable this sort of change. Innovation manage- ment requires care—too many controlling forces can stifle the initiative and dampen the results, and too little oversight can turn a best intentioned project into a science experiment that never delivers promised results. It is against this backdrop that Chap- ter 3 addresses Big Data adoption and planning considerations.

Given the nature of Big Data and its analytic power, there are many issues that need to be considered and planned for in the beginning. For example, with the adoption of any new technology, the means to secure it in a way that conforms to existing corporate standards needs to be addressed. Issues related to tracking the provenance of a dataset from its procurement to its utilization is often a new requirement for organizations. Managing the privacy of constituents whose data is being handled or whose identity is revealed by analytic processes must be planned for. Big Data even opens up additional opportunities to consider moving beyond on-premise environments and into remotely- provisioned, scalable environments that are hosted in a cloud. In fact, all of the above considerations require an organization to recognize and establish a set of distinct gov- ernance processes and decision frameworks to ensure that responsible parties under- stand Big Data's nature, implications and management requirements.

Organizationally, the adoption of Big Data changes the approach to performing business analytics. For this reason, a Big Data analytics lifecycle is introduced in this chapter. The lifecycle begins with the establishment of a business case for the Big Data project and ends with ensuring that the analytic results are deployed to the organization to gener- ate maximal value. There are a number of stages in between that organize the steps of identifying, procuring, filtering, extracting, cleansing and aggregating of data. This is all required before the analysis even occurs. The execution of this lifecycle requires new competencies to be developed or hired into the organization.

As demonstrated, there are many things to consider and account for when adopting Big Data. This chapter explains the primary potential issues and considerations.

Organization Prerequisites

Big Data frameworks are not turn-key solutions. In order for data analysis and analyt- ics to offer value, enterprises need to have data management and Big Data governance frameworks. Sound processes and sufficient skillsets for those who will be responsible for implementing, customizing, populating and using Big Data solutions are also neces- sary. Additionally, the quality of the data targeted for processing by Big Data solutions needs to be assessed.

Outdated, invalid, or poorly identified data will result in low-quality input which, regardless of how good the Big Data solution is, will continue to produce low-quality results. The longevity of the Big Data environment also needs to be planned for. A road- map needs to be defined to ensure that any necessary expansion or augmentation of the environment is planned out to stay in sync with the requirements of the enterprise.

Data Procurement

The acquisition of Big Data solutions themselves can be economical, due to the availabil- ity of open-source platforms and tools and opportunities to leverage commodity hard- ware. However, a substantial budget may still be required to obtain external data. The nature of the business may make external data very valuable. The greater the volume and variety of data that can be supplied, the higher the chances are of finding hidden insights from patterns.

External data sources include government data sources and commercial data markets. Government-provided data, such as geo-spatial data, may be free. However, most commercially relevant data will need to be purchased and may involve the continuation of subscription costs to ensure the delivery of updates to procured datasets.

Privacy

Performing analytics on datasets can reveal confidential information about organiza- tions or individuals. Even analyzing separate datasets that contain seemingly benign data can reveal private information when the datasets are analyzed jointly. This can lead to intentional or inadvertent breaches of privacy.

Addressing these privacy concerns requires an understanding of the nature of data being accumulated and relevant data privacy regulations, as well as special techniques for data tagging and anonymization. For example, telemetry data, such as a car's GPS log or smart meter data readings, collected over an extended period of time can reveal
an individual's location and behavior, as shown in Figure 3.1.

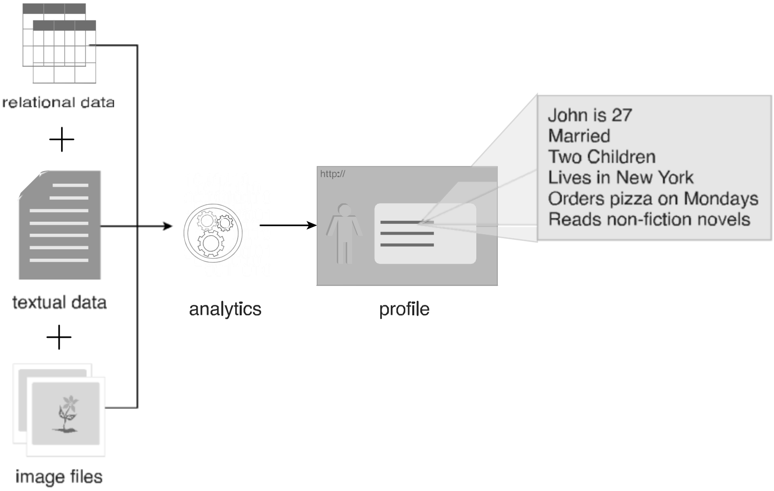


Figure 3.1
Information gathered from running analytics on image files, relational data and textual data is used to create John's profile.

Security

Some of the components of Big Data solutions lack the robustness of traditional enter- prise solution environments when it comes to access control and data security. Secur- ing Big Data involves ensuring that the data networks and repositories are sufficiently secured via authentication and authorization mechanisms.

Big Data security further involves establishing data access levels for different catego- ries of users. For example, unlike traditional relational database management systems, NoSQL databases generally do not provide robust built-in security mechanisms. They instead rely on simple HTTP-based APIs where data is exchanged in plaintext, making the data prone to network-based attacks, as shown in Figure 3.2.

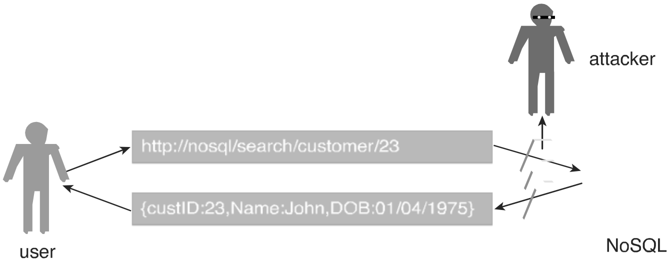

attacker
http://nosql/search/customer/23
{custID:23,Name:John,DOB:01/04/1975}
user
NoSQL

Figure 3.2
NoSQL databases can be susceptible to network-based attacks.

Provenance

Provenance refers to information about the source of the data and how it has been pro- cessed. Provenance information helps determine the authenticity and quality of data, and it can be used for auditing purposes. Maintaining provenance as large volumes of data are acquired, combined and put through multiple processing stages can be a com- plex task. At different stages in the analytics lifecycle, data will be in different states due to the fact it may be being transmitted, processed or in storage. These states correspond to the notion of data-in-motion, data-in-use and data-at-rest. Importantly, whenever Big Data changes state, it should trigger the capture of provenance information that is recorded as metadata.

As data enters the analytic environment, its provenance record can be initialized with the recording of information that captures the pedigree of the data. Ultimately, the goal of capturing provenance is to be able to reason over the generated analytic results with the knowledge of the origin of the data and what steps or algorithms were used to pro- cess the data that led to the result. Provenance information is essential to being able to realize the value of the analytic result. Much like scientific research, if results cannot be justified and repeated, they lack credibility. When provenance information is captured on the way to generating analytic results as in Figure 3.3, the results can be more easily trusted and thereby used with confidence.

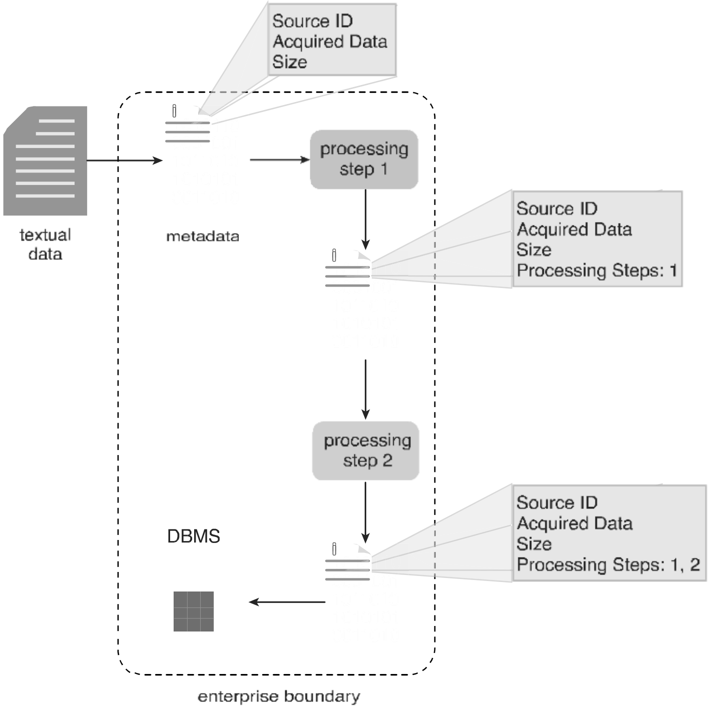


Figure 3.3
Data may also need to be annotated with source dataset attributes and processing step details as it passes through the data transformation steps.

Limited Realtime Support

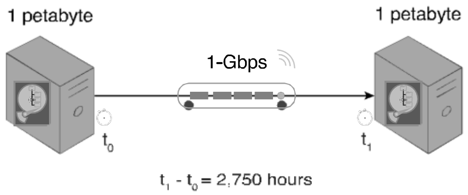
Dashboards and other applications that require streaming data and alerts often demand realtime or near-realtime data transmissions. Many open source Big Data solutions and tools are batch-oriented; however, there is a new generation of realtime capable open source tools that have support for streaming data analysis. Many of the realtime data analysis solutions that do exist are proprietary. Approaches that achieve near-realtime results often process transactional data as it arrives and combine it with previously summarized batch-processed data.

Distinct Performance Challenges

Due to the volumes of data that some Big Data solutions are required to process, per- formance is often a concern. For example, large datasets coupled with complex search algorithms can lead to long query times. Another performance challenge is related to network bandwidth. With increasing data volumes, the time to transfer a unit of data can exceed its actual data processing time, as shown in Figure 3.4.

Figure 3.4

Transferring 1 PB of data via a 1-Gigabit LAN connection at 80% throughput will take approximately 2,750 hours.



Distinct Governance Requirements

Big Data solutions access data and generate data, all of which become assets of the busi- ness. A governance framework is required to ensure that the data and the solution envi- ronment itself are regulated, standardized and evolved in a controlled manner.

Examples of what a Big Data governance framework can encompass include:

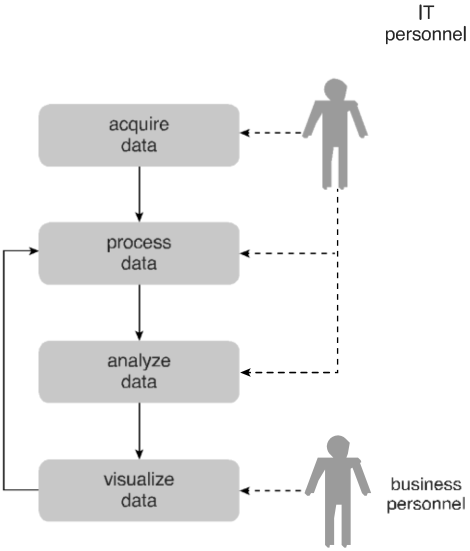
- standardization of how data is tagged and the metadata used for tagging
- policies that regulate the kind of external data that may be acquired
- policies regarding the management of data privacy and data anonymization
- policies for the archiving of data sources and analysis results
- policies that establish guidelines for data cleansing and filtering

Distinct Methodology

A methodology will be required to control how data flows into and out of Big Data solutions. It will need to consider how feedback loops can be established to enable the processed data to undergo repeated refinement, as shown in Figure 3.5. For example, an iterative approach may be used to enable business personnel to provide IT person- nel with feedback on a periodic basis. Each feedback cycle provides opportunities for system refinement by modifying data preparation or data analysis steps.

Figure 3.5

Each repetition can help fine-tune processing steps, algorithms and data models to improve the accuracy of results and deliver greater value to the business.



Clouds

As mentioned in Chapter 2, clouds provide remote environments that can host IT infrastructure for large-scale storage and processing, among other things. Regardless of whether an organization is already cloud-enabled, the adoption of a Big Data environ- ment may necessitate that some or all of that environment be hosted within a cloud. For example, an enterprise that runs its CRM system in a cloud decides to add a Big Data solution in the same cloud environment in order to run analytics on its CRM data. This data can then be shared with its primary Big Data environment that resides within the enterprise boundaries.

Common justifications for incorporating a cloud environment in support of a Big Data solution include:

- inadequate in-house hardware resources
- upfront capital investment for system procurement is not available
- the project is to be isolated from the rest of the business so that existing business processes are not impacted
- the Big Data initiative is a proof of concept
- datasets that need to be processed are already cloud resident
- the limits of available computing and storage resources used by an in-house Big Data solution are being reached

Big Data Analytics Lifecycle

Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes. To address the dis- tinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyz- ing and repurposing data. The upcoming sections explore a specific data analytics life- cycle that organizes and manages the tasks and activities associated with the analysis of Big Data. From a Big Data adoption and planning perspective, it is important that in addition to the lifecycle, consideration be made for issues of training, education, tooling and staffing of a data analytics team.

The Big Data analytics lifecycle can be divided into the following nine stages, as shown in Figure 3.6:

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

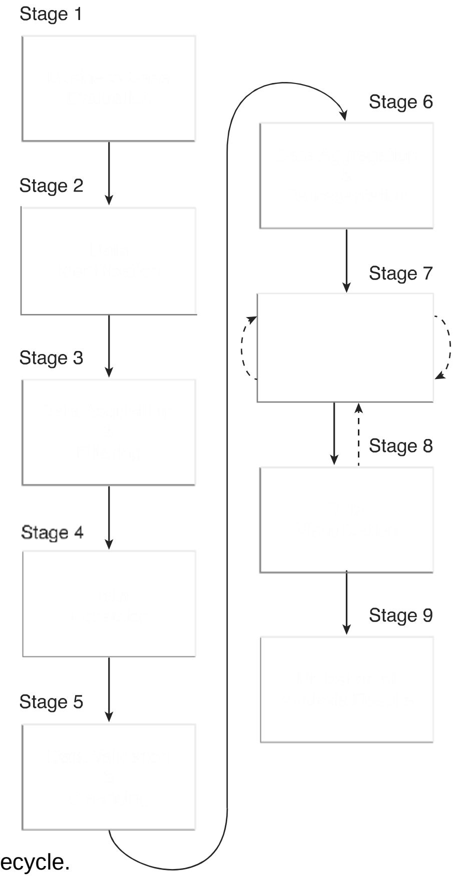


Figure 3.6
The nine stages of the Big Data analytics lifecycle.

Business Case Evaluation

Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis. The Business Case Evaluation stage shown in Figure 3.7 requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks.

An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle. The further identification of KPIs during this stage can help deter- mine assessment criteria and guidance for the evaluation of the analytic results. If KPIs

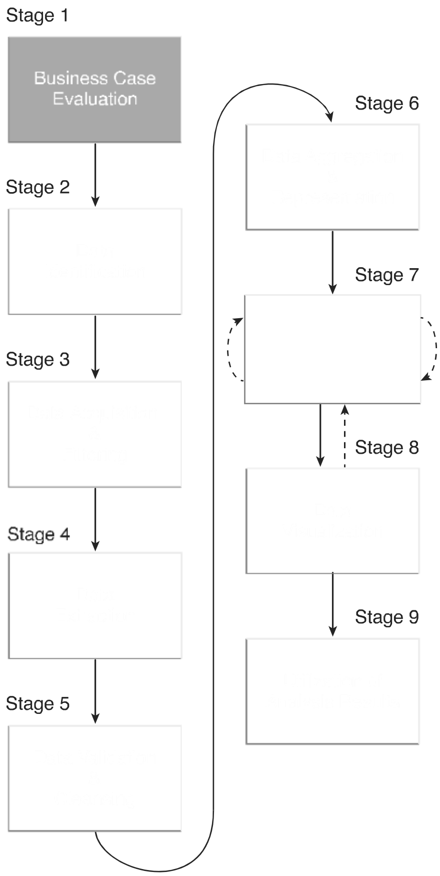


Figure 3.7
Stage 1 of the Big Data analytics lifecycle.

are not readily available, efforts should be made to make the goals of the analysis proj- ect SMART, which stands for specific, measurable, attainable, relevant and timely.

Based on business requirements that are documented in the business case, it can be determined whether the business problems being addressed are really Big Data prob- lems. In order to qualify as a Big Data problem, a business problem needs to be directly related to one or more of the Big Data characteristics of volume, velocity, or variety.

Note also that another outcome of this stage is the determination of the underlying budget required to carry out the analysis project. Any required purchase, such as tools, hardware and training, must be understood in advance so that the anticipated invest- ment can be weighed against the expected benefits of achieving the goals. Initial itera- tions of the Big Data analytics lifecycle will require more up-front investment of Big Data technologies, products and training compared to later iterations where these ear- lier investments can be repeatedly leveraged.

Data Identification

The Data Identification stage shown in Figure 3.8 is dedicated to identifying the data- sets required for the analysis project and their sources.

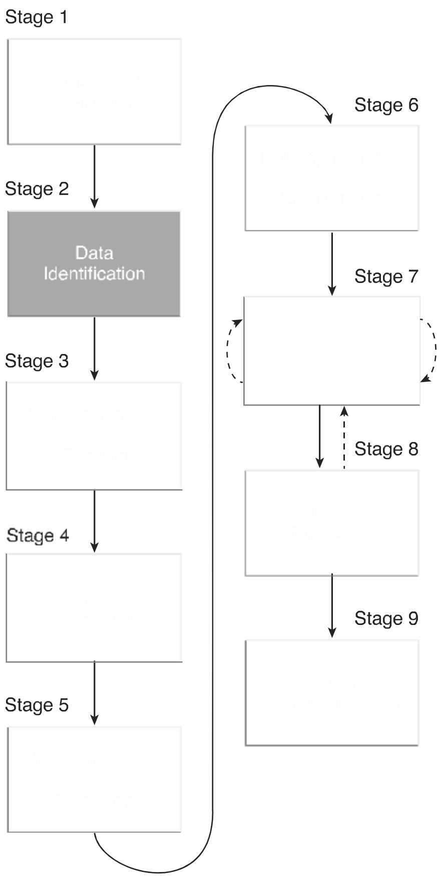
Identifying a wider variety of data sources may increase the probability of finding hid- den patterns and correlations. For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for.

Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.

In the case of internal datasets, a list of available datasets from internal sources, such as data marts and operational systems, are typically compiled and matched against a pre-defined dataset specification.

In the case of external datasets, a list of possible third-party data providers, such as data markets and publicly available datasets, are compiled. Some forms of external data may be embedded within blogs or other types of content-based web sites, in which case they may need to be harvested via automated tools.

Figure 3.8
Data Identification is stage 2 of the Big Data analytics lifecycle.

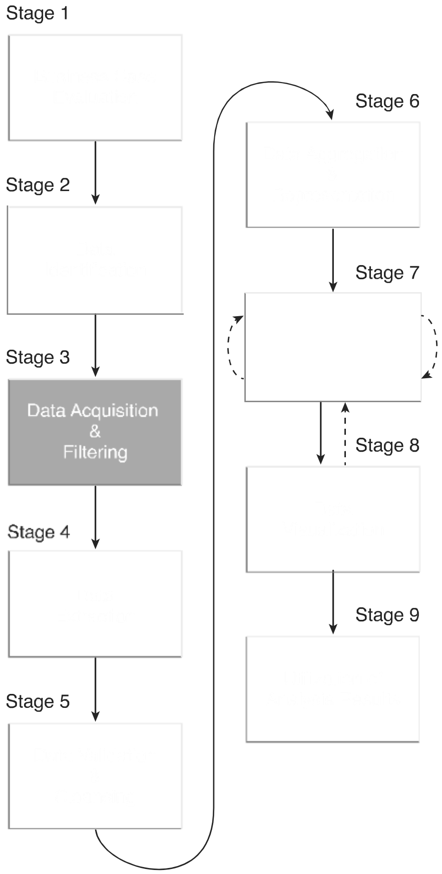


Data Acquisition and Filtering

During the Data Acquisition and Filtering stage, shown in Figure 3.9, the data is gath- ered from all of the data sources that were identified during the previous stage. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.

Depending on the type of data source, data may come as a collection of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter. In many cases, especially where external, unstructured data is con- cerned, some or most of the acquired data may be irrelevant (noise) and can be dis- carded as part of the filtering process.

Figure 3.9
Stage 3 of the Big Data analytics lifecycle.



Data classified as “corrupt” can include records with missing or nonsensical values or invalid data types. Data that is filtered out for one analysis may possibly be valuable for a different type of analysis. Therefore, it is advisable to store a verbatim copy of the original dataset before proceeding with the filtering. To minimize the required storage space, the verbatim copy can be compressed.

Both internal and external data needs to be persisted once it gets generated or enters the enterprise boundary. For batch analytics, this data is persisted to disk prior to analysis. In the case of realtime analytics, the data is analyzed first and then persisted to disk.

As evidenced in Figure 3.10, metadata can be added via automation to data from both internal and external data sources to improve the classification and querying. Examples of appended metadata include dataset size and structure, source information, date and time of creation or collection and language-specific information. It is vital that metadata be machine-readable and passed forward along subsequent analysis stages. This helps maintain data provenance throughout the Big Data analytics lifecycle, which helps to establish and preserve data accuracy and quality.

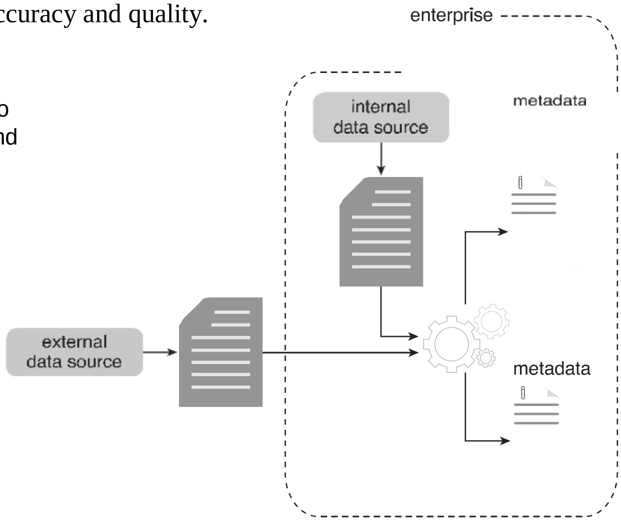


Figure 3.10
Metadata is added to data from internal and external sources.

Data Extraction

Some of the data identified as input for the analysis may arrive in a format incompat- ible with the Big Data solution. The need to address disparate types of data is more likely with data from external sources. The Data Extraction lifecycle stage, shown in Figure 3.11, is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

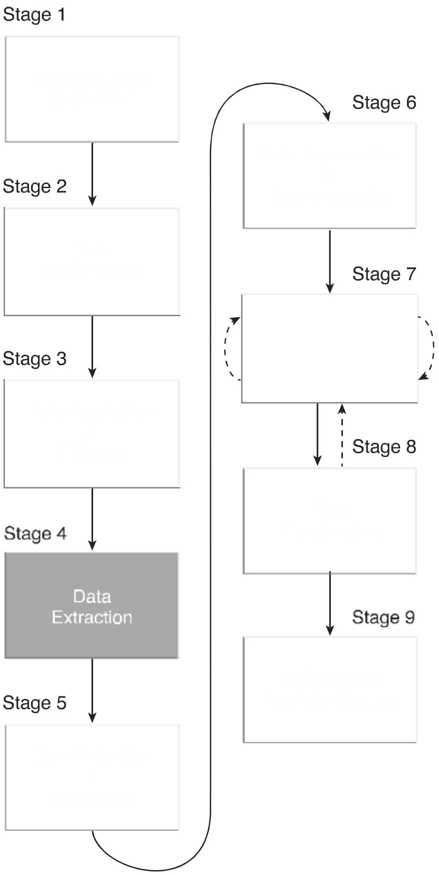
The extent of extraction and transformation required depends on the types of analyt- ics and capabilities of the Big Data solution. For example, extracting the required fields from delimited textual data, such as with webserver log files, may not be necessary if the underlying Big Data solution can already directly process those files.

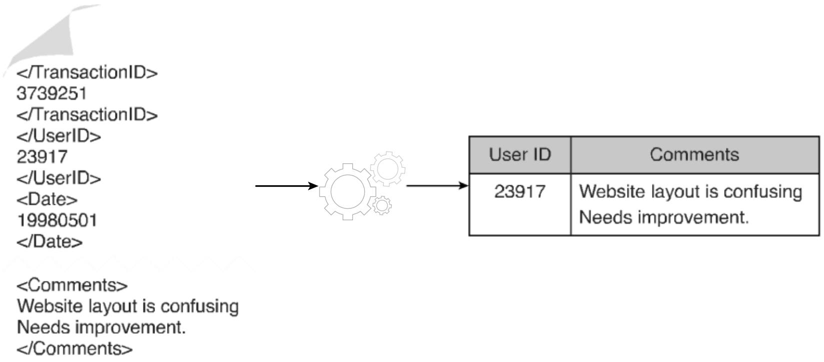
Similarly, extracting text for text analytics, which requires scans of whole documents, is simplified if the underlying Big Data solution can directly read the document in its native format.

Figure 3.12 illustrates the extraction of comments and a user ID embedded within an XML document without the need for further transformation.

Figure 3.11

Stage 4 of the Big Data analytics lifecycle.





```
</TransactionID>
3739251
</TransactionID>
</UserID>
23917
</UserID>
<Date>
19980501
</Date>

<Comments>
Website layout is confusing
Needs improvement.
</Comments>
```

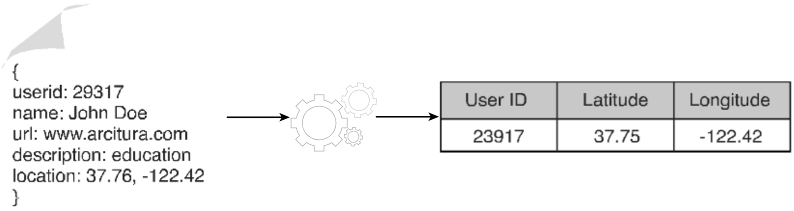
User ID	Comments
23917	Website layout is confusing Needs improvement.

Figure 3.12

Comments and user IDs are extracted from an XML document.

Figure 3.13 demonstrates the extraction of the latitude and longitude coordinates of a user from a single JSON field.

Further transformation is needed in order to separate the data into two separate fields as required by the Big Data solution.



```
{
userid: 29317
name: John Doe
url: www.arcitura.com
description: education
location: 37.76, -122.42
}
```

User ID	Latitude	Longitude
23917	37.75	-122.42

Figure 3.13

The user ID and coordinates of a user are extracted from a single JSON field.

Data Validation and Cleansing

Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, data input into Big Data analyses can be unstructured without any indication of validity. Its complexity can further make it difficult to arrive at a set of suitable validation constraints.

The Data Validation and Cleansing stage shown in Figure 3.14 is dedicated to establish- ing often complex validation rules and removing any known invalid data.

Big Data solutions often receive redundant data across different datasets. This redun- dancy can be exploited to explore interconnected datasets in order to assemble valida- tion parameters and fill in missing valid data.

For example, as illustrated in Figure 3.15:

- The first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against its corresponding value in Dataset A.
- If a value is missing, it is inserted from Dataset A.

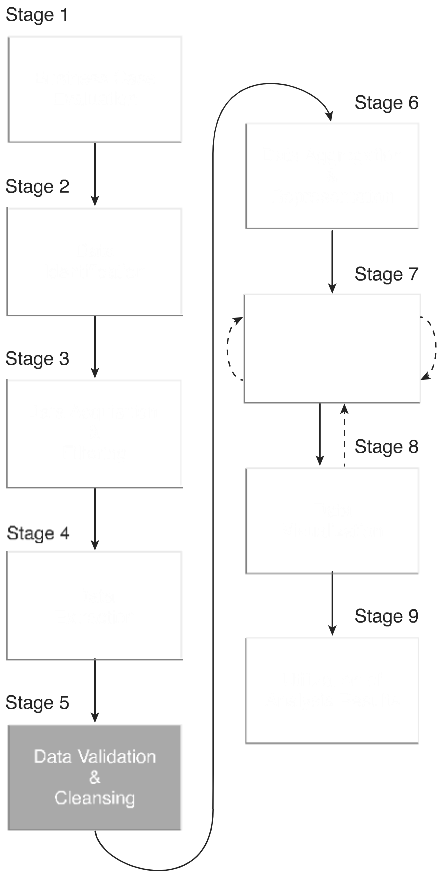


Figure 3.14
Stage 5 of the Big Data analytics lifecycle.

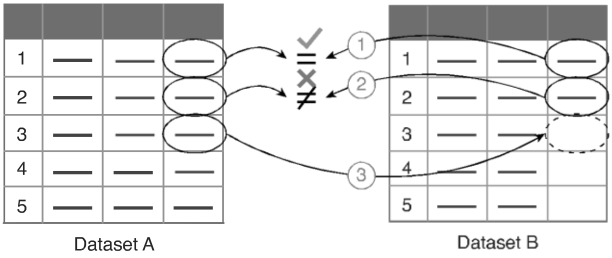
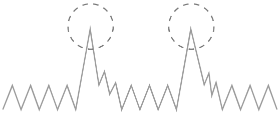


Figure 3.15
Data validation can be used to examine interconnected datasets in order to fill in missing valid data.

For batch analytics, data validation and cleansing can be achieved via an offline ETL operation. For realtime analytics, a more complex in-memory system is required to validate and cleanse the data as it arrives from the source. Provenance can play an important role in determining the accuracy and quality of questionable data. Data that appears to be invalid may still be valuable in that it may possess hidden patterns and
trends, as shown in Figure 3.16.

Figure 3.16
The presence of invalid data is resulting in spikes. Although the data appears abnormal, it may be indicative of a new pattern.



Data Aggregation and Representation

Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. Either way, a method of data reconcili- ation is required or the dataset representing the correct value needs to be determined.

The Data Aggregation and Representation stage, shown in Figure 3.17, is dedicated to integrating multiple datasets together to arrive at a unified view.

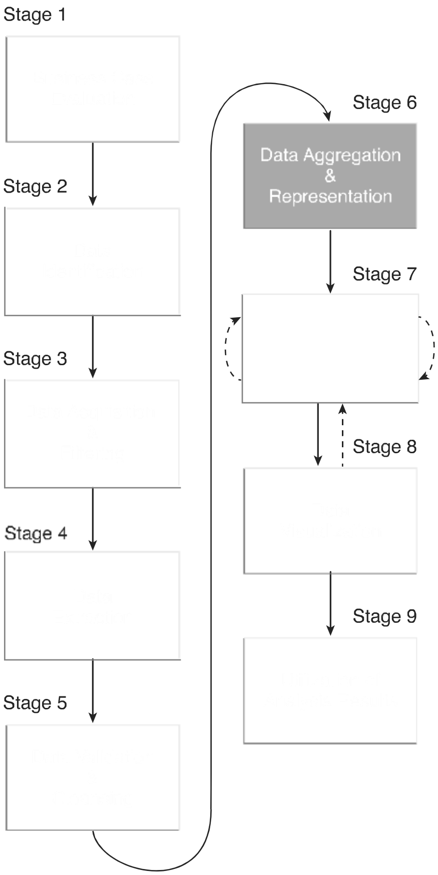
Performing this stage can become complicated because of differences in:

- *Data Structure* – Although the data format may be the same, the data model may be different.
- *Semantics* – A value that is labeled differently in two different datasets may mean the same thing, for example “surname” and “last name.”

The large volumes processed by Big Data solutions can make data aggregation a time and effort-intensive operation. Reconciling these differences can require complex logic that is executed automatically without the need for human intervention.

Future data analysis requirements need to be considered during this stage to help foster data reusability. Whether data aggregation is required or not, it is important to under- stand that the same data can be stored in many different forms. One form may be better suited for a particular type of analysis than another. For example, data stored as a BLOB would be of little use if the analysis requires access to individual data fields.

Figure 3.17
Stage 6 of the Big Data analytics lifecycle.



A data structure standardized by the Big Data solution can act as a common denomina- tor that can be used for a range of analysis techniques and projects. This can require establishing a central, standard analysis repository, such as a NoSQL database, as shown in Figure 3.18.

Figure 3.18
A simple example of data aggregation where two datasets are aggregated together using the Id field.

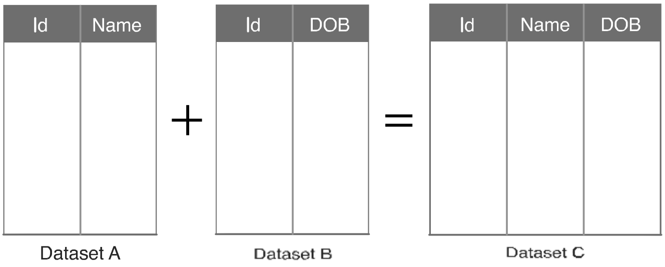


Figure 3.19 shows the same piece of data stored in two different formats. Dataset A con- tains the desired piece of data, but it is part of a BLOB that is not readily accessible for querying. Dataset B contains the same piece of data organized in column-based storage, enabling each field to be queried individually.

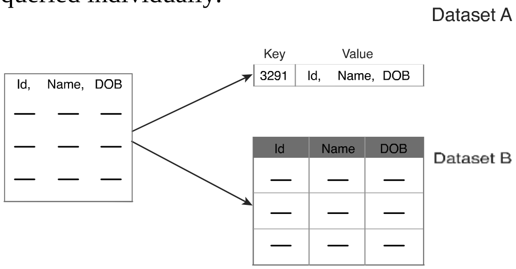


Figure 3.19
Dataset A and B can be combined to create a standardized data structure with a Big Data solution.

Data Analysis

The Data Analysis stage shown in Figure 3.20 is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

Depending on the type of analytic result required, this stage can be as simple as query- ing a dataset to compute an aggregation for comparison. On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.

Data analysis can be classified as confirmatory analysis or exploratory analysis, the lat- ter of which is linked to data mining, as shown in Figure 3.21.

Confirmatory data analysis is a deductive approach where the cause of the phenom- enon being investigated is proposed beforehand. The proposed cause or assumption is called a hypothesis. The data is then analyzed to prove or disprove the hypothesis and provide definitive answers to specific questions. Data sampling techiniques are typically used. Unexpected findings or anomalies are usually ignored since a predeter- mined cause was assumed.

Exploratory data analysis is an inductive approach that is closely associated with data mining. No hypothesis or predetermined assumptions are generated. Instead, the data

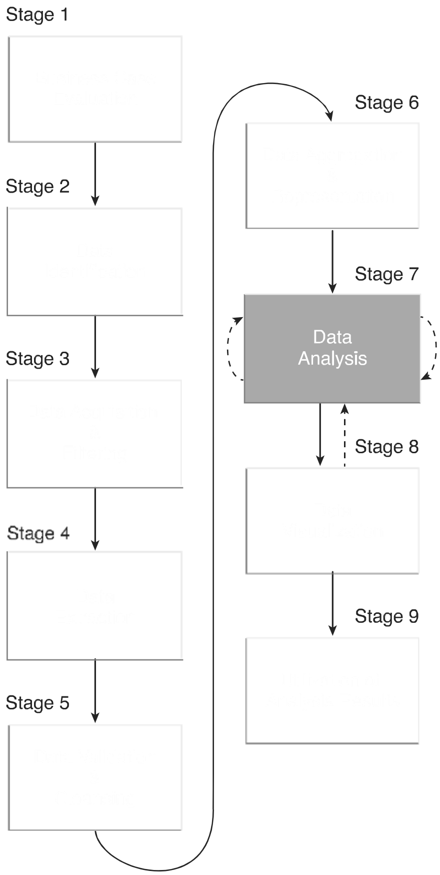


Figure 3.20
Stage 7 of the Big Data analytics lifecycle.

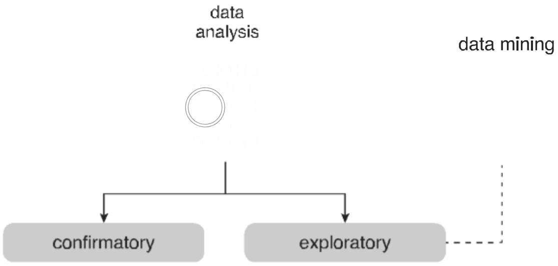


Figure 3.21
Data analysis can be carried out as confirmatory or exploratory analysis.

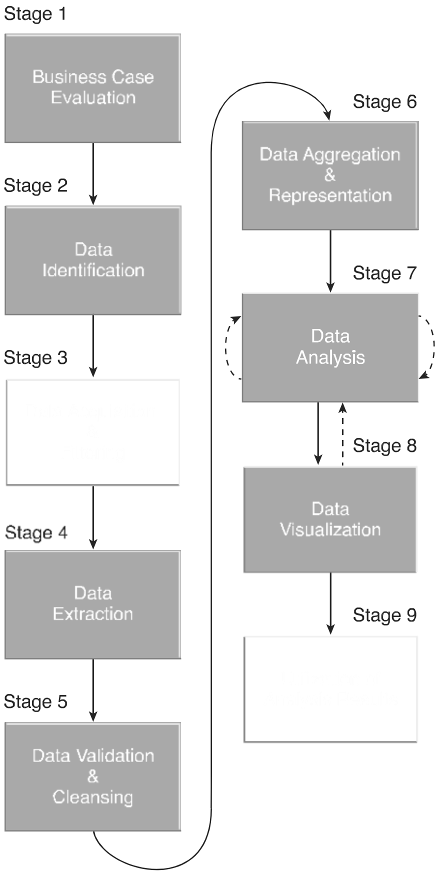
is explored through analysis to develop an understanding of the cause of the phenom- enon. Although it may not provide definitive answers, this method provides a general direction that can facilitate the discovery of patterns or anomalies.

Data Visualization

The ability to analyze massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts.

The Data Visualization stage, shown in Figure 3.22, is dedicated to using data visualiza- tion techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

Figure 3.22
Stage 8 of the Big Data analytics lifecycle.



Business users need to be able to understand the results in order to obtain value from the analysis and subsequently have the ability to provide feedback, as indicated by the dashed line leading from stage 8 back to stage 7.

The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated. Visual analysis techniques are covered later in this book.

The same results may be presented in a number of different ways, which can influence the interpretation of the results. Consequently, it is important to use the most suitable visualization technique by keeping the business domain in context.

Another aspect to keep in mind is that providing a method of drilling down to com- paratively simple statistics is crucial, in order for users to understand how the rolled up or aggregated results were generated.

Utilization of Analysis Results

Subsequent to analysis results being made available to business users to support business decision-making, such as via dashboards, there may be further opportuni- ties to utilize the analysis results. The Utilization of Analysis Results stage, shown in Figure 3.23, is dedicated to determining how and where processed analysis data can be further leveraged.

Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce "models" that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed. A model may look like a mathematical equation or a set of rules. Models can be used to improve business process logic and application system logic, and they can form the basis of a new system or software program.

Common areas that are explored during this stage include the following:

- *Input for Enterprise Systems* – The data analysis results may be automatically or manually fed directly into enterprise systems to enhance and optimize their behaviors and performance. For example, an online store can be fed processed customer-related analysis results that may impact how it generates product recom- mendations. New models may be used to improve the programming logic within existing enterprise systems or may form the basis of new systems.

- *Business Process Optimization* – The identified patterns, correlations and anomalies discovered during the data analysis are used to refine business processes. An example is consolidating transportation routes as part of a supply chain process. Models may also lead to opportunities to improve business process logic.
- *Alerts* – Data analysis results can be used as input for existing alerts or may form the basis of new alerts. For example, alerts may be created to inform users via email or SMS text about an event that requires them to take corrective action.

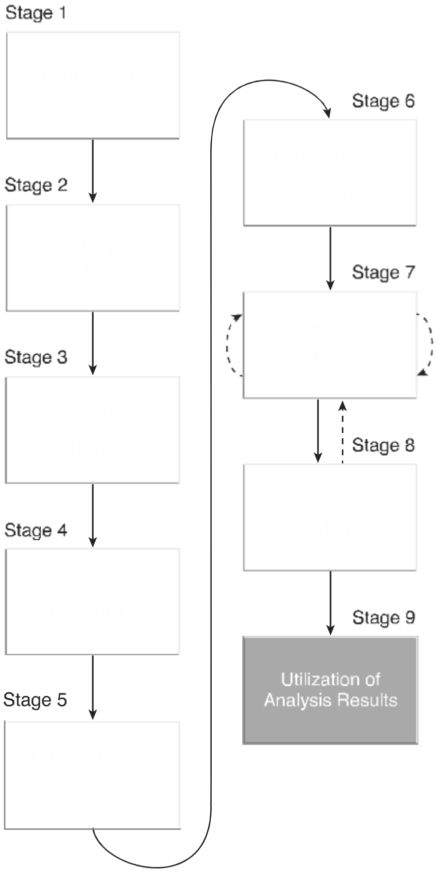


Figure 3.23
Stage 9 of the Big Data analytics lifecycle.

CASE STUDY EXAMPLE

The majority of ETI's IT team is convinced that Big Data is the silver bullet that will address all of their current issues. However, the trained IT members point out that adopting Big Data is not the same as simply adopting a technology platform. Rather, a range of factors first need to be considered in order to ensure successful adoption of Big Data. Therefore, to ensure that the impact of business-related factors is fully understood, the IT team sits together with the business managers to create a feasibil- ity report. Involving business personnel at this early stage will further help create an environment that reduces the gap between management's perceived expectations and what IT can actually deliver.

There is a strong understanding that the adoption of Big Data is business-oriented and will assist ETI in reaching its goals. Big Data's abilities to store and process large amounts of unstructured data and combine multiple datasets will help ETI compre- hend risk. The company hopes that, as a result, it can minimize losses by only accept- ing less-risky applicants as customers. Similarly, ETI predicts that the ability to look into the unstructured behavioral data of a customer and discover abnormal behavior will further help reduce loss because fraudulent claims can be rejected.

The decision to train the IT team in the field of Big Data has increased ETI's readiness for adopting Big Data. The team believes that it now has the basic skillset required for undertaking a Big Data initiative. Data identified and categorized earlier puts the team in a strong position for deciding on the required technologies. The early engagement of business management has also provided insights that allow them to anticipate changes that may be required in the future to keep the Big Data solution platform in alignment with any emerging business requirements.

At this preliminary stage, only a handful of external data sources, such as social media and census data, have been identified. It is agreed by the business personnel that a sufficient budget will be allocated for the acquisition of data from third-party data providers. Regarding privacy, the business users are a bit wary that obtain- ing additional data about customers could spark customer distrust. However, it is thought that an incentive-driven scheme, such as lower premiums, can be introduced in order to gain customers' consent and trust. When considering issues of security, the IT team notes that additional development efforts will be required to ensure that standardized, role-based access controls are in place for data held within the Big Data solution environment. This is especially relevant for the open-source databases that will hold non-relational data.

Although the business users are excited about being able to perform deep analytics through the use of unstructured data, they pose a question regarding the degree to which can they trust the results, for the analysis involves data from third-party data providers. The IT team responds that a framework will be adopted for adding and updating metadata for each dataset that is stored and processed so that provenance is maintained at all times and processing results can be traced all the way back to the constituent data sources.

ETI’s present goals include decreasing the time it takes to settle claims and detect fraudulent claims. The achievement of these goals will require a solution that pro- vides results in a timely manner. However, it is not anticipated that realtime data analysis support will be required. The IT team believes that these goals can be satis- fied by developing a batch-based Big Data solution that leverages open source Big Data technology.

ETI’s current IT infrastructure consists of comparatively older networking standards. Similarly, the specifications of most of the servers, such as the processor speed, disk capacity and disk speed, dictate that they are not capable of providing optimum data processing performance. Hence it is agreed that the current IT infrastructure needs an upgrade before a Big Data solution can be designed and built.

Both the business and IT teams strongly believe that a Big Data governance frame- work is required to not only help them standardize the usage of disparate data sources but also fully comply with any data privacy-related regulations. Further- more, due to the business focus of the data analysis and to ensure that meaningful analysis results are generated, it is decided that an iterative data analysis approach that includes business personnel from the relevant department needs to be adopted. For example, in the “improving customer retention” scenario, the marketing and sales team can be included in the data analysis process right from the selection of datasets so that only the relevant attributes of these datasets are chosen. Later, the business team can provide valuable feedback in terms of interpretation and applica- bility of the analysis results.

With regards to cloud computing, the IT team observes that none of its systems are currently hosted in the cloud and that the team does not possess cloud-related skill- sets. These facts alongside data privacy concerns lead the IT team to the decision to build an on-premise Big Data solution. The group notes that they will leave the option of cloud-based hosting open because there is some speculation that their internal CRM system may be replaced with a cloud-hosted, software-as-a-service CRM solution in the future.

Big Data Analytics Lifecycle

ETI's Big Data journey has reached the stage where its IT team possesses the neces- sary skills and the management is convinced of the potential benefits that a Big Data solution can bring in support of the business goals. The CEO and the directors are eager to see Big Data in action. In response to this, the IT team, in partnership with the business personnel, take on ETI's first Big Data project. After a thorough evalu- ation process, the "detection of fraudulent claims" objective is chosen as the first Big Data solution. The team then follows a step-by-step approach as set forth by the Big Data Analytics Lifecycle in pursuit of achieving this objective.

Business Case Evaluation

Carrying out Big Data analysis for the "detection of fraudulent claims" directly cor- responds to a decrease in monetary loss and hence carries complete business back- ing. Although fraud occurs across all the four business sectors of ETI, in the interest of keeping the analysis somewhat straightforward, the scope of Big Data analysis is limited to identification of fraud in the building sector.

ETI provides building and contents insurance to both domestic and commercial cus- tomers. Although insurance fraud can both be opportunistic and organized, oppor- tunistic fraud in the form of lying and exaggeration covers the majority of the cases. To measure the success of the Big Data solution for fraud detection, one of the KPIs set is the *reduction in fraudulent claims by 15%*.

Taking their budget into account, the team decides that their largest expense will be in the procuring of new infrastructure that is appropriate for building a Big Data solution environment. They realize that they will be leveraging open source tech- nologies to support batch processing and therefore do not believe that a large, ini- tial up-front investment is required for tooling. However, when they consider the broader Big Data analytics lifecycle, the team members realize that they should bud- get for the acquisition of additional data quality and cleansing tools and newer data visualization technologies. After accounting for these expenses, a cost-benefit analy- sis reveals that the investment in the Big Data solution can return itself several times over if the targeted fraud-detecting KPIs can be attained. As a result of this analysis, the team believes that a strong business case exists for using Big Data for enhanced data analysis.

Data Identification

A number of *internal* and *external* datasets are identified. Internal data includes policy data, insurance application documents, claim data, claim adjuster notes, inci- dent photographs, call center agent notes and emails. External data includes social media data (Twitter feeds), weather reports, geographical (GIS) data and census data. Nearly all datasets go back five years in time. The claim data consists of historical claim data consisting of multiple fields where one of the fields specifies if the claim was *fraudulent* or *legitimate*.

Data Acquisition and Filtering

The policy data is obtained from the policy administration system, the claim data, incident photographs and claim adjuster notes are acquired from the claims manage- ment system and the insurance application documents are obtained from the docu- ment management system. The claim adjuster notes are currently embedded within the claim data. Hence a separate process is used to extract them. Call center agent notes and emails are obtained from the CRM system.

The rest of the datasets are acquired from third-party data providers. A compressed copy of the original version of all of the datasets is stored on-disk. From a provenance perspective, the following metadata is tracked to capture the pedigree of each data- set: dataset’s name, source, size, format, checksum, acquired date and number of records. A quick check of the data qualities of Twitter feeds and weather reports sug- gests that around four to five percent of their records are corrupt. Consequently, two batch data filtering jobs are established to remove the corrupt records.

Data Extraction

The IT team observes that some of the datasets will need to be pre-processed in order to extract the required fields. For example, the tweets dataset is in JSON format. In order to be able to analyze the tweets, the *user id*, *timestamp* and the tweet *text* need to be extracted and converted to tabular form. Further, the weather dataset arrives in a hierarchical format (XML), and fields such as *timestamp, temperature forecast, wind speed forecast, wind direction forecast, snow forecast* and *flood forecast* are also extracted and saved in a tabular form.

Data Validation and Cleansing

To keep costs down, ETI is currently using free versions of the weather and the cen- sus datasets that are not guaranteed to be 100% accurate. As a result, these datasets need to be validated and cleansed. Based on the published field information, the team is able to check the extracted fields for typographical errors and any incorrect data as well as data type and range validation. A rule is established that a record will not be removed if it contains some meaningful level of information even though some of its fields may contain invalid data.

Data Aggregation and Representation

For meaningful analysis of data, it is decided to join together policy data, claim data and call center agent notes in a single dataset that is tabular in nature where each field can be referenced via a data query. It is thought that this will not only help with the current data analysis task of detecting fraudulent claims but will also help with other data analysis tasks, such as risk evaluation and speedy settlement of claims. The resulting dataset is stored in a NoSQL database.

Data Analysis

The IT team involves the data analysts at this stage as it does not have the right skill- set for analyzing data in support of detecting fraudulent claims. In order to be able to detect fraudulent transactions, first the nature of fraudulent claims needs to be analyzed in order to find which characteristics differentiate a fraudulent claim from a legitimate claim. For this, the *exploratory data analysis* approach is taken. As part of this analysis, a range of analysis techniques are applied, some of which are discussed in Chapter 8. This stage is repeated a number of times as the results generated after the first pass are not conclusive enough to comprehend what makes a fraudulent claim different from a legitimate claim. As part of this exercise, attributes that are less indicative of a fraudulent claim are dropped while attributes that carry a direct relationship are kept or added.

Data Visualization

The team has discovered some interesting findings and now needs to convey the results to the actuaries, underwriters and claim adjusters. Different visualization methods are used including bar and line graphs and scatter plots. Scatter plots are used to analyze groups of fraudulent and legitimate claims in the light of different factors, such as *customer age, age of policy, number of claims made* and *value of claim*.

Utilization of Analysis Results

Based on the data analysis results, the underwriting and the claims settlement users have now developed an understanding of the nature of fraudulent claims. However, in order to realize tangible benefits from this data analysis exercise, a model based on a machine-learning technique is generated, which is then incorporated into the existing claim processing system to flag fraudulent claims. The involved machine learning technique will be discussed in Chapter 8.

This page intentionally left blank

About the Authors

Thomas Erl

Thomas Erl is a top-selling IT author, founder of Arcitura Education and series editor of the *Prentice Hall Service Technology Series from Thomas Erl*. With more than 200,000 cop- ies in print worldwide, his books have become international bestsellers and have been formally endorsed by senior members of major IT organizations, such as IBM, Micro- soft, Oracle, Intel, Accenture, IEEE, HL7, MITRE, SAP, CISCO, HP and many others. As CEO of Arcitura Education Inc., Thomas has led the development of curricula for the internationally recognized Big Data Science Certified Professional (BDSCP), Cloud Certified Professional (CCP) and SOA Certified Professional (SOACP) accreditation pro- grams, which have established a series of formal, vendor-neutral industry certifications obtained by thousands of IT professionals around the world. Thomas has toured more than 20 countries as a speaker and instructor. More than 100 articles and interviews by Thomas have been published in numerous publications, including *The Wall Street Journal* and *CIO Magazine*.

Wajid Khattak

Wajid Khattak is a Big Data researcher and trainer at Arcitura Education Inc. His areas of interest include Big Data engineering and architecture, data science, machine learn- ing, analytics and SOA. He has extensive .NET software development experience in the domains of business intelligence reporting solutions and GIS.

Wajid completed his MSc in Software Engineering and Security with distinction from Birmingham City University in 2008. Prior to that, in 2003, he earned his BSc (Hons) degree in Software Engineering from Birmingham City University with first-class rec- ognition. He holds MCAD & MCTS (Microsoft), SOA Architect, Big Data Scientist, Big Data Engineer and Big Data Consultant (Arcitura) certifications.

Paul Buhler

Dr. Paul Buhler is a seasoned professional who has worked in commercial, government and academic environments. He is a respected researcher, practitioner and educator of service-oriented computing concepts, technologies and implementation methodologies. His work in XaaS naturally extends to cloud, Big Data and IoE areas. Dr. Buhler's more recent work has been focused on closing the gap between business strategy and process execution by leveraging responsive design principles and goal-based execution.

As Chief Scientist at Modus21, Dr. Buhler is responsible for aligning corporate strategy with emerging trends in business architecture and process execution frameworks. He also holds an Affiliate Professorship at the College of Charleston, where he teaches both graduate and undergraduate computer science courses. Dr. Buhler earned his Ph.D. in Computer Engineering at the University of South Carolina. He also holds an MS degree in Computer Science from Johns Hopkins University and a BS in Computer Science from The Citadel.

Index