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 distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data. The upcoming sections explore a specific data analytics lifecycle 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

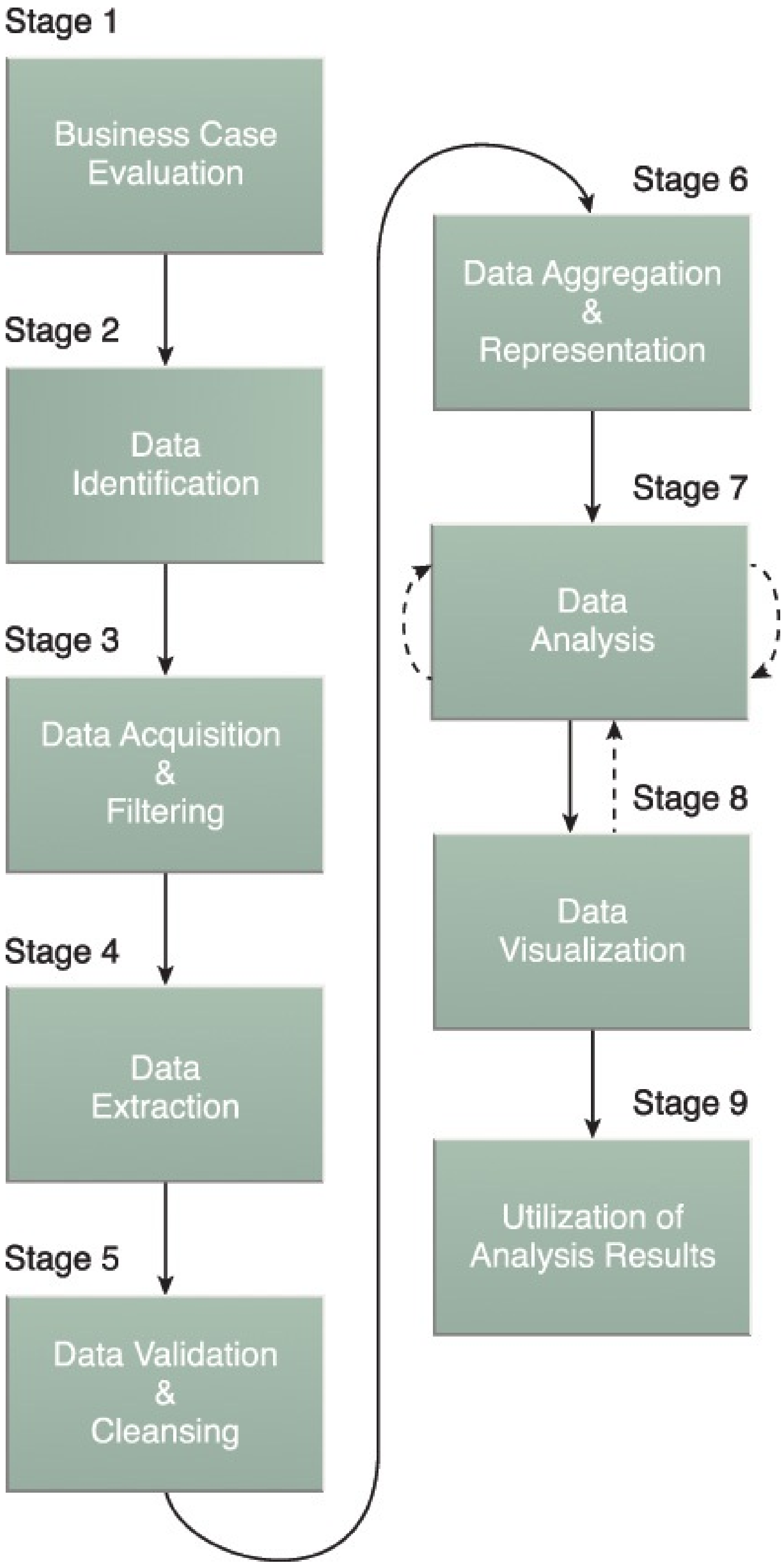


Figure 3.6 The nine stages of the Big Data analytics lifecycle.

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.

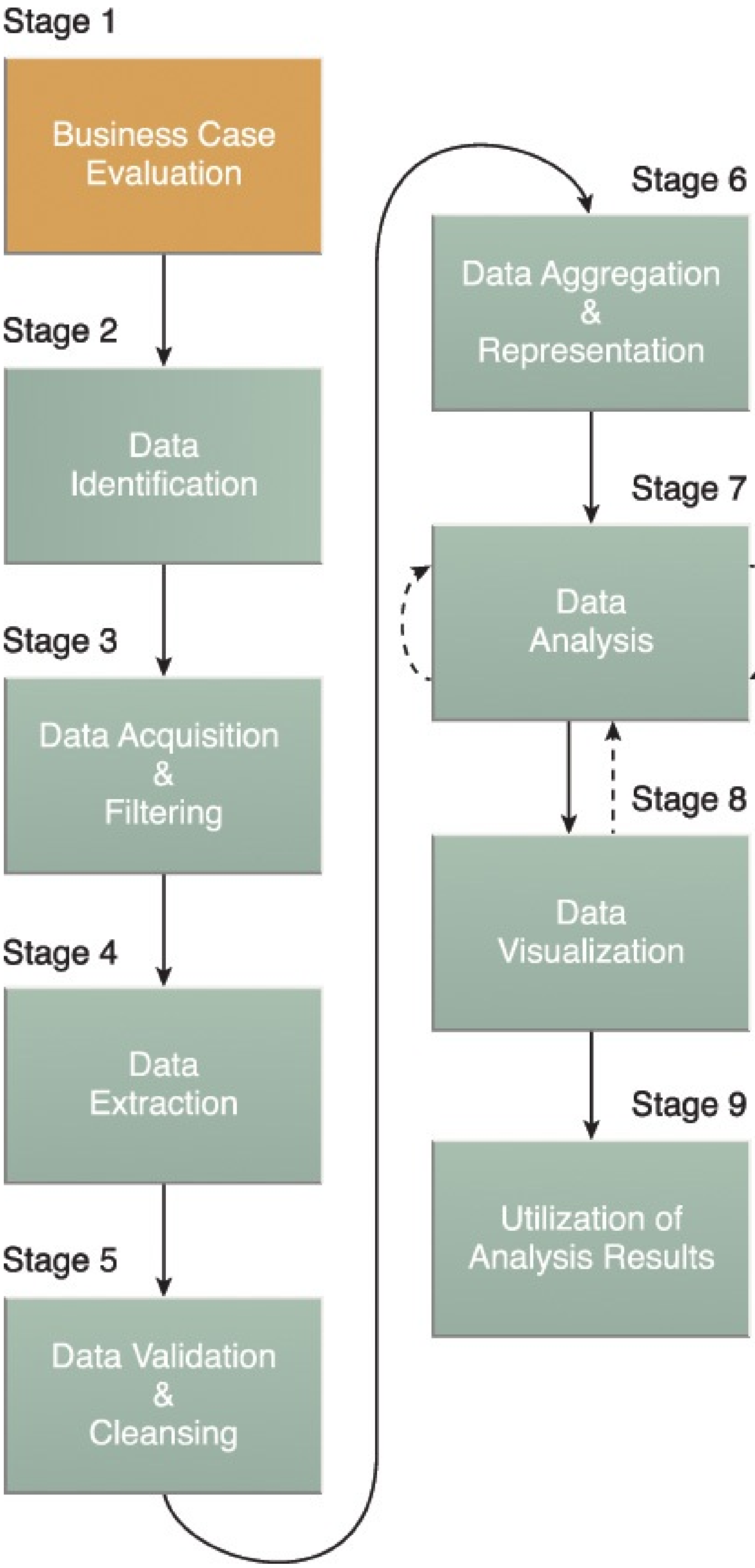


Figure 3.7 Stage 1 of the Big Data analytics lifecycle.

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 determine assessment criteria and guidance for the evaluation of the analytic results. If KPIs are not readily available, efforts should be made to make the goals of the analysis project 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 problems. 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 investment can be weighed against the expected benefits of achieving the goals. Initial iterations 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 earlier investments can be repeatedly leveraged.

Data Identification

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

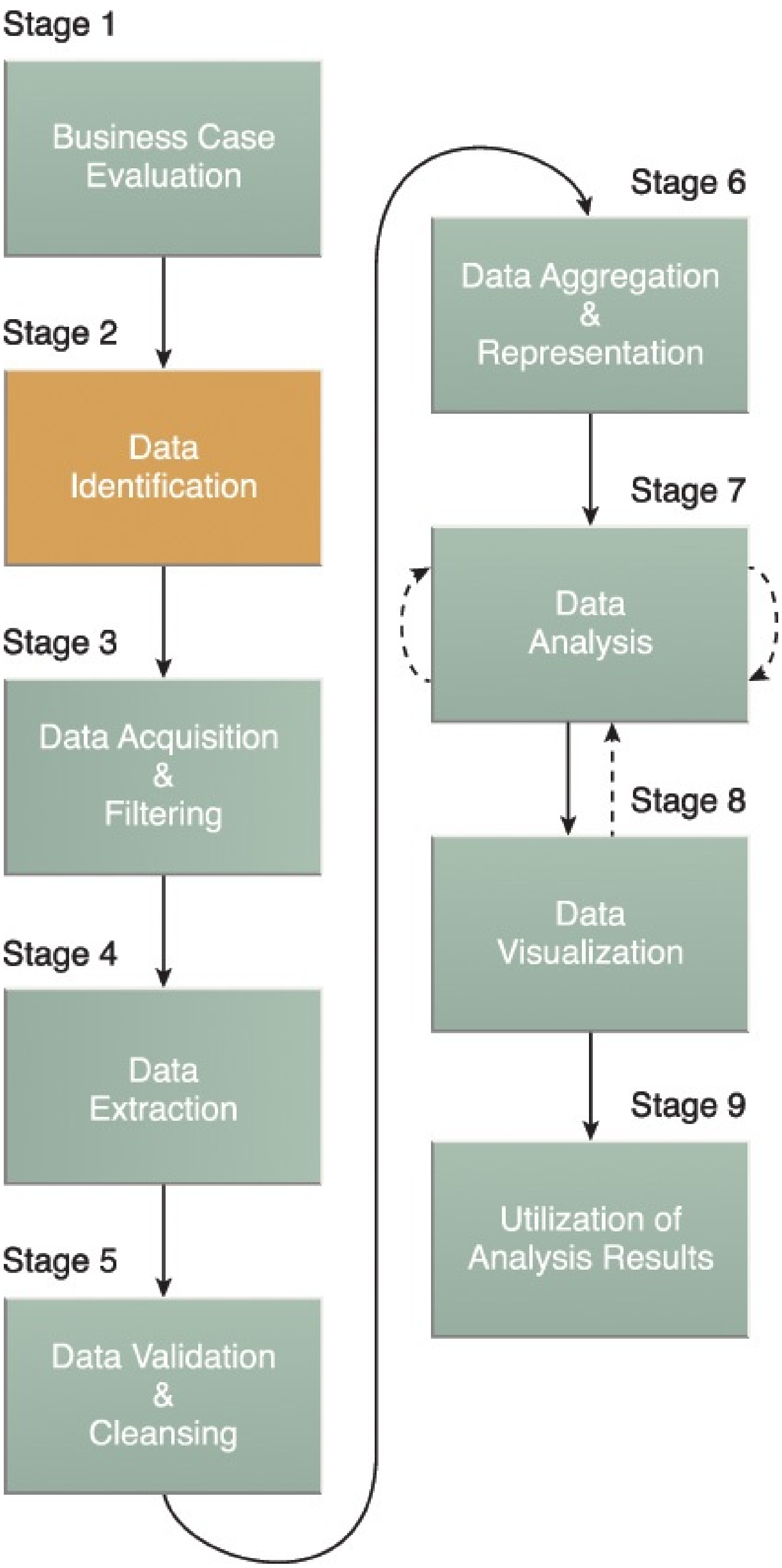


Figure 3.8 Data Identification is stage 2 of the Big Data analytics lifecycle.

Identifying a wider variety of data sources may increase the probability of finding hidden 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 predefined 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.

Data Acquisition and Filtering

During the Data Acquisition and Filtering stage, shown in Figure 3.9, the data is gathered 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.

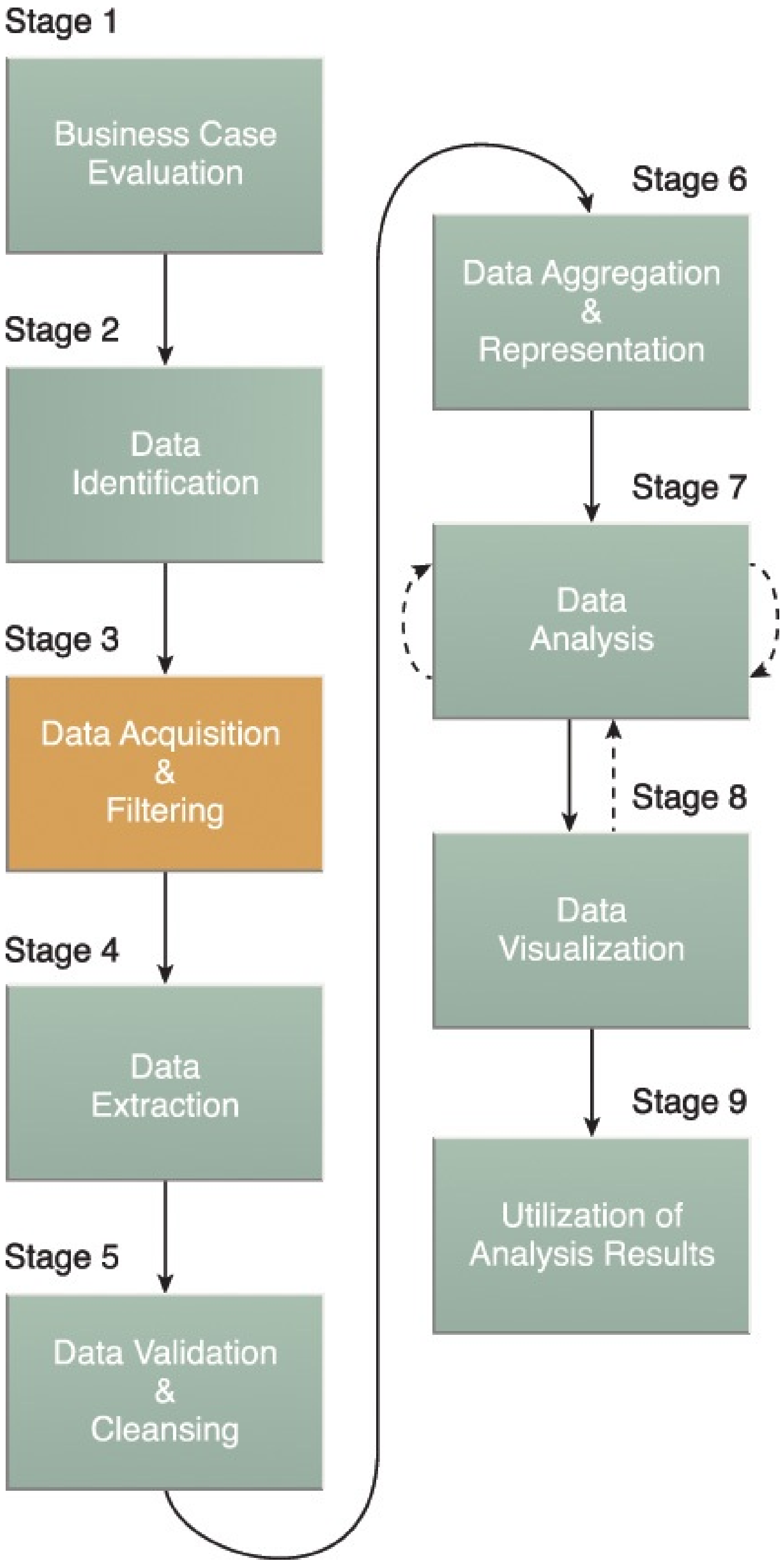


Figure 3.9 Stage 3 of the Big Data analytics lifecycle.

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 concerned, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process.

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.

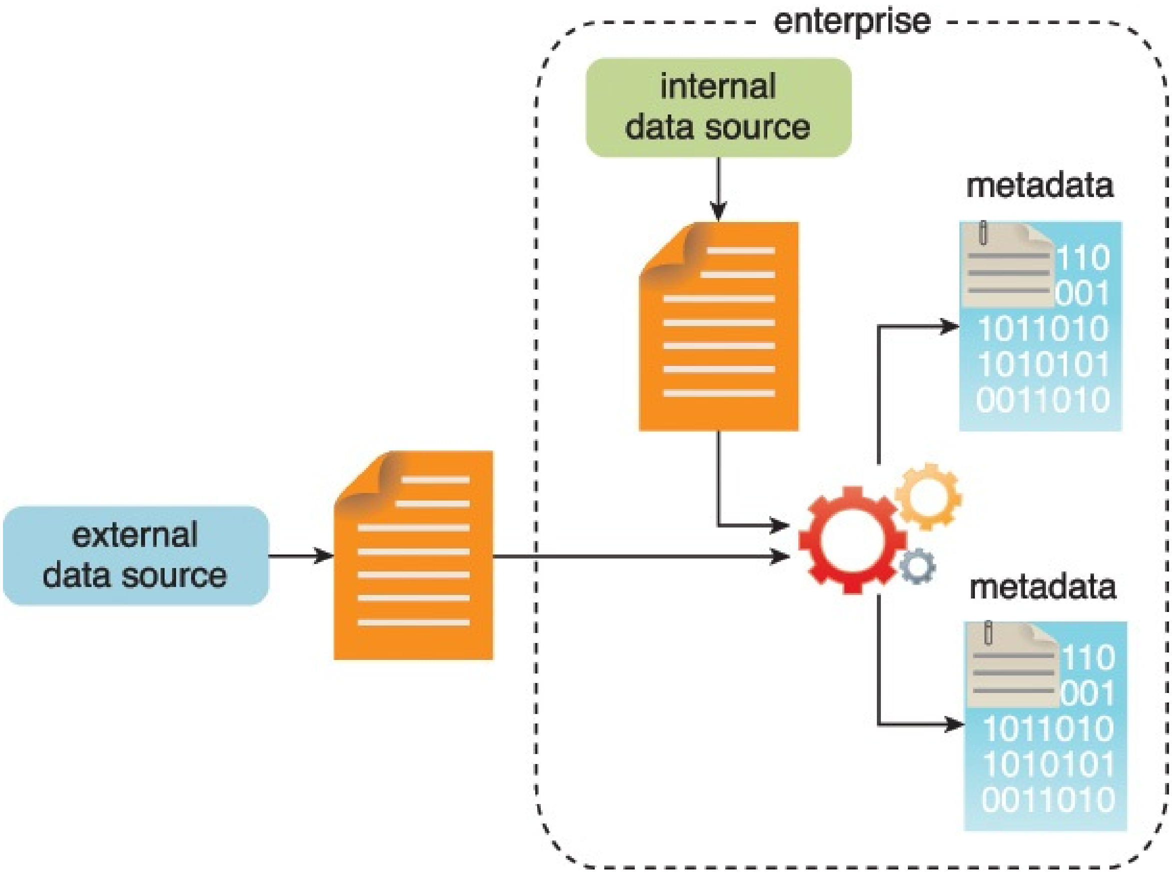


Figure 3.10 Metadata is added to data from internal and external sources.

Data Extraction

Some of the data identified as input for the analysis may arrive in a format incompatible 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.

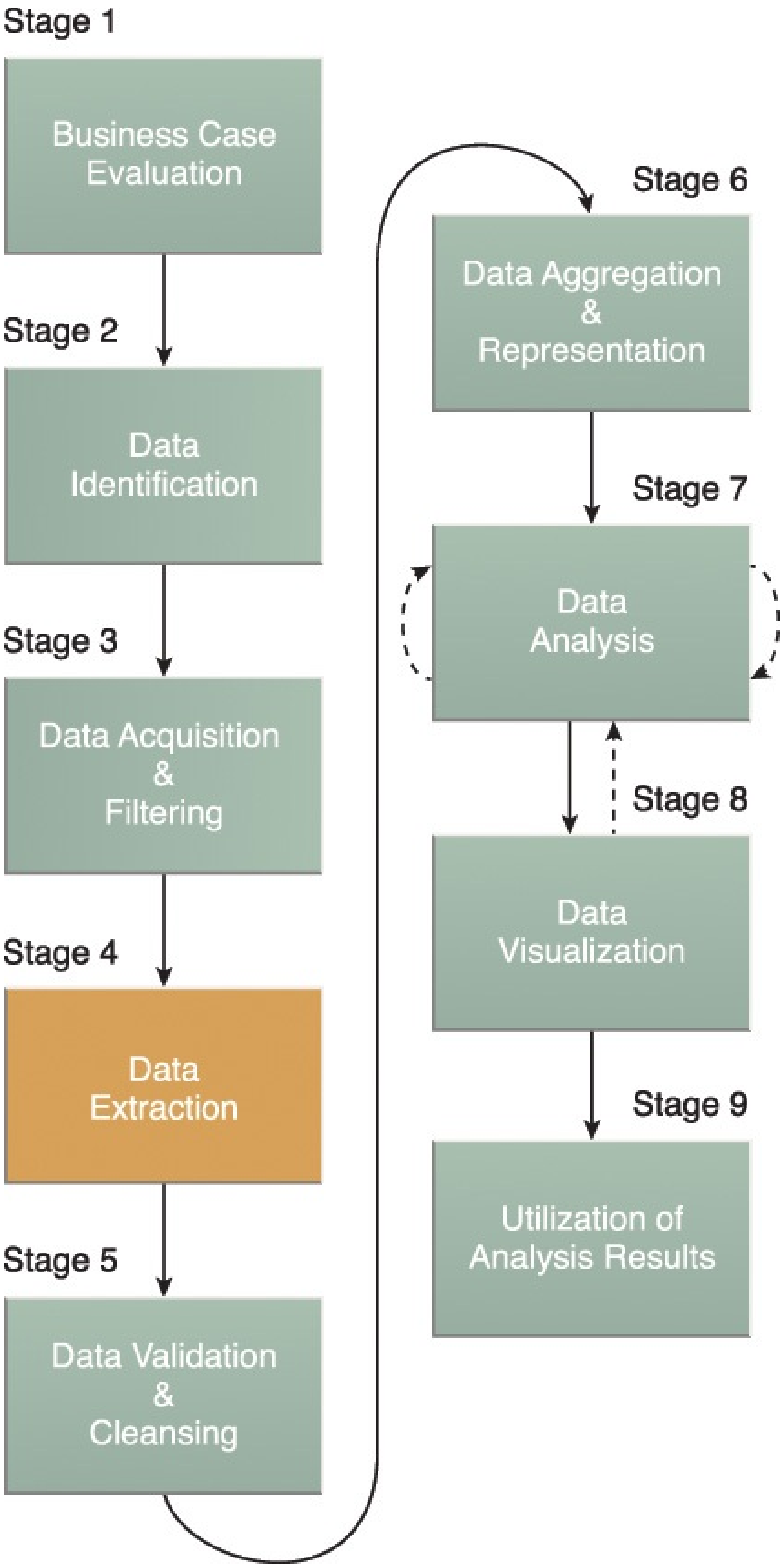
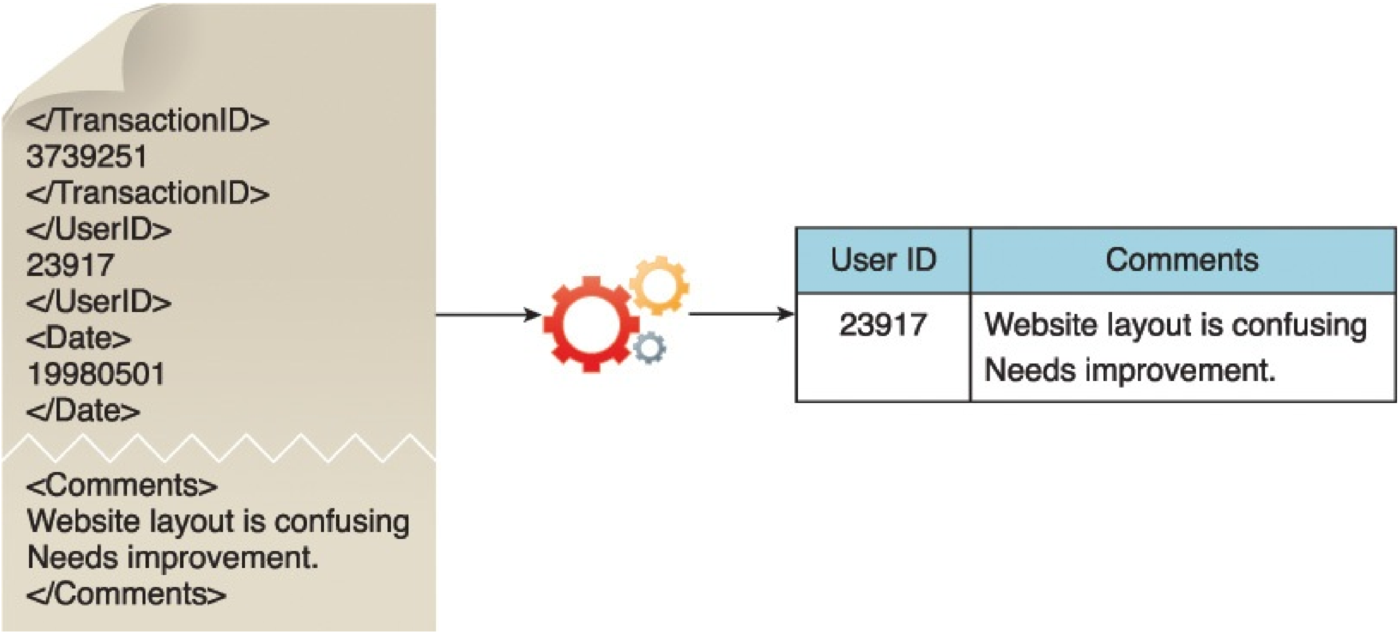


Figure 3.11 Stage 4 of the Big Data analytics lifecycle.

The extent of extraction and transformation required depends on the types of analytics 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.

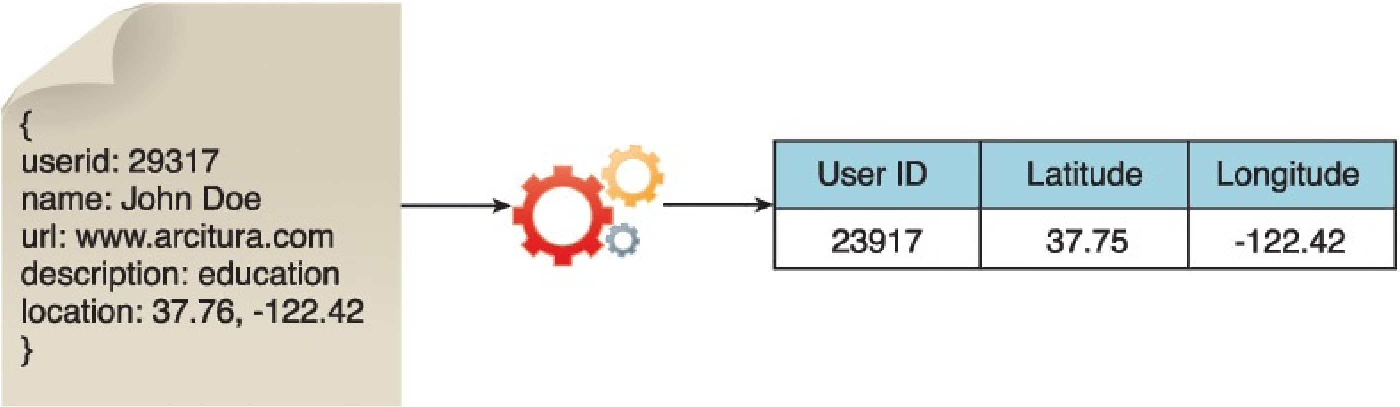


```
</TransactionID>
3739251
</TransactionID>
</UserID>
23917
</UserID>
<Date>
19980501
</Date>
<Comments>
Website layout is confusing
Needs improvement.
</Comments>
```

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.



```
{
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.

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

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 establishing often complex validation rules and removing any known invalid data.

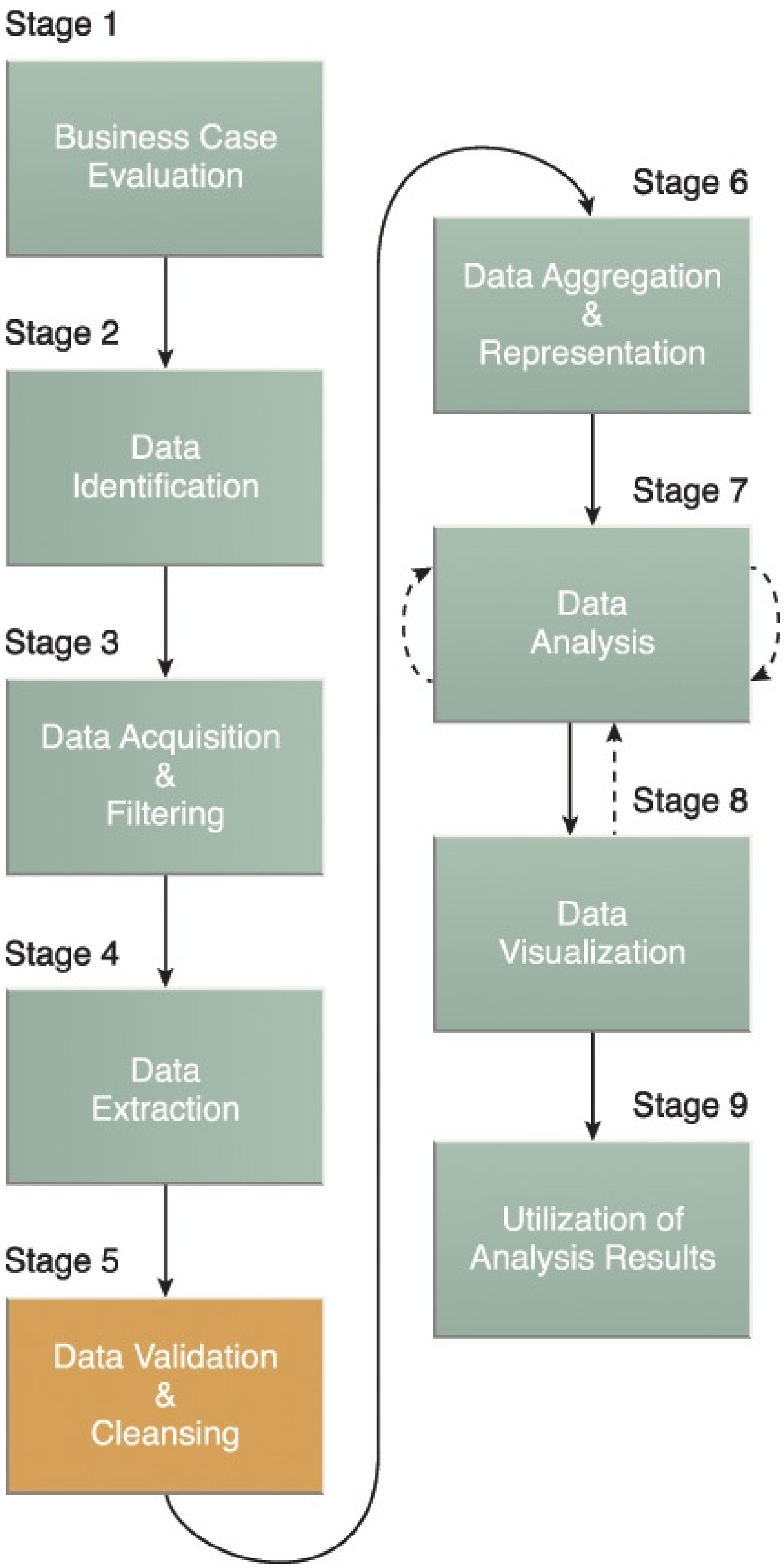


Figure 3.14 Stage 5 of the Big Data analytics lifecycle.

Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore interconnected datasets in order to assemble validation 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.

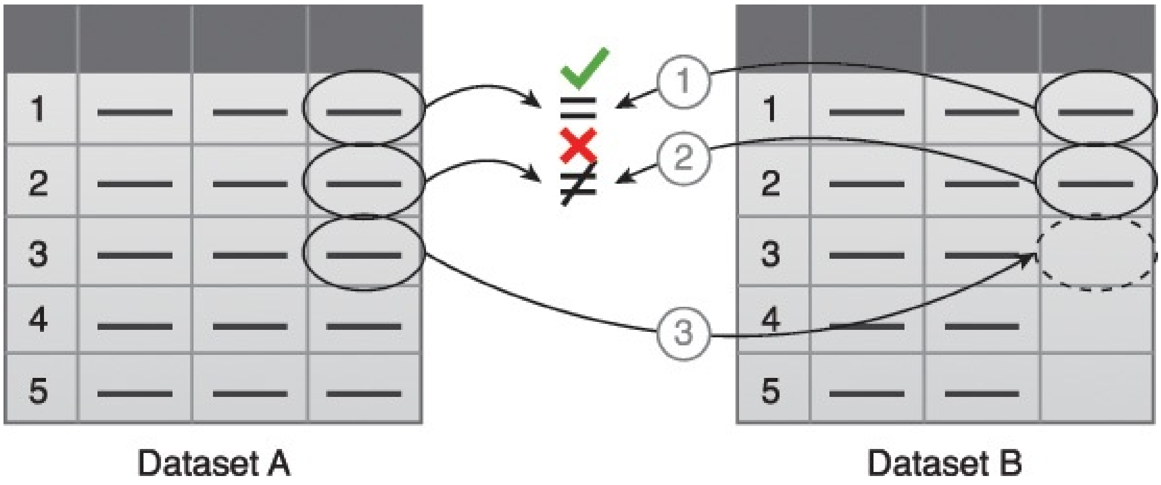


Figure 3.15 Data validation can be used to examine interconnected datasets in order to fill in missing valid data.

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.

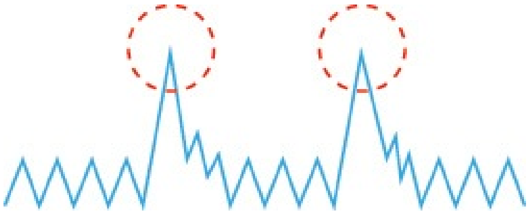


Figure 3.16 The presence of invalid data is resulting in spikes. Although the data appears abnormal, it may be indicative of a new pattern.

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 reconciliation 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.

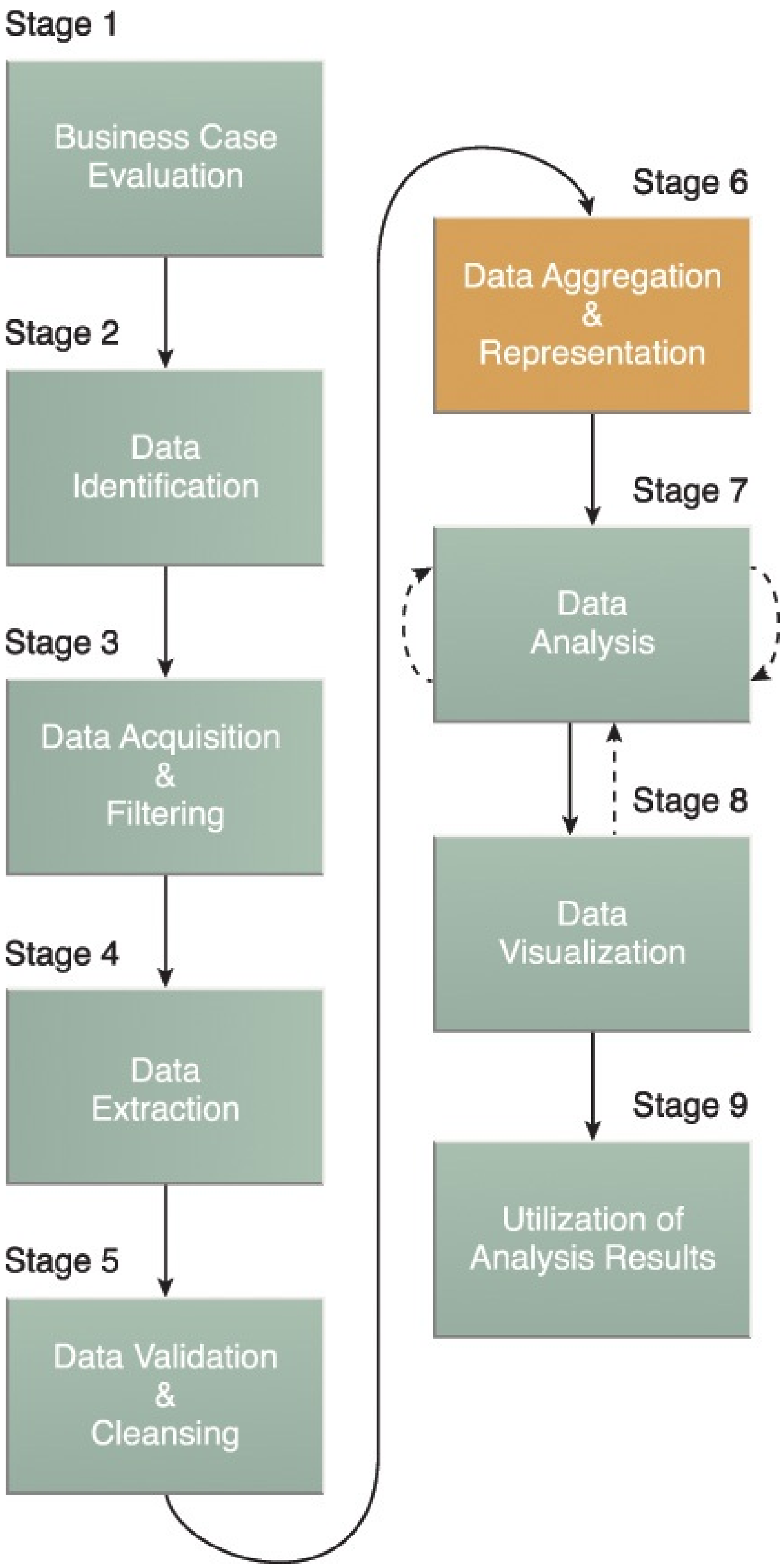


Figure 3.17 Stage 6 of the Big Data analytics lifecycle.

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 understand 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.

A data structure standardized by the Big Data solution can act as a common denominator 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.

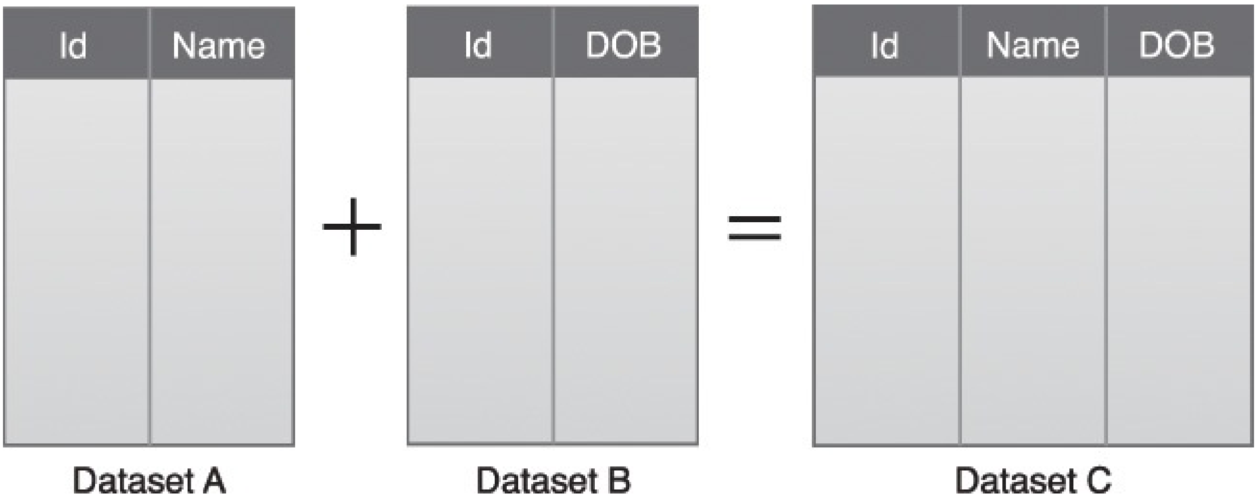


Figure 3.18 A simple example of data aggregation where two datasets are aggregated together using the Id field.

Figure 3.19 shows the same piece of data stored in two different formats. Dataset A contains 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.

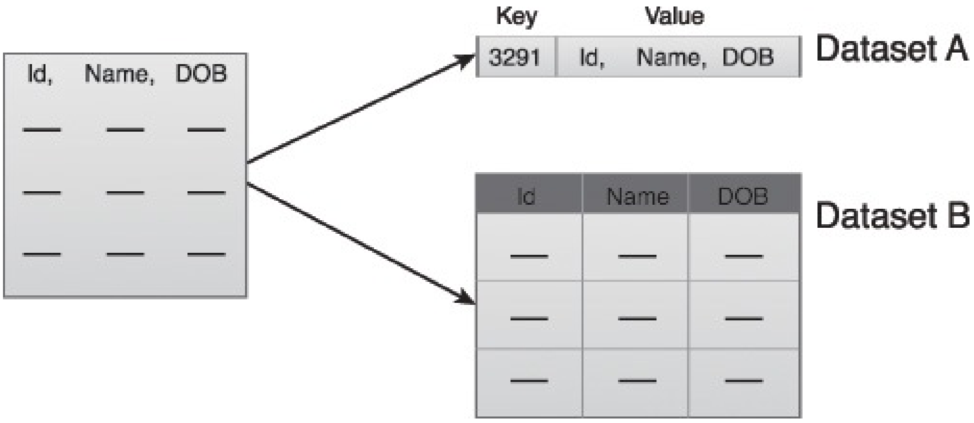


Figure 3.19 Dataset A and B can be combined to create a standardized data structure with a Big Data solution.

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.

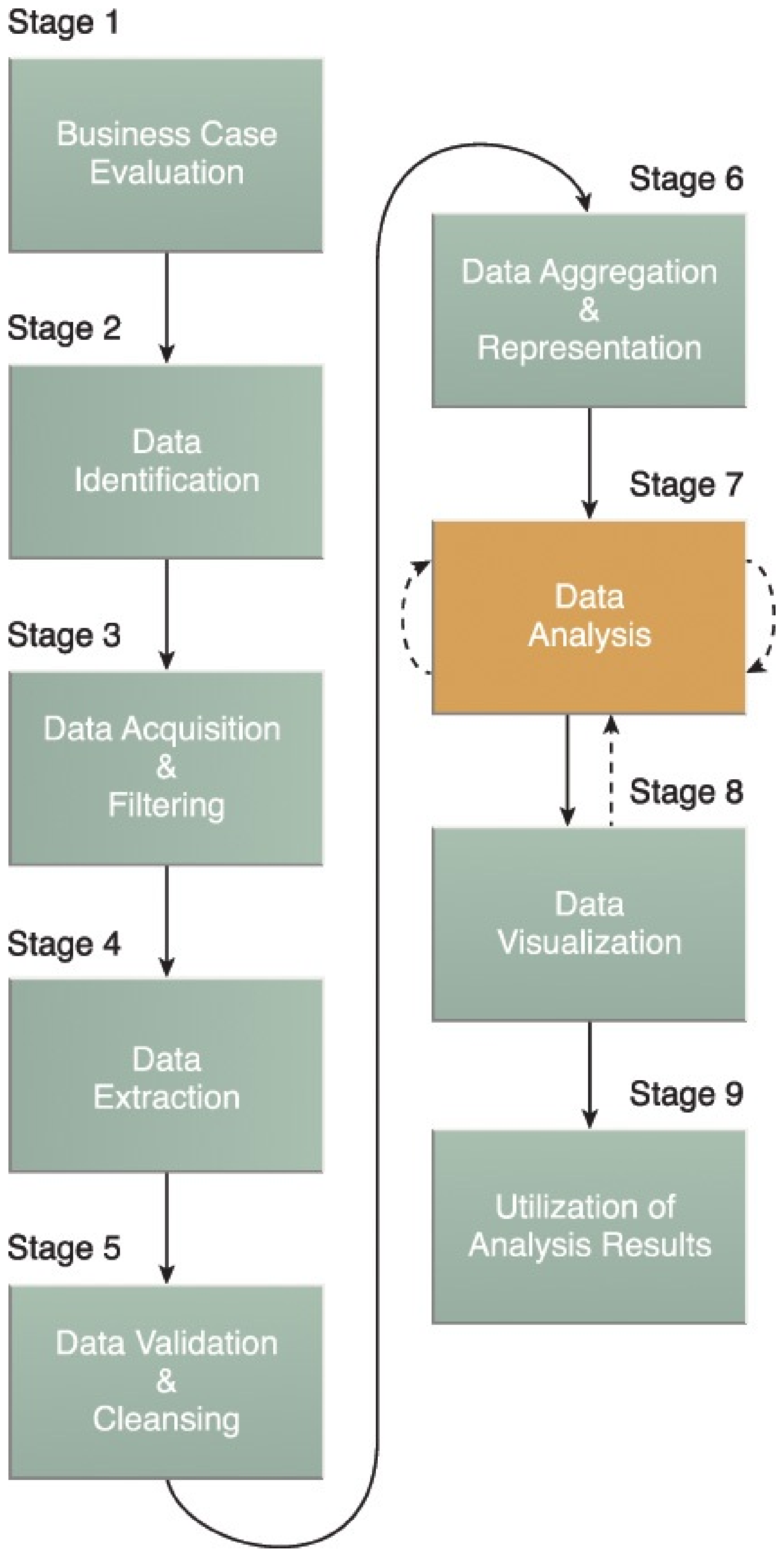


Figure 3.20 Stage 7 of the Big Data analytics lifecycle.

Depending on the type of analytic result required, this stage can be as simple as querying 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 latter of which is linked to data mining, as shown in Figure 3.21.

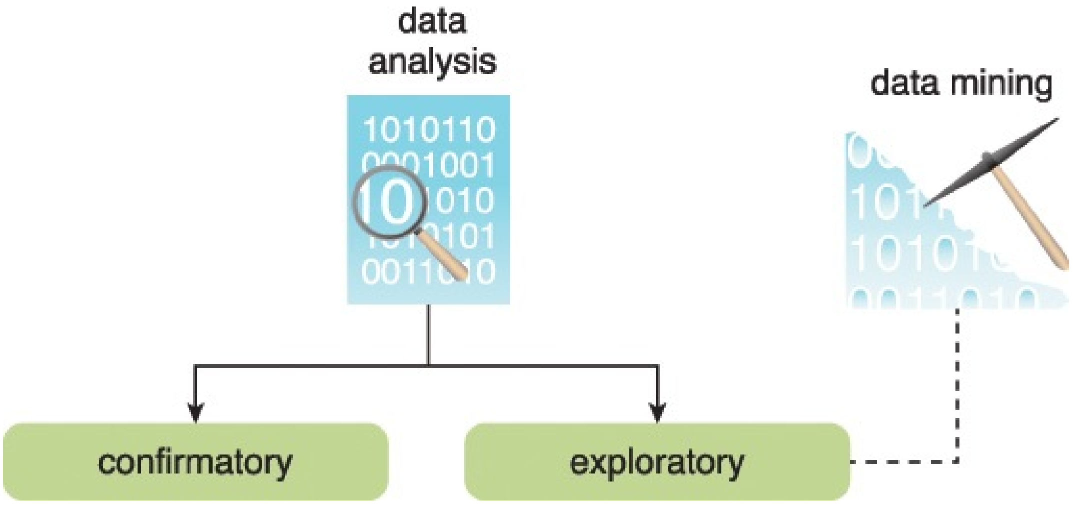


Figure 3.21 Data analysis can be carried out as confirmatory or exploratory analysis.

Confirmatory data analysis is a deductive approach where the cause of the phenomenon 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 predetermined 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 is explored through analysis to develop an understanding of the cause of the phenomenon. 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 visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

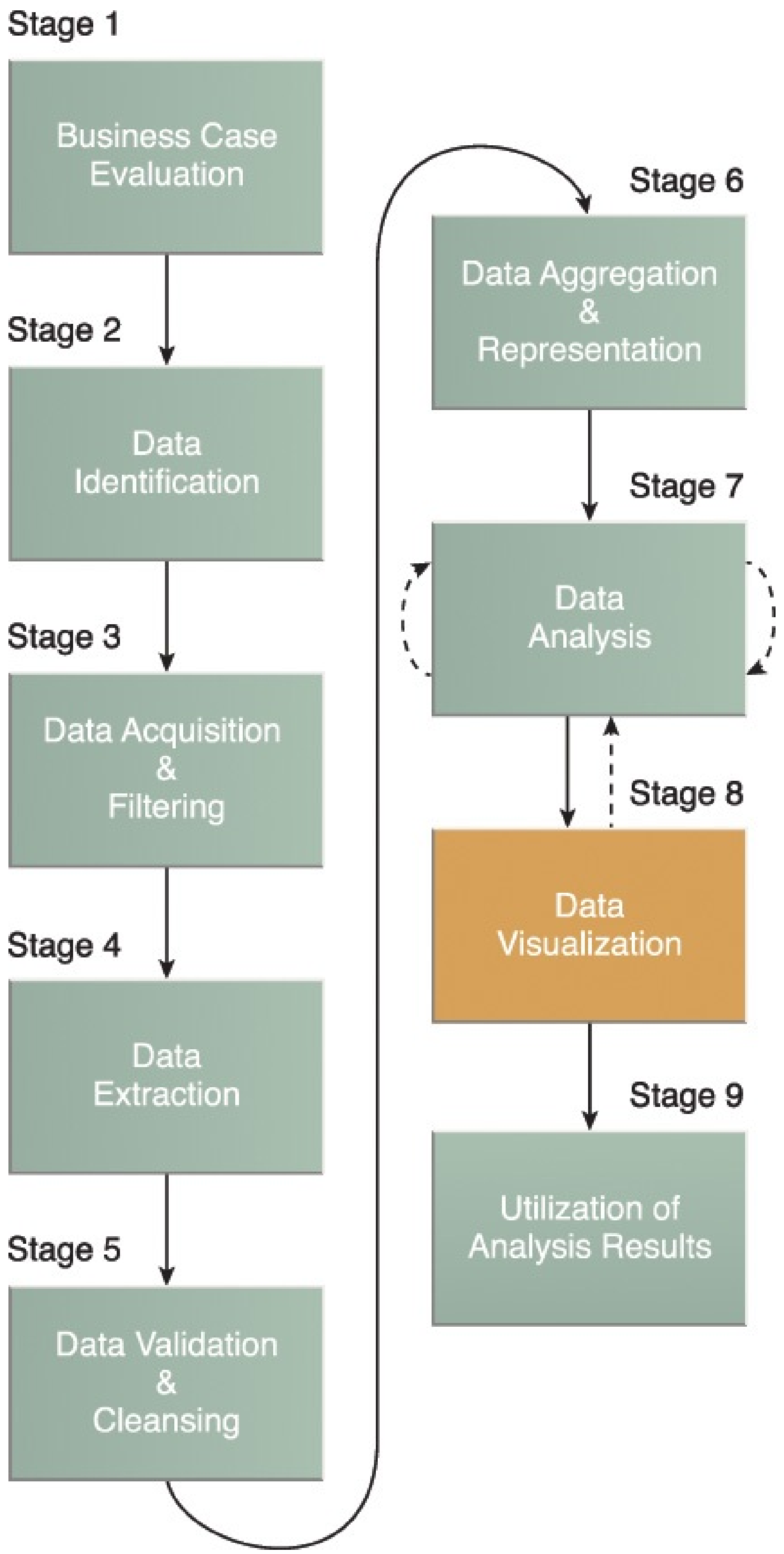


Figure 3.22 Stage 8 of the Big Data analytics lifecycle.

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 comparatively 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 opportunities 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.

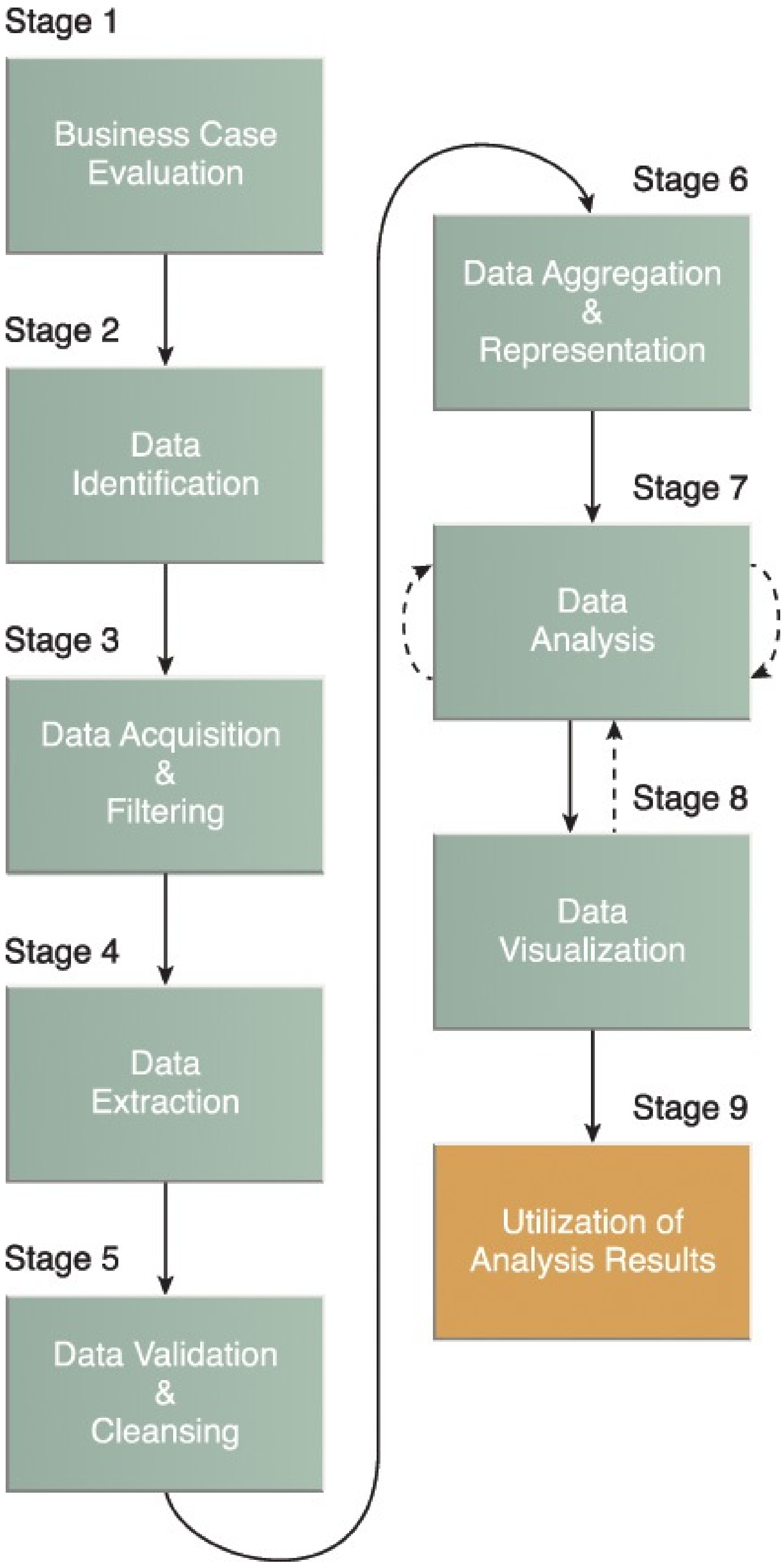


Figure 3.23 Stage 9 of the Big Data analytics lifecycle.

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 recommendations. 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.

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 feasibility 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 comprehend risk. The company hopes that, as a result, it can minimize losses by only accepting 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 obtaining 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 provides 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 satisfied 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 framework is required to not only help them standardize the usage of disparate data sources but also fully comply with any data privacy-related regulations. Furthermore, 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 applicability 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 skillsets. 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 necessary 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 evaluation 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 corresponds to a decrease in monetary loss and hence carries complete business backing. 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 customers. Although insurance fraud can both be opportunistic and organized, opportunistic 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 technologies to support batch processing and therefore do not believe that a large, initial up-front investment is required for tooling. However, when they consider the broader Big Data analytics lifecycle, the team members realize that they should budget for the acquisition of additional data quality and cleansing tools and newer data visualization technologies. After accounting for these expenses, a cost-benefit analysis 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, incident 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 management system and the insurance application documents are obtained from the document 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 dataset: 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 suggests 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 census 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 skillset 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.