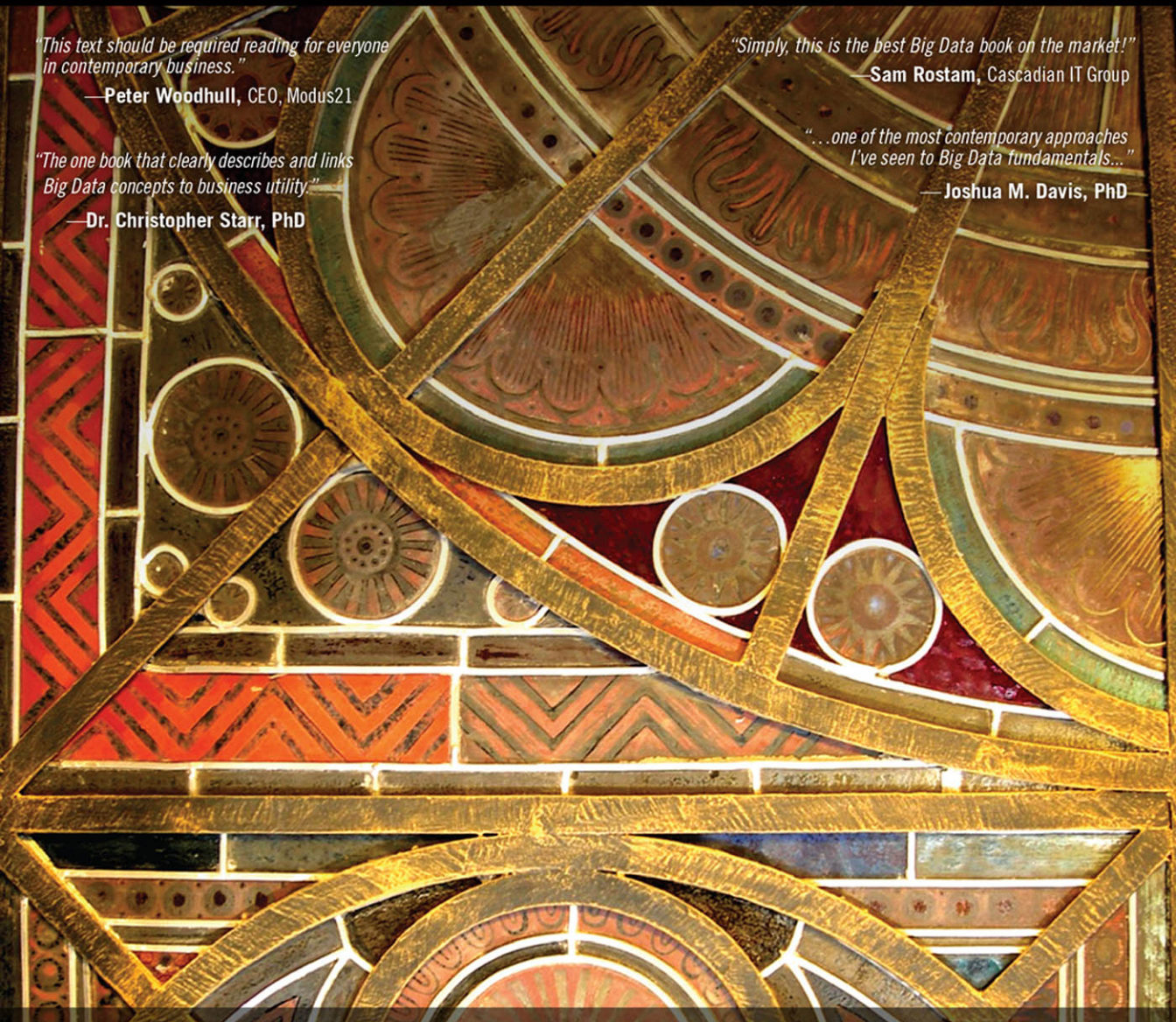
THE PRENTICE HALL SERVICE TECHNOLOGY SERIES FROM THOMAS ERL



"This text should be required reading for everyone in contemporary business."
—**Peter Woodhull,** CEO, Modus21

"The one book that clearly describes and links Big Data concepts to business utility."
—**Dr. Christopher Starr, PhD**

"Simply, this is the best Big Data book on the market!"
—**Sam Rostam,** Cascadian IT Group

"...one of the most contemporary approaches I've seen to Big Data fundamentals..."
—**Joshua M. Davis, PhD**

Big Data Fundamentals

Concepts, Drivers & Techniques

Co-authored and Edited by Best-Selling Author Thomas Erl
Co-authored by Wajid Khattak and Paul Buhler, PhD

PRENTICE HALL

ServiceTech PRESS

Big Data Fundamentals

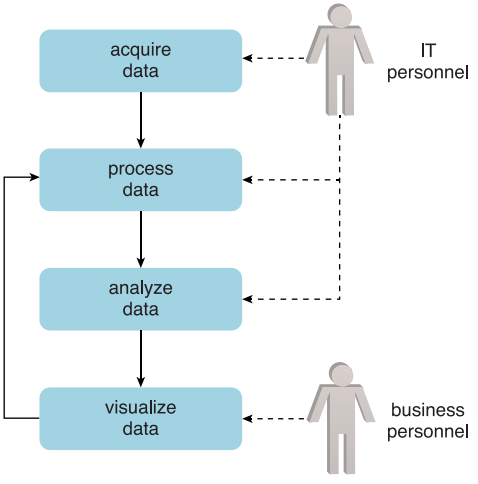


Figure 3.5
Each repetition can help fine-tune processing steps, algorithms and data models to improve the accuracy of results and deliver greater value to the business.

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 environment 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 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

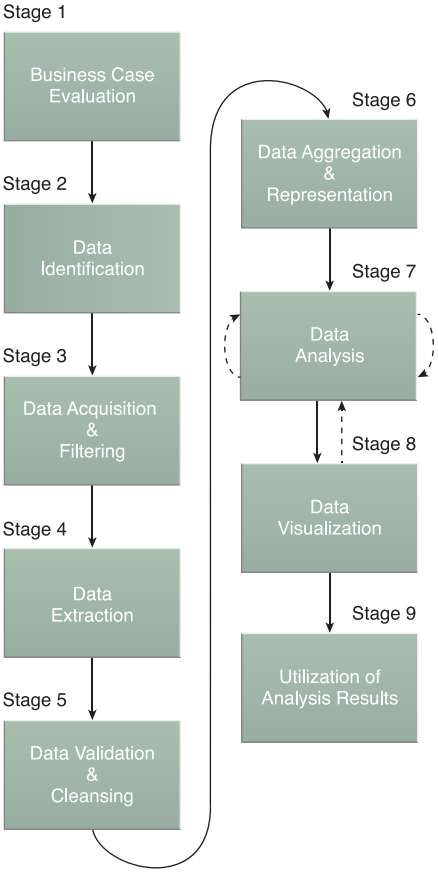


Figure 3.6
The nine stages of the Big Data analytics lifecycle.

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.

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

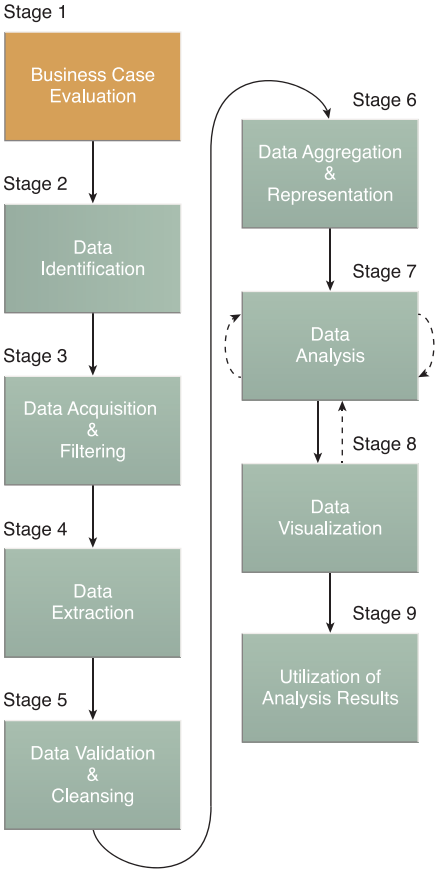


Figure 3.7
Stage 1 of the Big Data analytics lifecycle.

are not readily available, efforts should be made to make the goals of the analysis 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.

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 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.

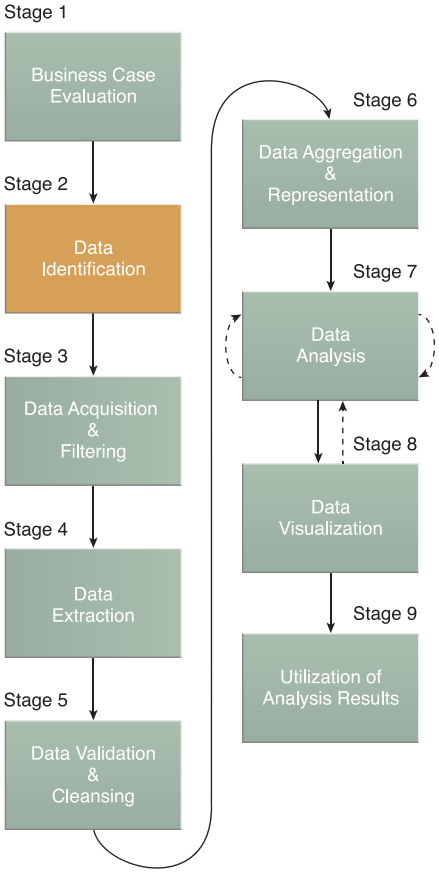


Figure 3.8
Data Identification is stage 2 of the Big Data analytics lifecycle.

Data Acquisition and Filtering

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

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.

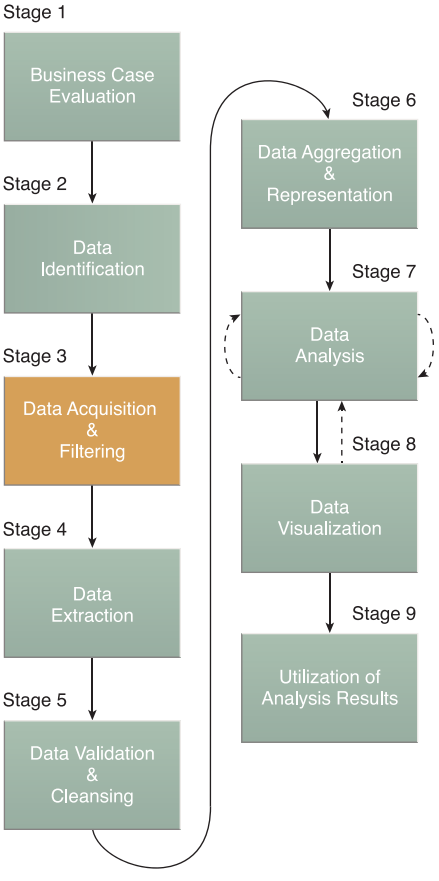


Figure 3.9
Stage 3 of the Big Data analytics lifecycle.

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.