Chapter 15
A Specification Framework for Big Data Initiatives

Anh D. Ta
Advanced Technology Consulting, LLC, USA

Marcus Tanque
Independent Researcher, USA

Montressa Washington
Independent Researcher, USA

ABSTRACT

Given the emergence of big data technology and its rising popularity, it is important to ensure that the use of this avant-garde technology directly addresses the enterprise goals which are required to maximize the Return-On-Investment (ROI). This chapter aims to address a specification framework for the process of transforming enterprise data into wisdom or actionable information through the use of big data technology. The framework is based on proven methodologies, which consist of three components: Specify, Design, and Refine. The recommended framework provides a systematic, top-down process to extrapolate big data requirements from high-level technical and enterprise goals. The framework also provides a process for managing the quality and relationship between raw data sources and big data products.

INTRODUCTION

This chapter addresses a specification framework for transforming the enterprise data into wisdom through the use of big data technology. The framework further integrates proven research results and methodologies to produce a holistic approach that is required for specifying, designing, and refining the big data implementation. The framework also defines a systematic, top-down process for deriving big data requirements from high-level functional enterprise goals. The framework also presents an analytical process for managing the quality and relationships of raw data sources, along with intermediary data products from big data solution. This chapter further addresses other

DOI: 10.4018/978-1-4666-8122-4.ch015

areas of big data initiatives (i.e., best practices, technical solutions and theories), as well as other sources required for further investigations.

BACKGROUND

Big Data

As more aspects of business operations are automated, the volume of data collected is increasing exponentially. This growth introduces the field of big data-analytics (or big data). This study is also intended to provide strategic and technical approaches for rapidly leveraging large volume of data required to support the enterprise decision making process. The theory of big data is more than working with a large volume of data, but rather about exploring innovative solutions to overcome the existing constraints of conventional methods. This approach, however, is also aimed for distributing the processing and leveraging data, where it resides versus transferring the data to the processing environments. Big data-analytics is defined by Gartner as "… high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information process for enhanced insight and decision making." (Gartner, 2013; Sicular, 2013) In today's era, data is only useful for the decision-making process, if converted into wisdom or actionable information. Generally, the conversion of data into wisdom should follow the following hierarchy of cognition, i.e., data, information, knowledge, and wisdom (Ackoff, 1989). The insights gained from big data initiatives are contingent on the size and scope of available data. To date, IT organizations are beginning to assess different ways to leveraging external data. Typically, this concept is referred to as *open data*. Open data is combined with existing proprietary data to improve business insights (Chui et al., 2014). International Data Corporation (International Data Corporation, 2013) predicts that big data technology and services market will grow at a 27% compound annual growth rate (CAGR) to a high of $32.4 billion through 2017. Such prediction is a result of the gradually ascending growth rate resulting in six times more than the overall information and communication technology (ICT) market. As with any new technology, the majority of available research and publications on big data, focus on either software product features or a wide-ranging of technical implementation details. On the contrary, how big data solution must be integrated into the enterprise Governance, Risk and Compliance (GRC) process and service delivery lifecycle is paramount.

Apache Hadoop

In an effort to define several terms and context of big data solutions referenced throughout this chapter, an overview of Apache Hadoop, a popular and de facto standard of an open source product for big data solution will be discussed throughout this study. IT experts agree that Hadoop allows end-users to overcome their traditional limitations of data analytics by distributing data across multiple machines, while using the pooled power and storage to surmount monolithic bottlenecks.

Initially Hadoop was designed to fix a scalability issue with Nutch project, an open source crawler and search engine. Thus hadoop architecture consists of two components namely the Hadoop Distributed File System (HDFS) for storage and MapReduce for computational capabilities. These components, however, are aimed to support the data partition and parallel computation of large datasets. Hadoop storage and computational capabilities are also intended to scale their performance by adding hosts to a Hadoop cluster, where the clusters can involve thousands of hosts and the volumes sizes can be in the petabytes of data.

HDFS is designed to be a scalable, fault-tolerant, distributed storage system that works closely with MapReduce. HDFS is optimized for high throughput by leveraging large block