**1. Introduction**

**1.1 Background**

Big Data is the common term used to describe the deluge of data in our networked, digitized, sensor-laden, information driven world. The availability of vast data resources carries the potential to answer questions previously out of reach. Questions like: How do we reliably detect a potential pandemic early enough to intervene? Can we predict new materials with advanced properties before these materials have ever been synthesized? How can we reverse the current advantage of the attacker over the defender in guarding against cybersecurity threats?

Within this context, on 29 March, 2013 The White House announced the Big Data Research and Development Initiative[[1]](#footnote-1). The initiative’s goals were to help accelerate the pace of discovery in science and engineering, strengthen the national security, and transform teaching and learning by improving our ability to extract knowledge and insights from large and complex collections of digital data.

Six Federal departments and their agencies announced more than $200 million in commitments – spread across 80+ projects – that aimed to significantly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data. The initiative also challenged industry, research universities, and non-profits to join with the Federal government to make the most of the opportunities created by Big Data.

Despite the widespread agreement on the opportunities inherent to Big Data, a lack of consensus on some important, fundamental questions continues to confuse potential users and hold back progress. What are the attributes that define Big Data solutions? How is Big Data different from the traditional data environments and related applications we have encountered thus far? What are the essential characteristics of Big Data environments? How do these environments integrate with currently deployed architectures? What are the central scientific, technological, and standardization challenges that need to be addressed to accelerate the deployment of robust Big Data solutions?

The NIST Big Data program was formally launched on 13 June, 2012 to help answer some of the questions surrounding Big Data and to support the federal government effort to incorporate Big Data as a replacement for, or enhancement to, traditional data analysis systems and models where appropriate.

[Need some transition verbiage here. How did the first conference lead to the BD-PWG?]

On 19 June, 2013 NIST hosted the Big Data Public Working Group (BD-PWG) kickoff meeting to begin addressing those questions. The Group was charged with developing a consensus definition, taxonomy, reference architecture, and technology roadmap for Big Data that can be embraced by all sectors.

These efforts will help define and prioritize requirements for *interoperability*, *portability*, *reusability*, and *extendibility* for Big Data analytic techniques and technology infrastructure in order to support secure and effective adoption of Big Data.

The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable Big Data stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner.

Within the BD-PWG, the following working groups were chartered in order to provide a technically-oriented strategy and standards-based guidance for the federal Big Data implementation effort:

* Definitions and Taxonomies
* General Requirements
* Security and Privacy Requirements
* Reference Architectures
* Technology Roadmap

**1.2 Objectives**

In general terms, a reference architecture provides “an authoritative source of information about a specific subject area that guides and constrains the instantiations of multiple architectures and solutions”[[2]](#footnote-2). Reference architectures generally serve as a reference foundation for solution architectures and may also be used for comparison and alignment purposes.

The broad goal of the Reference Architecture working group is to develop a Big Data open reference architecture that:

1. Provides a common language for the various stakeholders
2. Encourages adherence to common standards, specifications, and patterns
3. Provides consistency of implementation of technology to solve similar problem sets

The reference architecture is intended to facilitate the understanding of the operational intricacies in Big Data. It does not represent the system architecture of a specific Big Data system; instead it is a tool for describing, discussing, and developing system-specific architectures using a common framework of reference.

It provides a generic high-level conceptual model that is an effective tool for discussing the requirements, structures, and operations inherent to Big Data. The model is not tied to any specific vendor products, services or reference implementation, nor does it define prescriptive solutions that inhibit innovation.

The design of the NIST Big Data reference architecture serves the following objectives:

1. To illustrate and understand the various Big Data components, processes, and systems, in the context of an overall Big Data conceptual model;
2. To provide a technical reference for U.S. Government departments, agencies and other consumers to understand, discuss, categorize and compare Big Data solutions; and
3. To facilitate the analysis of candidate standards for interoperability, portability, reusability, and extendibility.

**1.3 How This Report Was Produced**

The approach for developing this document involved four steps:

1. The first step was to announce a Big Data Reference Architecture Working Group open to the public in order to attract and solicit a wide array of subject matter experts and stakeholders in government, industry, and academia.
2. The second step was to gather available Big Data architectures and materials representing various stakeholders, different data types, and different use cases.
3. The third step was to examine and analyze the Big Data materiel to better understand existing concepts of Big Data, what it is used for, its goals, objectives, characteristics, and key elements, and then document the these using the Big Data taxonomies model.
4. The fourth step was to develop an open reference architecture based on the analysis of Big Data material and the inputs from the other NIST Big Data working groups.

**1.4 Structure of This Report**

The remainder of this document is organized as follows:

Section 2 contains high level requirements relevant to the design of the Reference Architecture.

Section 3 represents a generic big data system comprised of technology-agnostic function blocks interconnected by interoperability surfaces.

Section 4 describes the main components of the generic system.

Section 5 describes the main components of the generic system interfaces.

Section 6 addresses security and privacy.

Section 7 contains the Big Data taxonomy.

Appendix A lists the terms and definitions appearing in the taxonomy.

Appendix B contains the acronyms used in this document.

Appendix C and D lists the references used in the document.

1. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal [↑](#footnote-ref-1)
2. Department of Defense Reference Architecture White Paper, June 2010 [↑](#footnote-ref-2)