**NIST Big Data Working Group (NBD-WD)**

**NBD-WD-2013/M0201 v1**

|  |  |
| --- | --- |
| **Source:** | **Technology Roadmap Subgroup** |
| **Status:** | **Informational** |
| **Title:** | **Data Services Abstraction Layer Expansion** |
| **Author:** | **Keith W. Hare**  **Senior Consultant, JCC Consulting, Inc.**  **Vice Chair, ANSI INCITS DM32.2, Databases**  **Convenor, ISO/IEC JTC1 SC32 WG3 Database Languages** |
| **Date:** | **August 30, 2013** |

# Introduction

As the NIST Big Data Working Group Technology Roadmap Subgroup proceeds, I have volunteered to identify “standardization and adoption priorities through an understanding of what standards are available or under development.” In this effort, I’ve been looking at the drafts of the Reference Architecture in hopes of identifying places where related standards and/or standards efforts exist, and places where they are missing.

While I have not reviewed all of the use cases, there are a couple of requirements that I have not seen clearly articulated:

1. The ability to locate and identify data resources without having to know beforehand the data location or even that the data exists.
2. The ability to identify the types and semantics of data available in those data sets.
3. The ability to integrate the analysis of multiple disparate data sets without having to have previous knowledge of the location and contents of the data sets.

For example, I might have information on occurrences of some diseases in a particular region. I would like to see if there is a correlation between the disease occurrence, temperature, and precipitation. I don’t immediately have temperature and precipitation data and so would like to have my analysis tool locate the appropriate data and complete the analysis.

This implies some sort of data registry that can be queried to find what data exists, and how individual elements are represented (Are temperatures in Fahrenheit, Celsius, or Kelvin, are precipitation measurements in English or metric units, etc.).

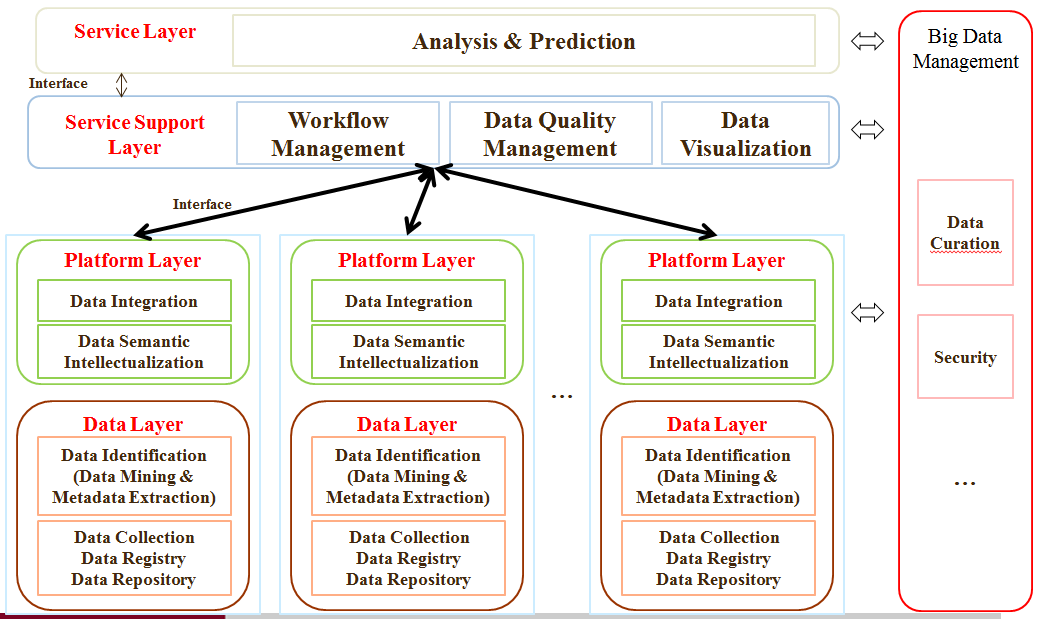
Once the data set has been located in the data registry, some sort of service layer is needed to pass queries and retrieve results in an implementation-agnostic manner.

This kind of capability would be useful both within organizations and between organizations.

All of this needs to occur within the bounds of the appropriate security and privacy requirements and restrictions.

The following conceptual architecture comes from

ISO/IEC JTC 1/SC32 N2386 “Next Generation Analytics & Big Data (A Reference Model for Big Data)”



The “Service Support Layer” is approximately equivalent to the “Data Service Abstraction” layer in M0126 v3“RA Proposal based from Earlier Proposals, Use Cases’ Requirements and Feedback”, page 2, while the “Platform/Data Layer pairs” correspond to instances of the “Data Manager” in the same diagram.

The benefit of showing multiple Platform/Data Layer instances is that it clearly shows that the goal is to be able to interact with more than one data source.