Taking Big Data Beyond the Hype

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) White Paper
Prepared for SAP

March 2013

EMA™
IT & DATA MANAGEMENT RESEARCH,
INDUSTRY ANALYSIS & CONSULTING

Table of Contents

More Than One Way to Execute on Big Data

Over the past year or two there has been much debate surrounding the topic of Big Data. Identifying a single definition or selecting a platform can seem impossible. The technology press has covered the trend as if it's a panacea and constantly combines the open source Big Data framework Hadoop with the topic causing confusion in the market. The origin of the term is even under debate, although many believe John R. Mashey, Chief Scientist at Silicon Graphics, coined it in the mid 1990s.[1] Mr. Mashey may have been first to use the term but he couldn't have foreseen the wide array of technology that would eventually enable it.[2] Journalists continue to cover the topic from a high level with inaccurate analysis that leads many to believe that the only way to deliver Big Data into the enterprise is by utilizing Hadoop technology. This is simply not the case.

Over the past several years, the landscape of data management has evolved well beyond traditional solutions. The drivers of this change are an unstoppable force causing IT and business to identify new ways to solve challenges in support of operational and analytic systems. Big Data presents an opportunity for companies to execute on strategies that were once thought impossible or at the very least impractical. At the heart of this evolution are four drivers creating a "perfect storm." Maturing Users and Applications, Economics, Technology Advances and Valuable Data Sources are shifting focus away from the Enterprise Data Warehouse (EDW) as the central hub of our data management ecosystem toward the practice of aligning data and workloads with platforms custom built to execute at higher levels. This shift is exactly why Big Data isn't just a Hadoop issue. In many cases, other platforms meet the capacity and performance criteria to process Big Data in a manner that better matches the specific needs of the enterprise.

Over the past several years, the landscape of data management has evolved well beyond traditional solutions.

Drivers of Change

Creating change within the data management ecosystem isn't an easy task. It requires drivers that are simple in definition but powerful as they relate to the critical initiatives of enterprise companies. Each of these drivers is timely and when combined causes significant change.

Creating change within the data management ecosystem isn't an easy task. It requires drivers that are simple in definition but powerful as they relate to the critical initiatives of enterprise companies.

1. **Maturing Users and Applications** – There are clear shifts converging on our analytic and operational environments and great demands are being put on traditional systems to support the maturing needs of end users. A greater complexity coupled with a larger, more diverse population of users is taxing the systems beyond their abilities.
2. **Economics** – Commodity hardware, low cost storage and memory are creating an opportunity to address projects that once were beyond the fiscal reach of most companies. The ability to add new purpose-built solutions to the data management landscape is affordable for many and driving a decentralization of the EDW in favor of solutions better suited to specific needs.

[1] Steve Lohr, 02/01/2013, New York Times, The Origins of "Big Data": Etymological Detective Story http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

[2] John Mashey, 1998, Big Data and the next Wave of InfraStress, UseNIX Technical Conference 1999 http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

3. **Technology Advancements** – Moore's law[3] is alive and well in the enterprise. The systems available today eclipse the scale and performance of those in which we invested just a few short years ago. This ongoing improvement arc is fueling our ability to address Big Data and create value from it.
4. **Valuable Data Sources** – For years, we have been forced to ignore data sources that could prove valuable to our work processes and analytic insights. The combination of the above drivers now makes it possible to execute Big Data strategies and expand how we accomplish that task. New data sources that include social data, machine generated data and sensor data are all important to powering more complex analysis of our businesses.

Defining Big Data

There are many ways to view Big Data, and defining it has turned into a hobby for many analysts. There is prolific use of "V" words (voracity, value, volume, etc.) that many apply in an attempt to define what Big Data is, but at the core, it simply represents data sets that can no longer be easily managed or analyzed with traditional or common data management tools, methods and infrastructures. This data can be highly structured or unstructured, be static in its origin or streaming. More importantly than defining the term is understanding what metrics to utilize when selecting the proper platform to leverage the data. In the end it's not about definitions, it's about strategies to empower analytics and operational workloads that are critical to your company.

Requirements to Consider

Response – The need for platforms to respond at new speeds and scale has opened the door for new ways to leverage data and provide insights to end users. This is especially true in the area of Big Data analytics where response rates are a key component to the value these platforms can deliver. Sub-second data delivery is not necessary for all applications and data driven scenarios but it's clear that real-time use cases are growing in importance and becoming more critical to many companies. Platforms such as Big Data Frameworks, analytic databases, and appliances are part of this evolution and are powering new solutions with improved response.

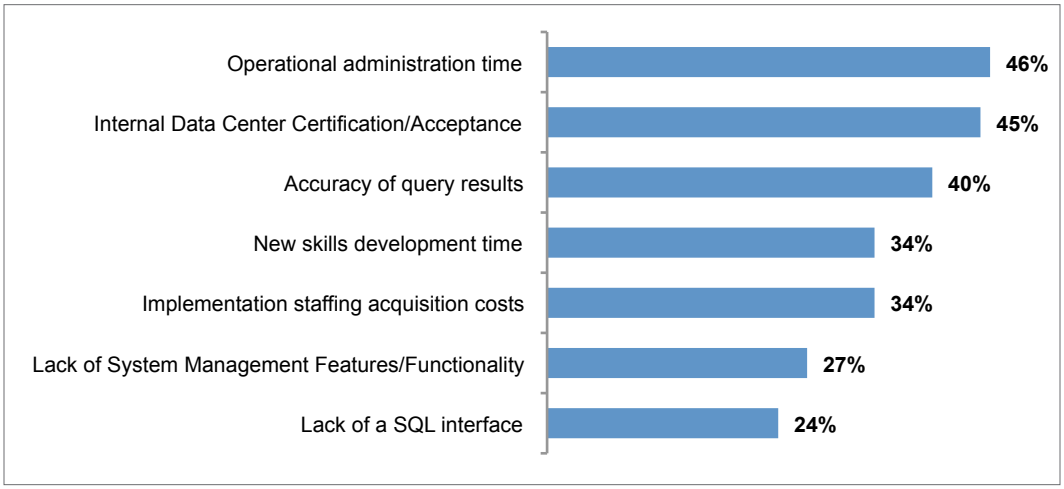
The need for platforms to respond at new speeds and scale has opened the door for new ways to leverage data and provide insights to end users.

Analytics / Workload Complexity – Addressing the requirements of complexity within analytic environments is getting more challenging, while running highly complex analytic models over massive data stores is becoming more commonplace. When coupling the ***workload complexity*** to the ***response,*** selecting the best platform can create powerful tools of differentiation. The ability to introduce new data types, such as social information or machine or process data can be leveraged to add even greater levels of insight and value.

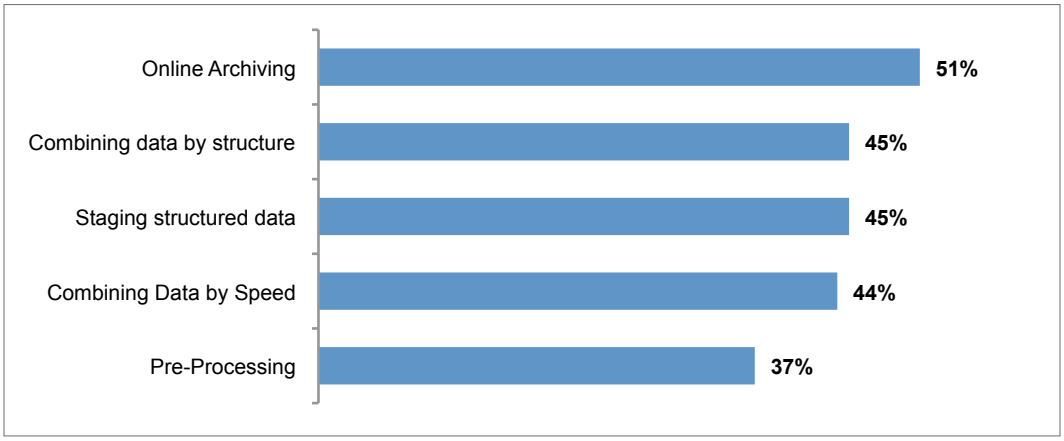
Economics – The economics of technology is the great equalizer and often can attribute to an early majority adoption of the technology. This has been especially true with Big Data. Many companies have identified needs to address ***response*** and ***workload complexity***, but the return on investment has slowed adoption. Big Data platforms are leveraging commodity hardware and often the software is free so it breaks through the economic barrier to adoption. Companies that plan to adopt Big Data should

[3] Moore's Law, Wikipedia http://en.wikipedia.org/wiki/Moore's_law

be warned that the barrier to entry is significantly reduced, but that doesn't mean it's cheap. Special skill sets are required, and lack of mainstream management tools create hidden costs that need to be taken into account before adopting this type of technology. There are alternatives such as analytic databases and appliances that can prove equally economical and leverage traditional skills sets already within your organization. ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) research shows that NoSQL environments such as Hadoop carry significant hurdles to adoption.[4]



Structure – Flexibility of structure is a growing decision point for selecting Big Data platforms. Big Data frameworks provide a level of flexibility not present in traditional data platforms. These systems can load and store data without requiring the time investment of designing and building complex data models. Analytics can be executed on these platforms without models and while running at speeds that eclipse many standard relational databases. Many users are employing "late binding" models to the data as they move it forward in the analytic process enabling a smaller set of data to be manipulated and leveraged. It is at this point that the data is often moved or accessed by another system designed to execute complex analytics or provide data to operational workflows. These platforms often utilize standard SQL and place the data into a relational data model for further analysis. EMA research has determined that Pre-Processing of data utilizing Hadoop is used by nearly 40% of users and is a common use case with the final analytic work executed on the best-suited platform.[5]



[4, 5] "Big Data Comes of Age" Enterprise Management Associates and 9sight Consulting, 2012, Shawn Rogers, John Myers and Dr. Barry Devlin

Load – Data loads are growing more complex and the sources are more diverse. Prompted by greater complexity and demand, Big Data adoption is driven by the need to provide flexibility. The power of Big Data platforms to load a mixture of data creates an opportunity to address both analytic and operational scenarios. Without this data to fuel these workloads, it would be impossible to execute against the growing demands of enterprise applications and analytic environments.

Including these five requirements in your planning for Big Data platforms will ensure that you select the best possible solution or combination with which to execute your Big Data strategy. It's critical to map out the analytic process to determine which platform(s) will deliver on the ***Response*** required; where best to execute ***complex workload***; which platform(s) presents the best ***Economic*** advantage; how best to leverage the data ***Structure,*** and lastly; how best to support the ***Load*** of the data.

Requirements at Work

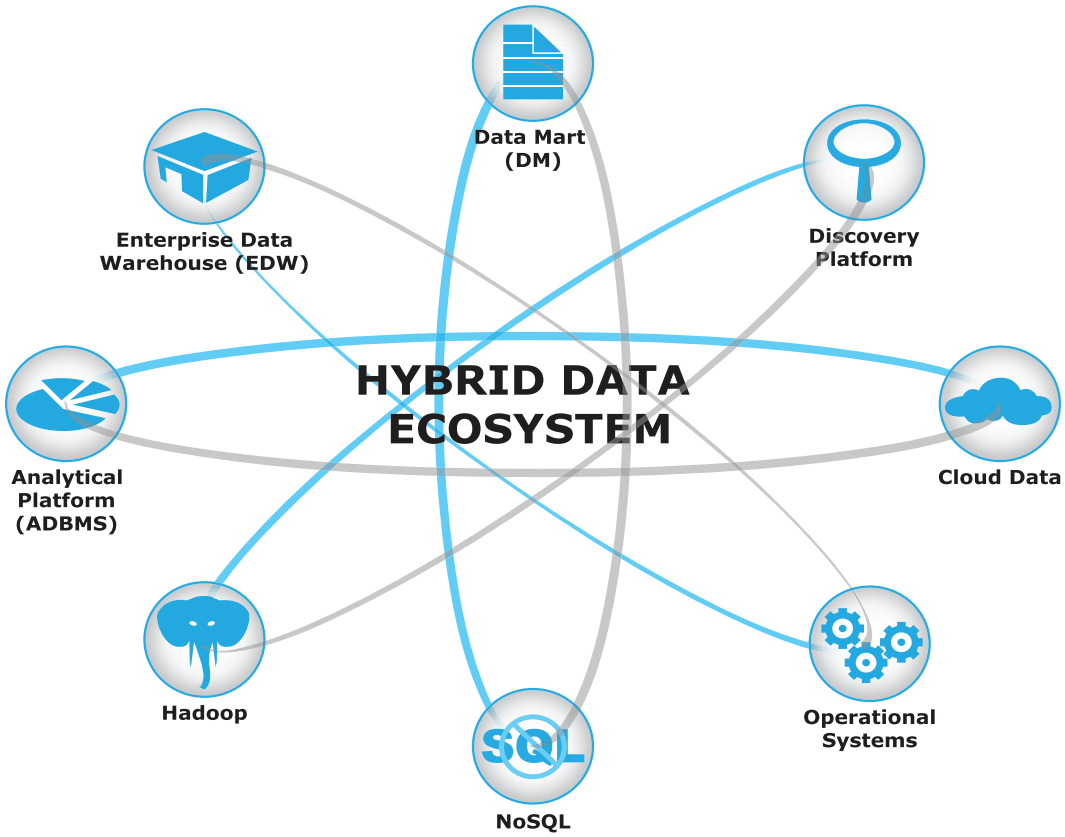
It is important to remember that most Big Data frameworks such as Hadoop don't support traditional SQL without additional modules such as Hive. In some instances, this is a great advantage of the platform as it is well designed to support textual analysis and programmatic processing such as MapReduce. But, for workloads that need to utilize counts, sums and group-by processing or connect several data sources, it is much easier to leverage a platform that utilizes relational models and supports standard SQL queries. This scenario brings into account both ***Structure*** and ***Workload Complexity***.

It is important to remember that most Big Data frameworks such as Hadoop don't support traditional SQL without additional modules such as Hive.

Another scenario puts these requirements to work against advanced analytics use cases. As illustrated in the chart above, on-line archiving is the leading use case for NoSQL platforms. EMA research shows 51% of NoSQL platform users are utilizing the system as warm data storage. This can be an effective strategy when compared to the ***Economics*** of traditional offline storage strategies, but for companies wanting to create value with advanced analytics and near-time operational applications (***Response***), there is a need to move the data beyond the NoSQL platform to a solution geared for this workload. Again, analytic platforms play a role in the ecosystem and can contribute to the process by leveraging in-memory technology or column-store schemas that are designed to serve solutions such as R and other advanced analytic tools.

Landscapes and Ecosystems

Throughout this EMA paper, many references have been made to the new or evolving data management landscape or ecosystem. Enterprise data warehouses continue to play an important role in this arena, but at the same time, they have migrated away from being the center of the data management universe creating a Hybrid Data Ecosystem that is more flexible and better prepared to meet the evolving needs of enterprise users and Big Data projects. At the center of the system are the requirements for determining which of the platforms or combination of platforms will serve the enterprise best. Big Data can be found across the entire field of solutions.



Enterprise Management Associates – Hybrid Data Ecosystem™

Components of the Hybrid Data Ecosystem

Operational systems: Business support systems such as website order entry applications, Point Of Sale (POS), Customer Relationship Management (CRM) or Supply Chain Management (SCM) applications. These platforms contain increasingly fine-grained information on transactions and demographics.

Enterprise data warehouse: Centralized analytical environments where corporate-level, reconciled and historical information of an organization is stored. These platforms have structured data organizations (schemas) based on time rather than present information.

Data mart: Often distributed analytical environments where a particular subject area or department level data set is stored for historical or other analysis. These platforms often have similar data organization to the enterprise data warehouse, but serve smaller user groups.

Analytical platform: Specifically architected and configured environments for providing rapid response times for analytical queries. These platforms are generally developed to support high-end analysis via tuned data structures like columnar data storage or indexing.

Discovery platform: Data discovery platforms support both standard SQL and programmatic API interfaces for iterative and exploratory analytics.

NoSQL: NoSQL data stores use non-traditional organizational structures such as key-value, wide-column, graph or document storage structures. These data stores support programming APIs and limited SQL variants for data access.

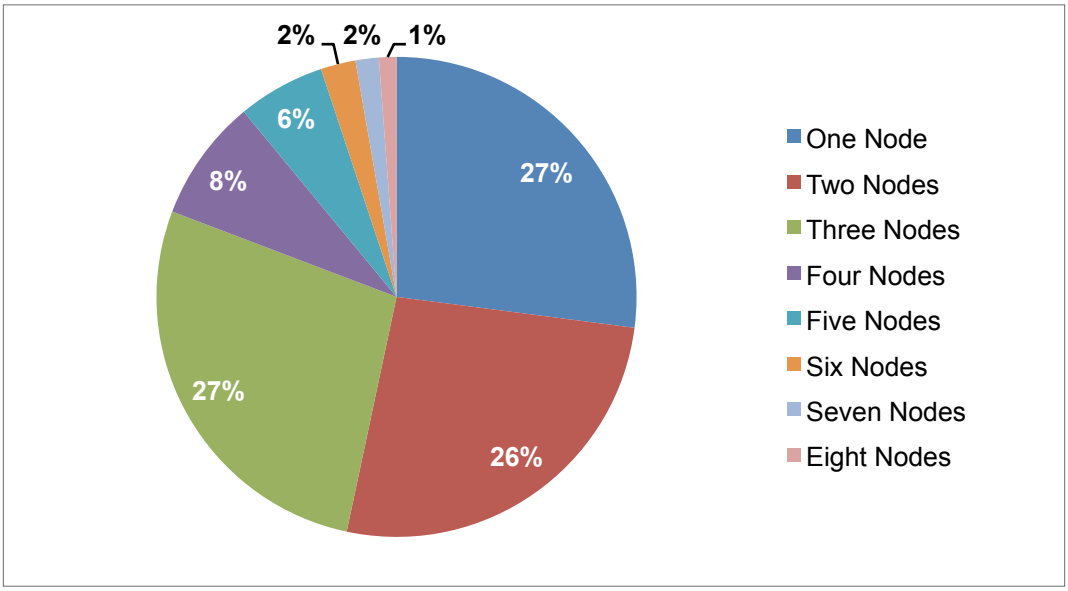
Hadoop: A specific variant of the NoSQL platform based on the Apache Hadoop Open Source project and its associated sub-projects. These platforms are based on Hadoop's Distributed File System (HDFS) storage and MapReduce processing framework.

Cloud: Cloud data sources make information available via standardized interfaces (APIs) and bulk data transfers. Examples are Dunn & Bradstreet D&B360, NOAA National Weather Service (NWS), API and social data aggregators.

Summary

Big Data brings new and exciting opportunities to companies who utilize the platforms available within the Hybrid Data Ecosystem. It properly applies the requirements of Response, Analytics, Complex Workloads, Economics, Structure and Load to their selection strategy. EMA research has determined that 53% of companies who are actively working on Big Data challenges already have two to three of the platforms. We found that 8% have four nodes and 6% have five within their ecosystem.

EMA research has determined that 53% of companies who are actively working on Big Data challenges already have two to three of the platforms.



Matching data and workloads to the best possible platform(s) helps IT and business to achieve their respective business goals while delivering greater business insights, quicker operational processes and overall better service through data. Companies who hold on to traditional views of the data landscape and refuse to leverage the best tools for the job will forever be mired in the hype of Big Data instead of enjoying the ability to be agile and more competitive within their markets.

About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help its clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals and IT vendors at www.enterprisemanagement.com or blogs.enterprisemanagement.com. You can also follow EMA on Twitter or Facebook.

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com

2626.021213