BIG DATA ANALYTICS
A PRACTICAL GUIDE

STEP 1:
GETTING YOUR DATA PLATFORM IN ORDER
INTRODUCTION

Everybody keeps extolling the virtues of Big Data analytics. This is fine. We love to hear stories of how Big Data analytics helps companies find the invisible needle in the haystack, double customer retention, triple cross-sell, and cut costs by 95%.

We get it. We WANT it. But we’re busy people. We don’t have a lot of time to decipher hieroglyphics. We need somebody to explain, in plain language, how to do it.

That’s what this white paper series is about. It’s a practical guide to planning, executing and benefiting from a Big Data analytics implementation. It’s in three parts:

- **Step 1**: Getting your Data Platform in Order
- **Step 2**: Building the Right Plan
- **Step 3**: Following your Plan to New Insights

In Step 1: Getting your Data Platform in Order, we’ll look at what makes Big Data different and lay out some of the things you’ll have to do to get ready to manage and work with it.
BIG DATA

WHAT MAKES BIG DATA DIFFERENT?

Big Data can mean a lot of things. Often, Big Data refers to a technology set or solution architecture. Sometimes the term is used to describe a business challenge. Other times, the term is used to refer to the data itself. And the definition keeps evolving. Here are a few definitions that all apply:

1. A collection of data assets that requires new forms of processing to enable enhanced decision making, to extract new insight or new discovery.
2. Data sets whose characteristics include high volume, high velocity, and a variety of data structures.
3. Data that cannot be processed using standard databases because it is too big, too fast moving or too complex for traditional data processing tools.
4. An evolving concept that refers to the growth of data and how to curate, manage, and process the data within performance goals.

We all know the amount of data created in the world is accelerating at a stunning pace. It’s estimated that:

- 35 zettabytes of data will be produced in 2020
- That’s 35 trillion terabytes
- 44 times the amount of data that was produced in 2009

Social networks, sensors, and machines are all contributing to this new growth of data. And much of this new data is unstructured, less ordered, and more interrelated than traditional data.

That means these new massive data sets can no longer be easily managed or analyzed with traditional or common data management tools, methods, and infrastructures. New techniques and tools are required to harness massive data sets to improve decision-making.

To meet these challenges, enterprises are exploring Big Data technologies to discover facts they didn’t know before. Big Data represents a significant paradigm shift in enterprise technology. It’s rapidly changing the traditional data analytics landscape across all industries.
NEW TOOLS FOR BIG DATA

Some of the new tools that have evolved to manage and analyze Big Data include:

- Big Data-focused technologies – Hadoop is the most important example
- NoSQL databases – Hbase, MongoDB
- Massively Parallel Processing (MPP) databases – Greenplum, Vertica, and Teradata Aster Data
- In-Memory databases – from vendors like SAP (HANA), IBM, Oracle, and Teradata

These are all taking a radically different approach to data processing, analytics and applications from traditional tools and technologies.

HADOOP

At the center of the Big Data movement is an open-source software framework called Hadoop, which uses large numbers of computing nodes to capture valuable insights from Big Data sets and facilitate analytics. Hadoop Distributed File System (HDFS) and MapReduce are the two core components of Hadoop. HDFS provides high-performance access to data across Hadoop clusters while MapReduce provides a powerful framework for processing data sets across clusters of nodes. The combination of HDFS and MapReduce provides a software framework for processing vast amounts of data in parallel on large clusters of commodity hardware. Hadoop is the de facto standard for the distributed storing, processing and analyzing of extremely large sets of data.

NOSQL

Standing for “Not Only SQL,” NoSQL represents database technologies that are horizontally scalable, non-relational, distributed, and often don’t adhere to the principles of traditional relational databases. NoSQL databases are especially useful for storing non-traditional data items (images, video, documents) in such a way that they can be indexed, queried and searched efficiently. Some examples of NoSQL databases are Hbase, MongoDB, Cassandra, CouchBase, and MarkLogic.

MASSIVELY PARALLEL PROCESSING (MPP) DATABASES

MPP databases use a “shared-nothing” massively parallel processing architecture to deliver Big Data Analytics and Discovery. These databases do not share any space or physical/virtual architecture. As a result queries can run more quickly and in parallel as they are broken up and executed by the distributed database segments. Some examples of MPP databases are Teradata, Greenplum, and Vertica.

IN-MEMORY DATABASES

In-memory databases rely on main-memory (server RAM) as opposed to disk-based storage to store and query data. In-memory technology as an enterprise solution is relatively new to the Big Data landscape, because until recently server memory was too expensive to use for large data sets. But as server memory costs have dropped, large databases can be run within server memory. This means that multiple terabytes of data can be analyzed almost in real time. SAP HANA has been the leader in this space, with new in-memory add-ons to traditional RDBMS systems coming from IBM, Oracle, and Teradata.
A Traditional EDW typically works with abstracted data that has been rolled up into a separate database for specific analytics. EDW databases are based on stable data models. They ingest data from enterprise applications like CRM, ERP and financial systems. Various Extract, Transform, Load (ETL) processes update and maintain these databases incrementally, typically on hourly, weekly and monthly schedules. A typical EDW runs from hundreds of gigabytes to multiple terabytes. For example Facebook runs an EDW (along with Big Data technology such as HDFS and MPP), that is around 50 terabytes in size.

ISSUES WITH TRADITIONAL DATA WAREHOUSING

A few EDW pain points are:

1. **Not so Real-Time Access.** Separating the database from operational data sources causes data availability issues. Batch window limitations also add to data latency.
2. **Ad-hoc Analysis.** The need to run ad-hoc analysis from time to time in addition to regular operational reporting degrades system response times.
3. **Expensive Change.** Changes to system and configuration are expensive due to rigid and inflexible design.
The explosive growth in the amount of data – from inputs such as sensor and machine data, transactional data, metadata, and social network data – has put tremendous pressure on the traditional EDW. New data collection processes are no longer centralized. Data is coming in all varieties and formats.

HDFS has emerged as an efficient scale out storage layer providing a low-cost landing zone for all such data that is at rest in the system. One of the key principles of Hadoop is the ability to store data “as is” and distill it to add the necessary structure only as needed.

Unlike an EDW, in which data is transformed into a specified schema when it is loaded into the warehouse, the “Schema On Read” principle eliminates the need for heavy ETL processing of data as it is ingested into the system. Analysts create the schema to suit the needs of their applications at the time they choose to analyze the data. This overcomes issues around the lack of structure and investing in data processing when there is questionable initial value of incoming data.

For organizations invested heavily in traditional data warehousing solutions, a hybrid solution that uses the best of both the technologies can be developed. Data warehousing platforms are very mature and offer an impressive list of features for the majority of data analytics needs. Sub-second access, aggregate join indexes, sparse join indexing, spatial indexing and materialized views are a few of the unique DW features.

Availability of third party tools especially BI tools such as Microstrategy, Tableau, IBM Cognos and others provide business users with direct access to data warehouse insights and are easy to implement against an EDW. And Big Data Solutions, whether they be Hadoop, MPP, NoSQL, In-Memory, or a combination of some or all of the above can work together with the traditional data warehouse in a single information supply chain.
As illustrated above, combining a Big Data Environment and underlying Analytics with the traditional EDW makes it possible to implement Big Data projects that efficiently use both traditional and non-traditional data sources. Traditional legacy systems can flow into the data warehouse environment through ETL processes. Meanwhile, unstructured, real-time and untraditional data sources can be accessed via the Big Data Environment. The Traditional Data Warehouse and the Big Data Environment can function as an integrated data platform, with data passing regularly between each.

Underlying analytics tools can pull any data, real-time, from source systems, create and validate a predictive and prescriptive analytics model through analytics iterations, and integrate this information into the data layer.

All data layer elements can be integrated and prepared for consumption through the Business Intelligence semantic and presentation layer and delivered via visualization and interactive dashboards and reports to the common user base.
Of course, not everyone’s data platform is going to look the same. You’ll need to put together the data platform that plays into the strengths of your organization, complements the existing technology footprint in place today, and utilizes the most effective tools to meet your data ingest and analysis needs. Typically this will be a dynamic combination of legacy and new technology, off-the-shelf and open source licensing and static and fluid data access methods.

Getting the right foundation in place – your Data Platform – will be critical to the success of all your Big Data initiatives. Making the effort to get the right tools and environment in place will be well worth it.

At DMI, we’ve successfully executed scores of Big Data projects for companies just like yours. Our Big Data Insights solutions drive enhanced business performance and millions in incremental revenue for companies like Christ Hospitals, Luxottica, Marzetti, Lane Bryant, McKesson, Lexmark, Teradata, and Vantiv. Our experts are ready to help you get your project underway.
ABOUT DMI

DMI IS A WORLD-LEADING PROVIDER OF SOLUTIONS AND SERVICES THAT LEVERAGE BIG DATA AND MOBILE TECHNOLOGIES TO ENHANCE BUSINESS PERFORMANCE.

Our Big Data Insights solutions deliver better insights for better decisions and better results. Our mobile solutions combine the award-winning user experience design that has made us one of the top creators of consumer apps with the deep middleware and engineering expertise that we’ve used to build and manage enterprise applications for the most demanding IT departments in the world. DMI mobility solutions improve business processes, tap new revenue streams, build customer loyalty, and increase employee productivity. And we offer a full range of Managed Services to securely set up, configure, and manage your mobile devices, apps and data.

THE PROOF

- Our Big Data Insights solutions drive enhanced business performance and millions in incremental revenue for companies like Christ Hospitals, Luxottica, Marzetti, Lane Bryant, McKesson, Lexmark, Teradata, and Vantiv.
- We’ve built more than 200 mobile apps—in the past 12 months alone—for world-leading organizations—Bacardi, Toyota, Vodafone, The National Guard, Novartis, Unilever, and Universal Studios.
- We offer brilliant creative and user experience: Our mobile app development group was named the Best Branded App Developer at the 2013 Mobile Entertainment Awards.
- We have 750,000 devices under management for 100+ clients, including many Fortune 500 companies—Johnson & Johnson, Sears, The Associated Press, Allergan, and more.
- Our Managed Mobility Services and BYOD Management services have received top ratings from both Gartner and Forrester.
- We deliver secure mobile and eCommerce solutions for more than 50 leading consumer and business brands.
- We provide 24 x 7 x 365 mobile service support. DMI is the one call our customers need to make to resolve any issue—devices, apps, infrastructure, even carriers.
- We offer a full range of security options that include Federal-grade hardware-based security, two-factor authentication, secure container, and sophisticated encryption solutions.
- With our expertise and economies of scale, we can provide mobility management at a higher service level and at, on average, 20%-40% lower cost than most companies can do on their own.
- Pervasive excellence is our commitment to quality service. DMI is one of only a handful of companies that is CMMI L3 appraised for both application development and services, as well as ISO 9001:2008, ISO 27001:2005, and ISO 20000-1:2011 certified. Our average D&B Open Ratings performance score from our clients is 95/100.