

## BIG DATA ANALYTICS

*Jasmine Zakir, Minot State University, jasminezakir@outlook.com*  
*Tom Seymour, Minot State University, tom.seymour@minotstateu.edu*  
*Kristi Berg, Minot State University, kristi.berg@minostatateu.edu*

### ABSTRACT

*Today Big Data draws a lot of attention in the IT world. The rapid rise of the Internet and the digital economy has fuelled an exponential growth in demand for data storage and analytics, and IT department are facing tremendous challenge in protecting and analyzing these increased volumes of information. The reason organizations are collecting and storing more data than ever before is because their business depends on it. The type of information being created is no more traditional database-driven data referred to as structured data rather it is data that include documents, images, audio, video, and social media contents known as unstructured data or Big Data. Big Data Analytics is a way of extracting value from these huge volumes of information, and it drives new market opportunities and maximizes customer retention. This paper primarily focuses on discussing the various technologies that work together as a Big Data Analytics system that can help predict future volumes, gain insights, take proactive actions, and give way to better strategic decision-making. Further this paper analyzes the adoption, usage and impact of big data analytics to the business value of an enterprise to improve its competitive advantage using a set of data algorithms for large data sets such as Hadoop and MapReduce.*

**Keywords:** Big Data, Analytics, Hadoop, MapReduce

### INTRODUCTION

Big Data is an important concept, which is applied to data, which does not conform to the normal structure of the traditional database. Big Data consists of different types of key technologies like Hadoop, HDFS, NoSQL, MapReduce, MongoDB, Cassandra, PIG, HIVE, and HBASE that work together to achieve the end goal like extracting value from data that would be previously considered dead. According to a recent market report published by Transparency Market Research, the total value of big data was estimated at \$6.3 billion as of 2012, but by 2018, it's expected to reach the staggering level of \$48.3 billion that's almost a 700 percent increase [29]. Forrester Research estimates that organizations effectively utilize less than 5 percent of their available data. This is because the rest is simply too expensive to deal with. Big Data is derived from multiple sources. It involves not just traditional relational data, but all paradigms of unstructured data sources that are growing at a significant rate. For instance, machine-derived data multiplies quickly and contains rich, diverse content that needs to be discovered. Another example, human-derived data from social media is more textual, but the valuable insights are often overloaded with many possible meanings.

Big Data Analytics reflect the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods. From businesses and research institutions to governments, organizations now routinely generate data of unprecedented scope and complexity. Gleaning meaningful information and competitive advantages from massive amounts of data has become increasingly important to organizations globally. Trying to efficiently extract the meaningful insights from such data sources quickly and easily is challenging. Thus, analytics has become inextricably vital to realize the full value of Big Data to improve their business performance and increase their market share. The tools available to handle the volume, velocity, and variety of big data have improved greatly in recent years. In general, these technologies are not prohibitively expensive, and much of the software is open source. Hadoop, the most commonly used framework, combines commodity hardware with open-source software. It takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analyzing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although attention to technology isn't sufficient, it is always a necessary component of a big data strategy. This paper discusses some of the most commonly used big data technologies mostly open source that work together as a big data analytics system for leveraging large quantities of unstructured data to make more informed decisions.

## REVIEW OF LITERATURE

Big Data is a data analysis methodology enabled by recent advances in technologies that support high-velocity data capture, storage and analysis. Data sources extend beyond the traditional corporate database to include emails, mobile device outputs, and sensor-generated data where data is no longer restricted to structured database records but rather unstructured data having no standard formatting [30]. Since Big Data and Analytics is a relatively new and evolving phrase, there is no uniform definition; various stakeholders have provided diverse and sometimes contradictory definitions. One of the first widely quoted definitions of Big Data resulted from the Gartner report of 2001. Gartner proposed that, Big Data is defined by three V's volume, velocity, and variety. Gartner expanded its definition in 2012 to include veracity, representing requirements about trust and uncertainty pertaining to data and the outcome of data analysis. In a 2012 report, IDC defined the 4<sup>th</sup> V as value—highlighting that Big Data applications need to bring incremental value to businesses. Big Data Analytics is all about processing unstructured information from call logs, mobile-banking transactions, online user generated content such as blog posts and tweets, online searches, and images which can be transformed into valuable business information using computational techniques to unveil trends and patterns between datasets.

Another dimension of the Big Data definition involves technology. Big Data is not only large and complex, but it requires innovative technology to analyze and process. In 2013, the National Institute of Standard and Technology (NIST) Big Data workgroup proposed the following definition of Big Data that emphasizes application of new technology; Big Data exceed the capacity or capability of current or conventional methods and systems, and enable novel approaches to frontier questions previously inaccessible or impractical using current or conventional methods. Business challenges rarely show up in the appearance of a perfect data problem, and even when data are abundant, practitioners have difficulties to incorporate it into their complex decision-making that adds business value. In 2012, McKinsey & Company conducted a survey of 1,469 executives across various regions, industries and company sizes, in which 49 percent of respondents said that their companies are focusing big data efforts on customer insights, segmentation and targeting to improve overall performance [10] An even higher number of respondents 60 percent said their companies should focus efforts on using data and analytics to generate these insights. Yet, just one-fifth said that their organizations have fully deployed data and analytics to generate insights in one business unit or function, and only 13 percent use data to generate insights across the company. As these survey results show, the question is no longer whether big data can help business, but how can business derive maximum results from big data.

### **Predictive Analytics**

Predictive Analytics is the use of historical data to forecast on consumer behavior and trends [18]. It is the use of past/historical data to predict future trends. This analysis makes use of the statistical models and machine learning algorithms to identify patterns and learn from historical data [25]. Predictive Analysis can also be defined as a process that uses machine learning to analyze data and make predictions [22].

Sixty seven percent of businesses aim at using predictive analytics to create more strategic marketing campaign in future, and 68% sight competitive advantage as the prime benefit of predictive analysis [17]. Broadly speaking, predictive analysis can be applied in ecommerce for product recommendation, price management, and predictive search. Typically a large e-commerce site offers thousands of product and services for sale. Navigating and searching for a product out of thousands on a website could be a major setback to consumers. However, with the invention of recommender system, an E-Commerce site/application can quickly identify/predict products that closely suit the consumer's taste [24].

Using a technology called Collaborative Filtering a database of historical user preferences is created. When a new customer access the ecommerce site, the customer is matched with the database of preferences, in order to discover a preference class that closely matches the customer taste. These products are then recommended to the customer [24]. Another technology that is used in ecommerce is the clustering algorithm. Clustering algorithm works by identifying groups of users that have similar preferences. These users are then clustered into a single group and are given a unique identifier.

New customers cluster are predicted by calculating the average similarities of the individual members in that cluster. Hence a user could be a partial member of more than one cluster depending of the weight of the user's average opinion [24]. Advanced analytics is defined as the scientific process of transforming data into insight for making better decisions. As a formal discipline, advanced analytics have grown under the Operational Research domain. There are some fields that have considerable overlap with analytics, and also different accepted classifications for the types of analytics [2].

## **BIG DATA TECHNOLOGIES**

### **Apache Flume**

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store. Flume deploys as one or more agents, each contained within its own instance of the Java Virtual Machine (JVM). Agents consist of three pluggable components: sources, sinks, and channels. Flume agents ingest incoming streaming data from one or more sources. Data ingested by a Flume agent is passed to a sink, which is most commonly a distributed file system like Hadoop. Multiple Flume agents can be connected together for more complex workflows by configuring the source of one agent to be the sink of another. Flume sources listen and consume events. Events can range from newline-terminated strings in stdout to HTTP POSTs and RPC calls — it all depends on what sources the agent is configured to use. Flume agents may have more than one source, but at the minimum they require one. Sources require a name and a type; the type then dictates additional configuration parameters.

Channels are the mechanism by which Flume agents transfer events from their sources to their sinks. Events written to the channel by a source are not removed from the channel until a sink removes that event in a transaction. This allows Flume sinks to retry writes in the event of a failure in the external repository (such as HDFS or an outgoing network connection). For example, if the network between a Flume agent and a Hadoop cluster goes down, the channel will keep all events queued until the sink can correctly write to the cluster and close its transactions with the channel. Sink is an interface implementation that can remove events from a channel and transmit them to the next agent in the flow, or to the event's final destination and also sinks can remove events from the channel in transactions and write them to output. Transactions close when the event is successfully written, ensuring that all events are committed to their final destination.

### **Apache Sqoop**

Apache Sqoop is a CLI tool designed to transfer data between Hadoop and relational databases. Sqoop can import data from an RDBMS such as MySQL or Oracle Database into HDFS and then export the data back after data has been transformed using MapReduce. Sqoop also has the ability to import data into HBase and Hive. Sqoop connects to an RDBMS through its JDBC connector and relies on the RDBMS to describe the database schema for data to be imported. Both import and export utilize MapReduce, which provides parallel operation as well as fault tolerance. During import, Sqoop reads the table, row by row, into HDFS. Because import is performed in parallel, the output in HDFS is multiple files.

### **Apache Pig**

Apache's Pig is a major project, which is lying on top of Hadoop, and provides higher-level language to use Hadoop's MapReduce library. Pig provides the scripting language to describe operations like the reading, filtering and transforming, joining, and writing data which are exactly the same operations that MapReduce was originally designed for. Instead of expressing these operations in thousands of lines of Java code which uses MapReduce directly, Apache Pig lets the users express them in a language that is not unlike a bash or Perl script.

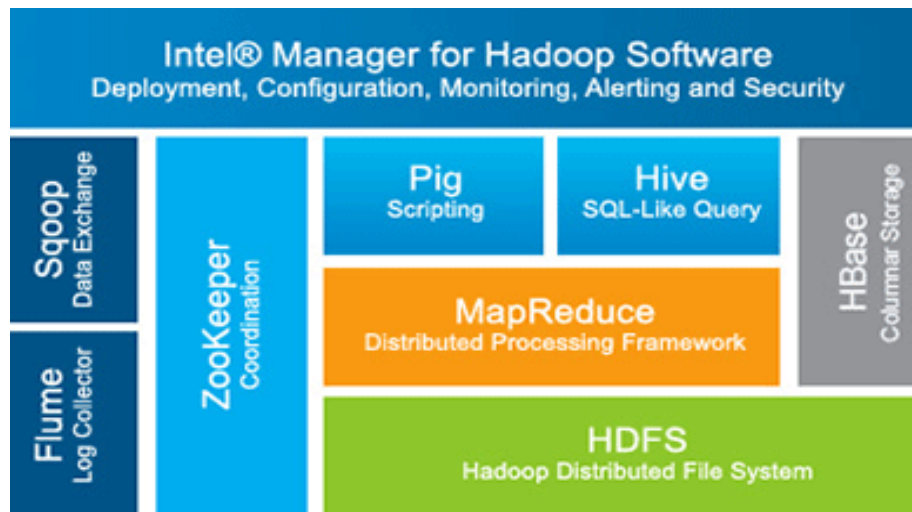
Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in 2007. Unlike SQL, Pig does not require that the data must have a schema, so it is well suited to process the unstructured data. But, Pig can still leverage the value of a schema if you want to supply one. PigLatin is relationally complete like SQL, which means it is at least as powerful as a relational algebra. Turing completeness requires conditional constructs, an infinite memory model, and looping constructs.

### Apache Hive

Hive is a technology developed by Facebook that turns Hadoop into a data warehouse complete with a dialect of SQL for querying. Being a SQL dialect, HIVEQL is a declarative language. In PigLatin, you specify the data flow, but in Hive we describe the result we want and hive figures out how to build a data flow to achieve that result. Unlike Pig, in Hive a schema is required, but you are not limited to only one schema. Like PigLatin and SQL, HiveQL itself is a relationally complete language but it is not a Turing complete language.

### Apache ZooKeeper

Apache Zoo Keeper is an effort to develop and maintain an open-source server, which enables highly reliable distributed coordination. It provides a distributed configuration service, a synchronization service and a naming registry for distributed systems. Distributed applications use ZooKeeper to store and mediate updates to import configuration information. ZooKeeper is especially fast with workloads where reads to the data are more common than writes. The ideal read/write ratio is about 10:1. ZooKeeper is replicated over a set of hosts (called an ensemble) and the servers are aware of each other and there is no single point of failure.



**Figure 1.** Intel Manager for Hadoop [3]

### MongoDB

MongoDB is an open source, document-oriented NoSQL database that has lately attained some space in the data industry. It is considered as one of the most popular NoSQL databases, competing today and favors master-slave replication. The role of master is to perform reads and writes whereas the slave confines to copy the data received from master, to perform the read operation, and backup the data. The slaves do not participate in write operations but may select an alternate master in case of the current master failure. MongoDB uses binary format of JSON-like documents underneath and believes in dynamic schemas, unlike the traditional relational databases. The query system of MongoDB can return particular fields and query set compass search by fields, range queries, regular expression search, etc. and may include the user-defined complex JavaScript functions. As hinted already, MongoDB practice flexible schema and the document structure in a grouping, called Collection, may vary and common fields of various documents in a collection can have disparate types of the data.

The MongoDB is equipped with the suitable drivers for most of the programming languages, which are used to develop the customized systems that use MongoDB as their backend player. There is an increasingly demand of using MongoDB as pure in-memory database; in such cases, the application dataset will always be small. Though, it is probably are easy for maintenance and can make a database developer happier; this can be a bottle neck for complex applications that require tremendous database management capabilities.

In order to efficiently address the challenges of Big Data, the leading vendor developed the Oracle NoSQL database. It was built by Oracle Berkeley DB team and the Berkeley DB Java Edition is the building block of Oracle NoSQL. Berkeley DB is a robust and scalable key-value store and used as the underlying storage for several popular data model such as Amazon Dynamo, GenieDB, MemcacheDB and Voldemort [28].

There are several other database systems that discern the strength of Berkeley DB and have attained greater scalability, throughput, and reliability with little tuning efforts. It is an efficient and a resilient transaction model that significantly eases the development process of applications, involving Big Data. It is a distributed, scalable yet simple key-value pair data model that fully supports the ACID transactions and JSON format and integrated with Oracle Database and Hadoop. It offers scalable throughput with bounded latency. The model very well accommodates the horizontal scaling with dynamic annexation of new capacity, citing high availability; the design architecture of Oracle NoSQL does not support single point of failure, and lucid load balancing. Actually, the goals of high availability, rapid failover in the event of a node failure, etc. are achieved by replicating the storage nodes.

### **Apache Cassandra**

Apache Cassandra is the yet another open source NoSQL database solution that has gained industrial reputation which is able to handle big data requirements. It is a highly scalable and high-performance distributed database management system that can handle real-time big data applications that drive key systems for modern and successful businesses. It has a built-for-scale architecture that can handle petabytes of information and thousands of concurrent users/operations per second as easily as it can manage much smaller amount of data and user traffic. It has a peer to peer design that offers no single point of failure for any database process or function, in addition to the location independence capabilities that equate to a true network-independent method of storing and accessing data, data can be read and written anywhere. Apache Cassandra is also equipped with flexible/dynamic schema design that accommodates all formats of big data applications, including structured, semi-structured, and unstructured data. Data is represented in Cassandra via column families that are dynamic in nature and accommodate all modifications online.

### **Apache Hadoop**

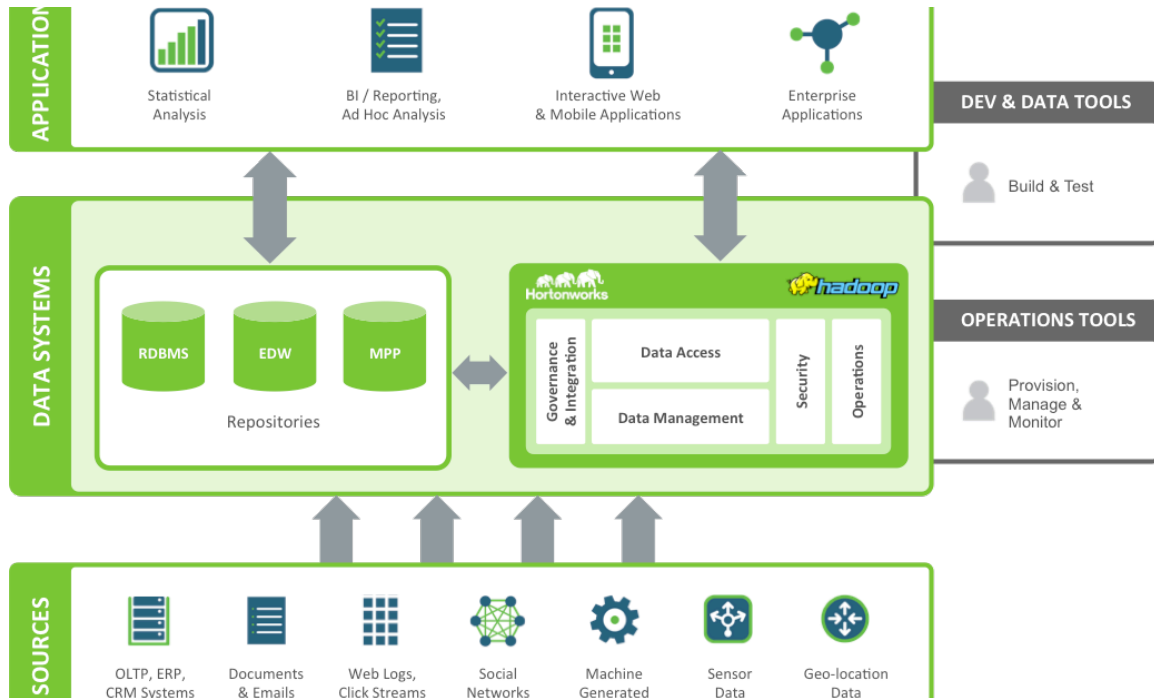
The Apache Hadoop software library is a framework that enables the distributed processing of large data sets across clusters of computers. It is designed to scale up from single servers to thousands of machines, with each offering local computation and storage. The basic notion is to allow a single query to find and collect results from all the cluster members, and this model is clearly suitable for Google's model of search support. One of the largest technological challenges in software systems research today is to provide mechanisms for storage, manipulation, and information retrieval on large amount of data. Web services and social media produce together an impressive amount of data, reaching the scale of petabytes daily (Facebook, 2012). These data may contain valuable information, which sometimes is not properly explored by existing systems. Most of this data is stored in a non-structured manner, using different languages and format, which, in many cases, are in compatible.

Parallel and distributed computing currently has a fundamental role in data processing and information extraction of large datasets. Over the last years, commodity hardware became part of clusters, since the x86 platform cope with the need of having an overall better cost/performance ratio, while decreasing maintenance cost. Apache Hadoop is a framework developed to take advantage of this approach, using such commodity clusters for storage, processing and manipulation of large amount of data. The framework was designed over the MapReduce paradigm and uses the HDFS as a storage file system. Hadoop presents key characteristics when performing parallel and distributed computing, such as data integrity, availability, scalability, exception handling, and failure recovery.

Hadoop is a popular choice when you need to filter, sort, or pre-process large amounts of new data in place and distill it to generate denser data that theoretically contains more information. Pre-processing involves filtering new data sources to make them suitable for additional analysis in a data warehouse. Hadoop is a top-level open source project of the Apache Software Foundation. Several suppliers, including Intel, offer their own commercial Hadoop distributions, packaging the basic software stack with other Hadoop software projects such as Apache Hive, Apache Pig, and Apache Sqoop. These distributions must integrate with data warehouses, databases, and other data



management products so data can move among Hadoop clusters and other environments to expand the data pool to process or query.



**Figure 2.** Data Architecture with Hadoop Integrated with existing data system [12]

MapReduce is the original massively scalable, parallel processing framework commonly used with Hadoop and other components such as the Hadoop Distributed File System (HDFS) and YARN. YARN can be described as a large-scale, distributed operating system for big data implementations. As Hadoop has matured, the batch-oriented, disk-intensive MapReduce's limitations have become more apparent as Big Data analytics moves to more real-time, streaming processing and advanced implementations such as the aforementioned machine learning.

### MapReduce

MapReduce is the model of distributed data processing introduced by Google in 2004. The fundamental concept of MapReduce is to divide problems into two parts: a map function that processes source data into sufficient statistics and a reduce function that merges all sufficient statistics into a final answer. By definition, any number of concurrent map functions can be run at the same time without intercommunication. Once all the data has had the map function applied to it, the reduce function can be run to combine the results of the map phases. For large scale batch processing and high speed data retrieval, common in Web search scenarios, MapReduce provides the fastest, most cost-effective and most scalable mechanism for returning results. Today, most of the leading technologies for managing "big data" are developed on MapReduce. With MapReduce there are few scalability limitations, but leveraging it directly does require writing and maintaining a lot of code.

### Apache Splunk

Splunk is a general-purpose search, analysis and reporting engine for time-series text data, typically machine data. Splunk software is deployed to address one or more core IT functions: application management, security, compliance, IT operations management and providing analytics for the business. The Splunk engine is optimized for quickly indexing and persisting unstructured data loaded into the system. Specifically, Splunk uses a minimal schema for persisted data – events consist only of the raw event text, implied timestamp, source (typically the

filename for file based inputs), source type (an indication of the general type of data) and host (where the data originated).

Once data enters the Splunk system, it quickly proceeds through processing, is persisted in its raw form and is indexed by the above fields along with all the keywords in the raw event text. Indexing is an essential element of the canonical “super-grep” use case for Splunk, but it also makes most retrieval tasks faster. Any more sophisticated processing on these raw events is deferred until search time. This serves four important goals: indexing speed is increased as minimal processing is performed, bringing new data into the system is a relatively low effort exercise as no schema planning is needed, the original data is persisted for easy inspection and the system is resilient to change as data parsing problems do not require reloading or re-indexing the data.

## **BIG DATA FRAMEWORK**

### **Apache Spark**

Apache Spark an open source big data processing framework built around speed, ease of use, and sophisticated analytics. It was originally developed in 2009 in UC Berkeley’s AMP Lab, and open sourced in 2010 as an Apache project. Hadoop as a big data processing technology has been around for ten years and has proven to be the solution of choice for processing large data sets. MapReduce is a great solution for one-pass computations, but not very efficient for use cases that require multi-pass computations and algorithms. Each step in the data processing workflow has one Map phase and one Reduce phase and you’ll need to convert any use case into MapReduce pattern to leverage this solution. Spark takes MapReduce to the next level with less expensive shuffles in the data processing. With capabilities like in-memory data storage and near real-time processing, the performance can be several times faster than other big data technologies.

Spark also supports lazy evaluation of big data queries, which helps with optimization of the steps in data processing workflows. It provides a higher-level API to improve developer productivity and a consistent architect model for big data solutions. Spark holds intermediate results in memory rather than writing them to disk, which is very useful especially when you need to work on the same dataset multiple times. It’s designed to be an execution engine that works both in-memory and on-disk. Spark operators perform external operations when data does not fit in memory. Spark can be used for processing datasets that larger than the aggregate memory in a cluster. Spark will attempt to store as much as data in memory and then will spill to disk. It can store part of a data set in memory and the remaining data on the disk. You have to look at your data and use cases to assess the memory requirements. With this in-memory data storage, Spark comes with a great performance advantage.

Spark is written in Scala Programming Language and runs on the Java Virtual machine. It currently supports programming languages like Scala, java, python, Clojure and R. Other than Spark Core API, there are additional libraries that are part of the Spark ecosystem and provide additional capabilities in Big Data analytics. Spark Streaming is one among the spark library that can be used for processing the real-time streaming data. This is based on micro based on micro batch style of computing and processing. Spark SQL provides the capabilities to expose the spark datasets over JDBC API and allow running the SQL like queries on Spark data using traditional BI and visualization tools. MLlib, GraphX are some other libraries from spark.

## **COMPETITIVE ADVANTAGES**

Thomas H. Davenport was perhaps the first to observe in his Harvard Business Review article published in January 2006 (“Competing on Analytics”) how companies who orientated themselves around fact based management approach and compete on their analytical abilities considerably out-performed their peers in the marketplace. The reality is that it takes continuous improvement to become an analytics-driven organization. In a presentation given at the Strata New York conference in September 2011, McKinsey & Company showed the eye opening; 10-year category growth rate differences (see Figure 7, below) between businesses that smartly use their big data and those that do not.

Amazon uses Big Data to monitor, track and secure 1.5 billion items in its inventory that are laying around 200 fulfillment centers around the world, and then relies on predictive analytics for its ‘anticipatory shipping’ to predict when a customer will purchase a product, and pre-ship it to a depot close to the final destination. Wal-Mart handles

more than a million customer transactions each hour [23], imports information into databases to contain more than 2.5 petabytes and asked their suppliers to tag shipments with radio frequency identification (RFID) systems [8] that can generate 100 to 1000 times the data of conventional bar code systems. UPS deployment of telematics in their freight segment helped in their global redesign of logistical networks [6]. Amazon is a big data giant and the largest online retail store. The company pioneered e-commerce in many different ways, but one of its biggest successes was the personalized recommendation system, which was built from the big data it gathers from its millions of customers' transactions.

The U.S. federal government collects more than 370,000 raw and geospatial datasets from 172 agencies and sub agencies. It leverages that data to provide a portal to 230 citizen-developed apps, with the aim of increasing public access to information not deemed private or classified. Professional social network LinkedIn uses data from its more than 100 million users to build new social products based on users' own definitions of their skill sets. Silver Spring Networks deploys smart, two-way power grids for its utility customers that utilize digital technology to deliver more reliable energy to consumers from multiple sources and allow homeowners to send information back to utilities to help manage energy use and maximize efficiency. Jeffrey Brenner and the Camden Coalition mapped a city's crime trends to identify problems with its healthcare system, revealing services that were both medically ineffective and expensive.

### CONCLUSION

Today's technology landscape is changing fast. Organizations of all shapes and sizes are being pressured to be data-driven and to do more with less. Even though big data technologies are still in a nascent stage, relatively speaking, the impact of the 3V's of big data, which now is 5v's cannot be ignored. The time is now for organizations to begin planning for and building out their Hadoop-based data lake. Organizations with the right infrastructures, talent and vision in place are well equipped to take their big data strategies to the next level and transform their businesses. They can use big data to unveil new patterns and trends, gain additional insights and begin to find answers to pressing business issues. The deeper organizations dig into big data and the more equipped they are to act upon what's learned, the more likely they are to reveal answers that can add value to the top line of the business. This is where the returns on big data investments multiply and the transformation begins. Harnessing big data insight delivers more than cost cutting or productivity improvement but it definitely reveals new business opportunities. Data-driven decisions always tend to be better decisions.

### REFERENCES

1. Apache Software Foundation. (2010). Apache ZooKeeper. Retrieved April 5, 2015 from <https://zookeeper.apache.org>
2. Chae, B., Sheu, C., Yang, C. and Olson, D. (2014). The impact of advanced analytics and data accuracy on operational performance: A contingent resource based theory (RBT) perspective, *Decision Support Systems*, 59, 119-126.
3. Chambers, C., Raniwala, A., Adams, S., Henry, R., Bradshaw, R., and Weizenbaum, N. (2010). Flume Java: Easy, Efficient Data-Parallel Pipelines. Google, Inc. Retrieved April 1, 2015 from <http://pages.cs.wisc.edu/~akella/CS838/F12/838-CloudPapers/FlumeJava.pdf>
4. Cisco Systems. Cisco UCS Common Platform Architecture Version 2 (CPA v2) for Big Data with Comprehensive Data Protection using Intel Distribution for Apache Hadoop. Retrieved March 15, 2015, from [http://www.cisco.com/c/en/us/td/docs/unified\\_computing/ucs/UCS\\_CVDs/Cisco\\_UCS\\_CPA\\_for\\_Big\\_Data\\_with\\_Intel.html](http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/Cisco_UCS_CPA_for_Big_Data_with_Intel.html)
5. DATASTAX Corporation. (2013, October). Big Data: Beyond the Hype - Why Big data Matters to you [White paper]. Retrieved March 15, 2015 from <https://www.datastax.com/wp-content/uploads/2011/10/WP-DataStax-BigData.pdf>
6. Davenport, T & Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, 90, 70-76.
7. Dhawan, S & Rathee, S. (2013). Big Data Analytics using Hadoop Components like Pig and Hive. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, 88, 13-131. Retrieved from <http://iasir.net/AIJRSTEMpapers/AIJRSTEM13-131.pdf>



8. Edwards, P., Peters, M. and Sharman, G. (2001). The Effectiveness of Information Systems in Supporting the Extended Supply Chain, *Journal of Business Logistics* 22 (1), 1-27
9. EMC Corporation. (2013, January). EMC Accelerates Journey to Big Data with Business Analytics-as-a-Service [White paper]. Retrieved from <http://www.emc.com/collateral/white-papers/h11259-emc-accelerates-journey-big-data-ba-wp.pdf>
10. EMC Corporation. Big Data, Big Transformations [White paper]. Retrieved from <http://www.emc.com/collateral/white-papers/idg-bigdata-umbrella-wp.pdf>
11. EMC Solutions Group. (2012, July). Big Data-as-a-Service [White paper]. Retrieved from <https://www.emc.com/collateral/software/white-papers/h10839-big-data-as-a-service-perspt.pdf>
12. Enterprise Hadoop: The Ecosystem of Projects. Retrieved from <http://hortonworks.com/hadoop/>
13. George, L. (2014, September). Getting Started with Big Data Architecture. Retrieved April 5, 2015, from <http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/>
14. IBM Corporation. IBM Big Data Platform. Retrieved from <http://www-01.ibm.com/software/in/data/bigdata/enterprise.html>
15. Intel Corporation. Big Data Analytics - Extract, Transform, and Load Big data with Apache Hadoop [White paper]. Retrieved April 3, 2015 from <https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf>
16. McClary, D. (2013, June). Acquiring Big Data Using Apache Flume. Retrieved March 3, 2015 from <http://www.drdoobs.com/database/acquiring-big-data-using-apache-flume/240155029>
17. Millard, S. (2013). Big Data Brewing Value in Human Capital Management – Ventana Research. Retrieved April 2, 2015 from <http://stephanmillard.ventanaresearch.com/2013/08/28/big-data-brewing-value-in-human-capital-management>
18. Mosavi, A. and Vaezipour, A. (2013). Developing Effective Tools for Predictive Analytics and Informed Decisions. Technical Report. University of Tallinn.
19. Oracle Corporation. (2013, March). Big Data Analytics - Advanced Analytics in Oracle Database [White paper]. Retrieved March 5, 2015 from <http://www.oracle.com/technetwork/database/options/advanced-analytics/advanced-analytics-wp-12c-1896138.pdf?ssSourceSiteId=ocomen>
20. Oracle Enterprise Architecture. (2015, April). An Enterprise Architect's Guide to Big Data - Reference Architecture Overview [White paper]. Retrieved from <http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf>
21. Penchikala, S. (2015, January). Big Data Processing with Apache Spark - Part 1: Introduction. Retrieved from <http://www.infoq.com/articles/apache-spark-introduction>
22. Puri, R. (2013). How Online Retailers Use Predictive Analytics To Improve Your Shopping Experience. Retrieved April 5, 2015 from <http://blogs.sap.com/innovation/analytics/how-online-retailers-use-predictive-analytics-to-improve-your-shopping-experience-0108060>
23. Sanders, N.R. (2014). Big Data Driven Supply Chain Management: A Framework for Implementing Analytics and Tuning Information into Intelligence, 1st Edition, Pearson, NJ
24. Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2002). Recommendation systems for large e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the fifth international conference on computer and information technology*, 1.
25. Shmueli, G. & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, 35(3), pp. 553-72.
26. Sorkin, S. (2011). Splunk Technical Paper: Large-Scale, Unstructured Data Retrieval and Analysis Using Splunk. Retrieved April 15, 2015 from <https://www.splunk.com/content/dam/splunk2/pdfs/technical-briefs/splunk-and-mapreduce.pdf>
27. The Bloor Group. IBM and the Big Data Information Architecture. Retrieved April 3, 2015 from <http://insideanalysis.com/wp-content/uploads/2014/08/BDIAVendor-IBMv01.pdf>
28. Tiwari, S. (2011). Using Oracle Berkeley DB as a NoSQL Data Store. Retrieved April 5 2015 from <http://www.oracle.com/technetwork/articles/cloudcomp/berkeleydb-nosql-323570.htm>
29. Transparency Market Report. (May, 2015). Big Data Applications in Healthcare likely to Propel Market to US\$48.3 Bn by 2018. Retrieved June 26, 2015, from <http://www.transparencymarketresearch.com/pressrelease/big-data-market.htm>
30. Villars, R. L., Olofson, C. W., & Eastwood, M. (2011, June). Big data: What it is and why you should care. IDC White Paper. Framingham, MA: IDC.

31. Wolpe, T. (2015, March). How Facebook is speeding up the Pesto SQL query engine. Retrieved April 3, 2015, from <http://www.zdnet.com/article/how-facebook-is-speeding-up-the-presto-sql-query-engine>
32. Zahari et al. (2010). Spark: Cluster Computing with Working Sets. Retrieved April 7, 2015, from [http://www.cs.berkeley.edu/~matei/papers/2010/hotcloud\\_spark.pdf](http://www.cs.berkeley.edu/~matei/papers/2010/hotcloud_spark.pdf)