Four Key Pillars To A Big Data Management Solution
EXECUTIVE SUMMARY ................................................................. 4
1. Big Data: a “Big” Term ............................................................. 4

EVOLVING BIG DATA USE CASES ............................................. 7
Recommendation Engines ............................................................ 7
Marketing Campaign Analysis ..................................................... 7
Customer Retention and Churn Analysis ...................................... 7
Social Graph Analysis ............................................................... 8
Capital Markets Analysis .......................................................... 8
Predictive Analytics ................................................................. 8
Risk Management ................................................................. 8
Rogue Trading .................................................................. 9
Fraud Detection ................................................................ 9
Retail Banking .................................................................. 9
Network Monitoring ........................................................... 9
Research and Development .................................................. 10

THE UNIQUE CHALLENGES OF BIG DATA ............................... 10
Limited Big Data Resources .................................................... 10
Poor Data Quality + Big Data = Big Problems .......................... 10
Project Governance ............................................................. 11

THE FOUR KEY PILLARS TO A BIG DATA MANAGEMENT SOLUTION ...... 12
1. Big Data Integration ........................................................... 12
2. Big Data Manipulation ....................................................... 12
3. Big Data Quality ............................................................. 13

TALEND AND BIG DATA: AVAILABLE TODAY .......................... 14
Talend Open Studio for Big Data ............................................. 14
Talend Enterprise Big Data .................................................... 14
Talend Platform For Big Data ................................................. 15

CONCLUSION ........................................................................... 15
## APPENDIX: TECHNOLOGY OVERVIEW

<table>
<thead>
<tr>
<th>Technology</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>MapReduce as a Framework</td>
<td>16</td>
</tr>
<tr>
<td>How Hadoop Works</td>
<td>16</td>
</tr>
<tr>
<td>Pig</td>
<td>17</td>
</tr>
<tr>
<td>Hive</td>
<td>17</td>
</tr>
<tr>
<td>HBase</td>
<td>17</td>
</tr>
<tr>
<td>HCatalog</td>
<td>18</td>
</tr>
<tr>
<td>Flume</td>
<td>18</td>
</tr>
<tr>
<td>Oozie</td>
<td>18</td>
</tr>
<tr>
<td>Mahout</td>
<td>18</td>
</tr>
<tr>
<td>Sqoop</td>
<td>18</td>
</tr>
<tr>
<td>NoSQL(Not only SQL)</td>
<td>18</td>
</tr>
</tbody>
</table>

## ABOUT TALEND

<table>
<thead>
<tr>
<th>Contact Us</th>
<th>19</th>
</tr>
</thead>
</table>
Executive Summary

Big data represents a significant paradigm shift in enterprise technology. Big data radically changes the nature of the data management profession as it introduces new concerns about the volume, velocity and variety of corporate data. This change means that we must also adjust technologies and strategies when fueling the corporation with valuable and actionable information. Big data enables companies to gain insight into new business opportunities (and threats) and stands to transform much of what the modern enterprise is today. Some clear conclusions that we can draw are as follows:

- Big data is addressing real market needs based on a new technological advance.
- While many are in an explorative phase looking at different use cases, several beneficial big data usage patterns have emerged.
- Data integration is crucial to managing big data, but project governance and data quality will be key areas of focus for big data projects.
- Big data projects will move from experimentation into a strategic corporate asset.
- Developer tools are needed to fuel adoption of these new technologies and reduce the current dependency on highly skilled developers. All major infrastructure and database vendors are bringing big data solutions to market.

1. Big Data: a “Big” Term

A clear definition of “big data” is elusive as what is “big” to one organization may not big to the next. It is not defined by a set of technologies; rather it defines a set of techniques and technologies. This is an emerging space and as we learn how to implement and extract value from this new paradigm, the definition shifts. While a distinct definition may be elusive, many believe that entire industries and markets will be affected and created as these capabilities enable new products and functionalities that were only just a dream or not even thought of.

1.1 The “Big” Data

As the name implies, big data is characterized by a size or volume of information, but this is not the only attribute to consider, there is also velocity and variety. Consider variety. Big data often relates to unstructured and semi-structured content, which may cause
issues for traditional relational storage and computation environments. Unstructured and semi-structured data appears in all kinds of places; for example web content, twitter posts and free form customer comments. There is also velocity or the speed at which data is created. With these new technologies, we can now analyze and use the vast amount of data that is created from website log files, social media sentiment analysis, to environmental sensors and video streams. Insight is possible where it was once not.

In order to get a handle on the complexities introduced by volume, velocity and variety, consider the following:

• Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data - the equivalent of 167 times the information contained in all the books in the US Library of Congress.
• Facebook handles 40 billion photos from its user base.
• Decoding the human genome originally took 10 years to process; now it can be achieved in one week
• Hortonworks, a Hadoop distribution, is managing over 42,000 Yahoo! machines serving up millions of requests per day.

These cutting-edge examples are becoming more and more common as companies realize that these huge stores of data contain valuable business information.

1.2 The “Big” Technology

In order to understand the impact of this new paradigm, it is important to have a basic understanding of the technologies and core concepts of big data. It is defined by an entire new set of concepts, terms and technologies. At the foundation of the revolution is a concept called MapReduce. MapReduce provides a massively parallel environment for executing computationally advanced functions in very little time.

Introduced by Google in 2004, it allows a programmer to express a transformation of data that can be executed on a cluster that may include thousands of computers operating in parallel.

At its core, it uses a series of “maps” to divide a problem across multiple parallel servers and then uses a “reduce” function to consolidate responses from each map and identify an answer to the original problem.
Many big data technologies are available as open source technologies, such as Hadoop, Pig, and Hive. Open source provides the benefits of standards-based interoperability, community development, cost savings over proprietary software, and continuous innovation.

A more detailed description of how the MapReduce technology works as well as a glossary of big data technologies can be found in the appendix of this paper.

1.3 The “Big” Paradigm Shift

Big data technologies have already revolutionized the way we live. Facebook, Groupon, Twitter, Zynga and countless other new business models have all been made possible because of this advance. This is a paradigm shift in technology that may end up bigger than the commercialization of the Internet in the late nineties. Entire industries and markets will be impacted as we learn how to use these capabilities to not only provide better delivery and advanced functionality in the products and services we provide today, but we will be able to deliver new capabilities that were once only a dream.

Take for instance the single view of the customer as delivered by master data management products. Solutions today rely on a somewhat static relational store to persist data and need to execute an algorithm in batch mode to create a single view. They are limited by current performance and storage limitations to use an explicit set of data. With Hadoop, these limitations are removed so that a single view of a customer can be created on the fly and incorporate more information such as transactional data. How would we use customer sentiment found in social media to expand the view of the customer?

This sort of advance may disrupt many existing markets. Consider ERP and data warehousing where big data is playing an important role in next generation data warehouse and analytics products. What about using big data technologies to replace an operational database? This type of radical thinking is not that far from reality as big data open source tools can be used to augment and in some respects, replace some of these functions and offer a new way to look at how we manage data today. We are at the genesis of massive change to technology that will spur massive societal change.

Big data changes everything.
Evolving Big Data Use Cases

Big data is an emerging space but some common industry use cases are developing that provide significant value today, for example:

**Recommendation Engines**

For years, organizations such as Amazon, Facebook and Google have used recommendation engines to match and recommend products, people and advertisements to users based on analysis of user profile and behavioral data. These large-volume data analysis problems were some of the first tackled by big data and have helped develop the technology into what it is today.

**Marketing Campaign Analysis**

The more information made available to a marketer, the more granular targets can be identified and messaged. Big data is used to analyze massive amounts of data that was just not possible with traditional relational solutions. Marketers are now able to better identify a target audience and identify the “right” person for the “right” products and service offerings. Big data allows marketing teams to evaluate large volumes from new data sources, like click-stream data and call detail records, to provide new insight into customer buying patterns and behavior.

**Customer Retention and Churn Analysis**

An increase in products per customer typically equates to reduced churn and many organizations have large-scale efforts to improve customer cross-sell and up-sell activities. However, analysis of customers and products across lines of business is often difficult as heterogeneous data formats and governance issues restrict these efforts. Some enterprises are able to load this data into a Hadoop cluster to perform wide scale analysis and identify patterns that indicate which customers are most likely to leave for a competing vendor or better yet, which customers are more likely to expand their relationship with the company. Action can then be taken to save or incent these customers.

---

Social Graph Analysis

There are users and there are “super” users in any social network or community and it is difficult to identify these key influencers within these groups. With big data, social networking data is mined to identify the participants that pose the most influence over others inside social networks. This helps enterprises ascertain the “most important” customers, who may or may not be the customers with the most products or spend as we traditionally identify with business analytics.

Capital Markets Analysis

Whether looking for broad economic indicators, specific market indicators, or sentiments concerning a specific company or its stocks, there is a wealth of data available to analyze in both traditional and new media sources. While basic keyword analysis and entity extraction have been in use for years, the combination of this old data with new sources such as Twitter and other social media sources provide great detail about public opinion in near real-time. Today, most financial institutions are using some sort of sentiment analysis to gauge public opinion about their company, market, or the economy as a whole.

Predictive Analytics

Within capital markets, analysts have used advanced correlation algorithms and probability calculations against current and historical data to predict market changes. The large amounts of historical market data, and the speed at which new data needs to be evaluated (e.g. complex derivatives valuations) make this a big data problem. The ability to perform these calculations faster and on commodity hardware makes big data a reliable substitute for the relatively slow and expensive legacy approach.

Risk Management

Advanced, aggressive organizations seek to mitigate risk with continuous risk management and broader analysis of risk factors across wider sets of data. Further, there is mounting pressure to increase the speed at which data is analyzed despite its growing volume. Big data technologies are growing in popularity to solve this issue as they parallelize data access and computation. Whether it is cross-party analysis or the integration of risk and finance, risk-adjusted returns and P&L require that growing amounts of data be integrated from multiple, standalone departments across the firm, and accessed and analyzed on the fly.
Rogue Trading

Deep analytics that correlate accounting data with position tracking and order management systems can provide valuable insights that are not available using traditional data management tools. In order to identify these issues, an immense amount of near real time data needs to be crunched from multiple, heterogeneous sources. This computationally intense function can now be accomplished using big data technologies.

Fraud Detection

Correlating data from multiple, unrelated sources raises the possibility of identifying fraudulent activities. For example, correlating credit card, debit card, smartphone, ATM, and online banking activities with web behavior analysis (either on the bank’s site or externally). In this context big data improves fraud detection.

Retail Banking

In retail banking, the ability to accurately assess the risk profile of an individual or a loan is critical to offering (or denying) services to a customer. Getting this right protects the bank and can create a very happy customer. With growing access to more complete information about customers, banks can target service offerings with a greater level of sophistication and certainty. Additionally, significant customer life events such as a marriage, childbirth, or a home purchase, might be predicted to help cross-selling and up-selling activities.

Network Monitoring

Big data technologies are used to analyze networks of any type. Networks, such as a transportation network, a communications network, a police protection network and even a local office network all can benefit from better analysis. Consider a local area network. With big data technologies massive amounts of data collected from servers, network devices and other IT hardware can allow administrators to monitor network activity and diagnose bottlenecks and other issues before they have an adverse effect on productivity.
Research and Development

Firms with a significant research and development staff, such as pharmaceutical manufacturers, use big data technologies to comb through enormous volumes of text-based research and other historical data to assist in the development of new products.

The Unique Challenges of Big Data

While big data presents a significant opportunity, it is not without its own set of challenges. It comprises a relatively new stack of technologies that is fairly complex to learn, there is no available set of tooling to fuel adoption and development, and there is a lack of knowledgeable big data resources available. In fact, most big data projects are just that, a project: so they have not yet been wrapped into the expected governance frameworks for enterprise grade project management and data governance. This will surely change. Let’s look at these challenges in more detail.

Limited Big Data Resources

The majority of developers and architects who “understand” big data are working for the original creators of big data technologies; companies like Facebook, Google, Yahoo and Twitter to name a few. There are others employed by numerous startups in this space like Hortonworks, Cloudera and MapR. The technology is still a bit complex to learn which restricts the rate at which new big data resources are available. Compounding this issue, in this nascent market are limited tools available to aid development and implementation of these projects.

Poor Data Quality + Big Data = Big Problems

Depending on the goal of a big data project, poor data quality can have a big impact on effectiveness. It can be argued that inconsistent or invalid data could have an exponential impact on analysis in the big data world. As analysis on big data grows, so too will the need for validation, standardization, enrichment and resolution of data. Even identification of linkages can be considered a data quality issue that needs to be resolved for big data.
**Project Governance**

Today, big data projects are typically a small directive from the CTO to “go figure this stuff out”. It is early in the adoption process and most organizations are trying to sort out potential value and create an exploratory project or SWAT team. Typically these projects go unmanaged. It is a bit of the “wild west”. As with any corporate data management discipline however, they will eventually need to comply with established corporate standards and accepted project management norms for the organization, deployment and sharing of project artifacts.

While there are challenges, the technology is stable. There is room for growth and innovation as the complete lifecycle of data management, including quality and governance can be shifted into this new paradigm. Big data technologies have rabid interest and the talent pool will catch up to support widespread adoption.
The Four Key Pillars to a Big Data Management Solution

Integration is the key enabler and code generation is the fuel.

To address the challenges outlined in the previous section, there are four key pillars to be considered when building a big data management solution: big data integration, big data manipulation, big data quality, and big data project management and governance. Talend, a leader in open source integration, provides these capabilities in an intuitive data management environment that simplifies big data development, deployment and governance.

1. Big Data Integration

Loading big data (large volumes of log files, data from operational systems, social media, sensors, or other sources) into Hadoop via HDFS, HBase, Sqoop or Hive is considered an operational data integration problem. Talend delivers an immediate solution for linking traditional resources, such as databases, applications and file servers directly to these big data technologies.

Talend provides an intuitive set of graphical components and workspace that allows for interaction with a big data source or target without the need to learn and write complicated code. A configuration of a big data connection is represented graphically and the underlying code is automatically generated and then can be deployed as a service, executable or stand-alone job. The full set of Talend data integration components (application, database, service and even a master data hub) are used so that data movement can be orchestrated from any source or into almost any target. Finally, Talend provides graphical components that enable easy configuration for NoSQL technologies such as MongoDB, Cassandra, Hive and HBase to provide random, real-time read/write, columnar-oriented access to big data.

2. Big Data Manipulation

There are a range of tools that enable a developer to take advantage of big data parallelization to perform transformations on massive amounts of data. These languages such as Apache Pig provide a scripting language to compare, filter, evaluate and group data within an HDFS cluster. Talend abstracts these functions into a set of components that allow these
scripts to be defined in a graphical environment and as part of a data flow so they can be developed quickly and without any knowledge of the underlying language.

3. Big Data Quality

Talend offers data quality functions that take advantage of the massively parallel environment of Hadoop. These data quality features provide explicit functions and tasks to profile and identify duplicate records across these huge data stores in moments not days. This is a natural extension of enterprise data quality and data integration solutions and best practices.

4. Big Data Project Management and Governance

While most of the early big data projects are free of explicit project management structure, this will surely change as they become part of the bigger system. With that change, companies will need to wrap standards and procedures around these projects just as they have with data management projects in the past. Talend offers a comprehensive set of functions for big data project management. With Talend, users can schedule, monitor and deploy any big data job while utilizing a common repository where developers can collaborate and share project metadata and artifacts. Further, Talend simplifies constructs such as HCatalog and Oozie.
Talend and Big Data: Available Today

Talend’s open source approach and flexible integration platform for big data enables users to easily connect and analyze data from disparate systems to help drive and improve business performance. Talend’s big data capabilities integrate with today’s big data market leaders such as Cloudera, Hortonworks, Google, EMC/Greenplum, MapR, Netezza, Teradata and Vertica, positioning Talend as a leader in the management of big data. Talend’s goal is to democratize the big data market just as it has with data integration, data quality, master data management, application integration and business process management.

Talend offers three big data products:

1. Talend Open Studio for Big Data
2. Talend Enterprise Big Data
3. Talend Platform for Big Data

Talend Open Studio for Big Data

Talend Open Studio for Big Data is a free open source development tool that packages our big data components for Hadoop, Hbase, Hive, HCatalog, Oozie, Sqoop and Pig with our base Talend Open Studio for Data Integration. It is released into the community under the Apache license. It also allows you to bridge the old with the new as it includes hundreds of components for existing systems like SAP, Oracle, DB2, Teradata and many others. You can download it at www.talend.com.

Talend Enterprise Big Data

Talend Enterprise Big Data extends the Talend Open Studio for Big Data product with professional-grade technical support and enterprise-class features. An organization will upgrade to this version to take advantage of advanced collaboration, monitoring and project management features.
Talend Platform For Big Data

The Talend Platform for Big Data addresses the challenges of big data integration, data quality and big data governance, simplifying the loading, extraction and processing of large and diverse data sets so you can make more informed and timely decisions. Data quality components allow you to do big data profiling, cleansing and matching using a massively parallel environment such as Hadoop. Advanced clustering features allow you to integrate at any scale.

Delivered on top of the Talend unified platform, Talend Platform for Big Data improves productivity across data management domains by sharing a common code repository and tooling for scheduling, metadata management, data processing and service enablement.

For more information about the functionality in each version of our products, please visit www.talend.com.

Conclusion

Big data represents a significant paradigm shift in enterprise technology enabling companies to gain insight into new business opportunities (and threats). While big data presents a significant opportunity, it is not without its own set of challenges. It comprises a relatively new stack of technologies that is fairly complex to learn, there is limited tooling to fuel adoption and development, and there is a lack of knowledgeable big data resources available. Talend’s open source approach and flexible integration platform for big data addresses these challenges enabling users to easily connect and analyze data from disparate systems to help drive and improve business performance.
Appendix: Technology Overview

MapReduce as a Framework

MapReduce enables big data technology such as Hadoop to function. For instance, the Hadoop File System (HDFS) uses these components to persist data, execute functions against it and then find results. The NoSQL databases, such as MongoDB and Cassandra use the functions to store and retrieve data for the respective services. Hive uses this framework as the baseline for a data warehouse.

How Hadoop Works

Hadoop was born because existing approaches were inadequate to process huge amounts of data. Hadoop was built to address the challenge of indexing the entire World Wide Web every day. Google developed a paradigm called MapReduce in 2004, and Yahoo! eventually started Hadoop as an implementation of MapReduce in 2005 and released it as an open source project in 2007. Much like any other operating system, Hadoop has the basic constructs needed to perform computing: It has a file system, a language to write programs, a way of managing the distribution of those programs over a distributed cluster, and a way of accepting the results of those programs. Ultimately the goal is to create a single result set.

With Hadoop, big data is distributed into pieces that are spread over a series of nodes running on commodity hardware. In this structure the data is also replicated several times on different nodes to secure against node failure. The data is not organized into the relational rows and columns as expected in traditional persistence. This lends to the ability to store structured, semi-structured and unstructured content.

There are four types of nodes involved within HDFS. They are:

• Name Node: a facilitator that provides information on the location of data. It knows which nodes are available, where in the cluster certain data resides, and which nodes have failed.

• Secondary Node: a backup to the Name Node

• Job Tracker: coordinates the processing of the data using MapReduce

• Slave Nodes: store data and take direction from the Job Tracker.
A Job Tracker is the entry point for a “map job” or process to be applied to the data. A map job is typically a query written in java and is the first step in the MapReduce process. The Job Tracker asks the name node to identify and locate the necessary data to complete the job. Once it has this information it submits the query to the relevant named nodes. Any required processing of the data occurs within each named node, which provides the massively parallel characteristic of MapReduce.

When the each node has finished processing, it stores the results. The client then initiates a "Reduce" job. The results are then aggregated to determine the “answer” to the original query. The client then accesses these results on the filesystem and can use them for whatever purpose.

**Pig**

The Apache Pig project is a high-level data-flow programming language and execution framework for creating MapReduce programs used with Hadoop. The abstract language for this platform is called Pig Latin and it abstracts the programming into a notation, which makes MapReduce programming similar to that of SQL for RDBMS systems. Pig Latin is extended using UDF (User Defined Functions), which the user can write in Java and then call directly from the language.

**Hive**

Apache Hive is a data warehouse infrastructure built on top of Hadoop (originally by Facebook) for providing data summarization, ad-hoc query, and analysis of large datasets. It provides a mechanism to project structure onto this data and query the data using an SQL-like language called HiveQL. It eases integration with business intelligence and visualization tools.

**HBase**

HBase is a non-relational database that runs on top of the Hadoop file system (HDFS). It is columnar and provides fault-tolerant storage and quick access to large quantities of sparse data. It also adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes. It was originally developed by Facebook to serve their messaging systems and is used heavily by eBay as well.
**HCatalog**

HCatalog is a table and storage management service for data created using Apache Hadoop. It allows interoperability across data processing tools such as Pig, Map Reduce, Streaming, and Hive and a shared schema and data type mechanism.

**Flume**

Flume is a system of agents that populate a Hadoop cluster. These agents are deployed across an IT infrastructure and collect data and integrate it back into Hadoop.

**Oozie**

Oozie coordinates jobs written in multiple languages such as MapReduce, Pig and Hive. It is a workflow system that links these jobs and allows specification of order and dependencies between them.

**Mahout**

Mahout is a data mining library that implements popular algorithms for clustering and statistical modeling in MapReduce.

**Sqoop**

Sqoop is a set of data integration tools that allow non-Hadoop data stores to interact with traditional relational databases and data warehouses.

**NoSQL (Not only SQL)**

NoSQL refers to a large class of data storage mechanisms that differ significantly from the well-known, traditional relational data stores (RDBMS). These technologies implement their own query language and are typically built on advanced programming structures for key/value relationships, defined objects, tabular methods or tuples. The term is often used to describe the wide range of data stores classified as big data. Some of the major flavors adopted within the big data world today include Cassandra, MongoDB, NuoDB, Couchbase and VoltDB.
About Talend

Talend is one of the largest pure play vendors of open source software, offering a breadth of middleware solutions that address both data management and application integration needs. To learn more, visit www.talend.com.

Contact Us

www.talend.com/contact
info@talend.com
partners@talend.com
sales@talend.com