INTERNET RESEARCH GROUP

Big Data, Big Traffic

And the WAN

Internet Research Group
January, 2012

About The Internet Research Group
www.irg-intl.com

The Internet Research Group (IRG) provides market research and market strategy services to product and service vendors. IRG services combine the formidable and unique experience and perspective of the two principals: John Katsaros and Peter Christy, each an experienced industry veteran. The overarching mission of IRG is to help clients make faster and better decisions about product strategy, market entry, and market development. Katsaros and Christy published a book on high tech business strategy *Getting It Right the First Time* – Praeger, 2005 www.gettingitrightthefirsttime.com.

Table of Contents

Big Data, Big Traffic, and the WAN

1. Executive Summary

Big Data – large-scale data storage and distributed computing used to analyze large datasets for business improvement -- is moving into the enterprise in a big way. Early enterprise adopters include JPMorgan Chase, Disney, CBS, British Sky Broadcasting, and Nokia. Right in the center of Big Data is Hadoop, an open-source software framework for large-scale data storage and distributed computing on commodity hardware clusters and standard Ethernet networking.

Big Data in multi-site enterprises generates Big Traffic: the movement of large datasets over WANs needed to support a Hadoop application before, during and after execution. This factor can complicate the running of a Hadoop cluster and may be unaccounted for in the prototype or proof of concept phases, resulting in Hadoop applications that, when put in production, fall short of the performance and scalability expected.

Big Traffic demands Big Planning. The Internet Research Group recommends that enterprise CTOs, IT architects and similar professionals explore Big Traffic early on when considering or planning Hadoop cluster deployments. Without an understanding of the role the WAN plays in enterprise Hadoop applications, the scalability and utilization of a Hadoop cluster may be impaired. With such understanding, enterprises can be far more confident that the promise of Big Data will be fulfilled.

2. Introduction

In a relatively short time the term "Big Data," and the disruptive technology it describes, have moved to center stage. Big Data -- large-scale data storage and distributed computing -- makes affordable the analysis of very large datasets for a wide range of use cases that previously were not possible by most organizations. For instance, a company can analyze customer behavior across multiple business units for a 360-degree view of customer activity and combine this with other data (such as Website activity) to suggest new services, or categorize consumer sentiment across multiple social networks on the Internet to improve campaign success rates. Other use cases include advanced fraud detection and data mining, and extract, transform, and load (ETL) operations in

data warehousing.

A typical Hadoop node comprises a server with generous RAM and a dozen or more high-capacity SATA drives; a Hadoop cluster may contain anywhere from one to thousands of nodes connected by conventional Ethernet. Because the framework takes care of all the details of job execution, programmers, business analysts and data scientists are freed up to focus on developing the use cases.

Hadoop was designed to be scalable by adding more nodes (more processing power plus more storage). Big websites and their huge depositories of user information were the pioneers of Big Data using Hadoop, which was in fact inspired by the development of MapReduce by Google. For instance, Yahoo! operates several clusters with over 40,000 Hadoop nodes and uses Big Data analytics on each visit to improve its ability to select content which will be of interest. The results expected include keeping users on their site longer, to increase interactivity, and to create additional value for visitors. Now the technology is moving into the enterprise.

However, enterprises looking to unlock their Big Data with Hadoop clusters are also discovering Big Traffic: the movement of large datasets over WANs. The execution of a Hadoop application results in a lot of data movement. This can happen in a single data center or, increasingly, across many data centers.

For instance, a financial services firm could use a Hadoop cluster in its main data center to provide a 360-degree view of its clients across all its divisions, such as the insurance, brokerage, and banking units. These are in different parts of the country (or world), yet all of them could have datasets which are needed to complete the analysis. Cisco, Arista, and others have proposed network optimization solutions for Hadoop clusters *within* a data center, but these do not address Big Traffic.

Big Traffic calls for Big Planning. The Internet Research Group recommends that enterprise CTOs, IT architects and similar professionals explore Big Traffic early on when considering or planning Hadoop cluster deployments. To aid that exploration, this white paper puts Big Data and Hadoop in perspective with an overview of their development, operation, typical applications, and benefits. It discusses how and why Big Traffic arises, and the consequences for enterprise Hadoop deployment.

3. Processing and Storing Large Datasets

In 2004, two Google engineers published a paper, *MapReduce: Simplified Data Processing on Large Clusters*, that described an application execution model for processing and generating large data sets. This paper helped set in motion a chain of events leading to the creation of Apache Hadoop. Hadoop is an open-source software framework that implements the MapReduce model and the

Hadoop Distributed File System (HDFS) along with supporting software systems. Along the way, Hadoop both boosted and got a boost from the burgeoning growth of unstructured data -- email, web click stream data, documents, multimedia, and even IT events (log data) in an explosion of file types. Compliance with regulations has played a role in launching this growth by requiring organizations to keep unstructured data longer. Companies also realized how valuable much of this data is and began building tools to analyze it to improve operations and get new customers. Just a few years ago, it would have been cost prohibitive for an enterprise to retain and analyze such large amounts of data. With HDFS, Hadoop offers as much as a 30-to-one reduction in the cost of storage due to its use of commodity disk drives rather than conventional enterprise storage systems. Figure 1 illustrates the expected growth of the unstructured data that is meat and potatoes to Hadoop applications, in comparison to the structured data that dominated enterprise computing in the past.

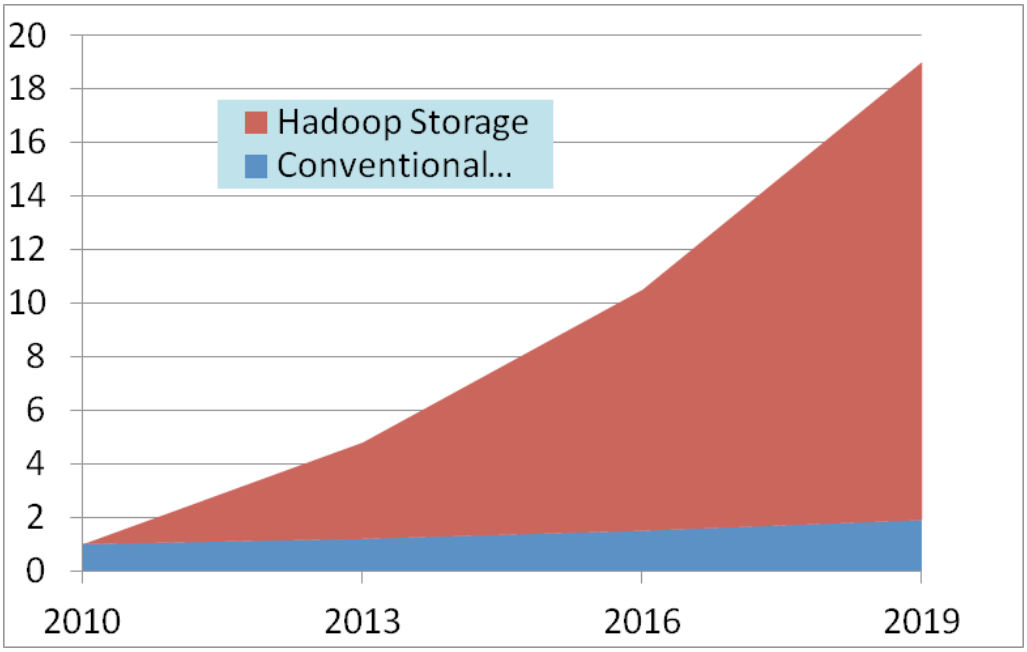

Figure 1: Growth of structured and unstructured data.

The confluence of these two developments – low-cost storage coupled with an affordable scalable execution environment – meant lift-off for Big Data. Today many organizations recognize the potential that this type of large-scale analytics – a substantial increase in what their current data warehouse products could have on their ability to grow their businesses and improve operations.

4. Hadoop Moves into the Enterprise

Adoption of Hadoop and related Big Data modeling tools has been enthusiastic,

for reasons that are easy to discern. Developers love it because MapReduce automatically parallelizes and executes applications written to this framework in clusters that can be very, very large. They only need to map business drivers to application functionality: the scalability and performance are built in. IT loves it because Hadoop is designed for commodity hardware and standard Ethernet networking, making its scalability and performance extremely cost-effective and easy to deploy. CXOs love it because it monetizes data they already own or have access to, but couldn't afford to store or analyze for business improvement or competitive advantage.

The commodity hardware for a typical Hadoop node comprises a server with generous RAM and 12 to 16 high-capacity (2-3 TB) SATA disk drives. A Hadoop cluster can contain anywhere from one to thousands of nodes, typically connected by conventional Ethernet networking. Increasing the number of nodes adds both storage capacity (more drives) and processing power (more cores and more RAM). A bigger Hadoop cluster buys the enterprise larger datasets, greater throughput, and larger jobs at a price that opens up new vistas of data analysis. Table 1 lists five of the applications that have turned up most often in our conversations with data analysts (data experts) working with or considering Hadoop for a data analysis project.

Table 1: Popular enterprise applications of Hadoop.

Use Case	Description
Common Data Platform	Gather data from different systems to develop a 360-degree view of customers (e.g., what should you offer to a customer, does a customer show signs of changing firms?).
ETL	Extract, Transfer and Load - Accelerate loading (shorter Load Window) and increase size of datasets, potentially a big boost for data warehouses.
Data Mining	Find new information by looking across both Hadoop and existing data warehouse data.
Fraud Detection	Analyze larger, more complex datasets for Advanced Fraud Detection.
Sentiment Analysis	Understand the company or product perception of customers or potential customers.

Hadoop was pioneered by large Internet websites such as Yahoo! and e-Bay, and it is there that one still finds the largest Hadoop clusters. For instance, Yahoo! has one of the largest Hadoop implementations, with over 40,000 nodes spread across many clusters. This investment enables Yahoo! to use clickstream

and other data to optimize user sessions, keep users on their site longer, increase interactivity with members, and create additional value for visitors. Other notable online companies that have adopted Hadoop include eBay, Facebook, Twitter and LinkedIn, all looking for similar benefits. One of the most popular use cases for such organizations is the "people you may know" analysis which uses Hadoop extensively.

We expect most enterprises and government agencies to use much smaller Hadoop clusters. For enterprise production environments, Hadoop today is generally run on clusters with 10 to 20 nodes, but this is growing. At Hadoop World 2011, the average size of clusters reported by attendees had risen to 120 nodes (up from 66 in 2010), and the average amount of data stored was up to over 200 petabytes (from 60 in the prior year). Enterprises that already have sizable Hadoop deployments include JPMorgan Chase, Disney, CBS, British Sky Broadcasting, and Nokia.

5. Enterprise Data Isolation and Big Traffic

Hadoop processing may involve a lot of data movement, which is handled automatically by the software framework using Hadoop with HDFS. Big Traffic issues results from the staging of Big Data processing and the purposing of the results where it is needed when multiple, geographically distributed data centers are involved, as well as from the propagation of data between clusters for the purpose of storage hierarchy management (e.g., keeping the most current and most valuable data in HDFS (Hadoop File System) with higher intrinsic replication) or to keep two clusters up to date with an incoming data stream.

Big Data analysis often exhibits a wave pattern as data is initially processed by one job and returns results to the appropriate application. The data communications between servers is done by HDFS – the file system. The wave action is staging data into the cluster and then purposing the results to other clusters or data warehouses. For example, a collection of ads may be reduced to a subset which has a high probability of interesting a user profile. This match may be refined when the data has been refreshed resulting in another reduction and modification of the original results.

This makes Big Data analytics a matter of both throughput capacity and intelligent data movement. Factors include:

- Which datasets are required by jobs queuing up for execution
- The policies for moving and securing data in transit
- What resources may be required as jobs execute, and
- The allocation of the completed datasets to execution servers

In the megascale websites that drove the initial adoption of Hadoop, this all takes place in vast, centralized data centers created by the economies of scale

and natural development of web-based businesses. Hadoop clusters in such an environment enjoy all the advantages of a high-speed networking fabric at multi-gigabit speeds, and companies such as Cisco and Arista have created network optimization solutions to further improve Hadoop communications within such large data centers. Furthermore, the data (clickstream, user, etc.) is right there.

This is rarely the case with enterprise deployments. There, the reality is one of what we call "data isolation," which makes this coordinated data movement difficult because of data storage and analytics decisions made long before Big Data was an option. Data isolation has several causes. Most common is simple geographic distribution, where the data needed is scattered across divisional silos distant from each other. In addition, the data collected by each division may never have been thought of as useful to other divisions, may have been considered of only short-term or imposed (i.e., regulatory) interest, or further handling of the data may have been too expensive.

For instance, a financial services company may have an insurance division, and perhaps its main data center, in San Francisco, while its brokerage operations are in New York, and its banking unit in Chicago. The economics of large web hosting data centers may result in the Hadoop cluster ending up in yet a fourth location. Wherever it is, a Hadoop cluster is a valuable resource, and because Hadoop applications are batch jobs, sharing one large cluster for multiple purposes yields better cluster utilization and thus a better return on investment.

But centralization creates Big Traffic concerns. For that financial services company, analyzing client behavior across all of those units requires using the WAN to stage datasets into the Hadoop cluster, propagation of data between clusters as the job executes and distributing the results across the WAN links. In some cases the WAN may not have enough throughput capacity to support these tasks. Other applications may involve types of data somewhat less sensitive to bandwidth limitations because of their incremental nature, such as log files, network alerts, clickstream, and location. But these too can feel the impact of Big Traffic in order to meet scheduling requirements.

To get the most from their Hadoop investment, most organizations will eventually want to run hundreds of Hadoop jobs daily, some running for only a few seconds while others may run for hours. The pressure to run more jobs leads to shrinking data movement windows for all jobs. If the data arrival rate slows due to the impact of Big Traffic (saturated WAN links and increased latency), then job execution slows as well. Fewer jobs can be run, scalability suffers, the utilization rate of the cluster drops, and with it the return on investment. These and other factors can create a lot of stress: shrinking windows, resource sharing, utilization rates are all part of the performance and scalability equation for Big Data.

6. Big Traffic Calls for Big Planning

Although the complexities of parallelization, data movement, and execution are all handled automatically by Hadoop, careful planning is required to get the most from a Hadoop cluster, especially where Big Traffic is involved.

Figure 2 shows the shape of Big Data adoption to date, which echoes the three phases that individual organizations must pass through as they move their Hadoop clusters to production.

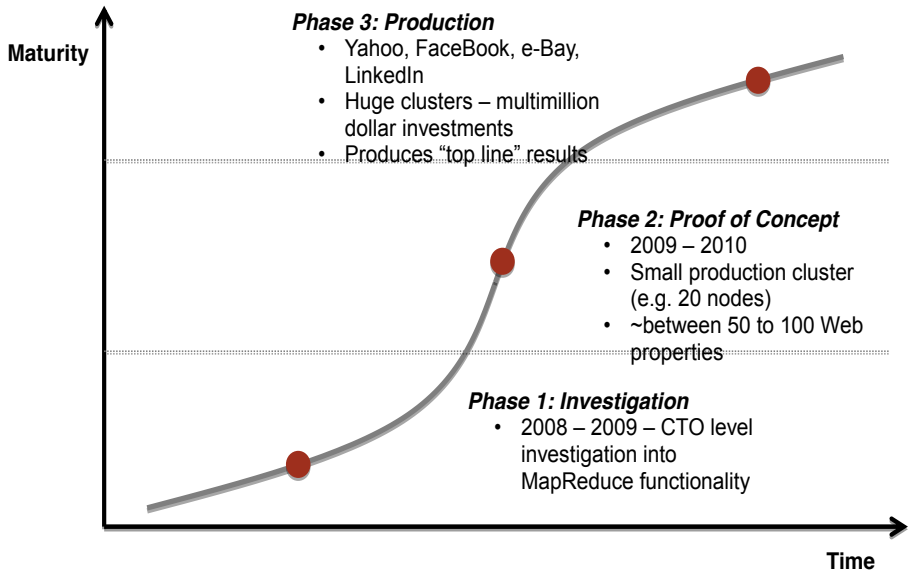

Figure 2: The three phases of Big Data adoption.

Like any technology, Big Data starts with education and investigation. How does Hadoop work? How well prepared is the organization, not just technically but in terms of process? What additional training or outside systems integration support might be needed? An important part of the educational phase is business case development. For any given use case, what combination of business improvement and cost savings will the investment deliver? Finding the best use cases takes a lot of research and study. It's partly an objective business analysis and partly a guess as to how better analytics can improve the business, and definitely an iterative process: the more you learn, the clearer the parameters become and the more focused the use cases.

Once the initial goals are understood, then the operations team can build a prototype cluster for development and testing purposes. Because Hadoop runs on commodity hardware, this can be quite inexpensive, making use of repurposed processor nodes and storage. The first prototype may actually occur before the main educational phase is finished, and may not need to be business-relevant. Its purpose is simply to acquaint the team with the technical ins and outs of Hadoop.

But at some point the organization will want to build an actual proof of concept, using real data. This is pretty much the inflection point at the bottom of phase two in the diagram above. It's here that problems arise if the Hadoop team has not taken data isolation into account. Unfortunately, because this stage generally takes place in a small cluster of 10 to 20 nodes inside a single data center, Big Traffic is often overlooked.

The result is that partway through phase two, when the implementation moves into production with real live data, and other divisions start yelling for a crack at the great new tools IT is delivering, cluster utilization and performance start to suffer. Just throwing more bandwidth at the problem is neither practical nor effective. Just as within the Hadoop cluster itself, attention to intelligent data movement on the WAN is critical.

The solution starts with education. Hadoop planners and implementation teams should pay attention to data isolation and Big Traffic from the very beginning. Without an understanding of the role the WAN plays in enterprise Hadoop applications, no use case will survive contact with the real world. With such understanding, enterprises can be far more confident that the promise of Big Data will be fulfilled.

7. Conclusion

For a variety of good business reasons, putting Big Data tools to work can make an organization more profitable and productive. While the path to adoption of Big Data analytic implementations isn't too difficult, knowing the right mix of processor, storage and communications and managing Big Traffic data movement can help guide the successful implementation of this new technology.