Hadoop and the Data Warehouse:
When to Use Which

Dr. Amr Awadallah, Founder and CTO, Cloudera, Inc.
Dan Graham, General Manager, Enterprise Systems, Teradata Corporation

cloudera

TERADATA.

Preface

This paper is written jointly by Cloudera Corporation and Teradata Corporation. It is intended for a technical audience who has a strong familiarity in one of the two topics: data warehouses or Hadoop/MapReduce. This paper does not address the positioning of Hadoop with Aster Data (now referred to as Teradata Aster), a company acquired by Teradata while this paper was being written. Both Teradata Aster and Hadoop have implementations of the MapReduce parallel processing framework. Read *MapReduce and the Data Scientist** to understand the architectural differences, which analytical workloads are best for each, and how Teradata Aster and Hadoop uniquely work together to solve big data problems for customers at the lowest total cost of ownership.

Our appreciation and thanks go out to Mike Olson, Omer Trajman, Ed Albanese, Randy Lea, Drew O'Brien, and Todd Sylvester for the time they spent contributing, editing, and helping in various ways.

Table of Contents

* "MapReduce and the Data Scientist", Colin White, 2012

Introduction

Once or twice every decade, the IT marketplace experiences a major innovation that shakes the entire data center infrastructure. In recent years, Apache Hadoop has emerged from humble beginnings to worldwide adoption - infusing data centers with new infrastructure concepts and generating new business opportunities by placing parallel processing into the hands of the average programmer.

As with all technology innovation, hype is rampant, and non-practitioners are easily overwhelmed by diverse opinions. Even active practitioners miss the point, claiming for example that Hadoop replaces relational databases and is becoming the new data warehouse. It is easy to see where these claims originate since both Hadoop and Teradata® systems run in parallel, scale up to enormous data volumes and have shared-nothing architectures. At a conceptual level, it is easy to think they are interchangeable. Of course, they are not, and the differences overwhelm the similarities. This paper will shed light on the differences and help architects identify when to deploy Hadoop and when it is best to use a data warehouse.

This paper is organized into these sections:

- Teamwork - Hadoop and the data warehouse
- Hadoop differentiators
- Data warehouse differentiators
- Shades of gray - where the technologies overlap

- - - - - - - - - -

BIG DATA

Who hasn't heard of Big Data lately? With hundreds of press articles and dozens of conferences exploring it, it is clear that the data deluge is upon us.

Nearly every discussion of Big Data begins with a debate over definitions. We offer this Wikipedia definition that was crafted by melding together commentary from analysts at The 451, IDC, Monash Research, a TDWI alumnus, and a Gartner expert.

"Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process within a tolerable elapsed time. Big Data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes in a single data set.

Examples include Web logs; RFID; sensor networks; social networks; Internet text and documents; Internet search indexing; call detail records; astronomy, atmospheric science, biological, genomics, biochemical, medical records; scientific research; military surveillance; photography archives; video archives; and large scale eCommerce."

Teamwork - Hadoop and the Data Warehouse

As customers, analysts, and journalists explore Hadoop and MapReduce, the most frequent questions are: “When should I use Hadoop, and when should I put the data into a data warehouse?” The answers are best explained with an example of a recently deployed big data source: Smart Meters

Smart meters are deployed in homes worldwide to help consumers and utility companies manage the use of water, electricity, and gas better. Historically, meter readers would walk from house to house recording meter read outs and reporting them to the utility company for billing purposes. Because of the labor costs, many utilities switched from monthly readings to quarterly. This delayed revenue and made it impossible to analyze residential usage in any detail.

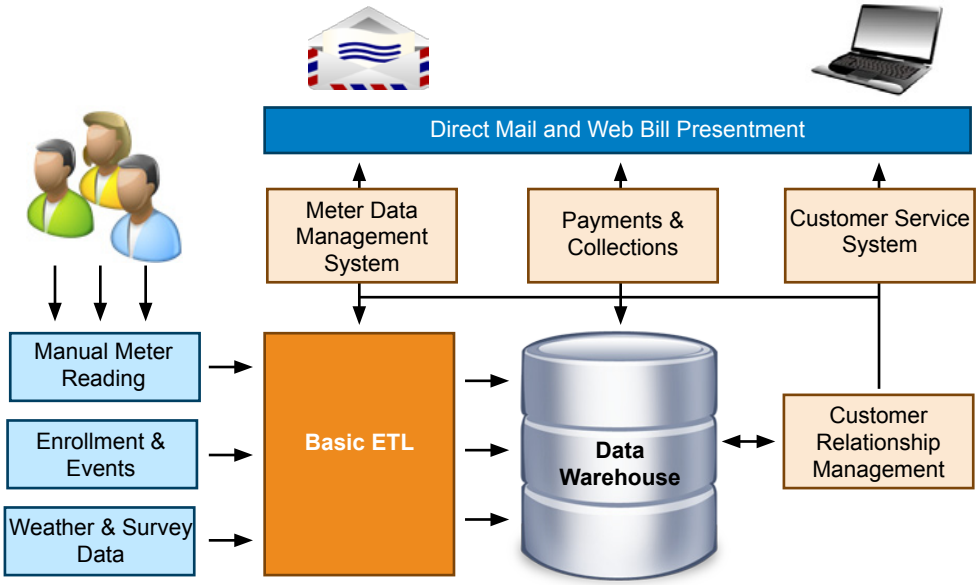

Figure 1. Before: Data flow of meter reading done manually

Consider a fictional company called CostCutter Utilities that serves 10 million households. Once a quarter, they gathered 10 million readings to produce utility bills. With government regulation and the price of oil skyrocketing, CostCutter started deploying smart meters so they could get hourly readings of electricity usage. They now collect 21.6 billion sensor readings per quarter from the smart meters. Analysis of the meter data over months and years can be correlated with energy saving campaigns, weather patterns, and local events, providing savings insights both for consumers and CostCutter Utilities. When consumers are offered a billing plan that has cheaper electricity from 8 p.m. to 5 a.m., they demand five minute intervals in their smart meter reports so they can identify high-use activity in their homes. At five minute intervals, the smart meters are collecting more than 100 billion meter readings every 90 days, and CostCutter Utilities now has a big data problem. Their data volume exceeds their ability to process it with existing software and hardware. So CostCutter Utilities turns to Hadoop to handle the incoming meter readings.

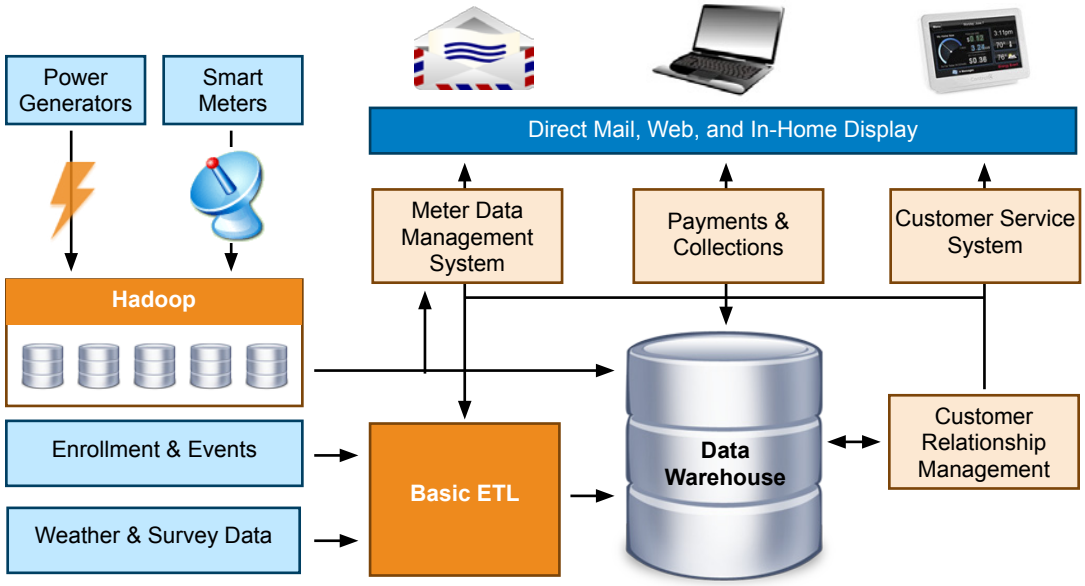

Figure 2. After: Meter reading every 5 or 60 minutes via smart meters

Hadoop now plays a key role in capturing, transforming, and publishing data. Using tools such as Apache Pig, advanced transformations can be applied in Hadoop with little manual programming effort; and since Hadoop is a low cost storage repository, data can be held for months or even years. Since Hadoop has been used to clean and transform the data, it is loaded directly into the data warehouse and MDMS systems. Marketing is developing additional offers for consumers to save money by using analysis of the trends by household, neighborhood, time of day, and local events. And now the in-home display unit can give consumers detailed knowledge of their usage.

Note that Hadoop is not an Extract-Transform-Load (ETL) tool. It is a platform that supports running ETL processes in parallel. The data integration vendors do not compete with Hadoop; rather, Hadoop is another channel for use of their data transformation modules.

As corporations start using larger amounts of data, migrating it over the network for transformation or analysis becomes unrealistic. Moving terabytes from one system to another daily can bring the wrath of the network administrator down on a programmer. It makes more sense to push the processing to the data. Moving all the big data to one storage area network (SAN) or ETL server becomes infeasible with big data volumes. Even if you can move the data, processing it is slow, limited to SAN bandwidth, and often fails to meet batch processing windows. With Hadoop, raw data is loaded directly to low cost commodity servers one time, and only the higher value refined results are passed to other systems. ETL processing runs in parallel across the entire cluster resulting in much faster operations than can be achieved pulling data from a SAN into a collection of ETL servers. Using Hadoop, data does not get loaded into a SAN just to then get pulled out of the SAN across the network multiple times for each transformation.

It should be no surprise that many Hadoop systems sit side by side with data warehouses. These systems serve different purposes and complement one another. For example:

A major brokerage firm uses Hadoop to preprocess raw click streams generated by customers using their website. Processing these click streams provides valuable insight into customer preferences which are passed to a data warehouse. The data warehouse then couples these customer preferences with marketing campaigns and recommendation engines to offer investment suggestions and analysis to consumers. There are other approaches to investigative analytics on clickstream data using analytic platforms. See “MapReduce and the Data Scientist” for more details.

An eCommerce service uses Hadoop for machine learning to detect fraudulent supplier websites. The fraudulent sites exhibit patterns that Hadoop uses to produce a predictive model. The model is copied into the data warehouse where it is used to find sales activity that matches the pattern. Once found, that supplier is investigated and potentially discontinued.

Increasingly, Hadoop and MapReduce solutions are being deployed in enterprises together with data warehouses. Figure 3 shows one perspective on data sources, data sharing, and the diversity of data experts who can access both systems.

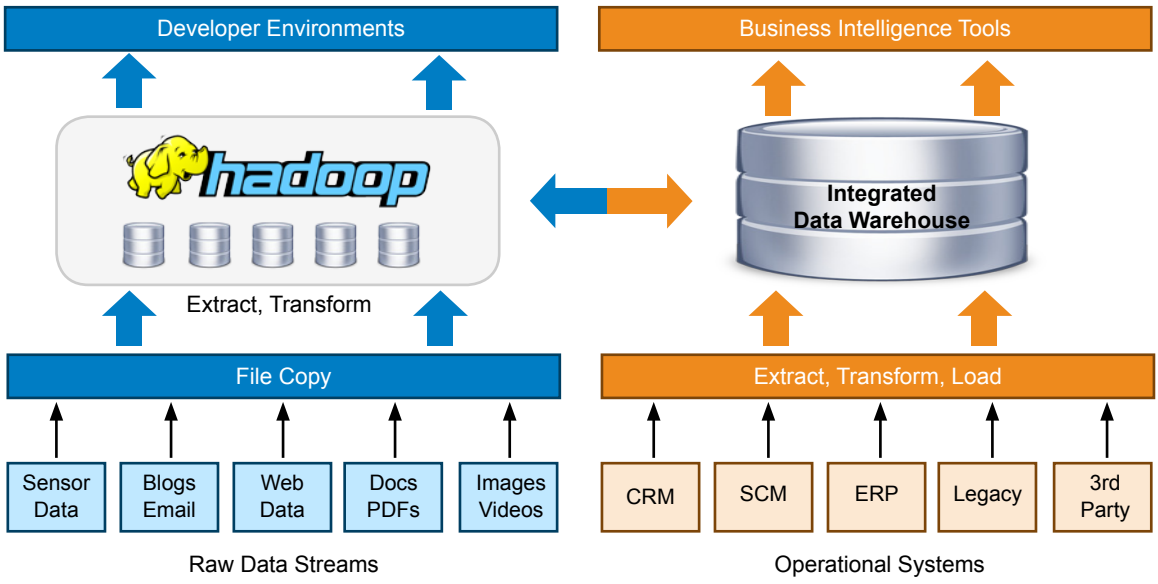

Figure 3. An enterprise data architecture

Complex Hadoop jobs can use the data warehouse as a data source, simultaneously leveraging the massively parallel capabilities of two systems. Any MapReduce program can issue SQL statements to the data warehouse. In one context, a MapReduce program is "just another program," and the data warehouse is "just another database." Now imagine 100 MapReduce programs concurrently accessing 100 data warehouse nodes in parallel. Both raw processing and the data warehouse scale to meet any big data challenge. Inevitably, visionary companies will take this step to achieve competitive advantages.

- - - - - - - - - -

Hadoop Differentiators

Hadoop is the result of new developments in compute and storage grid technologies. Using commodity hardware as a foundation, Hadoop provides a layer of software that spans the entire grid, turning it into a single system. Consequently, some major differentiators are obvious in this architecture:

- Hadoop is the repository and refinery for raw data.
- Hadoop is a powerful, economical and active archive.

Thus, Hadoop sits at both ends of the large scale data lifecycle -- first when raw data is born, and finally when data is retiring, but is still occasionally needed.

Hadoop as the Repository and Refinery

As volumes of big data arrive from sources such as sensors, machines, social media, and click stream interactions, the first step is to capture all the data reliably and cost effectively. When data volumes are huge, the traditional single server strategy does not work for long. Pouring the data into the Hadoop Distributed File System (HDFS) gives architects much needed flexibility. Not only can they capture 10s of terabytes in a day, they can adjust the Hadoop configuration up or down to meet surges and lulls in data ingestion. This is accomplished at the lowest possible cost per gigabyte due to open source economics and leveraging commodity hardware.

Since the data is stored on local storage instead of SANs, Hadoop data access is often much faster, and it does not clog the network with terabytes of data movement.

Once the raw data is captured, Hadoop is used to refine it. Hadoop can act as a parallel “ETL engine on steroids,” leveraging handwritten or commercial data transformation technologies. Many of these raw data transformations require the unraveling of complex freeform

WHAT ARE HADOOP AND MAPREDUCE?

Hadoop is an Apache open source project that provides a parallel storage and processing framework. Its primary purpose is to run MapReduce batch programs in parallel on tens to thousands of server nodes.

MapReduce refers to the application modules written by a programmer that run in two phases: first mapping the data (extract) then reducing it (transform).

Hadoop scales out to large clusters of servers and storage using the Hadoop Distributed File System (HDFS) to manage huge data sets and spread them across the servers.

One of Hadoop’s greatest benefits is the ability of programmers to write application modules in almost any language and run them in parallel on the same cluster that stores the data. This is a profound change! With Hadoop, any programmer can harness the power and capacity of thousands of CPUs and hard drives simultaneously.

More advantages of Hadoop include affordability (it runs on commodity hardware), open source (free download from Cloudera), and agility (store any data, run any analysis).

data into structured formats. This is particularly true with click streams (or web logs) and complex sensor data formats. Consequently, a programmer needs to tease apart the wheat from the chaff, identifying the valuable signal in the noise.

Let's look at another example. Web log entries have a "user agent" string that reveals which operating system and Internet browser was used when accessing the site. As part of the ETL process, structured columns are extracted from the click stream data. During the transformation, Internet Explorer 6, 7, 8; Firefox; and a few other browsers are identified. For unidentified browsers, that click stream may be placed into a category of "unknowns," or the field may simply be filled with "other." Next, the click streams are loaded into the data warehouse. When a new browser is released, such as Safari on the iPhone, these records all end up categorized as "other." A few months go by, and the business user complains that the "other" category is surging in size. Since the ETL developer did not have the transformation designed to identify the iPhone Safari browser, he must now add a new ETL rule. With the raw click streams held in HDFS for many months, the developer can rerun the extracts and transformations. Now the corrected data is added to the data warehouse, giving the business user insights into the growth of users with iPhones. Unlike traditional ETL tools, Hadoop persists the raw data and can be used to re-process it repeatedly in a very efficient manner.

In another example, an eCommerce retailer uses Hadoop to analyze graphic images depicting items for sale over the Internet. When a consumer wants to buy a red dress, their search may not match the tags used to identify each item's search terms. To assist the consumer in finding a match, Hadoop is used to analyze thousands of dress images, detecting the red prominence of the primary object in the graphic (JPGs, GIFs and PNGs). This requires enormously complex logic for the computer to "see" the dress and its primary colors as humans do. This is a necessary process since manufacturers do not always label their goods clearly for the distributors or identify keywords with which users are likely to search. Using Hadoop, millions of images are tagged with additional information to assist consumers with their search, increasing the chances that they find the item they were looking for and make a purchase. This type of analysis and processing would be too difficult and expensive to perform in a data warehouse using SQL.

Hadoop as the Active Archive

In a 2003 interview with ACM, Jim Gray claimed that hard disks can be treated as tape. While it may take many more years for magnetic tape archives to be retired, today some portions of tape workloads are already being redirected to Hadoop clusters. This shift is occurring for two fundamental reasons. First, while it may appear inexpensive to store data on tape, the true cost comes with the difficulty of retrieval. Not only is the data stored offline, requiring hours if not days to restore, but tape cartridges themselves are prone to degradation over time making data loss a reality and forcing companies to factor in those costs. To make matters worse, tape formats change every couple of years requiring organizations to either perform massive data migrations to the newest tape format or risk the inability to restore data from obsolete tapes.

Second, it has been shown that there is value in keeping historical data online and accessible. As in the click stream example, keeping raw data on spinning disk for a longer duration makes it easy for companies to revisit data when the context changes and new constraints need to be applied. Searching thousands of disks with Hadoop is dramatically faster and easier than spinning through hundreds of magnetic tapes. Additionally, as disk densities continue to double every 18 months, it becomes economically feasible for organizations to hold many years' worth of raw or refined data in HDFS. Thus, the Hadoop storage grid is useful both in the preprocessing of raw data and the long-term storage of data. It's a true "active archive" since it not only stores and protects the data, but also enables users to quickly, easily and perpetually derive value from it.

- - - - - - - - - -

Data Warehouse Differentiators

After nearly 30 years of investment, refinement and growth, the list of features available in a data warehouse is quite staggering. Built upon relational database technology using schemas and integrating Business Intelligence (BI) tools, the major differences in this architecture are:

- Data warehouse performance
- Integrated data that provides business value
- Interactive BI tools for end users

Data Warehouse Performance

Basic indexing, found in open source databases, such as MySQL or Postgres, is a standard feature used to improve query response times or enforce constraints on data. More advanced forms such as materialized views, aggregate join indexes, cube indexes and sparse join indexes enable numerous performance gains in data warehouses. However, the most important performance enhancement to date is the cost-based optimizer. The optimizer examines incoming SQL and considers multiple plans for executing each query as fast as possible. It achieves this by comparing the SQL request to the database design and extensive data statistics that help identify the best combination of execution steps. In essence, the optimizer is like having a genius programmer examine every query and tune it for the best performance. Lacking an optimizer or data demographic statistics, a query that could run in minutes may take hours, even with many indexes. For this reason, database vendors are constantly adding new index types, partitioning, statistics, and optimizer features. For the past 30 years, every software release has been a performance release.

Integrating Data: the Raison d'être

At the heart of any data warehouse is the promise to answer essential business questions. Integrated data is the unique foundation required to achieve this goal. Pulling data from multiple subject areas and

WHAT IS A DATA WAREHOUSE?

In the 1990s, Bill Inmon defined a design known as a data warehouse. In 2005, Gartner clarified and updated those definitions. From these we summarize that a data warehouse is:

1. Subject oriented: The data is modeled after business concepts, organizing them into subjects areas like sales, finance, and inventory. Each subject area contains detailed data.
2. Integrated: The logical model is integrated and consistent. Data formats and values are standardized. Thus, dates are in the same format, male/female codes are consistent, etc. More important, all subject areas use the same customer record, not copies.
3. Nonvolatile: Data is stored in the data warehouse unmodified, and retained for long periods of time.
4. Time variant: When changes to a record are needed, new versions of the record are captured using effective dates or temporal functions.
5. Not virtual: The data warehouse is a physical, persistent repository.

numerous applications into one repository is the raison d'être for data warehouses. Data model designers and ETL architects armed with metadata, data cleansing tools, and patience must rationalize data formats, source systems, and semantic meaning of the data to make it understandable and trustworthy. This creates a common vocabulary within the corporation so that critical concepts such as "customer," "end of month," or "price elasticity," are uniformly measured and understood. Nowhere else in the entire IT data center is data collected, cleaned, and integrated as it is in the data warehouse.

The payoff is well worth it. For example, consider a well-implemented product inventory subject area. Business questions, such as, "How much inventory will be obsolete next month in each location?" provide vital guidance each week. Similarly, an orders subject area in the data warehouse answers questions such as, "What products are on back order?" Using data models, the inventory subject area provides 25 primary answer sets (or reports) whereas the orders subject area offers 21. In isolation, the subject areas are powerful and when combined, they deliver 74 complex reports with answers to these questions. Through integration, the BI user can ask, "How will this large order affect current inventory levels?" The results of integrated subject areas are better strategic business decisions made hundreds or thousands of times per day.

Interactive BI tools

BI tools such as MicroStrategy, Tableau, IBM Cognos, and others provide business users with direct access to data warehouse insights. First, the business user can create reports and complex analysis quickly and easily using these tools. As a result, there is a trend in many data warehouse sites towards end-user self service. Business users can easily demand more reports than IT has staffing to provide. More important than self service however, is that the users become intimately familiar with the data. They can run a report, discover they missed a metric or filter, make an adjustment, and run their report again all within minutes. This process results in significant changes in business users' understanding the business and their decision-making process. First, users stop asking trivial questions and start asking more complex strategic questions. Generally, the more complex and strategic the report, the more revenue and cost savings the user captures. This leads to some users becoming "power users" in a company. These individuals become wizards at teasing business value from the data and supplying valuable strategic information to the executive staff. Every data warehouse has anywhere from two to 20 power users.

Query performance with BI tools lowers the analytic pain threshold. If it takes 24 hours to ask and get an answer, users only ask once. If it takes minutes, they will ask dozens of questions. For example, a major retailer was comparing stock-on-hand to planned newspaper coupon advertising. Initially they ran an eight-hour report that analyzed hundreds of stores. One power user saw they could make more money if the advertising was customized for stores by geographic region. By adding filters and constraints and selecting small groups of regional stores, the by-region query ran in two minutes. They added more constraints and filters and ran it again. They discovered that inventory and regional preferences would sell more and increase profits. Where an eight-hour query was discouraging, two-minute queries were an enabler. The power user was then willing to spend a few hours analyzing each region for the best sales, inventory, and profit mix. The lower pain threshold to analytics was enabled by data warehouse performance and the interactivity of the BI tools.

- - - - - - - - - -

Shades of Gray

Since there are workloads that run on both Hadoop and the data warehouse, it is best to match the requirements to each platform's ability to perform the task quickly, easily, and at the best cost over time. Data centers that are using both technologies are developing the skills to know "when to use which" intuitively. Organizations that are just now adopting Hadoop will have to cross train the Hadoop and the data warehouse teams to appreciate the other platform's strengths.

The data warehouse and Hadoop are not highly differentiated in the following categories. Either tool could be the right solution. Choosing the best tool depends on the requirements of the organization. Based on several requirements and a score of one to ten, the data warehouse might be a five and Hadoop a seven for the task or the reverse. In many cases, Hadoop and the data warehouse work together in an information supply chain and just as often, one tool is better for a specific workload.

Provisional Data

There are various provisional data sets that are inappropriate for a data warehouse. Provisional data sets are analyzed or used in isolation. They represent a single, relevant point in time as defined by the data scientist and do not derive their primary value through integration with other data sets. Examples of provisional data are scientific analysis of astronomy and physics data. Typically the data scientist has one massive data set to study, and once the results are found, the data may only be retained for a month or so. This kind of data does not fit into the data warehousing paradigm. Hadoop has been a boon to these kinds of research projects.

Other kinds of provisional data include Internet URLs and web pages. This is the search index problem that caused Google to originally conceive the MapReduce solution that led to the open source development of Hadoop. Search engine companies crawl the web gathering data, including URLs, tags, metadata and outbound links. Using that data, search engines perform massive MapReduce computations to rank and index URLs so that visitors to the search engine get the most relevant answer to every request. The challenge is that web pages and URL links change every day. Some sites have 10,000 web pages with 12-15 webmasters performing constant revisions. Consequently, the pages and URLs scanned yesterday may be gone tomorrow. Given the enormity of data collected, it's infeasible to save even half of it for more than a few days. Therefore, Hadoop is the only viable solution for these kinds of problems.

In contrast, consider a bank that grows through acquiring small regional banks. Prior to a merger, they get magnetic tapes of consumer accounts from the regional bank. They use the data to find overlapping clients, estimate profitability, and assess long-term loyalty. Using this information, the acquiring bank will negotiate for an equitable acquisition price -- but they only have four weeks to do the analysis. This makes it difficult to do data modeling, restructuring, and other

tasks needed to put the data into the data warehouse. Furthermore, the data will be thrown away and fully integrated into the data warehouse after the acquisition. Hadoop has the advantage of flexibility, time to value, and not being limited by governance committees or administrators. The data warehouse has the advantage of allowing the regional bank's accounts to be joined to existing accounts providing quick identification of overlapping consumers and comparisons of account quality to existing accounts. Which is the best platform since both Hadoop and the data warehouse can do the job? It depends on what the acquiring bank's requirements are and how they prioritize them. One solution is for Hadoop to store and refine the data, and then load some of the refined data into the data warehouse for further analysis and account comparisons.

Sandboxes and Data Labs

Historically, data mining and predictive analytics have been performed using SAS® software on small samples of data. These "sandboxes" are used by data scientists to explore the data. Predictive analytics are in their heyday following more than a decade running on small servers. It began when Teradata and SAS worked together to put SAS "procs" in-database so they could run in parallel as close to the disk data as possible. One credit card issuer was able to reduce the time to build predictive models from 14 weeks to two weeks and their analytic model scoring from 175 hours to 36 minutes. This produced a competitive advantage for a couple of years.

The value of in-database analytics is twofold: first, the data mining algorithm runs in parallel, providing fast results enabling many more explorations per day. Second, the data scientist no longer needs to settle for small samples of data. They can now crunch all the data with rapid, more accurate results. These are the same capabilities that MapReduce delivers in implementations within Hadoop. Hadoop MapReduce runs predictive analytics in parallel against enormous quantities of data. An obvious difference is that the data warehouse has clean integrated data where Hadoop often has raw data in quantity. Consequently, one way to choose between Hadoop and the data warehouse for data mining is based on the data itself.

Parallel exploratory analytics started with SAS in-database and expands with the arrival of Hadoop. Universities and the open source community have been developing data mining algorithms that leverage the MapReduce framework. An example is the Apache Mahout project which is a collection of open source machine learning algorithms developed for Hadoop. Data scientists now have access to a distributed computing platform where they are not limited by any governance committee or approved technology. They can use common programmatic languages -- including SAS -- to develop predictive analytic solutions and models. Most importantly, they do this in a discovery mode using raw data at large scale. With Hadoop there is no limitation on the data set size so analysis can be done at full scale instead of with a 5% sample. This achieves two benefits: first, the accuracy of the analysis increases dramatically, limited only by the skill of the data scientist. Second, anomalies (called outliers) become easier to identify; which is often critical in fraud detection and risk analysis.

Complex Batch versus Interactive Analysis

One simple way of deciding when to use which tool is to look at the type of workload and the user. Most business users can leverage interactive BI tools for building reports, dashboards, ad-hoc and iterative analysis. Nowhere is this more obvious than in the use of Online Analytical Processing (OLAP). Using pre-computed summaries, the business user can get hierarchical roll-up reports at the "speed-of-thought," an OLAP mantra meaning sub-second analysis. If the business requirement is interactive self-service analytics, the data warehouse is the best choice. In contrast, batch processing typically runs in minutes and is not typically invoked by business users.

Both Hadoop and data warehouses depend on running complex batch jobs to process massive amounts of data. When the application must run in parallel to achieve scalability and the program is highly complex, Hadoop has many advantages. However, as large scale processing becomes more complicated, a MapReduce programmer is required. In contrast, many complex applications run nightly in batch using the data warehouse. While these applications may be as sophisticated as any MapReduce program, they do not run in parallel. Thus, the requirement to run any language and any level of program complexity in parallel favors Hadoop. If the application does not need to run in parallel, the ease of SQL programming coupled with a parallel data warehouse is probably the best choice. Running parallel MapReduce programs issuing SQL to the data warehouse leverages the advantages of both subsystems.

Customer Churn and Recommendation Engines

Many data warehouses include components that detect probable customer defections (churn) and use a recommendation engine to persuade the consumer to stay. Using consumer profiles and predictive analytics, personalized offers are sent proactively to potential defectors to earn their continued loyalty. This application is mandatory in many industries such as telecommunications, eCommerce, retail, banking, and any consumer-based enterprise. Hadoop is ideal for pulling apart click streams from websites to find consumer preferences. The question now is, "Should we pull data out of the data warehouse for Hadoop to combine with the consumer preferences?" or "Should we put consumer preferences from Hadoop into the data warehouse for use by power users of BI tools?" The answer is that it depends on the business requirements. With Hadoop, many reports, churn results, and preferences can be refined down to a final result – so in some cases, Hadoop performs all of the functions. In cases when a data warehouse is present, it is probably already the system of record for persisting consumer preferences and using them with a campaign management subsystem to manage loyalty and multi-touch recommendations. In these cases, scrubbed data from Hadoop can be imported into the data warehouse. The question is not either-or, but how to use both tools as part of one application workload.

Text Parsing and Mining

Here is another case where Hadoop performs a critical part of the information supply chain and passes results to the data warehouse for further refinement. To begin with, relational databases have never been exceptional at parsing text. Relational operators are not geared for byte manipulations, although they can do it. Consider the click stream example where Hadoop is able to deploy multiple programming languages to tease apart a stream of text into useful structured data. So with text analysis, Hadoop is a good first step in finding keywords and performing sentiment analysis in blogs, syndicated online publications, emails and so on. Hadoop is not a text analytic system, so ontologies, industry-specific dictionaries, and sentiment algorithms must be acquired from universities, open source repositories, experts, or proprietary vendors. Hadoop provides the large-scale playing field; each vertical industry must provide the players.

Social media, such as blogs and email are mostly text. Once Hadoop refines this text, it is often stored in a data warehouse. Great value is derived when coupling a consumer's sentiments about a brand or product with the consumer's profile in the data warehouse. For instance, if we know a consumer "loves high definition TVs," then the data warehouse can pump out personalized offers for "HD TV for the bedroom" since the profile shows teenagers in the household.

When to Use Which

While there are certain use cases that are distinct to Hadoop or the data warehouse, there is also overlap where either technology could be effective. The following table is a good starting place for helping to decide which platform to use based on your requirements.

Requirement	Data Warehouse	Hadoop
Low latency, interactive reports, and OLAP	●	
ANSI 2003 SQL compliance is required	●	
Preprocessing or exploration of raw unstructured data		●
Online archives alternative to tape		●
High-quality cleansed and consistent data	●	
100s to 1000s of concurrent users	●	●*
Discover unknown relationships in the data	●	●
Parallel complex process logic		●
CPU intense analysis	●	●
System, users, and data governance	●	
Many flexible programming languages running in parallel		●
Unrestricted, ungoverned sand box explorations		●
Analysis of provisional data		●
Extensive security and regulatory compliance	●	
Real time data loading and 1 second tactical queries	●	●*

Figure 4. Requirements match to platforms

*HBase

Conclusions

Hadoop and the data warehouse will often work together in a single information supply chain. When it comes to Big Data, Hadoop excels in handling raw, unstructured and complex data with vast programming flexibility. Data warehouses also manage big structured data, integrating subject areas and providing interactive performance through BI tools. It is rapidly becoming a symbiotic relationship. Some differences are clear, and identifying workloads or data that runs best on one or the other will be dependent on your organization and use cases. As with all platform selections, careful analysis of the business and technical requirements should be done before platform selection to ensure the best outcome. Having both Hadoop and a data warehouse onsite greatly helps everyone learn when to use which.

ABOUT CLOUDERA

Cloudera, the leader in Apache Hadoop-based software and services, enables data driven enterprises to easily derive business value from all their structured and unstructured data. As the top contributor to the Apache open source community and with tens of thousands of nodes under management across customers in financial services, government, telecommunications, media, web, advertising, retail, energy, bioinformatics, pharma/healthcare, university research, oil and gas and gaming, Cloudera's products, depth of experience and commitment to sharing expertise are unrivaled.

+1 (650) 843.0595 | **cloudera.com**

cloudera

ABOUT TERADATA

Teradata is the world's largest company solely focused on data warehousing and integrated marketing management through database software, enterprise data warehousing, data warehouse appliances, and analytics. Teradata provides the best database for analytics with the architectural flexibility to address any technology and business need for companies of all sizes. Supported by active technology for unmatched performance and scalability, Teradata's experienced professionals and analytic solutions empower leaders and innovators to create visibility, cutting through the complexities of business to make smarter, faster decisions.

Simply put, Teradata solutions give companies the agility to outperform and outmaneuver for the competitive edge.

+1 (937) 242.4030 | **teradata.com**

TERADATA

© 2011 Cloudera, Inc. and Teradata Corporation. All rights reserved. Cloudera and the Cloudera logo are registered trademarks of Cloudera and/or its affiliates in the U.S. and worldwide. Teradata and the Teradata logo are registered trademarks of Teradata and/or its affiliates in the U.S. and worldwide.

0212 > EB-6448