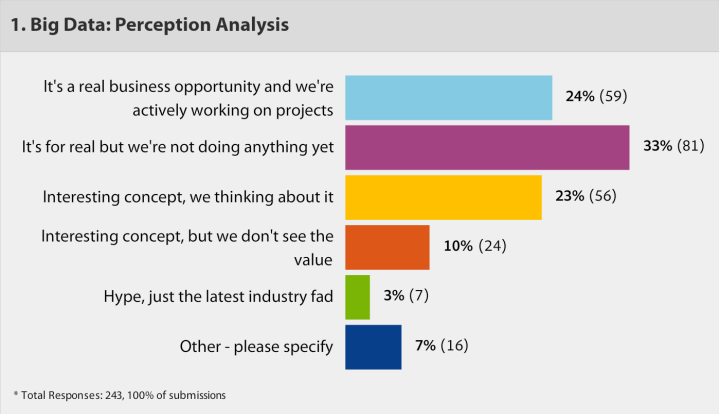
Big Data for the Rest of Us
Technical Whitepaper

TREASURE DATA
CLOUD DATA WAREHOUSING

Introduction

The importance of data warehousing and analytics has increased as companies seek to gain competitive advantage from their information assets. Indeed, there is a new breed of new company whose business model and operations are entirely driven by data, for example social media, online gaming and other Internet consumer businesses. For these companies, the ability to aggregate, analyze and act on massive amounts of data from disparate sources is their very lifeblood.



Treasure Data's May/June 2012 survey of business and IT users, showed widespread interest in Big Data solutions and acceptance of the Cloud as a deployment platform.

Big Data – defined by the "3 Vs" of volume, variety and velocity – is coming at companies from more sources, in more quantity and in greater variety than ever. As well as traditional data sources such as on-premise and web applications, data from sources such as mobile devices, social media applications and machine generated data is now threatening to overwhelm the ability of traditional data warehousing technologies and approaches. While the explosion of web and social media data may be the most obvious examples of this, even the humble residential smart meter can generate up to 400mb of raw data per year causing, for example, Pacific Gas & Electric to purchase an additional 1.2 Petabytes of disc storage to support a deployment of 700,000 smart meters.

Whatever their origin, these new sources generate volumes of data that have outgrown the capabilities of traditional data warehousing technologies and approaches. While emerging technologies such as Hadoop are designed to address and democratize "Big Data" analytics, they remain accessible only to larger enterprises with the technical resources and deep pockets required for success. And, as ex-Forrester Research analyst James Kobielus points out, despite the widespread interest in Hadoop, only 1% of enterprises are using Hadoop in production [1].

The Treasure Data Cloud Data Warehouse combines innovative technology with the economies of scale from Cloud computing to eliminate the cost and complexity barriers that make the benefits of leveraging Big Data understanding inaccessible. The Treasure Data Cloud Data Warehouse allows companies to deploy, load data and begin production analytics in days, not months—without the need to hire scarce and expensive technical resources.

This paper will outline the current problems facing Big Data projects; describe the Treasure Data Cloud Data Warehouse service / architecture, and discuss how this service benefits both the needs of fast-moving data-driven companies and traditional companies that want to use Cloud Big Data analytics alongside their existing analytics environment.

1 http://www.computerworld.com/s/article/9224587/Look_Before_You_Leap_Into_Hadoop

We are on the cusp of a tremendous wave of innovation, productivity and growth, all driven by big data.[2]

The Big Data Problem

Companies attempting to deploy Big Data analytic projects have found implementing and managing large scale analytics environments – whether traditional, Hadoop or custom Cloud-based - to be a major challenge. This is due to four critical issues:

SKILLS/RESOURCES

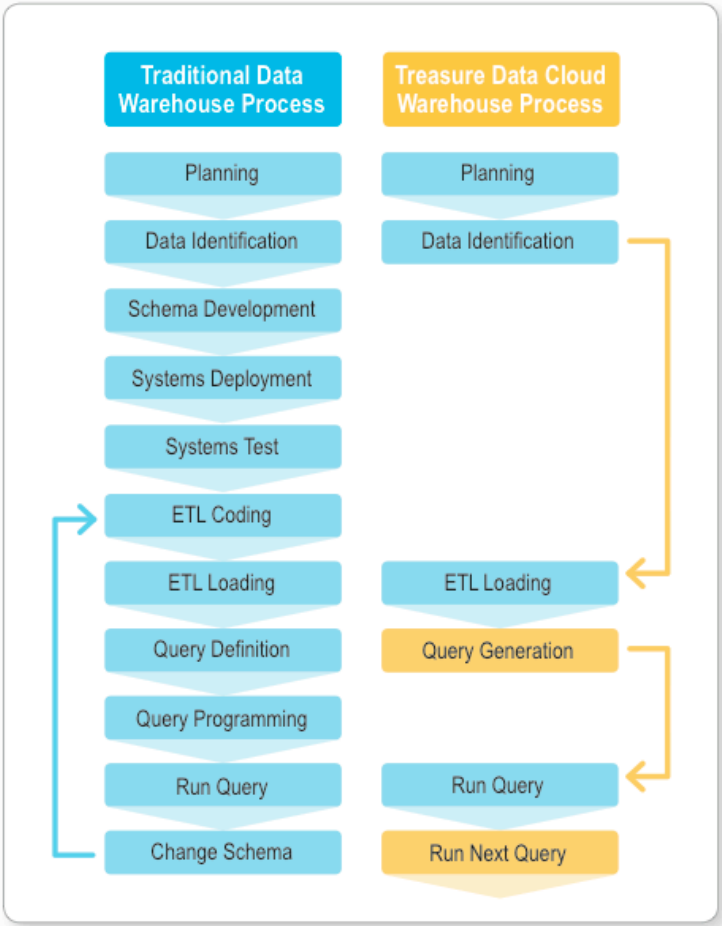
Big Data requires new and specialized skills that are in short supply. Traditional SQL savvy analysts are not well armed to meet current Big Data requirements. Developers are needed with proficiency in new languages and system design to build and maintain the data transfer and analytics code for high performance parallel processing systems. IT staff must also be skilled in deploying and managing the hardware, systems software and network to host the environment. If these skills exist in-house they are often committed to other critical projects. External hiring can take months and competitive salaries are high.

LONG IMPLEMENTATION TIME

Aside from staffing ramp-up time, traditional Big Data solutions require large upfront investments in proprietary systems that typically have a 6-12 month deployment-to-production delay. In the open source world, deployment and test of a multi-node Hadoop cluster is typically a 3-6 project including system deployment, custom code development, integration, test and tune—before the first query is run. As important as analytical understanding is, it is still an iterative process where insights drive new questions and dynamically changing data queries. In a resource constrained, handcrafted analytics environment, cycle time can easily exceed multiple days, or weeks, rendering this process impossible to achieve.

ETL AND SCHEMA DEPENDENCE

These two, often unconsidered, challenges can account for the majority of time and labor invested over the life of a Big Data initiative. ETL (extract, transform and load) includes moving bulk source data into the analytics engine as well as distributing subsets out to traditional warehouse platforms or BI tools. In the process, support for data formatting between the source and target platforms must be addressed. Current landscapes are populated with multiple scripting alternatives for different data types with manually created routines for both load and export processes, however the process is time consuming and results in an expanding code base that is difficult to maintain.



Most conventional data warehousing platforms are schema dependent, supporting an assumptive analytics model. In this model, data elements forecasted to yield insights are defined in advance, as well as structure of the data store schema. Performance considerations are important in initial design and the analyst must have knowledge of the underlying structure to insure query performance. When new columns are added to the table, the schema will need to be changed. Big Data analysis however, is largely non-assumptive meaning that the analyst seeks hidden patterns, relationships or events in the data that were not intuitively obvious from the outset. The user must be able to explore and query where the data takes them without the burden of performance consideration. In this model, schema dependence adds a significant administrative overhead that can become prohibitive.

MANAGEMENT OVERHEAD AND CHANGE

Operational overhead (the direct and indirect costs of people and processes) has emerged as one of the most significant barriers to Big Data projects. Systems management tools are immature or unavailable, and in an environment where complex systems and new techniques are deployed on premise, operational over head expense can quickly outpace upfront investment in hardware and software. Big Data environments are typically more dynamic than traditional warehousing and analytics environment which makes managing change in the ETL layer and scaling the data store an even greater – and more time and resource consuming - challenge than in traditional data warehouse solutions.

The Big Data problem creates two key barriers which must be overcome to allow mainstream adoption: (1) The cost to build, maintain, and support the infrastructure and systems, and (2) the time it takes to get answers to complex, cross tabular queries, which could be referred to as "Time-to-Answer." Today's Big Data initiatives are burdened with high up-front investment in systems and scarce, expensive human resource. Material ongoing management overhead is also seen as a burden for companies. Deployment time from initial investment to production analysis ranges from 6 months to over a year. And queue times between query response times are measured in days or weeks.

2"Big Data: The next frontier for innovation, competition and productivity" (McKinsey Global Institute, May 2011)

A Different Approach

The Treasure Data Cloud Data Warehouse merges the power of open-source Hadoop with patented technology innovation delivered as a Cloud service. This unique combination:

- Compresses time-to-answer complex analytical queries from months to days
- Reduces ramp up complexity and time of the “deployment-to-production” cycle
- Eliminates or reduces overhead costs
  - The upfront infrastructure costs to implement a big data warehouse
  - The need for specialized technical resources on staff

With Treasure Data you can take advantage of innovative data management and analytics technologies based on open source software to immediately load, store, manage and process large-scale data volumes. At the same time, Treasure Data avoids the cost and complexity typical of conventional data warehouse approaches so you can focus resources on business analytics. The following summarizes key architectural components and benefits:

BUILT-IN STANDARDIZED DATA COLLECTOR

Treasure Data provides TD-Agent, a fully supported Fluentd data collector. This lightweight daemon is installed on each data source and provides both batch and continuous data feed to the data store. The collector provides several advantages: Fluentd supports standard JSON format transformation for structured, semi-structured and unstructured data types. Output is machine readable as well as intuitive to humans. ETL routines are easily crafted, consistent, and reduce change management overhead.

The collector supports high performance parallel batch load to multiple concurrent targets as well a continuous feed to reduce subsequent load times and enable near-real time or event based analytics.

The collector supports high performance parallel batch load to multiple concurrent targets as well a continuous feed to reduce subsequent load times and enable near-real time or event based analytics. As important, the collector has been distributed as open-source software to provide broader adoption and innovation.

ENHANCED COLUMNAR STORE

The architecture extends the Hadoop stack to include an innovative columnar store to address the core performance and management issues of the Hadoop Distributed File System (HDFS) and conventional database engines. The data store provides greater efficiency in scanning target data at query run time and, unlike other implementations, it operates only on query-relevant columns rather than loading complete records for each query. These features dramatically reduce query time, processing overhead—and as a result, cost.

The database supports full schema independence with low overhead. This capability dramatically compresses both initial as well as update load cycles. Initial data load can consume hundreds of work hours over many elapsed weeks, depending on user familiarity with the data. Most of this time is spent identifying data and defining a format and schema. One schema change on a subsequent load can delay analysis by a week or more. Schema independence eliminates these delays, supports a non-assumptive analytics model and insulates the analyst from the need to understand the underlying data structure when generating queries. In some unique cases, the user may wish to instantiate a schema to gain incremental performance, this capability is also supported.

SQL – STYLE QUERY WITH PARALLEL EXECUTION

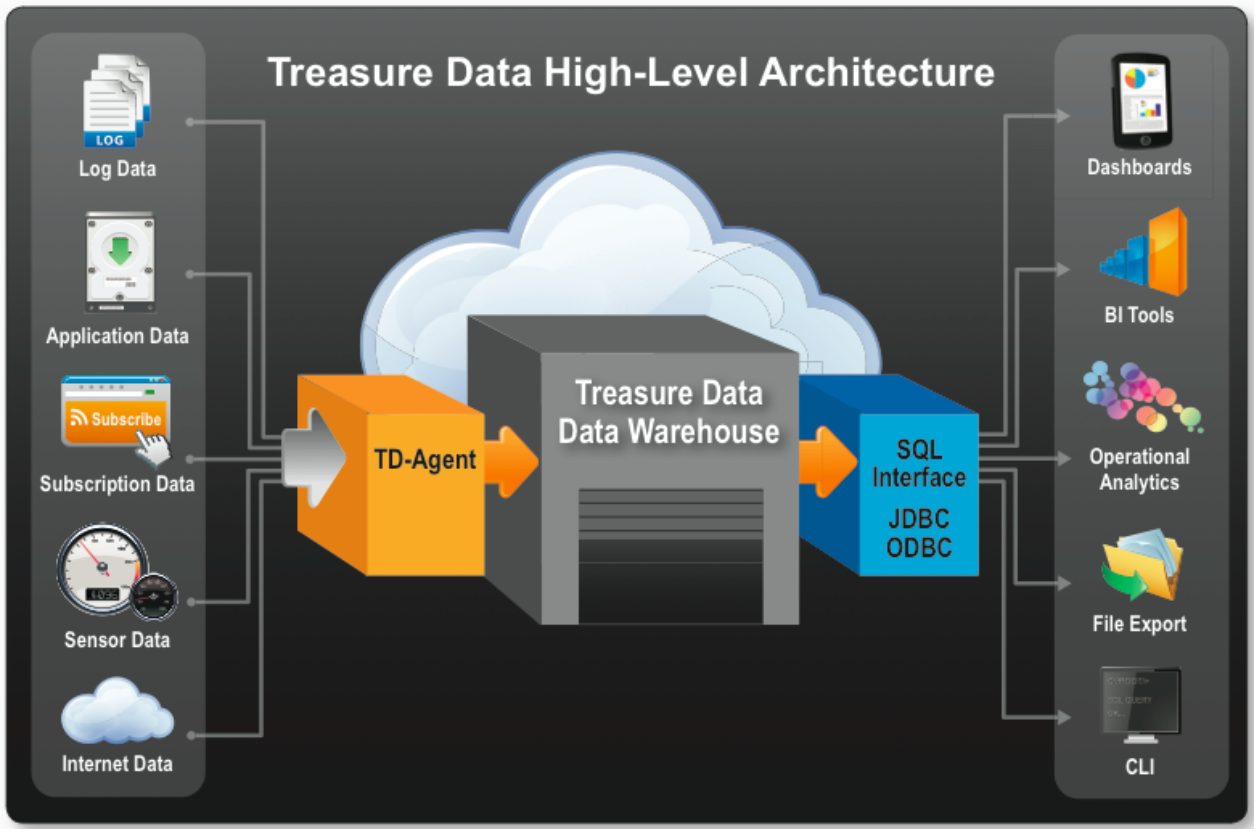
A familiar SQL style query interface is provided to abstract the complexity of Map Reduce programming. Statements are converted to Map Reduce format and executed at runtime to support full parallel processing. This enables staff with SQL capability to leverage the full power and flexibility of the traditional Hadoop environment without needing specialized programming skills. In addition this capability cuts the cycle time for query generation and supports an iterative dialog with the data. The need for business analysts and professional developers to generate and refine queries is eliminated.Existing investment in query intellectual property can be leveraged as well, transferring seamlessly to Treasure Data syntax.

EXPORT TO RDBMS/TRADITIONAL WAREHOUSE

A built-in export capability is provided for migrating data from the Treasure Data store to a traditional RDBMs or Data Warehouse. This enables efficient processing of large data volumes with Treasure Data both as a primary analytics engine as well as a pre-processing platform, using the results of Big Data analysis to better identify which data elements are appropriate for inclusion in the traditional Enterprise Warehouse query/reporting architecture, or Data Marts.

JDBC INTERFACE TO BI TOOLS

Treasure Data also provides a standard JDBC interface for data transfer to existing (or future) BI tools. Custom coding and maintenance to link these environments with the primary data store is eliminated. This capability enables Treasure Data business analysts to capitalize on a wide range of industry tools in a skills hierarchy where a professional analyst can broadly distribute the results/benefits of Big Data analysis and exploration to business colleagues. In turn, their colleagues can perform further t asks such as additional analysis, reporting, or KPI dashboard generation using their preferred business analytics tools and applications.



Cloud Service Delivery

The Treasure Data capabilities are delivered via a turnkey Cloud service delivery model. This approach extends the value of Treasure Data core technology and open source innovation to offer substantial technical and economic benefits:

IMMEDIATE DEPLOYMENT, FULLY ELASTIC

A production environment can be immediately deployed without the need to pull in house staff from other projects. Processing, storage and network resources are completely elastic, and can be scaled up or down as requirements dictate.

ROBUST ENVIRONMENT, TRANSPARENT SYSTEMS AND DATA MANAGEMENT

Treasure Data resides in an Amazon-hosted operational environment, a reliable and secure Cloud infrastructure platform that delivers 99.95% availability for on-demand computing, storage, and networking services to power your Cloud Data Warehouse. Extensive systems management and monitoring tools enable high-level control such as performance tuning and data protection for the service. The Cloud Data Warehouse is managed 24x7 by an operations staff highly familiar with all aspects of the system and service delivery best practices, eliminating the need to dedicate in-house operations staff and the operational overhead associated with an on-premise environment.

The Treasure Data Service is a subscription based, pay for use, offering that ties your investment in analytics more directly to your return.

CURATION AND SUPPORT

Treasure Data staff actively curates open source components (e.g. Apache Hadoop, etc.) for each system layer, and integrates and tests them in a complete systems environment to ensure that your Big Data solution is based on the latest software innovation and best practices. This integration is achieved while maintaining cross-layer compatibility. In addition, the Treasure Data dedicated support team is in place to proactively aid businesses in gaining the greatest return and productivity from the service.

With the Cloud Data Warehouse, upfront investment in systems and deployment are eliminated, as are downstream operations overhead and change management costs. Excess capacity cost is also eliminated in the on-demand model. The Treasure Data Service is a subscription based, pay for use, offering that ties your investment in analytics more directly to your return. Finally, the economies of scale inherent in Cloud computing, combined with open source software and the efficiencies of Treasure Data technology, results in a total cost structure for the Cloud Data Warehouse service which delivers a highly economical offering for the customer.

Conclusion

Meeting the challenges of the 3Vs of Big Data – volume, variety and velocity - have outpaced traditional warehousing technology and practices. Frustrated by the cost and pace of innovation associated with proprietary vendor solutions, many companies are turning to open-source alternatives to address their needs. While offering high performance, flexibility, and low cost, open source alternatives are inhibited by long deployment time, high complexity, skills shortages and overhead costs. The limitations are impractical for many data-driven organizations especially those in fast-paced Internet based markets. Treasure Data has developed a unique blend of Cloud computing, open-source software and technology innovation to remove these barriers and deliver a practical, accessible solution for data-driven companies seeking competitive advantage from Big Data and high performance analytics.

FOR MORE INFORMATION:

At Treasure Data we don't believe in business as usual. We're open source and Cloud natives and we are using the power of these movements to change data warehousing once and for all. Our Cloud Data Warehouse enables you to take advantage of Big Data in days not months. Sound too good to be true? Try it and prove it to yourself. Visit our website at http://www.treasure-data.com and start building your Cloud Data Warehouse today.