

HARNESSING THE VALUE OF BIG DATA ANALYTICS

By: Shaun Connolly, Vice President,
Corporate Strategy, Hortonworks
Steve Woledge, Sr. Director,
Product Marketing, Teradata

TERADATA®

How to Gain
Business Insight
Using MapReduce
and Apache
Hadoop with
SQL-Based Analytics

TABLE OF CONTENTS

- 2 Executive Summary
- 3 The Challenges of Converting Big Data Volumes into Insight
- 4 Clear Path to New Value
- 5 Choosing the Ideal Big Data Analytics Solution
- 7 Benefits of a Unified Data Architecture for Analytics
- 8 Choosing the Right Big Data Analytics Solution
- 9 Teradata Solutions for Big Data Analytics
- 10 Teradata, Aster, and Hadoop: When to Use Which Solution
- 12 Big Data Analytics in Action

EXECUTIVE SUMMARY

In business publications and IT trade journals, the buzz about “big data” challenges is nearly deafening. Rapidly growing volumes of data – from transactional systems like enterprise resource planning (ERP) software and non-transactional sources such as web logs, customer call center records, and video images – are everywhere. A tsunami of data, some experts call it.

Most companies know how to collect, store, and analyze their operational data. But these new multi-structured data types are often too variable and dynamic to be cost-effectively captured in a traditional data schema using only a standard query language (SQL) for analytics.

Some data scientists, business analysts, enterprise architects, developers, and IT managers are looking beyond these big data volumes and focusing on the analytic value they can deliver. These companies are searching for new analytic solutions that can transform huge volumes of complex, high-velocity data into pure business insight. They also seek new data-driven applications and analytics that can give them a competitive edge.

Leading organizations are exploring alternative solutions that use the MapReduce software framework, such as Apache Hadoop. While Hadoop can cost-effectively load, store, and refine multi-structured data, it is not well-suited for low latency, iterative data discovery or classic enterprise business intelligence (BI). These applications require a strong ecosystem of tools that provide ANSI SQL support, security, as well as high performance and interactivity.

The more complete solution is to implement a data discovery platform that can integrate Hadoop with a relational integrated data warehouse. New analytic discovery platforms like the Teradata Aster Big Analytics Discovery Solution combine the power of the MapReduce analytic framework with SQL-based BI tools that are familiar to analysts. The result is a unified solution that helps companies gain valuable business insight from new and existing data – using existing BI tools and skill sets as well as enhanced MapReduce analytic capabilities.

But which analytic workloads are best suited for Hadoop, the discovery platform, and an integrated data warehouse? How can these specialized systems best work together? What are the schema requirements for

different data types? Which system provides an optimized processing environment that delivers maximum business value with the lowest total cost of ownership? This paper answers these questions and shows you how to use MapReduce, Hadoop, and a unified data architecture to support big data analytics.

THE CHALLENGES OF CONVERTING BIG DATA VOLUMES INTO INSIGHT

What business value does data bring to your organization? If your company is like most, you wouldn't think of shifting production schedules, developing a marketing campaign, or forging a product strategy without insight gleaned from business analytics tools. Using data from transactional systems, your team reviews historical purchase patterns, tracks sales, balances the books, and seeks to understand transactional trends and behaviors. If your analytics practice is advanced, you may even predict the likely outcomes of events.

But it's not enough. Despite the value delivered by your current data warehouse and analytics practices, you are only skimming the surface of the deep pool of business value that data can deliver. Today there are huge volumes of interactional and observational data being created by businesses and consumers around the world. Generated by web logs, sensors, social media sites, and call centers, for example, these so-called "big data" volumes are difficult to process, store, and analyze.

According to industry analyst Gartner,¹ any effort to tackle the big data challenge must address multiple factors, including:

- ~ **Volume:** The amount of data generated by companies – and their customers, competitors, and partners – continues to grow exponentially. According to industry analyst IDC, the digital universe created and replicated 1.8 trillion gigabytes in 2011.² That's the equivalent of 57.5 billion 32GB Apple iPads.
- ~ **Velocity:** Data continues changing at an increasing rate of speed, making it difficult for companies to capture and analyze. For example, machine-generated data from sensors and web log data is being ingested

in real-time by many applications. Without real-time analytics to decipher these dynamic data streams, companies cannot make sense of the information in time to take meaningful action.

- ~ **Variety:** It's no longer enough to collect just transactional data – such as sales, inventory details, or procurement information. Analysts are increasingly interested in new data types, such as sentiments expressed in product reviews, unstructured text from call records and service reports, online behavior such as click streams, images and videos, and geospatial and temporal details. These data types add richness that supports more detailed analyses.
- ~ **Complexity:** With more details and sources, the data is more complex and difficult to analyze. In the past, banks used just transactional data to predict the probability of a customer closing an account. Now, these companies want to understand the "last mile" of the customer's decision process. By gaining visibility into common consumer behavior patterns across the web site, social networks, call centers, and branches, banks can address issues impacting customer loyalty before consumers decide to defect. Analyzing and detecting patterns – on the fly across and all customer records – is time-consuming and costly. Replicating that effort over time can be even more challenging.

Addressing the multiple challenges posed by big data volumes is not easy. Unlike transactional data, which can be stored in a stable schema that changes infrequently, interactional data types are more dynamic. They require an "evolving schema," which is defined dynamically – often on-the-fly at query runtime. The ability to load data quickly, and evolve the schema over time if needed, is a tremendous advantage for analysts who want to reduce time to valuable insights.

Some data formats may not fit well into a schema without heavy pre-processing or may have requirements for loading and storing in their native format. Dealing with this variety of data types efficiently can be difficult. As a result, many organizations simply delete this data or never bother to capture it at all.

¹ Source: "'Big Data' is Only the Beginning of Extreme Information Management," Gartner, April 2011

² Source: "Extracting Value from Chaos," John Gantz and David Reinsel, IDC, June 2011

CLEAR PATH TO NEW VALUE

Companies that recognize the opportunities inherent in big data analytics can take steps to unlock the value of these new data flows. According to Gartner, “CIOs face significant challenges in addressing the issues surrounding big data ... New technologies and applications are emerging... and should be investigated to understand their potential value.”³

Data scientists, business analysts, enterprise architects, developers, and IT managers are looking for alternative methods to collect and analyze big data streams. What’s needed is a unified data architecture that lets them refine raw data into valuable analytical assets. (See Figure 1.)

Specifically, they need to:

- ~ Capture, store, and refine raw, multi-structured data in a **data refinery platform**. This platform extends existing architectures that have been traditionally used to store data from structured information sources, such as transactional systems.
- ~ Explore and uncover value and new insights, quickly and iteratively, in an **analytic discovery platform**.
- ~ Provide IT and business users with a variety of **analytic tools and techniques** to discover and explore patterns.
- ~ Store valuable data and metadata in an integrated data warehouse so analysts and business applications can operationalize new insights from multi-structured data.



Figure 1. Architecture for Refining Big Data Volumes into Analytical Assets.

³ Source: CEO Advisory: 'Big Data' Equals Big Opportunity, Gartner, March 31, 2011

CHOOSING THE IDEAL BIG DATA ANALYTICS SOLUTION

To maximize the value of traditional and multi-structured data assets, companies need to deploy technologies that integrate Hadoop and relational database systems. Although the two worlds were separate not long ago, vendors are beginning to introduce solutions that effectively combine the technologies. For example, market leaders like Teradata and Hortonworks are partnering to deliver reference architectures and innovative product integration that unify Hadoop with analytic discovery platforms and integrated data warehouses.

What should companies look for to get the most value from Hadoop? Most importantly, you need a **unified data architecture** that tightly integrates the Hadoop/MapReduce programming model with traditional SQL-based enterprise data warehousing. (See Figure 2.)

A unified data architecture is based on a system that can capture and store a wide range of multi-structured raw data sources. It uses MapReduce to refine this data into usable formats, helping to fuel new insights for the business. In this respect, Hadoop is an ideal choice for capturing and refining many multi-structured data types with unknown initial value. It also serves as a cost-effective platform for retaining large volumes of data and files for long periods of time.

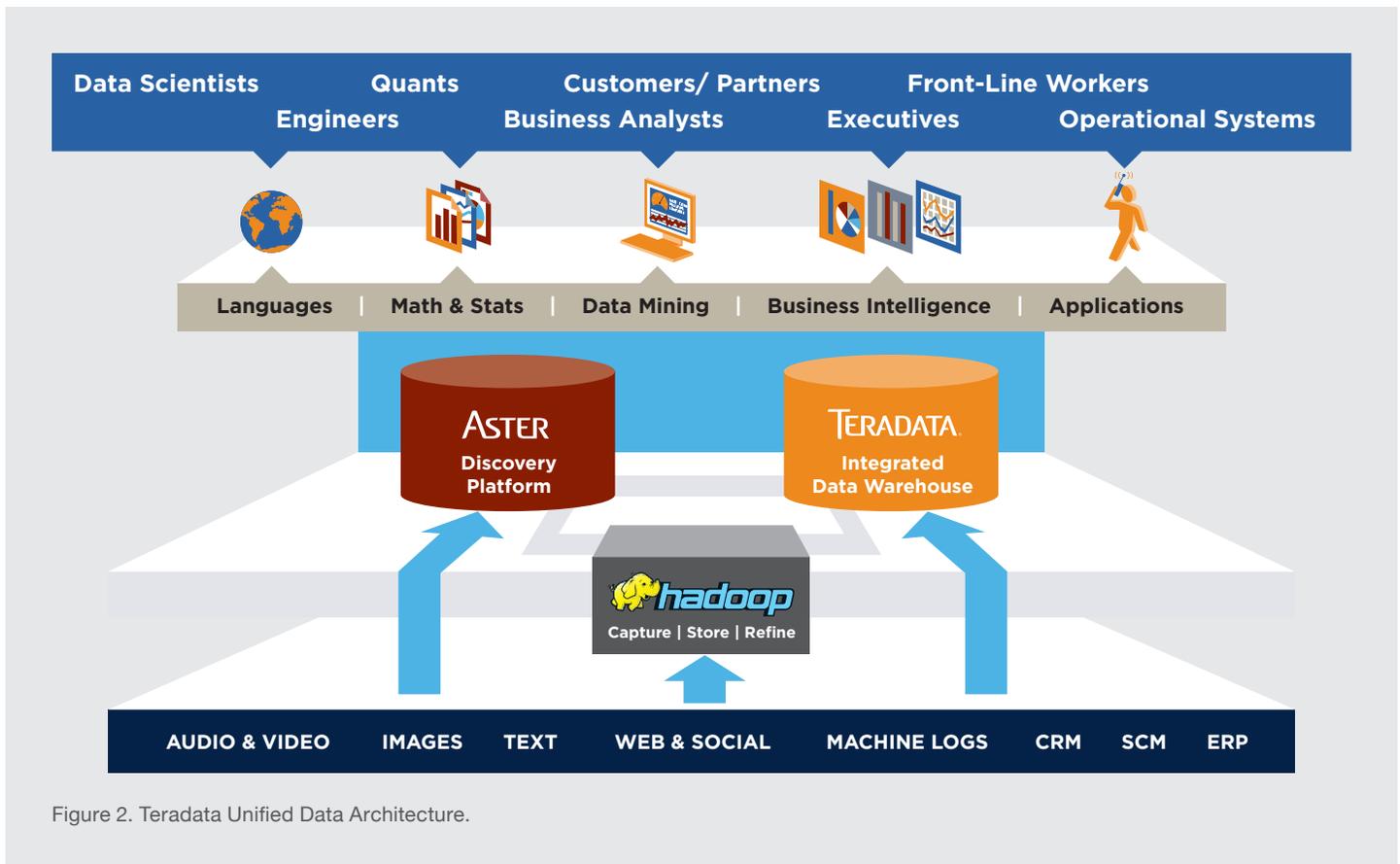


Figure 2. Teradata Unified Data Architecture.

A unified data architecture also preserves the declarative and storage-independence benefits of SQL, without compromising MapReduce's ability to extend SQL's analytic capabilities. By offering the intuitiveness of SQL, the solution helps less-experienced users exploit the analytical capabilities of existing and packaged MapReduce functions, without needing to understand the programming behind them. With this architecture, enterprise architects can easily and cost-effectively incorporate Hadoop storage and batch processing strengths together with the relational database system.

A critical part of a unified data architecture is a discovery platform that leverages the strengths of Hadoop for scale and processing while bridging the gaps around BI tool support, SQL access, and interactive analytical workloads. SQL-MapReduce helps bridge this gap by providing a **distinct execution engine** within the discovery platform. This allows the advanced analytical functions to execute automatically, in parallel across the nodes of the machine cluster, while providing a standard SQL interface that can be leveraged by BI tools.

Some products include a **library of prebuilt analytic functions** – such as path, pattern, statistical, graph, text and cluster analysis, and data transformation – that help speed the deployment of analytic applications. Users should be able to write custom functions as needed, in a variety of languages, for use in both batch and interactive environments.

Finally, an **interactive development tool** can reduce the effort required to build and test custom-developed functions. Such tools can also be used to import existing Java MapReduce programs.

To ensure that the platform delivers relevant insights, it must also offer enough **scalability** to support entire data sets – not just data samples. The more data you can analyze, the more accurate your results will be. As data science expert Anand Rajaraman recently wrote on the Datawocky blog, "...Adding more, independent data usually beats out designing ever-better algorithms to analyze an existing data set."⁴

MAPREDUCE AND HADOOP: A PRIMER

How do technologies such as MapReduce and Hadoop help organizations harness the value of unstructured and semi-structured data?

MapReduce supports distributed processing of the common map and reduction operations. In the map step, a master node divides a query or request into smaller problems. It distributes each query to a set of map tasks scheduled on a worker node within a cluster of execution nodes. The output of the map steps is sent to nodes that combine or reduce the output and create a response to the query. Because both the map and reduce functions can be distributed to clusters of commodity hardware and performed in parallel, MapReduce techniques are appropriate for larger datasets.

Apache Hadoop consists of two components: Hadoop MapReduce for parallel data processing and the Hadoop Distributed File System (HDFS) for low-cost, reliable data storage. Hadoop, the most popular open-source implementation of the MapReduce framework, can be used to refine unstructured and semi-structured data into structured formats that can be analyzed or loaded into other analytic platforms.

To support rapid iterations in the analytical discovery processes, the solution also must offer **high performance** and ease of analytic iteration. Look for standard SQL and BI tools that can leverage both SQL and MapReduce natively. By leveraging relational technology as the data store, analysts receive the performance benefit of a query optimizer, indexes, data partitioning, and simple SQL statements that can execute instantaneously.

⁴ "More data usually beats better algorithms," Anand Rajaraman, Datawocky, March 24, 2008, <http://anand.typepad.com/datawocky/2008/03/more-data-usual.html>.

In sum, a unified data architecture blends the best of Hadoop and SQL, allowing users to:

- ~ Capture and refine data from a wide variety of sources
- ~ Perform necessary multi-structured data preprocessing
- ~ Develop rapid analytics
- ~ Process embedded analytics, analyzing both relational and non-relational data
- ~ Produce semi-structured data as output, often with metadata and heuristic analysis
- ~ Solve new analytical workloads with reduced time to insight
- ~ Use massively parallel storage in Hadoop to efficiently store and retain data

WHY NOT REPLACE ANALYTICAL RELATIONAL DATABASES WITH HADOOP?

Analytical relational databases were created for rapid access to large data sets by many concurrent users. Typical analytical databases support SQL and connectivity to a large ecosystem of analysis tools. They efficiently combine complex data sets, automate data partitioning and index techniques, and provide complex analytics on structured data. They also offer security, workload management, and service-level guarantees on top of a relational store. Thus, the database abstracts the user from the mundane tasks of partitioning data and optimizing query performance.

Since Hadoop is founded on a distributed file system and not a relational database, it removes the requirement of data schema. Unfortunately, Hadoop also eliminates the benefits of an analytical relational database, such as interactive data access and a broad ecosystem of SQL-compatible tools. Integrating the best parts of Hadoop with the benefits of analytical relational databases is the optimum solution for a big data analytics architecture.

BENEFITS OF A UNIFIED DATA ARCHITECTURE FOR ANALYTICS

Using Hadoop with an analytic discovery platform and integrated data warehouse can help you meet the challenges of gaining insight from big data volumes.

With a blended solution, you can combine the developer-oriented MapReduce platform with the SQL-based BI tools familiar to business analysts. Providing the best of both worlds, this type of solution lets you use the right tool for the job – that is, SQL for structured data and MapReduce processing for large-scale procedural analytics that would otherwise require complex, multi-pass SQL technologies. Business users can easily and intuitively perform analytics processes that would otherwise be difficult or impossible.

This ease of use in turn enables extended use of the data. Data scientists and analysts alike can manage and analyze both relational and non-relational data, inside and outside the integrated data warehouse. They can also perform iterative, rich, big data analytics with greater accuracy and effectiveness.

Because a blended solution offers higher performance than SQL-only analytics, users can gain insights faster. What's more, native integration of SQL and MapReduce lets users perform analysis without changing the underlying code, so they can dig deeper for insight.

Unifying these best-of-breed solutions also offers a lower total cost of ownership (TCO) than individual tools. Many software-only products come with deployment options that use commodity hardware and offer linear, elastic scaling. Appliance-based solutions deliver high value in a prepackaged form.

By unifying these solutions into a single reference architecture, companies can unlock value from big data volumes without needing to retrain personnel or hire expensive data scientists. By protecting your existing investment in relational database technology and user skill sets, blended solutions also are kind to your budget. And unlike typical open source offerings, a blended solution supports corporate compliance, security, and usability requirements with greater rigor.

CHOOSING THE RIGHT BIG DATA ANALYTICS SOLUTION

As big data challenges become more pressing, vendors are introducing products designed to help companies effectively handle the huge volumes of data and perform insight-enhancing analytics. But selecting the appropriate solution for your requirements need not be difficult.

With the inherent technical differences in data types, schema requirements, and analytical workloads, it's no surprise that certain solutions lend themselves to optimal performance in different parts of the unified big data architecture. The first criteria to consider should be what type of data and schema exist in your environment.

Possibilities include:

- ~ **Data that uses a stable schema (structured)** – This can include data from packaged business processes with well-defined and known attributes, such as ERP data, inventory records, and supply chain records.
- ~ **Data that has an evolving schema (semi-structured)** – Examples include data generated by machine processes, with known but changing sets of attributes, such as web logs, call detail records, sensor logs, JSON (JavaScript Object Notation), social profiles, and Twitter feeds.
- ~ **Data that has a format, but no schema (unstructured)** – Unstructured data includes data captured by machines with a well-defined format, but no semantics, such as images, videos, web pages, and PDF documents.

DATA TASK	POTENTIAL WORKLOADS
Low-cost storage and retention	<ul style="list-style-type: none"> ~ Retains raw data in manner that can provide low TCO-per-terabyte storage costs ~ Requires access in deep storage, but not at same speeds as in a front-line system
Loading	<ul style="list-style-type: none"> ~ Brings data into the system from the source system
Pre-processing/prep/ cleansing/constraint validation	<ul style="list-style-type: none"> ~ Prepares data for downstream processing by, for example, fetching dimension data, recording a new incoming batch, or archiving old window batch.
Transformation	<ul style="list-style-type: none"> ~ Converts one structure of data into another structure. This may require going from third-normal form in a relational database to a star or snowflake schema, or from text to a relational database, or from relational technology to a graph, as with structural transformations.
Reporting	<ul style="list-style-type: none"> ~ Queries historical data such as what happened, where it happened, how much happened, who did it (e.g., sales of a given product by region)
Analytics (including user-driven, interactive, or ad-hoc)	<ul style="list-style-type: none"> ~ Performs relationship modeling via declarative SQL (e.g., scoring or basic stats) ~ Performs relationship modeling via procedural MapReduce (e.g., model building or time series)

Table 1. Matching Data Tasks and Workloads.

Semantics can be extracted from the raw data by interpreting the format and pulling out required data. This is often done with shapes from a video, face recognition in images, and logo detection. Sometimes formatted data is accompanied by metadata that can have a stable schema or an evolving schema, which needs to be classified and treated separately.

Each of these three schema types may include a wide spectrum of workloads that must be performed on the data. Table 1 lists several common data tasks and workload considerations.

TERADATA SOLUTIONS FOR BIG DATA ANALYTICS

To help companies cost-effectively gain valuable insight from big data volumes, Teradata recently introduced the Teradata Aster Big Analytics Discovery Solution. This analytic discovery platform helps companies to bridge the gap between esoteric data science technologies and the language of business. It combines the developer-oriented MapReduce platform with the SQL-based BI tools familiar to business analysts.

This unified, intuitive software product gives business analysts with ordinary SQL skills the ability to work like data scientists, asking questions and getting valuable insight from huge stores of data. Using this product, business analysts can quickly identify patterns and trends using a variety of techniques, such as pattern and graph analysis. For the first time, they can rapidly and intuitively explore massive volumes of multi-structured digital data from a wide variety of sources. And companies can unlock value from big data without needing to retrain personnel or hire expensive data scientists.

The Teradata Aster Big Analytics Discovery Solution includes Aster Database, a suite of prepackaged analytic functions and apps and an integrated development environment for easy development of custom SQL-MapReduce

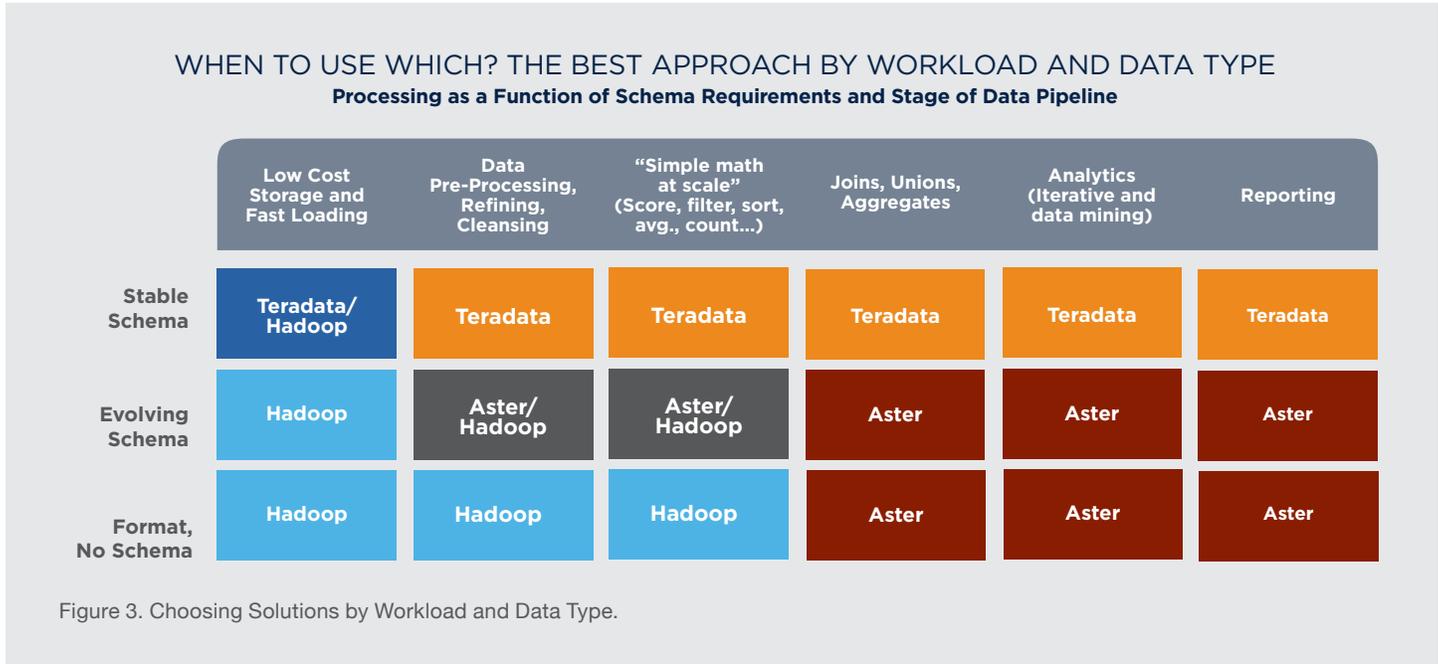
DEMOCRATIZING ANALYTICS

It's no secret that leading companies use data and analytics to drive competitive advantage. Some organizations build data science teams to mine big data volumes, using enhanced analytic techniques and tools that can expose hidden patterns in consumer behavior, preferences, fraud, and other business trends.

A new class of analytic discovery platform tools extends these valuable practices from highly skilled, highly paid developers and data scientists to analysts and business users. These platforms give users their choice of tools – whether they prefer SQL, BI tools, statistical packages (such as R or SAS), or programming languages.

By extending the use of these tools to a broader constituency of users, data discovery platforms help democratize the power of data science throughout the business. Instead of confining data discovery to data engineers – who may lack the business context of the problems they are asked to solve – the data discovery platform brings powerful analytics tools to the entire business community.

functions. This solution includes an embedded analytics engine that supports an array of common programming languages, allowing analysts to build and easily modify sophisticated algorithms without additional training or investment. Analysts can integrate insights from the discovery platform into the data warehouse. These insights are then available for ongoing strategic and operational analysis. Business analysts can ask and answer even more differentiated business questions.



TERADATA, ASTER, AND HADOOP: WHEN TO USE WHICH SOLUTION

Figure 3 offers a framework to help enterprise architects most effectively use each part of a unified data architecture. This framework allows a best-of-breed approach that you can apply to each schema type, helping you achieve maximum performance, rapid enterprise adoption, and the lowest TCO.

The following use cases demonstrate how you can apply this framework to your big data challenges.

STABLE SCHEMA

Sample applications: Financial analysis, ad-hoc/OLAP queries, enterprise-wide BI and reporting, spatial/temporal, and active execution (in-process, operational insights).

Characteristics: In applications with a stable schema, the data model is relatively fixed. For example, financial reporting and analysis is conducted much the same way throughout the fiscal quarter or year. Transactions collected from point-of-sale, inventory, customer relationship management, and accounting systems are known and

change infrequently. This business requires ACID (atomicity, consistency, isolation, durability) property or transaction guarantees, as well as security, well-documented data models, extract, transform, and load (ETL) jobs, data lineage, and metadata management throughout the data pipeline – from storage to refining through reporting and analytics.

Recommended Approach: Leverage the strength of the relational model and SQL. You may also want to use Hadoop to support low-cost, scale-out storage and retention for some transactional data, which requires less rigor in security and metadata management.

Suggested Products: Teradata provides multiple solutions to handle low-cost storage and retention applications as well as loading and transformation tasks. With this architectural flexibility, Teradata products help customers meet varying cost, data latency, and performance requirements. For example:

- ~ Customers that want to store large data volumes and perform light transformations can use the Teradata Extreme Data Appliance. This platform offers low-cost data storage with high compression rates at a highly affordable price.

- ~ For CPU-intensive transformations, the Teradata Data Warehouse Appliance supports mid-level data storage with built-in automatic compression engines.
- ~ Customers that want to minimize data movement and complexity and are executing transformations that require reference data can use the Teradata Active Enterprise Data Warehouse.

This appliance provides a hybrid, multi-temp architecture that places cold data on hard disks and hot data on solid-state storage devices. With Teradata Database, customers can dynamically and automatically compress cold data, driving higher volumes of data into the cold tier.

EVOLVING SCHEMA

Sample applications: Interactive data discovery, including web click stream, social feeds, set-top box analysis, sensor logs, and JSON.

Characteristics: Data generated by machine processes typically requires a schema that changes or evolves rapidly. The schema itself may be structured, but the changes occur too quickly for most data models, ETL steps, and reports to keep pace. Company e-commerce sites, social media, and other fast-changing systems are good examples of evolving schema. In many cases, an evolving schema has two components – one fixed and one variable. For example, web logs generate an IP address, time stamp, and cookie ID, which are fixed. The URL string – which is rich with information such as referral URLs and search terms used to find a page – varies more.

Recommended Approach: The design of web sites, applications, third-party sites, search engine marketing, and search engine optimization strategies changes dynamically over time. Look for a solution that eases the management of evolving schema data by providing features that:

- ~ Leverage the back end of the relational database management system (RDBMS), so you can easily add or remove columns
- ~ Make it easy for queries to do “late binding” of the structure
- ~ Optimize queries dynamically by collecting relevant statistics on the variable part of the data
- ~ Support encoding and enforcement of constraints on the variable part of the data

Suggested Products: Teradata Aster is an ideal solution for ingesting and analyzing data in an evolving schema. The product provides a discovery platform, which allows evolving data to be stored natively without pre-defining how the variable part of the data should be broken up.

Teradata Aster also allows the fixed part of the data to be stored in a schema and indexed for performance. With this feature, analysts can define structure of the variable component at query run time. This task happens as part of the SQL-MapReduce analytic workflow in a process called “late data binding” or “schema on read.” The system handles this processing behind the scenes, allowing the analyst to interpret and model data on-the-fly, based on different analytic requirements. Analysts never need to change data models or build new ETL scripts in order to break out the variable data. This feature reduces cost and saves time, giving analysts the freedom to explore data without being constrained by a rigid schema.

Hadoop can also ingest files and store them without structure, providing a scalable data landing and staging area for huge volumes of machine-generated data. Because Hadoop uses the HDFS file system for storage instead of a relational database, it requires additional processing steps to create schema on-the-fly for analysis. Therefore, Hadoop can slow an iterative, interactive data discovery process.

However, if your process includes known batch data transformation steps that require limited interactivity, Hadoop MapReduce can be a good choice. Hadoop MapReduce enables large-scale data refining, so you can extract higher-value data from raw files for downstream data discovery and analytics. In an evolving schema, Hadoop and Teradata Aster are a perfect complement for ingesting, refining, and discovering valuable insights from big data volumes.

NO SCHEMA

Sample applications: Image processing, audio/video storage and refining, storage, and batch transformation and extraction.

Characteristics: With data that has a format, but no schema, the data structure is typically a well-defined file format. However, it appears less relational than non-relational, lacks semantics, and does not easily fit into the notion of traditional RDBMS rows and columns. There is often a need to store these data types in their native file formats.

Recommended Approach: Hadoop MapReduce provides a large-scale processing framework for workloads that need to extract semantics from raw file data. By interpreting the format and pulling out required data, Hadoop can discern and categorize shapes from video and perform face recognition in images. Sometimes format data is accompanied by metadata, which can be extracted, classified, and treated separately.

Suggested Products: When running batch jobs to extract metadata from images or text, Hadoop is an ideal platform. You can then analyze or join this metadata with other dimensional data to provide additional value. Once you've used Hadoop to prepare the refined data, load it into Teradata Aster to quickly and easily join the data with other evolving- or stable-schema data.

BIG DATA ANALYTICS IN ACTION

Since its recent release, the Teradata Aster analytic discovery solution has already helped dozens of customers realize dramatic business benefit through enhanced insight. The following examples illustrate how a company can use the Teradata Aster analytic discovery solution, Teradata integrated data warehouse technology, and Hadoop to deliver new business insight from big data analytics.

CUSTOMER RETENTION AND PROFITABILITY

Banks and other companies with retail operations know that keeping a customer satisfied is far less costly than replacing a dissatisfied customer. A unified data architecture can help companies better understand customer communications and take action to prevent unhappy consumers from defecting to a competitor. (See Figure 4.)

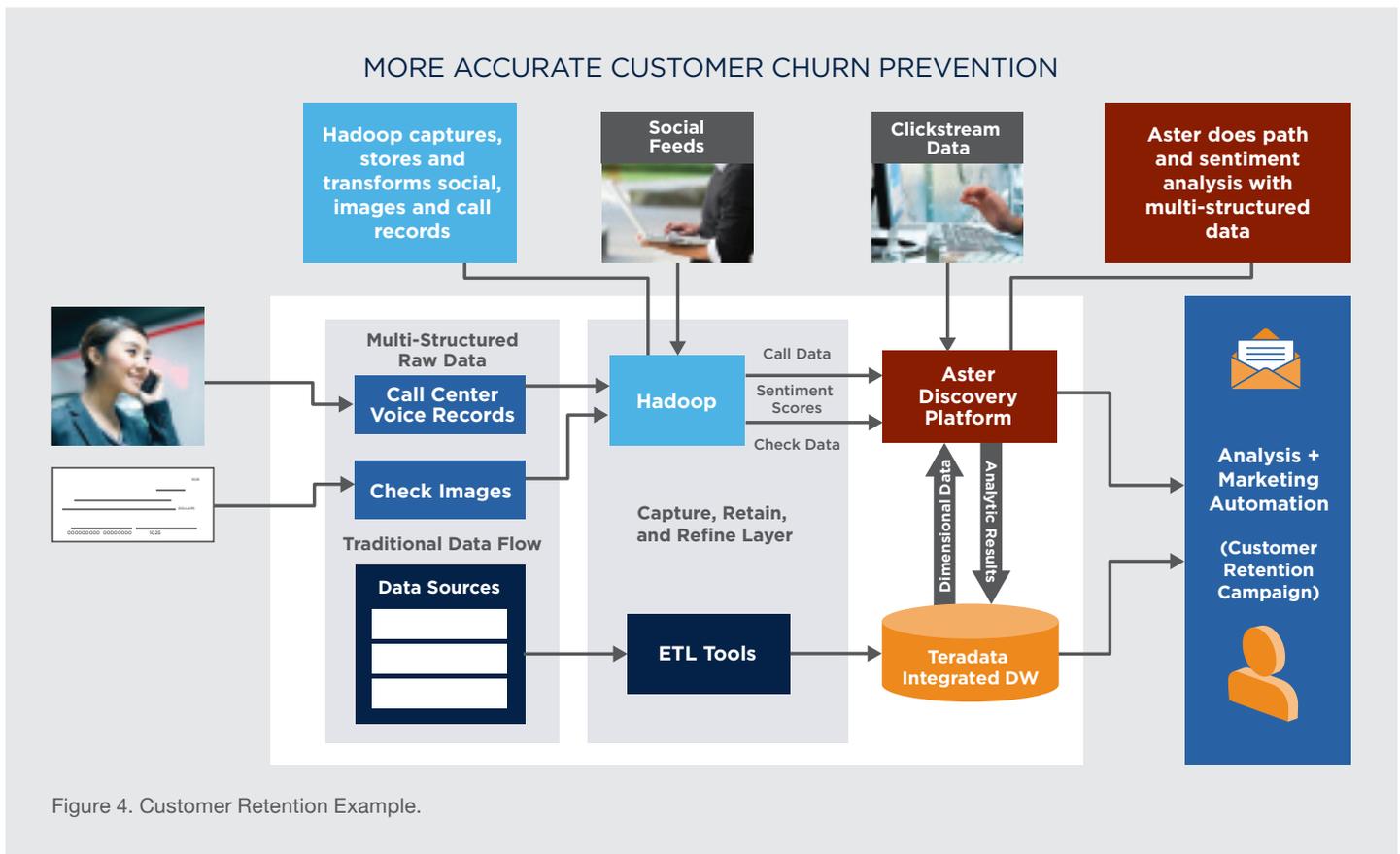


Figure 4. Customer Retention Example.

For example, assume that a customer, Mr. Jones, calls a bank's contact center to complain about an account fee. The bank collects interactive voice response information from the call center, storing this unstructured data in the data discovery platform. The bank also uses its Teradata integrated data warehouse to store and analyze high-resolution check images.

Using Hadoop, analysts can efficiently capture these huge volumes of image and call data. Then they can use the Aster-Hadoop adaptor or Aster SQL-H™ - method for on-the-fly data access of Hadoop data at query runtime - to merge the unhappy customer data from call center records with the check data.

By using Aster nPath - one of the SQL-MapReduce-enabled functions in the Teradata Aster solution - an analyst can quickly determine whether Mr. Jones may be about to switch over to the new financial institution. The analyst identifies the unhappy sentiment data from Mr. Jones' call to the contact center. In addition, the analyst notes that one of the customer's deposited checks is drawn on the account of another bank, with the note "brokerage account opening bonus." The analyst can recommend that a customer support agent reach out to Mr. Jones with an offer designed to prevent him from leaving.

Furthermore, the analyst can use these tools to reveal customers with similar behavior to that of Mr. Jones. Marketing and sales personnel can proactively approach these dissatisfied customers, making offers that save those relationships, too.

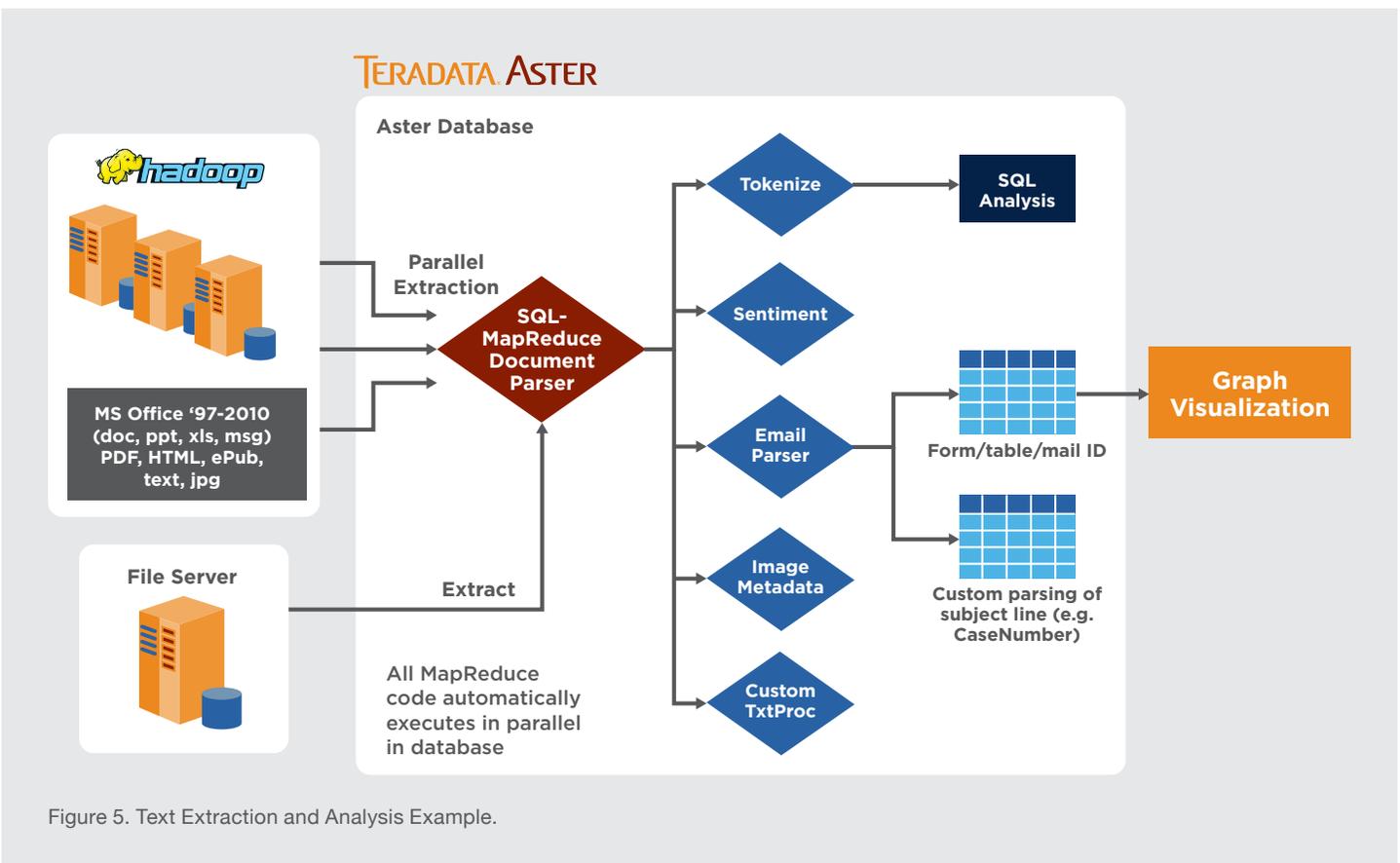


Figure 5. Text Extraction and Analysis Example.

TEXT EXTRACTION AND ANALYSIS

Applications such as e-discovery, sentiment analysis, and search rely on the ability to store, process, and analyze massive amounts of documents, text, and emails. In their native formats, it is very difficult to analyze these data types. Huge data volumes further complicate analysis efforts.

The Teradata Aster analytic discovery platform includes features that support text extraction and analysis applications. Hadoop's HDFS is ideal for quickly loading and storing any type of file in its native format. Once stored, these files can be processed to extract the relevant data and structure it for analysis.

Next, analysts can use SQL-MapReduce functions for tokenization, e-mail parsing, sentiment analysis, and other types of processing. These features allow businesses to

identify positive or negative consumer sentiments or look for trends or correlations in email communications. New insights can be combined with other information about the customer in the integrated data warehouse. Analyzing this data can help companies identify customers likely to churn or to identify brand advocates who might be open to a marketing affiliation program to help drive awareness and sales.

FOR MORE INFORMATION

For more information about how you can bring more value to the business through a unified data architecture, contact your Teradata or Teradata Aster representative or visit us on the web at Teradata.com or TeradataAster.com.



10000 Innovation Drive Dayton, OH 45342 teradata.com



TERADATA.

THE BEST
DECISION
POSSIBLE

The Best Decision Possible and SQL-H are trademarks, and Teradata and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide. Apache Hadoop and Hadoop are registered trademarks of the Apache Software Foundation. Teradata continually improves products as new technologies and components become available. Teradata, therefore, reserves the right to change specifications without prior notice. All features, functions, and operations described herein may not be marketed in all parts of the world. Consult your Teradata representative or Teradata.com for more information.