



Managing Big Data with Hadoop & Vertica

A look at integration between the Cloudera distribution for Hadoop and the Vertica Analytic Database

Copyright Vertica Systems, Inc. October 2009

Cloudera and Vertica

Relational database management systems such as Vertica excel at analytic processing for big volumes of structured data including call detail records, financial tick streams and parsed weblog data. Vertica is designed for high speed load and query when the database schema and relationships are well defined. Cloudera's Distribution for Hadoop, built on the popular open source Apache Software Foundation project, addresses the need for large-scale batch processing of unstructured or semi-structured data. When the schema or relationships are not well defined, Hadoop can be used to employ massive MapReduce-style processing to derive structure out of data. The Cloudera Distribution simplifies installation, configuration, deployment and management of the powerful Hadoop framework for enterprise users.

Each can be used stand alone – Vertica for high-speed loads and ad-hoc queries over relational data, Cloudera's Distribution for general-purpose batch processing, for example from log files. As described in this whitepaper, combining Hadoop and Vertica creates a nearly infinitely scalable platform for tackling the challenges of big data.

The Challenges of Big Data

Big data is by no means a new phenomenon. In domains as diverse as telecommunications, financial services, healthcare, life sciences, retail and web applications, large data volumes have been a normal part of daily operations for at least a decade. For years, companies have been trying to tackle increasing data volumes, initially with monolithic data management systems storing data in single company-wide enterprise warehouses, and then with distributed data marts and federated data models composed of physically separated databases. The classic data processing pipeline – capturing operational data in large volumes, and then transforming it – still holds today, but the kind of data captured, the amount that must be stored and the complexity of the processing and analysis required to digest it have all changed dramatically.

For the data-driven enterprise, new data management tools developed by the community and offered by commercial vendors are addressing these growing data challenges. These solutions are innovating in taking a scale-out approach as data volumes grow. Scale-out architectures allow additional capacity and processing to be added incrementally, in modular units such as server blades. Unlike scale-up architectures that require a large central machine, scale-out has no theoretical limit.

Scalable processing frameworks such as the Hadoop MapReduce framework and massively parallel analytic databases such as Vertica address distinct, but complementary, problems for managing large data. Both employ shared-nothing architectures where each processing node is a self-contained unit with independent processing, memory and storage. Each of

these nodes, which communicate over standard networks, handles a subset of the total data. In combination, they provide a single system image for accessing and querying data. These are scalable systems since users can simply add more nodes to increase storage capacity and improve response time.

Hadoop is well suited to batch processing on large distributed data sets. Hadoop operates on unstructured data such as images, and semi-structured data such as log files. It is a nearly infinitely scalable solution for collecting data from multiple sources, cleaning the data and storing it for long-term use. Hadoop excels at processing tasks – graph construction and traversal, complex aggregation and summarization, natural language processing, machine learning and other analyses – that SQL handles poorly. Hadoop uses a Java API and has support for various programming languages that can express the Map and Reduce stages of an algorithm. Hadoop is able to push these operations out to the nodes that store the data, exploiting inherent parallelism in the storage fabric and delivering extraordinary scalability.

The Vertica Analytic Database is designed to capture large volumes of structured data and to provide answers to complex analytic questions in real time. Because Vertica runs on a grid of shared-nothing machines, it can scale to large numbers of nodes, all operating in parallel. Vertica communicates using SQL over standard protocols such as JDBC for Java and ODBC for other languages. As a data source for tools such as Hadoop, Vertica provides near limitless structured storage and a bridge between large batch data processing and high speed data analytics.

Each of Vertica and Hadoop is able to store and process data on its own. Vertica excels at capturing structured data with a well-defined schema, and Hadoop handles high-volume and unstructured feeds where schema need not be enforced until processing time. Each delivers analysis and processing services. Vertica concentrates on high-speed interactive analyses, and Hadoop delivers exhaustive batch-style complex processing.

In combination, the two offer unmatched flexibility and power. Cloudera and Vertica have integrated the two products so that data can be moved easily between them, and so that Hadoop's MapReduce operators can work on data stored directly in the Vertica parallel RDBMS. Users can work with large-scale unstructured data directly in Hadoop, explore relational data in real time using Vertica, and exploit the complex algorithmic processing power of Hadoop's MapReduce over relational data stored natively in Vertica.

Where Data Comes From

In the past decade the scope of data challenges has expanded dramatically, with more people using the detailed information that flows through an enterprise and across the internet. From basic network data to detailed application logs, click stream data and distributed user data, businesses of all size are capturing and analyzing an exponentially

growing amount of information. In order to gain a competitive edge and then remain competitive, business users require immediate and widespread access to this data.

For example, the increased availability of detailed financial data, from basic tick stores to option feeds and mortgage data has spurred a cottage industry of financial analytic specialists. This has only intensified since the financial crisis of 2008. Pharmaceutical companies and health care providers are collecting detailed records from research and development through to individual patient health monitors after a drug has been released to market. Manufacturers embed monitoring systems in their products to gather operational statistics and collect them to correlate with environmental metrics and system performance.

Unstructured Data	Semi-structured Data
	<pre> 10.10.11.156 - user [15/Nov/2008:13:45:26 -0619] "GET /ad.gif HTTP/1.0" 200 394 "http://www.mysite.com/page.html" "Mozilla/5.02 [en] (WinXP; I ;Nav)" 10.10.11.159 - user [15/Nov/2008:13:45:28 -0619] "GET /ad.gif HTTP/1.0" 200 394 "http://www.mysite.com/page.html" "Mozilla/5.02 [en] (WinXP; I ;Nav)" 10.10.11.152- user [15/Nov/2008:13:45:29 -0619] "GET /ad.gif HTTP/1.0" 200 394 "http://www.mysite.com/page.html" "Mozilla/5.02 [en] (WinXP; I ;Nav)" </pre>

Analysts commonly expect worldwide data growth to approach 5 exabytes (an exabyte is over 1 thousand petabytes and over 1 million terabytes) of structured data and 8 exabytes of unstructured data in the next two years as measured in disk consumption (HDS, IDC http://blogs.hds.com/hu/2007/12/the_changing_enterprise_data_profile_idc.html).

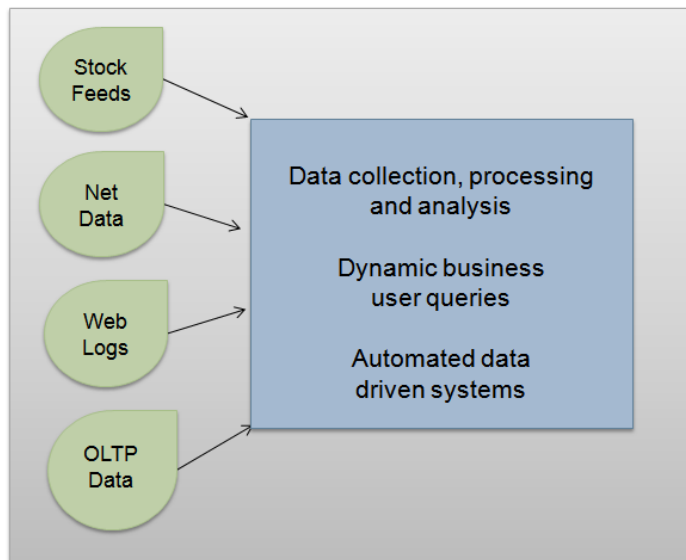
With advanced data analysis tools such as Hadoop and Vertica putting this data in the hands of sophisticated business users, access to data has become a critical component to daily business operations. Rapid access to, and sophisticated processing of, data provide a competitive advantage and introduce a new dependency within organizations. The dependency on data for everything from targeted advertising, split second trading algorithms and both tactical and strategic business decisions manifests in new requirements for availability, scalability, efficiency and security.

Paradigms of Data Management

The explosion of data collection and analysis has led to adoption of this new generation of tools. The scale-out architecture and high performance processing and analytics meet

demanding business requirements for handling growing data volumes. While the classic tools for warehousing, aggregating and aging data do not accommodate, the new model for data mining and business-driven analytics is built around data stores warehoused on tens, hundreds or thousands of commodity servers managing distributed file systems, massively parallel processing and highly optimized distributed analytic databases.

Businesses receive data from a wide variety of sources at a non-stop pace. Data is collected from grids of web servers, clusters of OLTP databases or networks of switches. This data is collected as it streams in, either in flat files across massive grids or as structured data directly into distributed databases.



Data originates in multiple forms from multiple sources and is combined to address diverse business requirements.

These raw data feeds, composed of detailed events, are analyzed on the fly as the data arrives. The data may be queried in real time to watch for important events that demand immediate action. It can also be stored with the full collection of historical information from the stream and processed to identify complex patterns. The classic click stream, for example, is composed of individual requests across a batch of web and application servers. Each request contains identifiers such as the referring page, requesting client and optional data such as cookies that can be used to recreate user sessions. The transformation process known as “sessionization” must handle billions of clicks per day and extract structures that describe user access to the site, pages visited and advertisements displayed, and construct a logical session that describes the logical path that each user followed while browsing the site.

Financial data has well-defined record structure, yet there are many patterns to be discovered across sets of data. Using Hadoop for data processing, analysts stream large

volumes of historical data from the file system, or directly from Vertica, to identify trends and correlations and to test complex trading algorithms. Quantitative analysts can apply complex patterns to billions of rows to identify trading strategies, refine ad targeting algorithms or explore multiple iterations in everything from machine design to drug delivery.

The Vertica cluster is optimized for data density, high-speed predicates, moving aggregates and relational analytics. These queries run in seconds and can stream out terabytes of data in parallel to multiple hadoop clusters. Each Hadoop cluster is sized independently for the scale and timeliness of computation. The results can stream back into Vertica and the storage is automatically optimized.

Tools that Scale

The classic toolset for managing operational data for analysis included a complex batch processing tool for extracting or capturing data, transforming it and loading into a monolithic database. While these tools have advanced to handle a variety of data processing tasks, they are generally limited in scalability and not well-suited to handle the large data volumes of today and tomorrow. The classic batch processing tools are flexible and can be parallelized but do not easily scale to even tens of nodes, never mind to hundreds or thousands. Similarly, monolithic databases can scale up with additional hardware but hit a wall at each order of magnitude increase in data size.

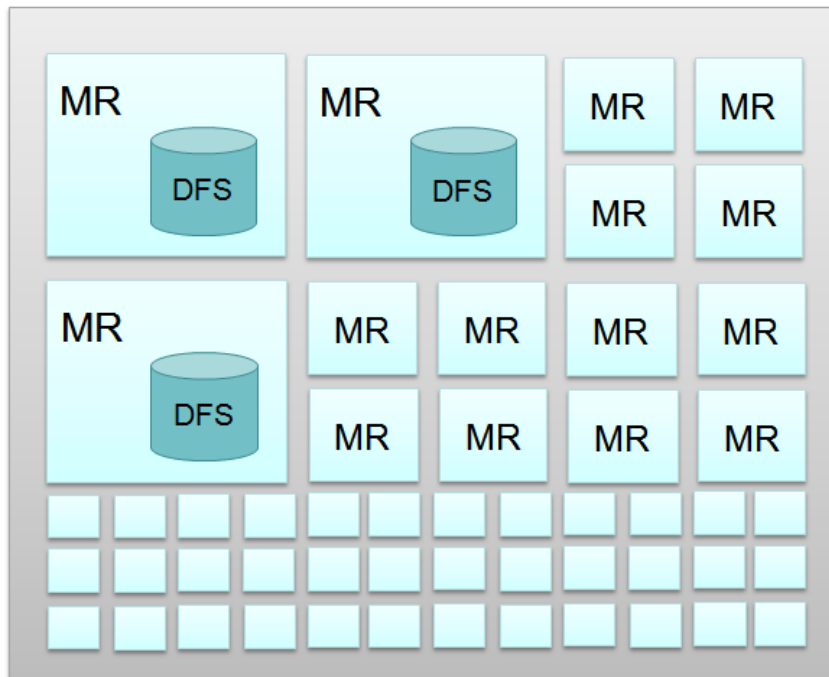
This ever-expanding world of data management presents new challenges, and technologists have adopted new tools to meet them. The four tenets of business-critical data management are availability, scalability, efficiency and security. While previous tools have been able to address some subset of those demands, modern tools are able to handle all four of them, now and into the future. A combination of the Hadoop MapReduce framework and Distributed File System, deployed and managed using Cloudera's tools, together with Vertica, a very fast massively parallel analytic database, offers businesses the tools that are critical for success.

Similar to classic ETL, Hadoop is able to extract or capture large batches of data, transform unstructured or semi-structured data and perform aggregations or logical data consolidation. In addition to processing data as it flows from its source to the database, Hadoop can transform data from any Vertica database back into that same database. This is useful for storing a simple structured representation, such as in an operational data store, and normalizing it into a multi-dimensional model for more flexible *ad hoc* analysis.

Hadoop is a scalable MapReduce framework that allows developers to create complex jobs that can be parallelized across tens, hundreds or even thousands of nodes. Written in Java, Hadoop can process data stored either in HFDS, the Hadoop Distributed File System or in a database such as Vertica. For example, web log data that is stored on HFDS can be

sessionized using a series of MapReduce jobs, with the final results stored in Vertica for ad hoc usage analysis.

Using Cloudera's Distribution for Hadoop, the same basic map and reduce functions can be prototyped on a single machine and deployed in scale on thousands to reduce computation time. Hadoop also includes a basic SQL interface called Hive to facilitate answering simple questions, and a high level language called Pig for defining transformations. Like other ETL solutions, MapReduce can be invoked in batch to process sets of incoming data. Hadoop can read data both from numerous data sources including HDFS and any JDBC compatible source.

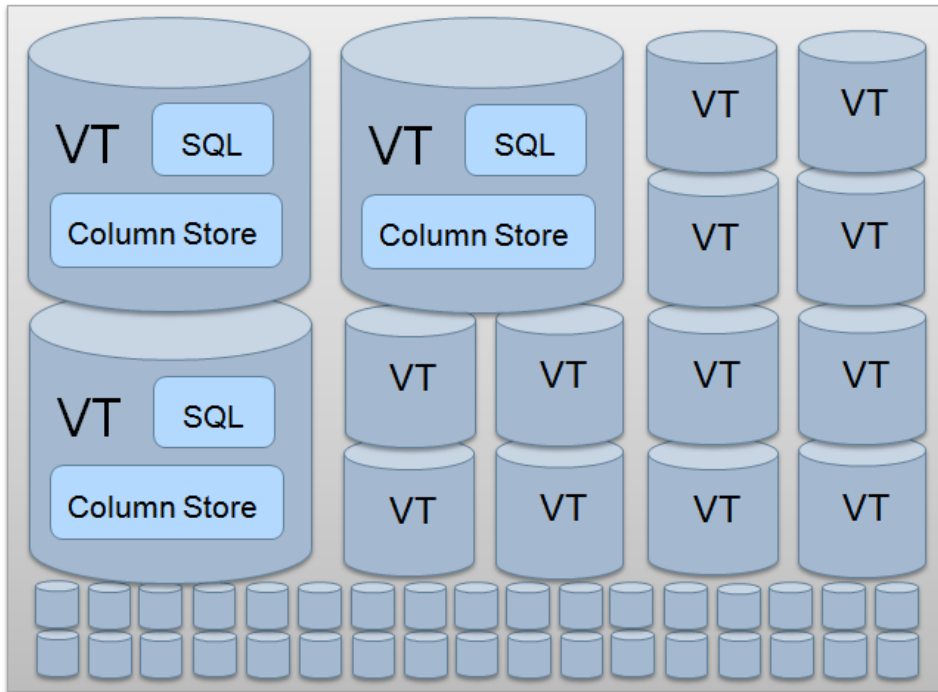


Map Reduce is a scalable framework for large data processing.

HDFS is a distributed file system that works in conjunction with Hadoop. It provides a fault tolerant storage system capable of managing distributed data across many nodes in a cluster. Both Hadoop and HDFS are optimized for file streaming, including large sequential writes and reads. HDFS stores multiple copies of every file for redundancy, and Hadoop's MapReduce implementation can take advantage of these copies to service multiple readers and to more efficiently run data processing workloads.

Vertica is a massively parallel analytic database, designed from scratch to meet large data load and query requirements. A Vertica database is composed exclusively of query-optimized structures, similar conceptually to a database composed entirely of materialized views (with no base tables). These structures scale out by distributing data across dozens of nodes and compressing data in columns. This unique combination of technologies enables

users to store near limitless amounts of structured data. Vertica provides a standard SQL interface to users, as well as compatibility with popular ETL, reporting and business intelligence tools.



Vertica is a high performance distributed SQL analytics database.

Vertica is also modeled on append often, read intensive usage. It supports very fast append to all nodes and real-time access to data from any node. This makes Vertica the perfect bridge between massive data processing in Hadoop and real-time analytics via any number of front end BI tools.

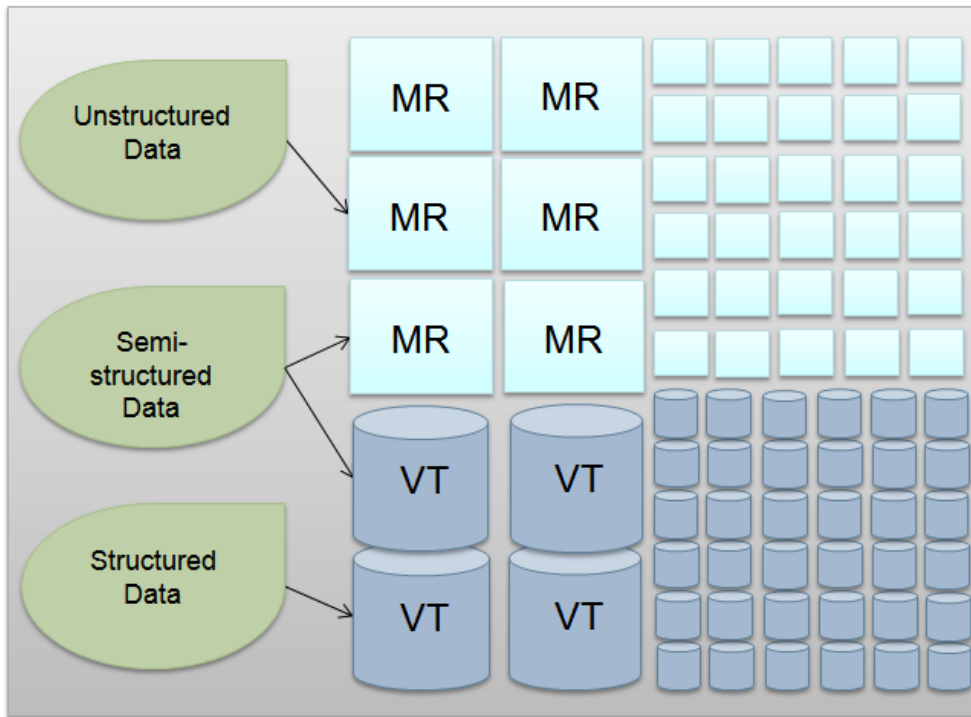
Scalable Big Data Lifecycle Management

These new tools that store and process large amounts of information allow businesses to extract much more value from their data. In large data centers there are numerous common scenarios where data flows through a number of tools as it gets processed and analyzed. As data volumes grow and become more available to users it increases the need for sophisticated analysis.

Consider a web-based application serving millions of users and collecting billions of web click events per day. This web data is logged into distributed file systems such as HDFS and processed in batches across thousands of machines using Hadoop. This process of converting semi-structured log data into structured relations of events, page views, sessions and users requires massive compute power. The requirements for data processing grow as a site becomes more popular attracting more users and handling increasing web traffic. The

output of this processing is stored in a Vertica cluster, represented in a highly efficient format, optimized for *ad hoc* analysis and real-time dashboards. As data grows the Vertica cluster scales out to improve load speeds, manage more data and accommodate more users running more complex analyses.

Hadoop handles complex procedural analysis of unstructured data stored in HDFS. Click events are compiled into page views and user sessions so that business analysts can track how effective the website is and so that they can mine the data for targeted advertising. This data is combined with existing structured data that is streamed out of Vertica. Hadoop runs MapReduce jobs to construct multi-dimensional detail data for analysis. For example, log data may reference existing user information or detailed product information. Combining these data sets exploits the scalability of both Hadoop and Vertica – Hadoop with hundreds of nodes capable of distributing the processing and Vertica on dozens of nodes optimized for data density and high speed analytics.



Unstructured and some semi-structured data is processed by Hadoop then loaded into the Vertica database. Structured and some semi-structured data streams directly into Vertica. Data streams between the database and MapReduce for processing.

The resulting structured data can be streamed directly into Vertica where it can be queried in real time. Vertica is used to drive up-to-date dashboards and for *ad hoc* analysis. For example, questions about user behavior can be reported on using any SQL front end tool concurrently by hundreds of users. Users issue *ad hoc* queries to explore different

dimensionality of the data, looking for user behavior by browsing patterns, previous requests or even geographic origin and time of day.

The Data Driven Enterprise

The modern competitive business is increasingly data-driven, and the data volumes and complexity of analysis required are growing exponentially. Scalable tools that can address these demands are available today and are being adopted by the most competitive organizations across many industries. Large-scale processing frameworks that use Hadoop are able to run complex algorithms and scale out as data demands. MPP analytic databases such as Vertica are used to warehouse and analyze hundreds of terabytes – even petabytes – of data, accessed by tens, hundreds or thousands of business analysts as well as automated systems.

With these tools widely available, and with compute and storage prices dropping, any organization that does not adopt such techniques and hone an analytic edge is at a disadvantage. Those that do will find emerging trends more quickly, respond to shifting market conditions faster and optimize critical functions that require timely data. For more information on Cloudera's Distribution for Hadoop and supporting products and services, go to <http://www.cloudera.com>. To learn about the MPP Vertica Analytic Database, go to <http://www.vertica.com/Hadoop>.