Understanding How Sensage Compares/Contrasts with Hadoop

1. How does Sensage's approach to managing large, distributed data systems compare/contrast with Hadoop in terms of storage, processing and querying?

From a storage perspective, it's important to look at data loading, data compression, distribution, columnar orientation, random access, serial access, big data capacity, wide/sparsely populated data and replication.

Sensage uses data loaders to partition data across clusters for optimal retrieval and compression and avoids normalizing data to ensure subsequent analysis is applied to complete datasets. Sensage's approach to data distribution also helps avoid hotspots where a small number of systems perform the majority of the work because all of the data needed by the query happens to reside on those computing platforms.

Hadoop relies on its Map Reduce function that operates on "in-place" data; this means data is collapsed into smaller chunks (aka horizontal partitioning) and distributed across the cluster without giving it any structure. Hadoop performs this function without the need for defining schema or requiring data transformation during the loading process.

Sensage achieves average compression rates of 10:1 and often gets compression rates higher depending on the cardinality of the data while Hadoop's compression rates vary.

When it comes to handling high volumes of distributed data, Sensage and Hadoop are adept at retrieving significant amounts of data across clusters of computing devices. In the case of Sensage, its Massively Parallel Processing Architecture is purpose-built to manage, store and analyze exceptionally high volumes of data in distributed environments.

Hadoop "stitches" data files together in favor of using a columnar or row-oriented database approach. On the other hand, Sensage leverages a columnar database architecture for event data and it gives structure to the unstructured/polystructured data. This enables faster access to the qualifying data. When dealing with large volumes of data or complex queries, columnar-based storage is highly scalable and responsive.

While Hadoop leverages H-Base, another columnar data store, there are latency and complexity issues to overcome.[1]

Unlike Hadoop, Sensage supports random data access (sparse data scans). Because Hadoop is based on linear access, it is not efficient at handling wide and sparsely populated data. In contrast, Sensage provides specialized "sparse query" optimization.

With respect to replication, Hadoop provides automatic duplication of data on multiple nodes for high-availability. And so does Sensage by default; data duplication is provided on multiple nodes to ensure highly available data.

From a processing perspective, both Sensage and Hadoop provide capabilities to store and stream extremely large datasets in batch mode. The two approaches differ in alternate ways to retrieve data faster. Hadoop has no indexing capabilities; and while Sensage has no index overhead on disk space, it has special access path structures to optimize data access.

[1] Apache Hadoop. Not derivative products like Hadapt. Because, Hadapt uses postgres in each node and loads data into Postgres instead of HDFS. http://dbmsmusings.blogspot.com/2010/03/distinguishing-two-major-types-of_29.html

Another key distinction between Sensage and Hadoop is in the area of data integrity. Hadoop has no native capabilities to ensure data integrity while Sensage does, which is critical when storing, retaining and analyzing petabytes of data.

Unlike Hadoop, which does not provide transactional protection, Sensage supports data persistence through atomicity – which enforces an "all or nothing" rule with respect to database modifications. In other words, if one part of the transaction fails, the entire transaction fails. Again, this is crucial to data fidelity.

When it comes to repeatable processing and data access, Hadoop is good for "one-off" processing such as Web analytics where repeated access is not needed once statistical information is gathered. In contrast, Sensage is designed to handle repeatable processing for data that requires "repeat visits".

Hadoop data has to be parsed every time a query is executed in order to analyze the data before returning results unless system administrators put additional effort into creating a pre-parsed format and corresponding readers. Sensage does not incur such overhead making query execution time faster than Hadoop+Hive SQL.

From a querying perspective, Sensage provides a number of built-in optimizers that ensure multiple data access paths are available. This includes sparse query optimizations, for example. Hadoop, on the other hand, has none.

Providing "joins" for queries is handled differently between Hadoop and Sensage. With Hadoop, system administrators have to write complex functions to achieve query joins. And while HIVESQL provides a SQL interface, it contains limited SQL operators. Sensage provides native query joins through its Open Access Extension (OAE) layer offering SQL 92 compatibility to deliver the necessary query translation. This enables seamless data access data to Sensage's Event Data Warehouse.

The same is true for "intersections", complex analysis (group by and order by), random and serial access, batch analysis, user-defined functions and procedures, and data analysis. These require writing complex functions in Hadoop while Sensage easily accommodates these key capabilities.

In addition, cached query results, drill downs, data partitioning, integration with third-party tools and pre-computed analytics are not available within Hadoop. In stark contrast, these functions are all readily available with Sensage.

2. How is Sensage complementary with Hadoop?

For customers that want to employ Hadoop, Sensage has the capability to provide seamless data integration through a data warehouse connector, such as those developed by companies including Persistent. Customers likely to want a Sensage connector to Hadoop are those that need to overcome Hadoop's latency issues that persist every time data is accessed. With a Hadoop-Sensage connector, users can push data for selective sources and process the data within Sensage using ETL tools such as Pentaho Kettle. Users familiar with ETL tools and Sensage can accomplish this.

3. What are the pros/cons of using Hadoop for event data?

Hadoop's advantages include no load overhead, the ability to shard (horizontal partitioning) data across thousands of nodes, and the ability to support multiple copies of data for fault tolerance. For those organizations that only need to process data once and discard it (such as Google analytics), Hadoop is a good fit.

On the flip side, every time a user wants to analyze data, MapReduce programs have to be written/reused to parse the data before the analysis can be completed. Because multiple files need to be opened every time the data is processed, latency is very high. In addition, in order for MapReduce to achieve its optimal efficiency, hundreds and thousands of nodes need to be employed. That said, nodes with disk drives running at different speeds pull performance levels down to the slowest device in the food chain.

As an open source tool, Hadoop support and expertise is fractional compared to well-established resources developing, servicing and supporting DBMS-based systems.

4. Are there any instances where customers should forward data from Sensage to Hadoop?

Sensage pioneered the concept of processing unstructured data in a structured manner 11 years ago. Moving structured data from Sensage into a format to process unstructured data is fraught with disadvantages that include: (1) loss of data access path optimization, (2) query latency and (3) unnecessary duplication of data.

5. What access control policies does Hadoop support?

Relative to Sensage, the security control mechanisms in Hadoop support weak permission settings. As a distributed file system, Hadoop relies on permission settings such as “chmod”, “chown”, and “chgrp” and quota settings such as “fileQuota” and “diskSpaceQuota”. This is in sharp contrast to Sensage’s use of role-based access controls.[2] OS level controls can be changed by anyone who has access to OS administration. Sensage’s role-based security provides more controlled and granular access to data depending on each user’s role. In addition, Sensage audits data access for users to analyze their own access patterns and provides control where needed.

6. For organizations considering Hadoop, are they more likely to use the open source version or something integrated like Cloudera?

This is a vexing problem for organizations considering Hadoop because there are so many “flavors” beyond the original Apache version. Companies considering Hadoop and its derivatives will need to evaluate alternatives from a laundry list of vendors that include Cloudera, CouchBase, Hortonworks (a Yahoo spin-off), Hadapt, Asterdata, Apache Cassandra, Amazon, IBM, Greenplum, MapR, Datameer, Zettaset, Hstreaming, Outerthought, and more.

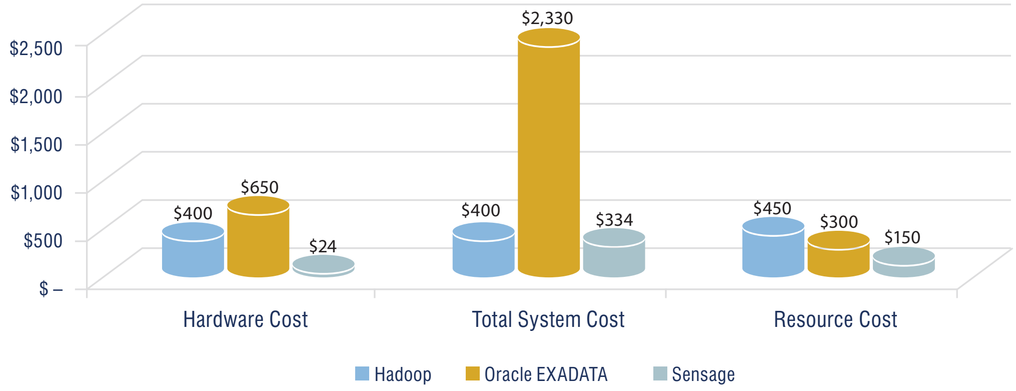
Organizations adopting the pure open source version of Hadoop will need to be well staffed with the appropriate technical resources (which are relatively scarce given how nascent Hadoop is with respect to production-scale deployment).[3]

7. While Hadoop is a “free” open source system, are there real hard dollar costs that users should be aware of?

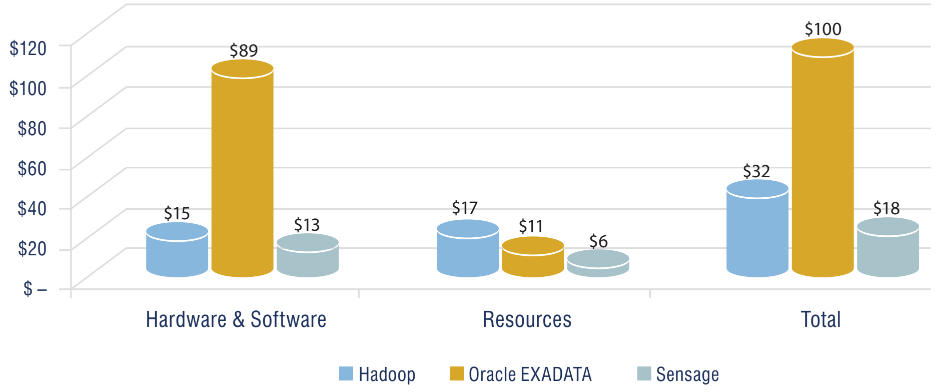
Absolutely. For starters, in order to derive the maximum value from the pure open source version of Hadoop, organizations will need to have a dedicated development team. And those technical resources, which are relatively scarce given how young Hadoop is, are commanding compensation commensurate with demand. In addition, Hadoop is a complex distributed file system, and its complexity requires highly experienced technical talent to derive value from the system. Plus, Hadoop is less efficient than DBMS-based systems running an MPP-based architecture like Sensage. Translation: more hardware is required to support Hadoop.

[2] http://www.defcon.org/images/defcon-17/dc-17-presentations/defcon-17-jason_schlesinger-cloud_security.pdf
[3] Recommended reading: “The Forrester Wave™: Enterprise Hadoop Solutions, Q1 2012 – February 2, 2012”

For a hard dollar analysis between Hadoop and Sensage, consider the following hardware, software and resource cost for a 168 TB data store over three years[4]:



Hardware, software and resource cost amortization over three years in $K for 168 TB data.



Hourly cost amortization over three years for 168 TB.

8. How does Hadoop's hardware footprint compare to the Sensage solution?

The short answer is that Sensage can perform more data processing with fewer nodes than Hadoop. In fact, Sensage delivers superior data processing capabilities compared to any Massively Parallel Processing database.

Here's why. Hadoop's scalability is best realized over a large number of nodes. It does not scale efficiently with a fewer number of nodes. While it can run on inexpensive commodity servers to drive down the cost per gigabyte compared to more expensive proprietary servers and storage systems, it involves additional hard dollar costs to operate and maintain. This additional cost basis includes developers, system administrator overhead and application development time. When it all adds up, Hadoop's total cost of ownership is greater than Sensage running on more expensive hardware platforms.

[4] Based on "http://www.itworld.com/data-centerservers/238091/road-hadoop-part-2-how-much-less-would-you-pay"

9. If an organization’s data is growing faster than it can be transformed for relational or structured formats, is Hadoop a good fit?

The short answer is “potentially”.

10. If a company’s data uses complex data types, such as deep nested XML or program-defined data structures that are expensive to convert to relational file structures, is Hadoop a good fit? And what about complex data objects such as BLOBs and data types that vary from row to row?

Hadoop has the potential to be a good fit for these environments.

11. What about environments where data source schema is well known or there is adequate time to model source schema?

Hadoop is not a fit. From a TCO perspective, companies are better off paying a small overhead at the load time, convert it to Sensage and get all benefits of Sensage.

12. What if you need to query data with varying query depths many times during the lifespan of retained data (e.g. 1 to 5 years)?

Hadoop is not a fit. For organizations requiring repeated access to data during extended retention periods (like 2 to 5 years), Hadoop has to parse, process and return data analysis results for each query. Data parsing for each query request will be very slow compared to data parsed once and stored in columnar data stores.

13. What if you need to analyze data in real-time or near real-time?

Hadoop is not a fit because it cannot do it by itself. Some organizations are trying to combine Hypertable and HBase to do some amount of real-time analysis. This approach assumes users can wait until data is converted from Hadoop to Hypertable/HBase.[5] That said, Hstream is making an effort to address this issue using Hadoop.

14. What if process, governance and compliance are requirements for an organization?

Hadoop is not a fit. That’s because data retention is an important criterion for compliance, and with Hadoop, this is a complex and manually intensive management challenge. The bottom line is that organizations cannot afford to delete data before the retention period expires. Conversely, they cannot afford to keep the data beyond the required retention period. Sensage has facilities to manage this automatically.

15. What about organizations with very large data retention requirements and a need to access up to hundreds of petabytes gathered over long retention timeframes?

Hadoop is not a fit. [See answers for questions 12 and 14.]

[5] http://gigaom.com/2009/09/20/getting-closer-to-real-time-with-hadoop/

16. For those organizations that want to program processes for accessing and analyzing data or want to do drill downs across a finite set of data, is Hadoop appropriate?

Generally speaking, yes.

17. What about organizations that need to do in-database analytics, statistical modeling and curve fitting?

Sensage and Hadoop do a good job addressing these needs.

18. What about companies with large datasets that don't have budgetary latitude for on-line storage?

Hadoop is not an optimal solution because on-line storage is more expensive than archival storage. Sensage is the only vendor who can move data from on-line storage to low-cost archival storage automatically while providing transparent access to on-line and archived data in a single query.

About Sensage, Inc.

Sensage®, Inc. helps organizations collect, store, analyze and interpret complex information to identify new threats, improve cyber-security defenses, and achieve industry and regulatory compliance.

Sensage serves our customers' most advanced Security Information and Event Management (SIEM), log management, Call Detail Record (CDR) retention and retrieval and Continuous Controls Monitoring (CCM) use cases. Hundreds of customers worldwide leverage patented Security Intelligence solutions from Sensage to effectively identify, understand and counteract insider threats, advanced persistent threats, cyber threats, fraud and compliance violations.

Combining powerful data warehousing with scalable, clustered multiprocessing and robust analytics, Sensage solutions handle all event data types, scale to petabytes, minimize storage costs and perform sophisticated data analysis. Sensage has achieved Federal Common Criteria and FIPS 140-2 Certification. Sensage partners include Cerner, Cisco, EMC, McAfee and SAP. For more information, visit www.Sensage.com, follow us on Twitter: @Sensage, and watch for us on www.youtube.com/Sensagetv.

SENSAGE

Sensage, Inc.
1400 Bridge Pkwy.
Suite 202
Redwood City
CA 94065
www.sensage.com