

RainStor Big Data Analytics on Hadoop

The Only Enterprise Database Running Natively on Hadoop

- Highest Compression Lowers Operating Cost
- Faster Query & Analysis
- Enterprise-Grade and Simple to Manage

RainStor's Big Data Management database provides proven compliant retention and on demand access of multi-structured data across the enterprise, on any combination of servers and storage, at the lowest possible TCO. RainStor's Big Data Analytics Edition is optimized for organizations leveraging Hadoop's commodity server and storage architecture, and looking to increase the performance and productivity of their analysis, while benefiting from a significant reduction in the number of nodes within their cluster.

RainStor Data Analytics can be deployed on any commercial Hadoop distribution including those provided by Cloudera, Hortonworks and MapR. It requires no change to the type of hardware and storage selected for the Hadoop cluster, nor any modification of Pig and MapReduce functions. Additionally it adds enterprise-grade capabilities such as direct SQL 92 access via standard ODBC/JDBC, as well as features like security, data lifecycle expiry, schema versioning, easy backup, recovery and high availability resulting in a powerful system that is secure, flexible and easy to manage.

The Real Cost of Deploying Hadoop

Data is growing exponentially and traditional enterprise technologies such as relational databases and data warehouses are unable to keep pace with changing business analysis requirements. Apache Hadoop's open source platform has quickly become a preferred technology to cost-effectively collect and analyze volumes of multi-structured data at network speed using commodity servers and low cost storage. The economics of Hadoop compared with traditional technologies are clear: free open-source software, lower priced servers and storage in a scale-as-you go model. While it may appear that Hadoop can be quickly adopted at low initial cost, the ongoing operating cost of a cluster with a significant number of nodes quickly outweighs the original purchase price. Furthermore since Hadoop is in the early phases of enterprise maturity, other additional costs must be factored in to access, integrate and protect data to the standards expected by enterprise IT.

“ *Hadoop gives organizations the ability to scale for Big Data analytics but the data actually grows as it's replicated across nodes. Reducing the size of data slated for retention makes enormous sense. The combination changes the class of hardware and storage required, making the economics even more attractive.* ”

Merv Adrian,
VP Research, Gartner

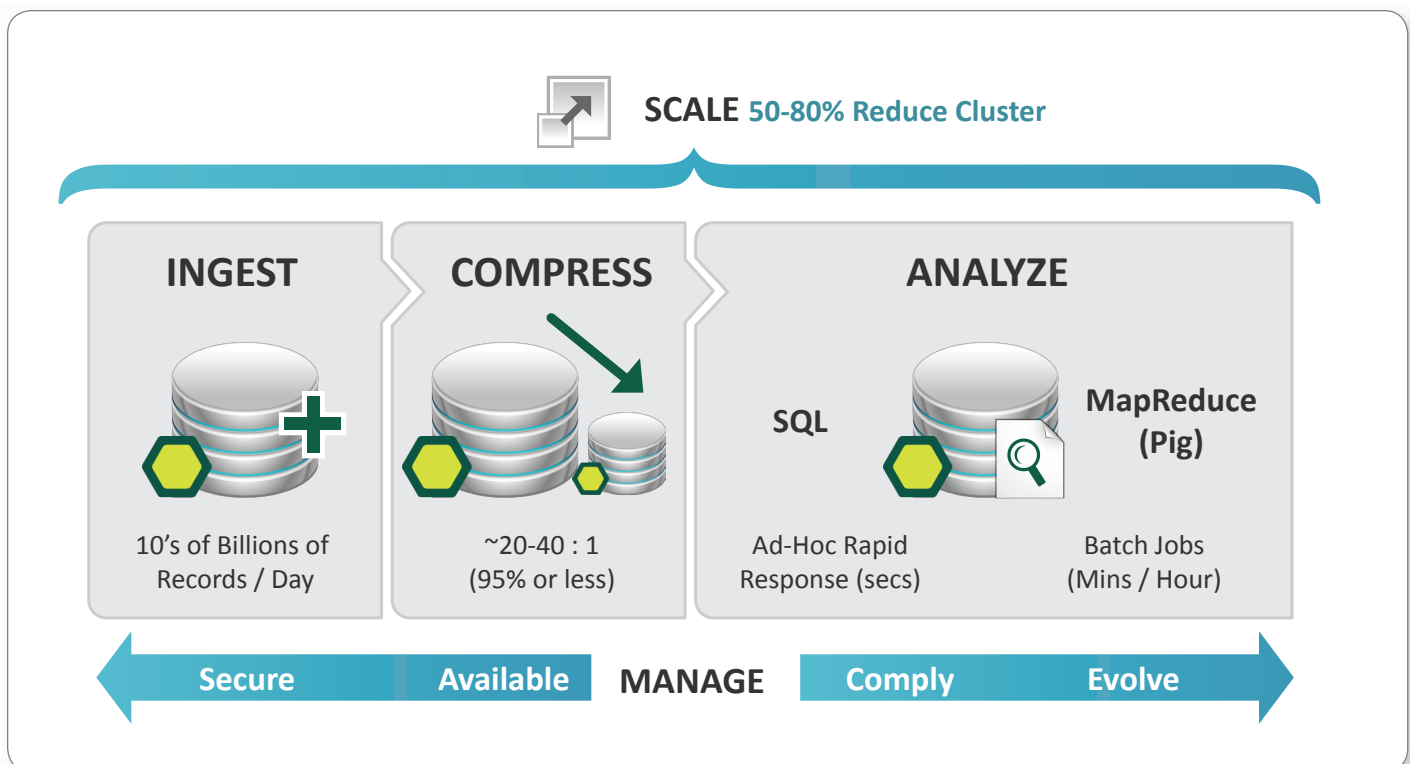
Unique Compression that Reduces Storage and Magnifies Performance

Compression is mandatory for any large Hadoop cluster to keep pace with raw data growth further compounded by Hadoop's replication factor (typically 3x) that protects the ongoing integrity of the cluster. To date LZ0 and Gzip have been the preferred methods for compressing data on Hadoop. Both forms of binary compression can yield rates between 3 to 7 times over raw, but come with numerous drawbacks including performance penalties upon access due to re-inflation. In contrast, enterprise databases and new generations of data warehouses have been able to extend their environments to offer higher rates of compression through new techniques such as columnar storage. However unlike RainStor, none of these run natively on Hadoop. With a large-block, HDFS and MapReduce friendly architecture, it combines the same network ingestion capabilities with patented value and de-duplication upon load, with compression rates as high as 40x over raw. Most importantly, RainStor does not require re-inflation upon access.

Hadoop nodes are typically added to meet growing data storage needs and since each node comes with corresponding CPUs, there could be excess processing power per node that may be underutilized. When data is loaded into RainStor, it is de-duplicated into a compressed file format, which contains considerably more records per block stored on HDFS. This magnifies disk I/O and data transfer bandwidth, minimizing hotspots to improve the efficiency and utilization of CPU cores on each node. So while RainStor compression reduces the number of nodes needed for storage, its efficiencies combined with previously underutilized CPUs can handle the same processing workload with fewer nodes. This means lower purchasing and operating cost of the Hadoop cluster with no compromise.

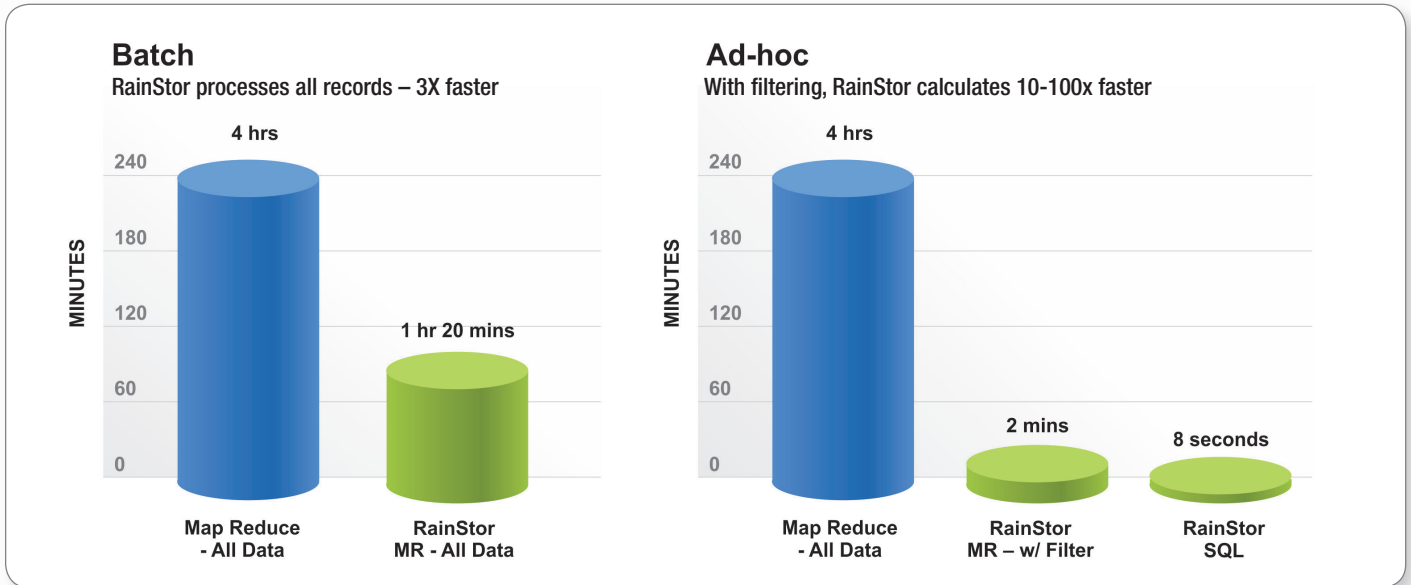
RainStor Big Data Analytics on Hadoop Delivers:

- Multi-Structured Data Management, Native on Hadoop and HDFS
- Ingestion at Network Speed
- Highest Compression (40x) with No Re-Inflation
- Node and Cluster TCO Reduction (50 -80%)
- Access via SQL 92 Query, ODBC/JDBC, Native Pig and MapReduce Analysis
- Eliminate or Reduce Need to Move Data Out of Hadoop
- Faster Access with Compression Magnifier and Intelligent Filtering (10-100X)
- Data Lifecycle Management Including Expiration and Purge
- Schema Versioning
- Simple Enterprise-Grade Security, Compliance, Scalability and Resilience



Built-in SQL Access and Faster MapReduce

Calculate average daily trading price on 1.5 billion stock trades.



Hadoop does not offer native SQL access, which means that existing investments and skills in BI tools such as IBM Cognos, SAP Business Objects and MicroStrategy cannot be used to access data. This has resulted in Hadoop being used to transform and aggregate high volume individual transactions, then transferring results back into the traditional enterprise data warehouse, or using connectors provided by vendors to access small subsets of the data. This can create further complexity and integration cost while propagating more copies of the same data within the enterprise. To solve this conundrum, other open source projects such as Hive (to provide SQL style accessibility) and HBase (to provide database-like features) can be used as well as tools that provide data visualization against data stored in Hadoop.

In contrast, once data is loaded into RainStor it can be accessed using standard SQL 92 and all enterprise BI tools through ODBC/JDBC. RainStor Data Analytics also provides full support for Pig and MapReduce jobs. RainStor's compression becomes a hardware and bandwidth performance multiplier allowing MapReduce jobs to run more than three times faster compared with processing the exact same number of records in Gzip-compressed CSV files on HDFS. RainStor also has built-in filtering which uses metadata gathered upon load to pinpoint and retrieve only the data that will contribute to the result of the MapReduce job or SQL Query. This can result in 10 to 100 times faster processing depending on the exact query type or job being run.

Simple Enterprise-grade Manageability

RainStor also includes built-in security for authentication and authorization to data, immutable guarantees, full auditing of database configuration and data access. It also supports data lifecycle management features such as expiration and purge of datasets, data access across evolving schema versions, as well as fast and easy incremental backup and recovery, geo-replication and high availability. Setting up and deploying RainStor Data Analytics on Hadoop takes minutes and ongoing administration requires no specialized skills and minimal care and feeding.

About RainStor

For more information, visit:
www.rainstor.com

Email us at:
info@rainstor.com

Join the conversation at:
twitter.com/rainstor

HQ:
45 Belden Place, 2nd floor
San Francisco, CA 94104

UK Office:
8 Pullman Court
Great Western Road
Gloucester GL1 3ND
United Kingdom