



# RainStor 4

## Big Data Retention and Retrieval for Compliance and Data Governance across Industry Sectors

RainStor is a specialized repository for the retention of historical data, kept online and therefore accessible for ongoing queries which enable regulatory compliance and high standards of data governance to be achieved. RainStor’s innovative technology solution stores and manages big data volumes in a much more efficient and cost-effective way.

RainStor’s patented technology solution uses sophisticated field and pattern de-duplication techniques to deliver 40:1 compression and higher of RDBMS, XML, and log data. RainStor can also be queried directly using SQL without the need to restore or re-inflate the data. The data within RainStor is stored in containers that can be easily managed using standard file systems and storage platforms which requires very little resource bandwidth to set-up and maintain, further reducing overall total cost of ownership.

The combination of RainStor’s Massively Parallel Processing (MPP) architecture, and support for commodity hardware and the cloud, delivers low-cost scalability. RainStor has been architected to ingest extremely large data volumes which uniquely sets it apart from traditional RDBMS's and is therefore well suited to a number of industries that experience extreme data growth.

Furthermore, RainStor has been architected to require no design, index creation, tuning, or ongoing maintenance. In summary, RainStor enables organizations to preserve historical data online with minimal effort, delivering substantial storage, hardware, and operational savings.

### RainStor 4 - Key Technology Capabilities

---

**Reduce** – Extreme compression using patented value and pattern de-duplication algorithms

---

**Retain** – Configurable rules for compliance and retention including automatic purge

---

**Retrieve** – Perform queries using SQL 92 or standard BI tools via ODBC/JDBC

---

**Ingestion** – at massive big data volumes in the tens of billions of records per day

---

**Flexible** - Supports all structured data sources – enterprise-wide data retention platform

---

**Low Administration** – No specialist skills needed. Very low maintenance

---

**Cloud or On-premise** – Runs on any hardware environment and cloud platforms

---

## Rapid Data Growth

Today's data management budgets are measured in millions of dollars, and with data growing at between 40% and 100% each year, there is a clear need to control these spiraling costs. Historical data that is retained for business, legal, or regulatory purposes is rapidly becoming a key area of focus in reducing storage, hardware, and operational costs.

## Storage Costs are Still High

Although the cost of raw disk storage is falling, operational factors continue to drive up the cost of the database terabyte. Consider mirroring for high availability, replication for disaster recovery, disk dedicated to snapshots, tape for backups; all compounded by cloning for development and test purposes. With these factors in play, it is not unusual for one database terabyte to map on to at least six terabytes of disk capacity.

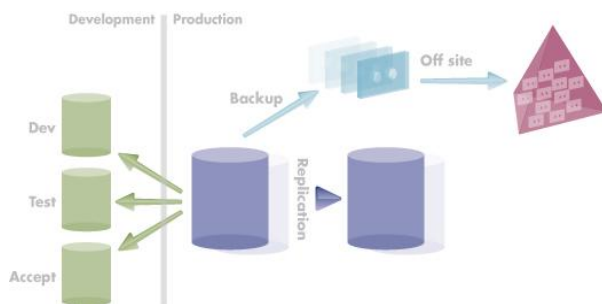


Figure 1: Database Data Proliferation

## Not all Data is Equal

Databases and data warehouses that have been in production for a number of years not only store active or 'hot' data, but are now

increasingly weighed down by inactive or 'cold' data that is rarely, if ever, referenced.

The key to managing both active and inactive data efficiently is to match the solution to the data class. While relational databases are ideally suited for managing active transactional data, and data warehouses are designed for high performance analytics, the cost-effective management of historical data demands a new class of database.

## Specialized Hardware

Many traditional databases leverage SMP architectures that require expensive servers, while many data warehouses are delivered as appliances using specialized hardware.

Neither of these solutions enables organizations to leverage commodity hardware or the cloud to retain and retrieve historical data at the lowest possible cost while satisfying scalability requirements.

## RainStor

The RainStor solution enables simple and cost-effective retention of historical structured data, such as inactive or static data from operational databases and data warehouses, or ever increasing volumes of data being generated by logs and security events from networks.

### RainStor – How It Works

Historical data is landed in a staging area in XML, BCP, or delimited formats and then automatically imported into RainStor. The historical data could be inactive data that has been removed from a production database using native export utilities, ILM, or ETL tools,



or log data that has been captured from a network. The import process can be performed on a one-off, ad hoc, or scheduled basis and can be configured according to business needs.

With a copy of the historical data now stored and immediately accessible in RainStor, the original records are removed from the staging area.

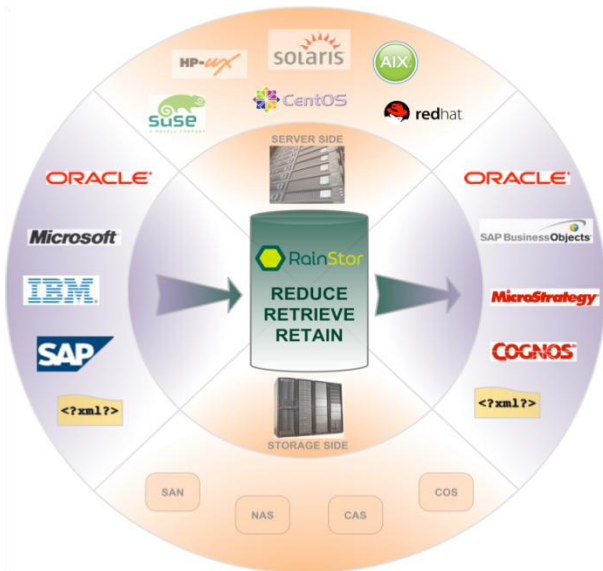


Figure 2: RainStor – Enterprise Flexibility & Database Agnostic

The data stored in RainStor typically benefits from 40:1 compression and remains online and instantly accessible. All the data, or subsets of the data, can also be exported in XML or delimited formats to facilitate import into other data management solutions.

### Data Compression and Reduction

The data repository automatically generated by RainStor is a “pattern” database. By only storing the unique field or pattern values contained within each record that is imported, RainStor de-duplicates the data, resulting in extreme compression. This can be optionally

boosted on a table-by-table basis by selecting additional byte-level compression.

### Accessing Data

The data in RainStor, while highly compressed and reduced in size, remains directly accessible using standard SQL over ODBC or JDBC through a range of BI or reporting tools. RainStor supports analysis using Business Objects, Crystal Reports, COGNOS, and Microstrategy.

Data within the repository is automatically partitioned as part of the import process and stored as a series of data files. RainStor’s optimizer automatically exploits this partitioning to efficiently process queries; using partition elimination, column elimination, and parallelization optimizations.

### Database Support

RainStor supports structured data from any RDBMS, XML, or log file source.

RainStor’s support for Oracle, MS SQL Server, and Sybase is further extended to include coverage for those platforms’ proprietary SQL language extensions. This additional support ensures that queries can be ported from other environments to RainStor without the need to rewrite into standard database-neutral SQL.

### Managing Schema Change

RainStor supports schema change without impacting the integrity of previously stored data, even when tables or columns are dropped or data types are changed. This is achieved using versioned table headers which project new schema definitions onto historical data to apply changes virtually and ensure consistency. RainStor supports native database DDL for definition of new schemas



as data is added to the repository incrementally over time.

### **Cross-schema Query**

RainStor's versioned approach to schema change ensures that queries written against the current schema are able to access data stored under multiple historical schemas in a consistent manner.

### **Point-in-time History View**

In addition, RainStor supports export and query of data as at a selected point in time. By setting a time context, users are able to limit their logical view of data to include an exact view of what would have been visible had they run the query at a specific point in time. Tables or columns created since then are masked from that view, while tables or columns dropped since then remain visible.

### **Assuring Information Integrity**

RainStor ensures the integrity of data by maintaining audit logs that track every action and event; from import, through data retrieval, and finally data disposition. The data files themselves are written once and never updated. MD5 fingerprints are generated at write time and recorded in the audit logs to enable tamper detection. RainStor also integrates with CAS storage devices for enhanced protection and security if required.

### **Full Control Enables Compliance**

RainStor's expiry of data from the repository is based on retention policies that also support legal hold. On expiry, qualifying data may be electronically shredded and removed from disk. Additionally, a unique record can be expired at a specific time or extended for

expiry at a future date. The ability to tag a group of records for future ease of search and adding history to specific records all enable a greater level of control over historical data. This allows the end-user to retain or purge data at the record level in order to meet specific compliance regulations.

### **Low-cost Scalability**

As data volumes within the repository grow, the underlying hardware and storage platforms can be easily extended to meet additional demand.

By leveraging an MPP architecture and shared-nothing approach at the server layer, RainStor scales across servers and cores during data import and query. RainStor has also been designed to run on commodity hardware, including the cloud, supporting a variety of chip architectures, to deliver low-cost scalability without the need for specialized hardware or skills. In addition, the availability of the RainStor virtual software appliance supports a vast array of operating systems and platform requirements.

Due to the extreme data compression delivered by RainStor, the underlying file system has been virtualized, supporting both distributed and shared file systems without being I/O bound. RainStor supports SAN, NAS, DAS, CAS, or cloud storage platforms, and additional storage can be provisioned on demand using existing management software.

### **Data Ingestion**

As data volumes grow, the amount of data that needs to be ingested on a regular basis by RainStor also grows. In the telecommunications sector typical daily call data record (CDR) and SMS text volumes of



one billion and up are not uncommon. Unlike traditional RDBMS's, RainStor can ingest very large volumes of data very quickly.

### Summary

Companies of all sizes must retain the growing volumes of data for longer time periods in order to satisfy business and regulatory requirements. This "Big Data" problem is often associated with unstructured information such as emails, documents, images, and video. However, most business-critical information is stored within production databases, data warehouses, and log files as structured data that is complex to manage and access.

Keeping this critical data in production systems is costly and an inefficient use of expensive infrastructure resources. RainStor provides a new class of data repository with the sole purpose of reducing the data footprint, keeping it online for access and query and doing so with significantly lower costs.

## Supported Platforms

<b>RainStor runs on:</b>
Red Hat Enterprise Linux 4 64-bit
Red Hat Enterprise Linux 5 64-bit
CentOS 4.5
CentOS 5.3
Solaris 10 for SPARC
IBM AIX
Windows x86

<b>Data sources support:</b>
All major RDBMS (including Oracle, MS SQL Server, Sybase, IBM DB2, MySQL)
Logs, Security Events, CDR, IPDR
XML, BCP and delimited files

<b>Native SQL extension support:</b>
Oracle Enterprise 9.2 and above
Sybase ASE 12.5 and above
MS SQL Server 2005 and above

<b>BI platform support:</b>
Business Objects XI
Crystal Reports XI
COGNOS 8
Microstrategy 8



## RainStor Features and Benefits

<b>RainStor Feature</b>	<b>Benefit</b>
<b>Storage Efficiency</b>	
Data compression	Data compressed by 40:1 delivering >97% storage savings
<b>Information Access</b>	
Query in place	Query data in-place without need to recover or re-inflate
Index free	Repository generated automatically with no design or maintenance
<b>Database Support</b>	
Native data type support	Preservation of data formats for Oracle, MS SQL Server, and Sybase
Proprietary SQL extensions	Run any query against the repository without rewrites
<b>Change Management</b>	
Schema change preservation	Transparent SQL query across historical schemas with no overhead
Point-in-time query	Retrieve data from the repository as it existed at any point in time
<b>Integrity</b>	
Read-only	Lock down historical data with immutable data model
Audit trail	Conserve chain of custody information from import through expiry
Tamper detection	Use MD5 fingerprinting and checks to insure against file tampering
<b>Open Standards</b>	
Simple import/export	Standard import/export formats, including delimited and XML
Extensive SQL support	Support for standard SQL-92, native Oracle, SQL Server, and Sybase
ODBC & JDBC connectivity	Integration with industry-standard BI and reporting tools
<b>Scalability</b>	
MPP server architecture	Low-cost and linear scalability across commodity servers and cores
Virtualized file system	Manage data on SAN, NAS, DAS, CAS, or cloud storage
Backup & recovery	Exploit industry-standard solutions for backup & recovery

For more information about RainStor visit [www.rainstor.com](http://www.rainstor.com) or email [info@rainstor.com](mailto:info@rainstor.com)