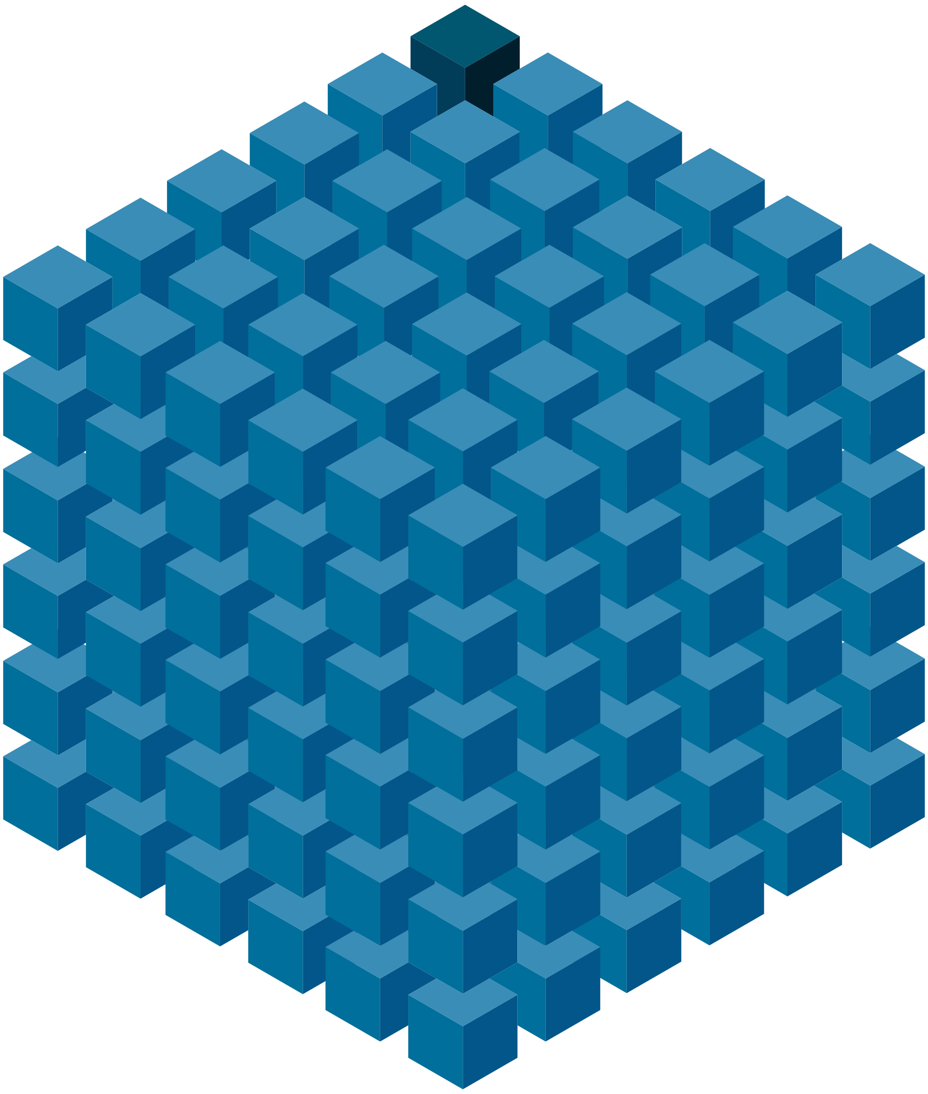
White paper | April 2016

Duplicated data
in the enterprise space



Executive summary

The impact of duplicate data cannot be underestimated. From excessive running costs to potentially fatal mistakes, duplicate or incorrect information cannot be tolerated in the modern data-driven enterprise.

For some organisations the cloud may seem like a lifeline when it comes to coping with accelerating data growth and the need for more storage. But if the underlying issues causing data duplication are not addressed, the problem is simply replicated into the hosted environment where it will continue unchecked.

The data-driven enterprise needs to address the issue of duplicate data now or jeopardise any strategic advantage they may otherwise expect to realise from their Big Data and IoT programs.

Contents

Accelerating data growth is compounding the duplicate data problem

The issue of corporate data growth is not new, but CTOs still struggle to keep up. But demand continues to accelerate, with global data storage needs expected to top 40 Zettabytes by 2020[1], a 14% increase on previous estimates[2] . The only thing that analysts can agree on is that the total volume will continue to at least double every two years for the foreseeable future.

Digital data created in 2020 forecasted at

35

zettabytes; cloud computing will manage data growth – TechTarget

The rise of Big Data disciplines and strategies has contributed to growth over the past few years, as businesses begin to realise the potential benefits of analysing large, unstructured datasets. This has led to the creation of 'data lakes' to consolidate data for analysis.

Following directly on from Big Data is the emergence of practical applications of the Internet of Things (IoT). IoT sensors have the potential to further accelerate data growth, feeding into the corporate data lake to provide context for other Big Data-enabled activities.

In the midst of all this data creation however, many organisations have downgraded the importance of de-duplicating data. And the eventual effect of this lax approach could be significant.

The impact

As the cost of storage has continued to fall, CTOs have found it easier to expand capacity to accommodate growth. However, falling costs coupled with a new focus on unstructured data means that traditional data management routines are being ignored in favour of simply adding more disks to cope.

The impact of duplicate data varies from organisation to organisation. Every business that continues to accumulate business is paying for unnecessary storage - up to 40% of storage is wasted on duplicate data[3]. For certain sectors however, the potential impact is far greater.

The question of speed is one that the enterprise continues to wrestle with as they roll out Big Data strategies. The insights generated by analysis is valuable, but being able to deliver results in real time is invaluable. The larger the unstructured data lake is though, the longer it takes to mine for actionable insights. And every duplicate item of data adds to the time required to complete an operation, losing any time-related advantage businesses may hope to realise.

In effect, businesses with a high proportion of duplicate data are crippling their Big Data programs from the outset.

When bad data kills

According to the American Health Information Management Association (AHIMA), US hospitals have a duplicate patient record incidence rate of 10%[4]. This not only places systems under additional strain, but threatens clinical outcomes for patients – with potentially fatal results.

One industry expert, Victoria Wheatley of medical software supplier QuadraMed, suggests that duplicate database records cost healthcare providers between $5 and $10, based on clerical time alone[5]. A 2011 study based on data stored by the Children's Medical Center of Dallas placed the figure closer to $96 per record[6].

But should one of those duplicate records result in a clinical error, the total cost will run into many thousands of dollars – repeated tests and treatment delays caused by problems verifying the correct patient record average around $1100 each[7]. A sobering thought when considering an average 10% rate of duplication.

Clearly not every industry endangers lives with duplicate records, but the impact is potentially huge regardless. Clearly ignoring the problem and adding capacity to delay data cleansing operations is not an option.

The cloud offers no quick fixes

Cloud computing has brought new levels of cost-effective flexibility to organisations of all sizes, allowing them to scale resources depending on demand for greater cost management. And with the promise of 'infinite' capacity, many businesses are using cloud storage to expand capacity, rather than purchase more onsite arrays.

> ***"We are being sold this ad hoc world, where we are told we can throw everything in this one bucket and not to worry it will find itself. But it does not actually work. We have stuff sitting on our servers, but we do not actually have a proper record of it, we do not know how to trace it and we do not really know how it got there in some cases. We have lost the audit trail."***
>
> **Alan Pelz-Sharpe, research director at 451 Research[8].**

Infinite storage breeds bad habits

On the face of it, expanding into the cloud makes perfect sense. Why make capital purchases when a supplier can pick up the bill instead? But the infinite storage model could actually be making things in worse in the case of duplicate data.

The ability to commission additional storage space automatically means that there are no longer any 'check' in the system that would normally require some kind of technical review when limits are reached. This could lead to a significant increase in data duplication, simply because little is done to manage it.

Businesses may no longer be bearing the capital cost of storing duplicate data onsite, but the utility billing model of cloud services means that they are still paying for it. After all, the more resources consumed regardless of its validity, the more the service user is billed.

Consumerized computing increases data sprawl

The ease with which cloud services can be bought and configured presents another problem for the CTO. The IT department is increasingly marginalised when it comes to making software purchasing decisions. 86% of cloud applications in use across the enterprise have never been officially sanctioned[9], creating all manner of problems for the CIO and CTO who are tasked with integrating such services after the purchase has completed.

86%
of cloud applications in use across the enterprise have never been officially sanctioned

Without back-end integration, these unsanctioned cloud applications exist as isolated silos. Not only is the data stored in each inaccessible for use in other operations, like Big Data analysis, but they also help to propagate duplicate data by their very existence.

Even forward thinking organisations that have implemented automated file deduplication appliances cannot properly manage information spread across multiple hosted platforms. In the silo scenario, there is no way to avoid the need for multiple copies of information – the issue is actually a direct result of poor planning.

What are the options when dealing with duplicate data?

The longer duplicate data is left unchecked, the worse the problem becomes. There will never be a better time than the present to work on data deduplication – but what are the available options?

In-house data deduplication

Staffed by talented engineers, most enterprises have the basic skills in-house to develop a routine capable of comparing and removing duplicate files. In many cases initial efforts will be relatively successful, if costly. Research performed by the Rand Institute for Civil Justice found that reviewing corporate data costs an average of $18,000 per gigabyte[10].

Deduplication is an ongoing process however, and any tools developed in-house to assist with the task need to be updated as the corporate infrastructure develops and changes. Each revision adds significantly to the cost of performing deduplication, reducing the savings realised from cleaning up datastores.

Automated, in-line data deduplication

Deduplication appliances like EMC's Data Domain are designed to capture and manage duplicated files at the point of backup. Manufacturer specifications suggest that such appliances are able to reduce the amount of required archive space by up to 30x[11].

An appliance provides an ongoing method of reducing data duplication, but it also represents a major capital investment in terms of hardware purchase and ongoing maintenance. And as operating data capacity grows, appliances will also need to be upgraded – according to the OEM's service and warranty roadmap.

It is also worth noting that these appliances are only capable of deduplicating data passing through them (in-line processing), doing nothing to solve the problem of disparate cloud-based datasets.

3rd party deduplication service

Outsourcing data deduplication to a reputable service provider offers several benefits. Involving a third party allows businesses to benefit from expert experience and ensures their own engineers are not taken away from day-to-day duties.

In many cases, the use of a third party deduplication service involves no additional capital spend. This model tends to be best suited to returning data to a 'clean' baseline state, in readiness for the deployment of an automated deduplication appliance or software solution, or in preparation for moving key systems into the cloud.

In the event that an organisation is able to migrate data back from disparate SaaS services, deduplication and record updates will be essential to ensure data can be reintegrated into the corporate data lake.

Conclusion

The drive to acquire and process increasing amounts of data means that businesses will need to tackle the issue of duplicate data now. Continuing to postpone data cleansing activities will result in a technical challenge of vast complexity and cost that few organisations are prepared for.

The value of corporate data – particularly for Big Data analytics – lies in its accuracy. But if between 10% and 40% of the average corporate data lake consists of duplicate data, any insights drawn from that information is likely to be biased, or downright inaccurate.

Reducing the amount of duplicate data held will be a crucial step towards building a Big Data program that creates a genuine strategic advantage.

> ***"Ultimately, poor data quality is like dirt on the windshield. You may be able to drive for a long time with slowly degrading vision, but at some point, you either have to stop and clear the windshield or risk everything."***
>
> **Ken Orr – The Good, The Bad, and The Data Quality[12].**

A de-duplicated data set also consumes far less of your valuable capacity, allowing you to get additional value from your existing hardware assets.

To learn more about extending the usable lifespan of your storage, and to reduce your support and maintenance costs by 40% or more, please **contact CDS**.

References

1. Data to grow more quickly says IDC's Digital Universe study – Computer Weekly
http://www.computerweekly.com/news/2240174381/Data-to-grow-more-quickly-says-IDCs-Digital-Universe-study

2. Digital data created in 2020 forecasted at 35 zettabytes; cloud computing will manage data growth – TechTarget
http://searchstorage.techtarget.com/news/1511342/Digital-data-created-in-2020-forecasted-at-35-zettabytes-cloud-computing-will-manage-data-growth

3. Study Finds Improved Copy Data Management Could Save Federal Government $16.5 Billion Wasted On Redundant Data – MeriTalk
https://www.meritalk.com/wp-content/uploads/2015/12/MeriTalk_Consolidation_Aggravation_Press_Release.pdf

4. How To Measure Duplicate Rates – For The Record Vol. 25 No. 7 P. 18
http://www.fortherecordmag.com/archives/0413bonusp18.shtml

5. Ibid.

6. The Risk of Duplicate Patient Records – Gallagher Healthcare Practice
http://engorgement/media/72413/Duplicate-Patient-Records_healthcare.pdf

7. Ibid.

8. Analyst highlights data management problems posed by the cloud – CloudPro
http://www.cloudpro.co.uk/saas/5189/cloud-storage-exacerbates-data-duplication-claims-451-analyst

9. Cloud Adoption and Risk Report – CipherCloud research
http://pages.ciphercloud.com/Cloud-Adoption-and-Risk-Report-landing-page.html

10. Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery - Rand Institute for Civil Justice
http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf

11. EMC Data Domain DD9500 product specifications
https://store.emc.com/gb/Product-Family/DATA-DOMAIN-PRODUCTS/EMC-Data-Domain-DD9500/p/DDM-DataDomain-DD9500?PID=EMC_PRD-DD9500-34FF_HEAD

12. The Good, The Bad, And The Data Quality – Ken Orr
https://www.cutter.com/article/good-bad-and-data-quality-430461

Multi-vendor | Multi-Platform | Multi-System | Single Point of Contact

EMC® NetApp® HDS® IBM® StorageTek® Cisco® Sun Microsystems® Dell®

Global Headquarters

Computer Data Source, Inc.
275 Industrial Way West
Eatontown
NJ 07724
USA

+1 732 542 7300
Toll Free: +1 866 237 8008
Fax: +1 732 542 7397
Asset Recovery: +1 732 542 7300
Email: sales@cds.net

DACH

Computer Data Source, Inc.
Deutschland GmbH
Platz der Einheit 1
60327 Frankfurt am Main
Deutschland

+49 69 975 39725
Asset Recovery: +1 732 542 7300
Email: salesemea@cds.net

Canada

Computer Data Source
Canada, Corp.
3780 14th Avenue, Unit 106
Markham, Ontario L3R 9Y5
Canada

+1 905 474 2100
Fax: +1 905 474 2101
Asset Recovery: +1 732 542 7300
Email: salescanada@cds.net

EMEA

Computer Data Source
Europe, Ltd.
Rawdon House, Bond Close
Basingstoke, RG24 8PZ
United Kingdom

+44 1256 362 983
Fax: +44 1256 476 969
Asset Recovery: +1 732 542 7300
Email: salesemea@cds.net

APJ

Computer Data Source Pty Ltd
685 Burke Road, Suite 208
Camberwell, Melbourne
Victoria 3124
Australia

+61-3-9006-1720
Email: jmoshovelis@cds.net