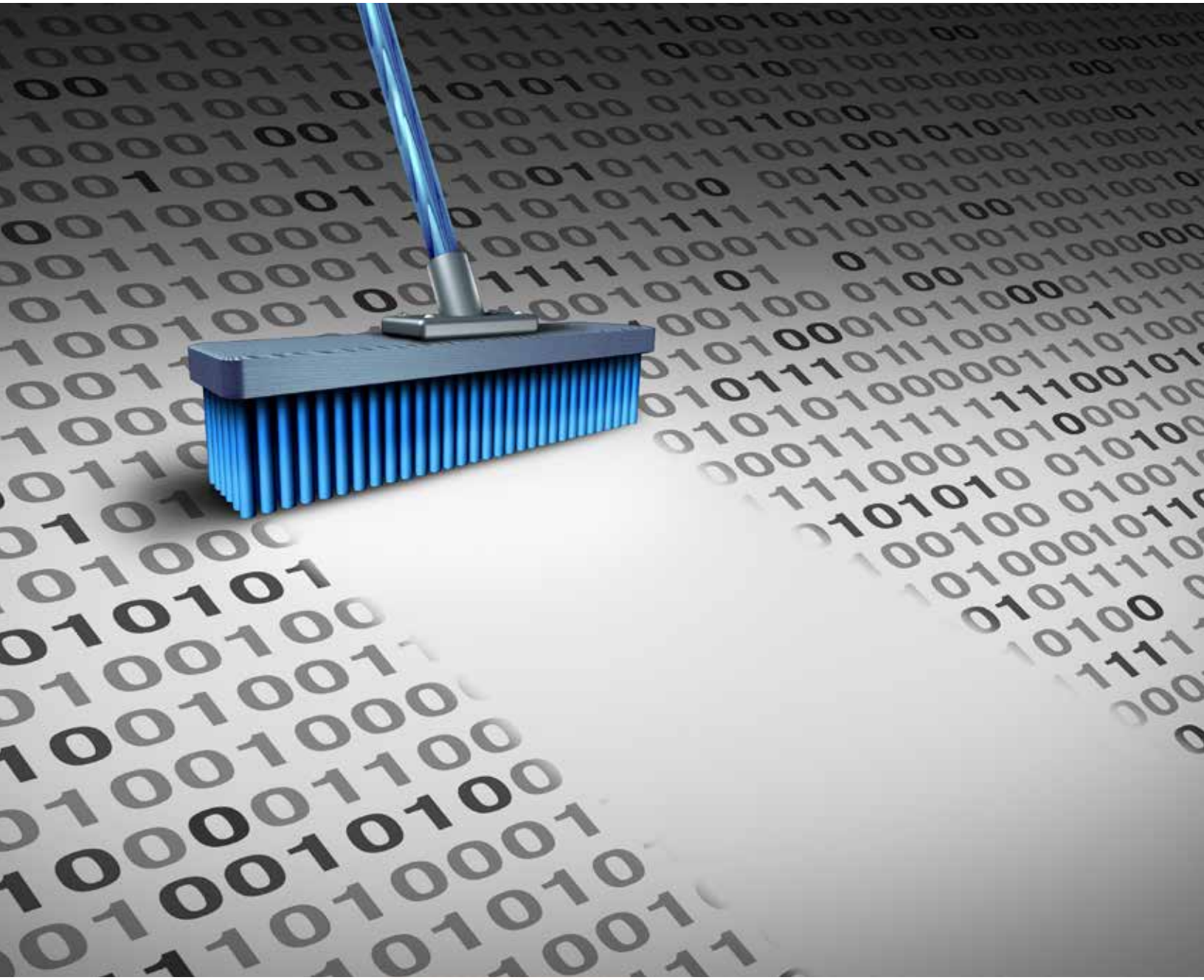
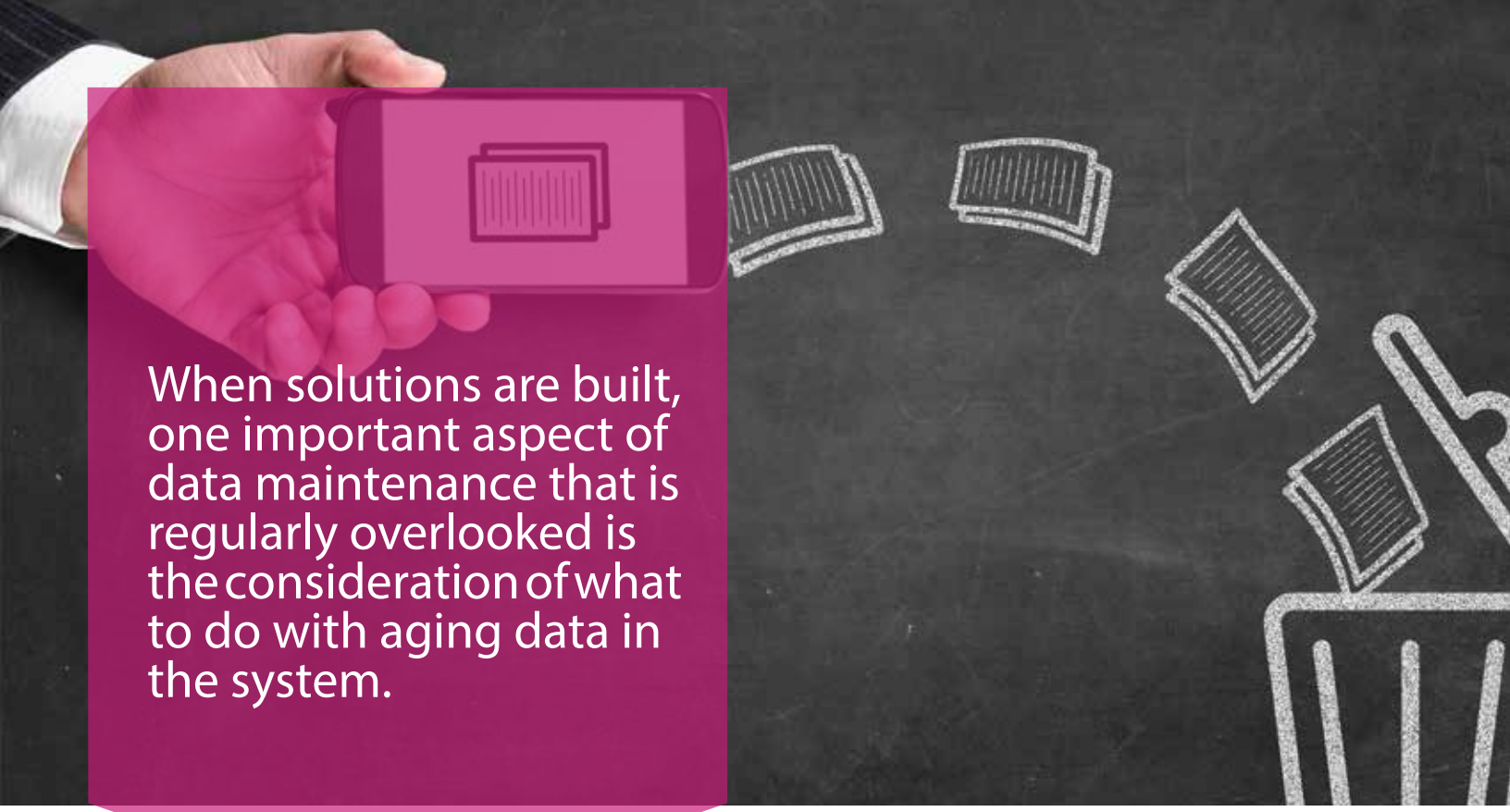
Mphasis
The Next Applied



The Art of Deleting Data (from Salesforce)

Whitepaper by
Martin Prosser
Principal Customer Consultant



When solutions are built, one important aspect of data maintenance that is regularly overlooked is the consideration of what to do with aging data in the system.

Regulations have a large influence over what information needs to be retained and for how long. Pension providers need to keep a full audit of all instructions for the lifetime of the pension (easily 60+ years). Conversely, there are many laws and directives across the globe dictating that personal confidential information should only be kept for as long as is necessary to perform the purpose for what it was originally collected.
For example: The USA's 2012 snappily titled publication of "CONSUMER DATA PRIVACY IN A NETWORKED WORLD: A FRAMEWORK FOR PROTECTING PRIVACY AND PROMOTING INNOVATION IN THE GLOBAL DIGITAL ECONOMY" [ref 7] and the European Union's equally snappily titled "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data" [ref 8].

But what about data that may not have any regulatory guidance as to how long it must be retained? For example, how long should the full details of a customer complaint be retained? If a customer leaves, does their customer history still hold any value? But, what if the customer comes back after a year or ten years? Does their old customer history still hold true? After leaving, the customer may have changed address, changed their name and possibly changed their mobile number. Would we still be able to correctly associate them with their previous customer history? Some of this could be deemed personal data that can be used to identify the individual, in which case it will need to be deleted after a period of time. How long is that time period and how much value would the remaining customer data hold? Similarly, for how long should a CRM system store details of contacts from customer/supplier companies, and keep records of interactions with those people (e.g. meeting invites, notes from telephone calls, etc.)?

For the data we can retain, these are just some of the many questions that businesses typically ignore because data storage is cheap and keeps on getting cheaper. If data storage was free and application performance time was never a problem, and data maintenance efforts never increased, then there would be an argument to keep all the data because "it might come in handy one day". Even with cheap storage, the more data you have, the longer the systems will take to access that data. Also, when you are using cloud-based solutions (e.g. Salesforce), there may be storage limits imposed upon you. Perhaps that data storage isn't so cheap after all.

An option for the Hoarder

Many commentators routinely quote about data (especially Big Data) that 90% of the world's data was created in the last two years. Marc Benioff (Salesforce CEO) used the phrase at Dreamforce 2014. The phrase is regularly attributed to IBM, dating back to an IBM Watson paper in 2012 [Ref 6]. Storing all the data is great if you've got somewhere to put it (e.g. Google, Facebook, and Salesforce – to enable Einstein to do its stuff), but most of us either don't have the necessary deep pockets or, in the case of Salesforce environments, the space. For most Salesforce licenses, data storage has a finite capacity. If you want to stay within the Salesforce limits and not spend more money on Salesforce data storage, removal of data that has little or no utility becomes essential.

A mechanism to help adhere to Salesforce data limits, yet still be able to access the archived data, is to store it in an accessible location (e.g. Amazon Simple Storage Service (S3), Google Cloud Storage, Microsoft Azure or Heroku/Postgres) and use the Salesforce External Objects capability to view/access the external/archived data from within Salesforce. This option provides the comfort factor of being able to access the beloved data while not impacting the Salesforce data limits. There will still be a need for a defined set of criteria when the data accessed in this manner is either finally transferred to cold storage (e.g. Amazon Glacier, Google Cloud Storage Nearline, etc.), or deleted forever.

De-clutter Drivers

The Collins dictionary definition of utility is: "the quality of practical use; usefulness" [Ref 11]. In economics, utility is also used as a measure of the preferences people make between some sets of goods and/or services. The goods/services provide benefits in terms of happiness or satisfaction, which are not easily measurable.

In Economics there is a recognized law of diminishing marginal utility [ref 12], where repeated consumption of units of a single commodity reduces the total utility value for that commodity. There are parallels between the attempts of economics to measure the utility of goods/services and the attempts to measure the utility of data. For data, is the accumulation of more and more of it increasing its utility or actually diluting it? The simple answer is to keep all of it and move the stuff into long-term archival storage. This still has an associated cost, which will keep increasing if there isn't an archive delete policy.

Data archiving is different from merely taking full backups. Archival data needs to be stored in a manner that can be subsequently searched for specific records and/or files. Backup storage tends to be stored in a more compressed format and only used when wanting to completely restore all the backed-up data.

So, before thinking about what data should be archived and/or removed, let us first examine the behavior of data storage within Salesforce.

The measurement of storage within Salesforce is divided into two categories: File Storage and Data Storage. File storage includes files in attachments, Salesforce CRM Content, Chatter files (including user photos), the Documents tab, the custom File field on Knowledge articles, and Site.com assets. Data storage includes over twenty of the common objects, plus all custom objects.

The clinical details of the storage constraints can be found at the following Salesforce pages.

https://help.salesforce.com/HTViewHelpDoc?id=limits_storage_allocation.htm

These details can be summarized as follows -

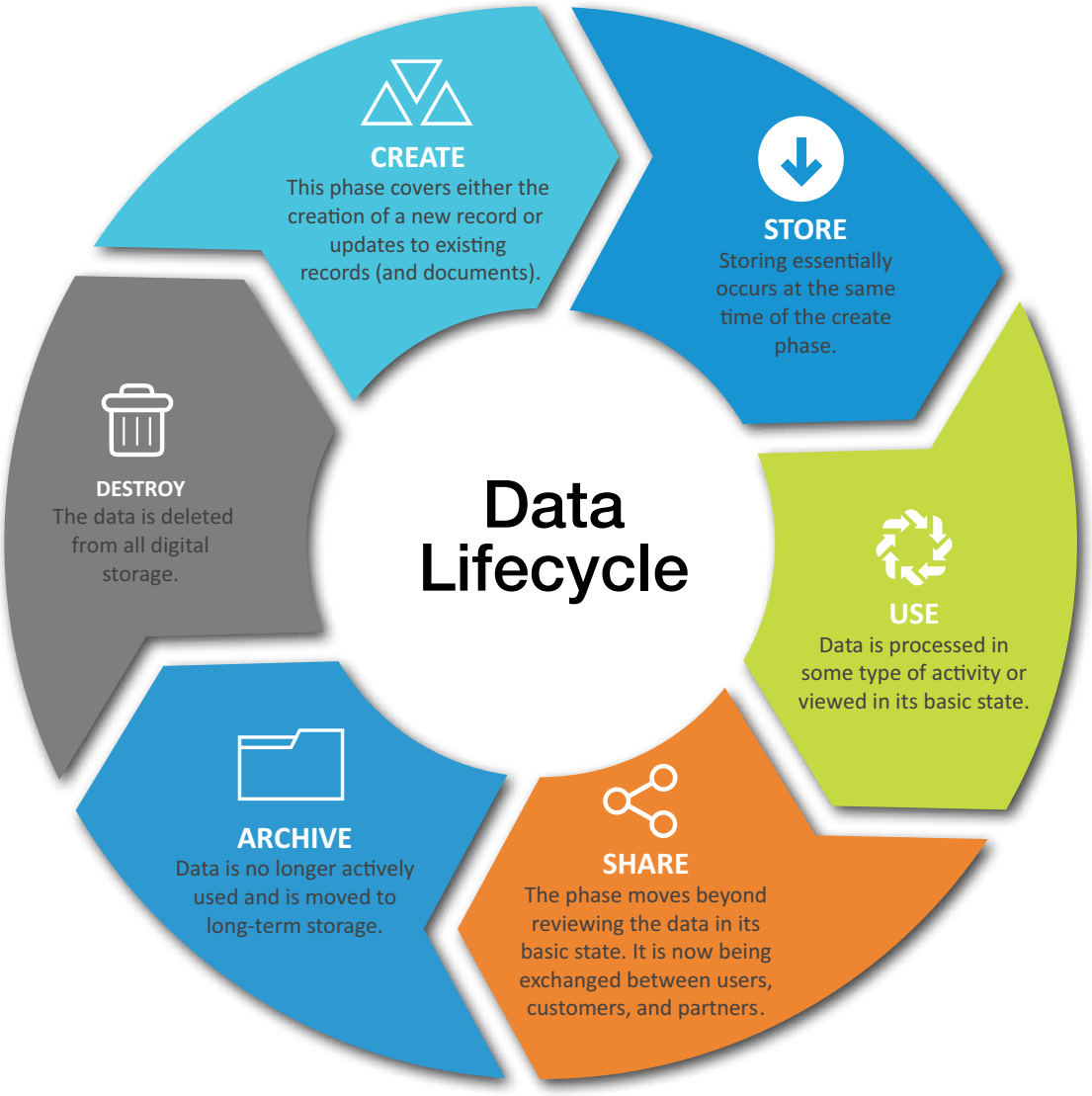
- For data storage, the following license types allocate a maximum storage, which is calculated as the greater of either a fixed 1GB, or 20MB per user-license - manager, group, professional, enterprise, performance, and unlimited editions.
- For file storage, contact manager, group, professional, enterprise, performance, and unlimited editions are allocated 10GB of file storage per org.
- Orgs are allocated additional file storage based on the number of standard user licenses. In enterprise, performance, and unlimited editions, orgs are allocated 2GB of file storage per user-license.
- The Contact Manager, Group and Professional edition orgs are allocated 612MB per standard user license. This total storage is split to include 100MB per user license and 512MB per license for the salesforce crm content feature license. The additional 512MB file storage is allocated irrespective of whether Salesforce CRM content is or is not enabled. An org with fewer than ten users is permitted to use a fixed total of 1GB per user file storage.

- Irrespective of the amount of data stored within individual records, to calculate data storage, Salesforce uses generic record lengths. There are some exceptions, but typically each object is assumed to consume 2KB of storage space per record [ref 1]. The main record types (accounts, contacts, leads, opportunities, cases, tasks, events and custom objects) all follow the 2KB rule.

Therefore, using the above data storage limits, a system with 51 or fewer users can store about half a million records before encountering data limits (this is roughly where the 1GB limit is exceeded by the 20MB per user limit). For systems with more than 51 users, the total record capacity constraint becomes about 10,000 records per user.

Data Lifecycle Management

The book The Official (ISC) Guide to the CCSP CBK [ref 10] identifies six stages of data in the secure cloud (which is also applicable to data in general) as being: Create, Store, Use, Share, Archive and Destroy. Their order and description is below.



Based on the above definitions, the usefulness duration of data can be viewed as an output of its lifecycle. After the record has been created, it enters an operational/useful state. The length of time for which the record is useful depends on many factors, including length of any associated processes and factors such as the frequency of similar record creation. High volume record creation usually means the lifetime of an individual record is limited before it is lost in the noise of all the other records (e.g. a social media or a chatter post). Conversely, where particular types of records are rarely created, their relevance tends to be longer lasting (e.g. currency codes). Once the record is no longer contributing to ongoing business operations, it moves into the reference state. It is still used in reporting or customer enquiries (e.g. a sale that has been paid, dispatched and delivered). After a period of time the record is no longer used and can be removed from the primary system, but kept for regulatory and/or legal purposes. This is where it moves into the archive state. Eventually (scores of years in some cases), the record is no longer required and can be deleted/destroyed.

With cloud models, compared to more traditional application solutions, the movement of data to an archive location requires some form of API access to the cloud application instead of direct access to the database layer. The use of the API is effectively another interface required to perform the data archive phase, and possibly the destroy phase capability. These last two phases of the data lifecycle tend to be overlooked in any (cloud) application solution. The interfaces for archive and deletion are usually not included in any business process or user story.

Now the difficult bit

The utility of data depends on the purpose or role the data will play within a business operation. Essentially we are asking, what is the reason for its creation and why do we need to keep it? There are two competing forces we are trying to balance, which are the diminishing returns we obtain from data the longer we keep it versus the quantity of data we are keeping and the associated costs of keeping and maintaining that data.

We have an understanding of how Salesforce calculates its permitted data storage and have examined the phases data typically takes during its journey from creation to destruction. We turn our attention to the difficult task of identifying the conditions for each type of data that define how long the data must exist, what data is nice to have and can stay around for a while, and what data has no longer any utility and has become nothing more than clutter. From this, a data retention (and removal) policy should be created. Deleting data according a defined policy is permissible as recognized by the USA Government. This was confirmed in 2005 during the Enron scandal when Andersen successfully overturned a conviction where they had instructed employees to destroy data as part of the document retention policy [ref 9].

Mphasis has been involved in successful partnerships with clients where such policies are created as part of the implementation of new systems. We delivered a CRM solution to a global recruitment firm who generates nearly 30 million activity records per month. As part of the delivery setup, they also included automated routines to delete the records that met the deletion criteria. They ensured their data would effectively plateau at a given size, and not grow any further.

There are six steps a company should follow in order to follow the path to enlightened data storage and the all-important data retention policy.

1. Determine where all the data is held. The IT department will have those answers. We are primarily looking at salesforce, so we already have a good idea where the data of interest is held.

2. Confirm the size of the data. Again, the IT department can quickly confirm how much data is held in salesforce and how close to the current limits they are.

3. Bring together the various business data owners and legal team to explain the content of the data they use and how they use it to achieve their operational and legal goals. This is the bit where the it team can check when reports/dashboards were last used [ref 2] and provide the evidence about what is and isn't being used in the system (how often has the business demanded something and then soon stopped using it?).

4. Classify the different types of data held and set the rules under which the data should be either deleted or archived-off the system (either into read only, accessible storage or into cold storage).

5. Implement the policy.

6. Setup periodic review of the policy to ensure it stays relevant as the company and external influences change.

By ensuring solution design also focuses on data archive and removal policies, the business is pro-actively putting in place the necessary steps to keep its application data relevant.

Enabling the Policy

Once the business rules, defining when data no longer provides utility, have been met (e.g. the length of time a record has been retained but not accessed) the data can be removed from the application. Once removed, the same question is asked - how long to retain the data either in an archive model, or in a separate data warehouse collection?

With the business rules defined for data retention within Salesforce, the use of technology can be considered. Typical retention rules are age-related, e.g. removing activities that were created more than x years ago, or deleting customers with whom there has not been any interaction for y years. The activity removal is quite straightforward - run a query filtering records where the creation date is older than a given date. Contact removal is trickier as the contact record itself may not change (and therefore the last modified date remains old), but activity with this person may be ongoing and frequent. Therefore associated records, such as financial transactions, case requests, etc. all need to be checked before setting a flag on the contact record to indicate it is ready for archive/removal from the system. Then decision about archive/removal of all the associated records (activities, cases, chatter, opportunities, etc.) need to be considered. Activity removal fits nicely within a regular batch run. Updating the contact flag is trickier. Its update could be built into record triggers of associated records as they are created or updated, but this may cause trigger cascade effects in Salesforce. Or the update could be performed within a regular batch containing the necessary querying logic to check the associated records, which in turn could place heavy query loads on the system. The detail of each business solution helps the selected approach. You probably wouldn't want to leave the task of identification and removal to a person as a manual exercise, but it is always an option.

Assuming the data pruning approach will be automated, the filtering rules will need to be created and applied to the selected technology. Within Salesforce there are a few options available:

- Data Loader
- SOAP API (for fewer than 50,000 records)
- BULK API (for greater than 50,000 records)

The quantity of records to extract/delete and their content may influence the choice of extraction technology. For example, where document extraction is required, the use of BULK API provides a challenge due to its lack of support for document extraction. Data Loader does provide document extraction capability, but a little bit of effort is required to convert extracted documents from their base64 extracted format back into their native binary format.

There will be cases where not all of the removed data will be gone forever, because the intention may be to move it from Salesforce into longer term storage, utilizing technology mentioned earlier. Where required, the data can still be accessed from Salesforce via the use of Salesforce External Objects.

Use of AWS and Heroku technologies to host externally held data will become more appealing to companies in light of the December 2016 announcement of AWS and Salesforce extending their global strategy alliance [ref 5]. The announcement states Heroku applications and associated resources running in AmazonVirtual Private Clouds will work together without using a public network connection. Coupling this with the announcement that Salesforce will Leverage AWS Infrastructure for core services in Canada (generally available in mid-2017), with others expected to follow, the combination of the three offerings looks to be even more compelling, especially as connectivity into Salesforce using Heroku Connect does not consume any Salesforce API limits [refs 3 and 4].

When companies decide on a new solution utilizing Salesforce, either a brand new venture with Salesforce or choosing to extend an existing implementation, the challenge of data volume growth and the planning of when data needs to be retired, or moved to other technologies, need to be considered during early solution design. All the necessary functionality doesn't need to be in place for day one, but it may be more cost effective to introduce the capability during solution build instead of retrofitting it after the solution has been rolled out. Encouraging proper and effective data pruning of a solution from day one will provide business benefit in keeping the Salesforce application healthy.

References

1. Salesforce record sizes: https://help.salesforce.com/apex/HTViewSolution?id=000193871
2. Salesforce identifying old reports/dashboards: https://help.salesforce.com/HTViewSolution?id=000004237
3. Heroku Connect API change July 2015: https://developer.salesforce.com/blogs/developer-relations/2015/07/heroku-connect-now-free-salesforce-api-calls.html
4. HerokuConnect API counting: https://devcenter.heroku.com/articles/heroku-connect#salesforce-integration
5. Salesforce AWS announcement: http://investor.salesforce.com/about-us/investor/investor-news/investor-news-details/2016/AWS-and-Salesforce-Extend-Global-Strategic-Alliance/default.aspx
6. IBM Watson: http://som.yale.edu/sites/default/files/files/IBM%20Watson.pdf
7. USA Consumer Data Privacy: https://www.whitehouse.gov/sites/default/files/privacy-final.pdf
8. Europe Data Privacy: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML
9. Andersons vs USA: https://www.law.cornell.edu/supct/html/04-368.ZS.html
10. Book: The Official (ISC) Guide to the CCSP CBK
    Publisher: Sybex
    Author: Steven Hernandez
11. Book: The Collins Dictionary And Thesaurus In One Volume
    Publisher: HarperCollins Publishers
    Managing Editor: William T McLeod
12. Economics, Marginal Utility: http://www.managedstudy.com/micro/law_of_diminishing_marginal_utility.htm

Martin Prosser

Principal Customer Consultant

Martin is Principal Customer Consultant within the CRM architecture and Cloud Center of Excellence.

He has experience of over 25 years in designing and delivering CRM solutions across a range of technologies and vertical with focus on secure and efficient solutions providing gain savings for clients.

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. To know more, please visit www.mphasis.com

For more information, contact: marketinginfo@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655

UK
88 Wood Street
London EC2V 7RS, UK
Tel.: +44 20 8528 1000

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
DoddanakundhiVillage, Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

www.mphasis.com

VAS 18/01/19 US LETTER BASIL 4259