By Patrick Cuba

What plans are in place after you meet your demise? This subject is not a comfortable discussion to have for most What happens to your belongings, your estate, all that you leave behind, who will to be notified, and what of those who depend on you?
Graveyards paint a picture of dread in horror films and usually the protagonist is not at the cemetery by choice. Whether it was a death by natural causes or the consequences of unfortunate events; those that are affected are left behind to pick up the pieces.
This is no different for our data that we hold. Data will naturally decay through aging, but the question arises is whether the data should be preserved and displayed if its presence obscures or even threatens the business? A contingency plan is needed in the event that a data vault artefact or data in the data vault is no longer needed, superseded by a more current view of the business or whether its presence is perceived as a risk to the business operationally or even legally.
In essence, data vault governance needs a funeral plan for the dearly departed hub, link or satellite or the data therein.
To illustrate this, let's explore two distinct levels of contingency:

- We do nothing
- We do something

Do nothing.

Figure 1 Dr Malcolm Crowe doesn't see the irony (yet)

With no provision in place we leave it up to the data community to "discover" that a data feed or table has become deceased. This is synonymous to thinking that for the last week your business has been operating on data that could be invalid or at least out of date -- and the business were not informed. And if we continue to do nothing how long before the same operational debt is "discovered" again? In the discipline of data vault this is akin to a feed no longer loading to a hub, link or satellite and the dimensions that rely on those artefacts are thus inadvertently affected. Was it intentional and no one was informed? Was it unintentional and there are no operational procedures in place to alert or prevent it?

"Trust, like reputation, is hard to earn, but easy to lose"

Do something

Unintentional disruptions of fresh data is (should be) catered for through operational procedures triggered as (at least) alerts and accompanied by a set workflows to recover from that disruption. Intentional disruption of data has other consequences and considerations not enclosed in an operating manual. The latter disruption relates to the retiring of a data source that may or may not be usurped by another data source for instance.

Figure 2 Steve was not amused when he learnt the truth

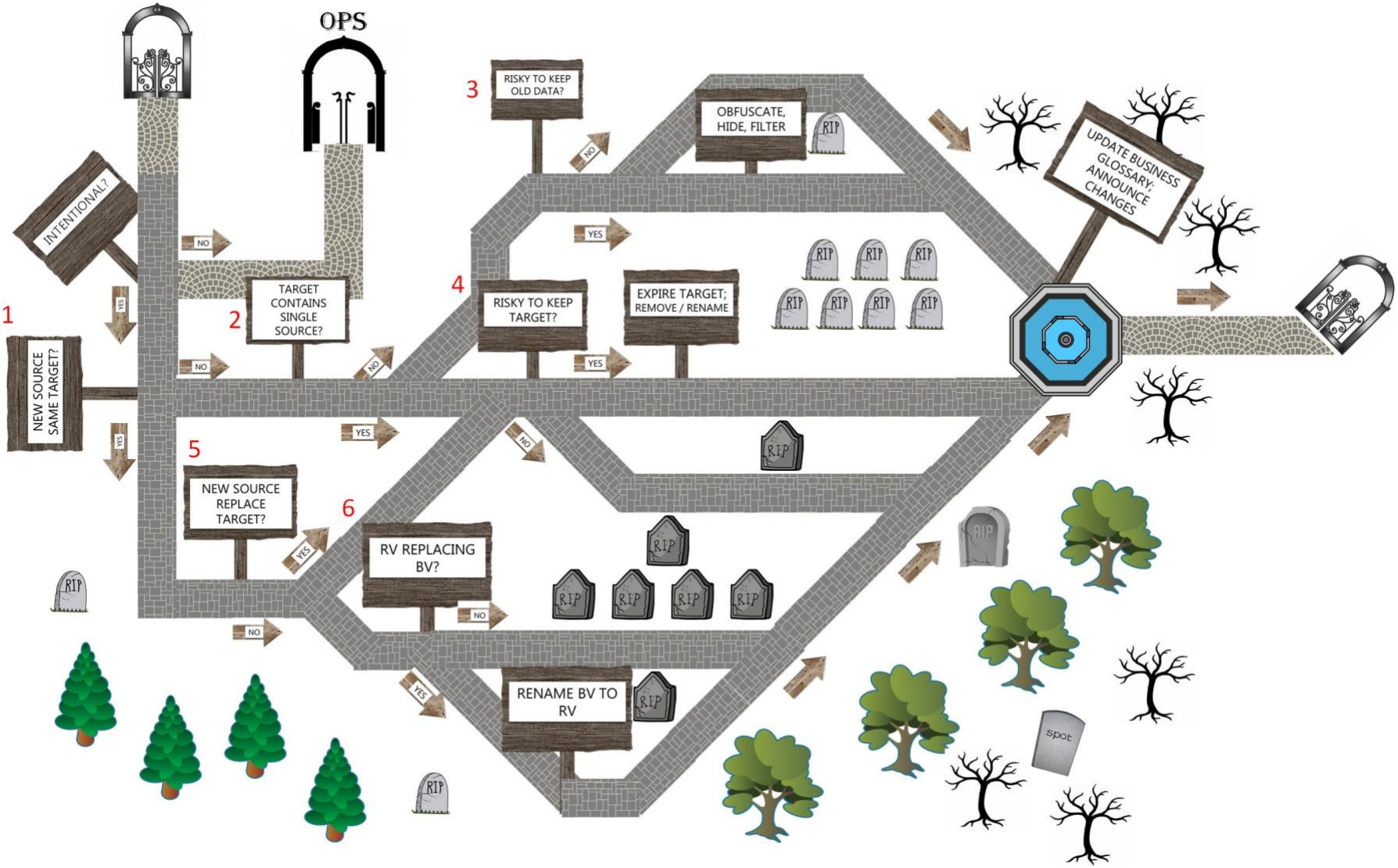
All who use the data must be informed, all analytics based on these data sources must be considered and the real value of tools to track and record data lineage is exposed.

Information about the change is not only needed to be spread vertically through data lineage and lines of business but horizontally across scrum teams as well.

"The misinformed are misaligned"

Governance, data modelling practices and methods of broadcasting information must be established early, although here we will only discuss what happens to the data vault artefacts on death row…

How do you decide what path to take? That depends on the context and organizational preferences; let’s take a stroll through the cemetery below… We have summarized our actions below the image in a table



#	Signboard	Choices and considerations
1	NEW SOURCE SAME TARGET?	Intended to establish if the reason for retiring an existing data vault table is because we want to add a new data source to an existing hub, link or satellite - NO, go to (2) - YES, go to (5)
2	TARGET CONTAINS SINGLE SOURCE?	Does the hub, link or satellite store the data feed from a single source system's data? Satellites are usually modelled as single source entities as it is rare to find two or more source files that would supply the same columns names. Hubs and Links will only contain a unique list of keys and relationships respectively and thus determining the number of source feeds to these data vault table structures requires the use of lineage tools to determine if all sources are indeed active. - NO, target table contains more than one source feed, go to (3) - YES, go to (4)
3	RISKY TO KEEP OLD DATA?	The old data in an existing hub, link or satellite could have consequences if left in there. Could the data in that data vault table be used incorrectly? Should the data in the target table be removed due to regulation? - NO, go to fountain (end) - YES, follow due process to remove or obfuscate the data depending on the context; then go to fountain (end).
4	RISKY TO KEEP TARGET?	Does keeping the existing hub, link or satellite pose a risk to the analytics and the business if not removed or obfuscated? It is important to establish all the usage this data vault table is used in before acting on it. Marts, views and established dashboards could be affected and therefore a migration (if possible) will need to take place. - NO, go to fountain (end) - YES, follow due process to either rename the target table, move to another schema; go to fountain
5	NEW SOURCE REPLACE TARGET?	Intended to establish if the new source is a full snapshot or a delta load. - NO, new source is a continuation of the existing target data and it fits the same grain; go to (6) Rarely is this possible with a satellite table but with hubs and links this is more common. - YES, new source replaces every record; go to (4)
6	RV REPLACING BV?	In the case of technical debt being successfully pushed to source (that is a win) ideally the business vault process and historized content is no longer needed if it is superseded by the raw vault process and data. - NO; go to fountain - YES, rename the business vault table into a raw vault table; go to fountain (end) Arguably the same could be said for (5) but a new snapshot should be loaded directly to a new raw vault table and therefore the need to establish and possibly blend a raw and business vault table is not needed.

#	Signboard	Choices and considerations
		You could also argue that keeping raw vault and business vault separate is better practice, but the source of the records can easily be distinguished in a blended data vault table by inspecting the record's record source column value. Of course, this is merely a suggestion and your practice and preference may differ.
end	BROADCAST	Using all the established channels, governance tools and et al; get approval and notify the wider community so all are up to date and are working on correct analytics

Then there is GDPR…

Figure 3 Jud Crandall has a sombre conversation with Louis Creed

Data retention within GDPR (one of six principles) must have a clear retention period beyond which it should be deleted and Personally Identifiable Information must be anonymised. Further; a person can request that their data be erased from the enterprise (right to be forgotten) and the enterprise must respond within one month. Strategies to minimise the impact of these updates include (but limited to) splitting satellites into non-PII and PII satellites; tokenizing and obfuscation of PII data before being loaded into data vault and data vault includes structures that can help prevent accidental reanimation of PII data (preventing the data from reappearing in data vault-think about record tracking satellites).
However, with GDPR data retention can still be justified in certain legal circumstances such as fraud detection and prevention and the answer always is; it depends.

Aging
Declared dead in abstentia may be declared despite the absence of direct proof of the person's death. This can be somewhat true in data too in terms of business keys and relationships. Implemented as a business rule we can assume that a business entity or relationship has not appeared in the file from source then a business rule can declare the entity is dead. Typically a key not seen in two months may be reported as missing at first but upon agreement the entity can be marked as dead.

What if the business entity reappears?
In all the above scenarios we never delete data in data vault but we use data vault satellite structures to record when a business entity or relationship was inserted, updated and deleted and when the last time they were seen. For these we can look towards…

- Status Tracking Satellites,
- Record Tracking Satellites and
- Effectivity Satellites

Figure 4 Is that you Wilson?

And in the case of GDPR these same structures can be used to prevent accidental reappearance of an entity

Writing an obituary

On Sunday, 29 July 2019, LINK_BV_CUSTOMER_ACCOUNT affectionately known as “the business customer table” has passed away deliberately to be replaced by LINK_CUSTOMER_ACCOUNT who will now become known as “the raw customer table”. LINK_BV_CUSTOMER_ACCOUNT will leave nothing behind and be moved to a secure location so as to no longer influence and prey on those that are still amongst the living. I am sure you as the business community will welcome LINK_CUSTOMER_ACCOUNT as your own and find all your reports and dashboards unaffected by this change and rest well in the knowledge that we have moved the technical debt to the source where it belongs. We shall not speak of thee again but take a nostalgic glance at the archived reports and think back to the days where “the business customer table” held sway.
Rest in Peace dear technical debt.

This article is written with a dash of humour but contains some elements of consideration on maturing data vault structures and content as the business matures.