PhUSE EU Connect 2018

Paper RW04

RWE: what we have observed

Peter Grolimund, Teradata, Zurich, Switzerland

ABSTRACT

Many customers are using data-platforms to perform advanced analytics on so called Real World Evidence data (RWE-data). It became somehow a key asset of many processes. As a provider of platforms in these area for quite several customers, we have seen different approaches, from a technology perspective, as well from a methodology setting. There are as well major differences in regard of the data-integration or data-management approach which is a key enabler to use the data efficiently. This presentation does highlight the different settings, and what different actions, use pattern or cultural approaches can leverage additional value. It should highlight some upcoming trends and approaches. There is additional information about the convergence of different domains within the 'health-care' industry stakeholders as e.g. among health-insurance, health-care providers and the pharmaceutical industry.

INTRODUCTION

Over several years RWE data analytics has proven that it can improve many steps in the development process and other business domains. In several white papers provided by data providers, service providers or the user community the potential has been highlighted.

More and more biostatistician, programmers, data-scientists and whatever roles are involved in the analytical process, have developed know-how, skills, methods and approaches to turn the data into real value driving solutions. It has become part of many business processes.

On the technology side many different approaches have been proven to be a suitable way to handle the larger data volumes. Technologies on owned infrastructure, on the cloud and mixed approaches are available or have been implemented.

Based on our experience with our own customers, the approaches documented and shared, we still believe, that there are further steps of improvements which could be leveraged, and we are going to highlight a few ones and will give a short outlook what might be next. What needs to be understood is that there will most probably no perfect solution able to cover each organization.

LICENSING AND THE GOVERNANCE

By no surprise, large organization naturally have so 'called' data and organizational silo's and because of the size, complexity this is sometimes almost unavoidable. However, it can as we know, add significant costs to the overall organization and as the cost pressure is increasing in many company, it might look a bit surprising to bring up this point. It is about licensing the same data, several times without even knowing, this is happening. Typical settings are licenses purchased by affiliates (country organizations, specific organizations etc.). Indeed, the needs might differ and even the legal setting might induce this overlap. However, we recommend the implementation of a clear governance around the licensing, the availability of the data (inventory/metadata) to increase the transparency across the organization. Some companies have realized this potential and have even established a central contact point, where RWE-data is purchased. As the licensing costs can be significant, a proper governance and the related effort to do so, can easily be covered by the saving of unnecessary licenses. What shouldn't be underestimated is indeed another side effect of a highly de-centralized data acquisition policy, the lack of the use or missed opportunities of existing data. Data which might exist in an organization is not known to other groups, which would generate significant additional benefit, if they would know and have access to this kind of data. Linked to this aspect is another flavor of governance. We call it analytics of analytics. This is the analysis of the metadata produced by the usage of the data (logs, what kind of tools and how the data is used). This can either highlight difficulties in the usage, or as well the lack of expected value which might then allow a re-assessment of the licensing model. Cost pressure might lead to the tracking of the generated business value out of the data-license investment.

Another hidden 'duplication' of licensing we observed by embedded data. Some data-enrichment or curation providers include some commercial data sources and are cleaning or enriching the data with other sources. This called curated data-sets, with an improved usage level, do sometimes include licenses of data, which is already available in the original format. It is not easy to uncover such situations, but some data-profiling might bring this to the daylight and additional saving or renegotiations might be meaningful.

Another point is linked to the licensing practice. Experienced with a lot of advance data analytics situation in different industries, we always recommend to license, collect and store the raw (most granular) data within an analytical platform. The reason is obvious. In many cases any aggregation might already include a bias, and even more important, new analytical cases or drill downs are not possible. In some cases, data gets licensed in different

aggregation levels across the organization. The most granular might be licensed for a specific geography or domain, it it's full granularity. Head-quarters more interested on a higher aggregation level, are going to license the data on the level they are looking for. An analytical platform should be planned in such a way, that the aggregation levels, which could be considered business-views, on a higher than the most granular level, should be provisioned easily. This can help to avoid additional costs.

If an organization can handle the governance, the licensing practice and the integration in an efficient way, literally millions of licensing cost might result as saving costs.

GOVERNANCE OF ANALYTICS AND DATA VERSIONING

In a GxP compliant environment, as e.g. primary use of clinical trial data, the need of having a well-functioning Governance of the analytical process is a must. In the context of RWE data analytics, there might be the need to apply GxP compliant processes. Sometimes the outcome of analytical activities can be used for as part of the primary analytics, as e.g. control group data generated out of RWE data because of ethical concerns to treat patients with placebo. Or the study will be completely virtual and used as part of the submission process.

There are many ways technically but as well process wise to apply the same governance process as for the primary purpose use. But there are still many companies which hesitate to apply the same architecture and processes, as a lot of the analytical work performed on RWE data, has much more an explorative or data-science based approach is never going to be used as part of a GxP relevant process. Some user communities indeed are frequently concerned to apply the most rigorous regulatory constrain, if not needed, as the flexibility of the system might be hampered.

However, new working practices as using version control management approaches like Git-Hub even in any software development process, or "Dockerizations" based approaches can help to minimize those risks and remove the concerns. But not only the analytical programs, but as well the data needs to be versioned to allow a re-construction of former process. Even the costs of storing data has significantly decreased, the approach to store always full data-sets can become costly and the governance of the archived sets might increase on itself the costs. To keep the version control on cohorts, or subsets of the original data means, over time a lot of redundancy will be generated which increases the data volume significantly.

We have seen approaches to archive the 'analytical-data-sets' on so called 'cheap' storage system e.g. data-lakes built on top of open Hadoop-technologies or simple 'object stores' as e.g. S3. This kind of approaches do work but they have their pro- 's and con's. As RWE-data is still mainly structured data, the version control could be as well realized by using so called 'temporal' concepts. This would imply that the data is loaded incrementally (only changed data compared to its previous version) and the versions are controlled by applying time-stamps to the view. We are fully aware, that some data-providers format do prevent such approaches but in general, such a concept might at least help to minimize the volume. Additionally, the analytics of the changes over time might give some additional insights in regard of the growth rates, change patterns etc.

In summary we see that the overall data volume generated by redundancy might easily reach a factor of 4 and above. Without considering the challenges imposed by handling the volume, the governance of accessing the data and keeping it under control demands sometimes complex frameworks and processes around.

THE INFRASTRUCTURE

We have seen many different approaches in leveraging the infrastructure for performing analytics on the rather high-volume RWE Data. The approach was in many case driven by economic drivers, by IT-policy approaches (e.g. moving to the cloud) and as well of course based on existing approaches and skills.

Of course, many of those solutions deliver the expected functionality and capabilities, but there are as well many differences which might have an impact on the efficiency of the usage of the data.

One of the key observations is, that the infrastructure is often tailored to the existing analytical skills and technologies in use. Said that, it means, there is a strong tendency to use still file-based flavored settings even the infrastructure might leverage a relational database.

It is a special point we are going to highlight one of the key elements of our approach; the scalability and performance. Many might argue, that there are not many time-critical analytical processes running on RWE-data as no SLA-based response times is required. But lack of performance can lead to different negative consequences.

One is the data-model quality. Long running times for complex analytics will lead to a reduced optimization of the underlying model approaches or parameters. This means, the model results might show reduced quality and hence the full potential cannot be used.

Another effect is the externalization of the data into other platforms and hence increasing data-copies, losing control of the usage and governance of the data. This adds cost and effort as the data-gets moved. Moving data is in addition not adding real value.

An interesting example of such a situation is the commonly use of cohorts. In many cases we have seen copies of data out of the RDMBS system to accomplish the cohorts. Or the cohort is a file extract of an RDMB system and the analytics is going to happen on this extract. Beside the effort and cost related to this kind of working practice, it does as well limit the usage. Our experience has shown, the change of cohorts, adjusting inclusion and exclusion criteria and rerun models is an essential part of explorative analytics. In some cases, it might be even interesting to perform metadata-driven changes of the cohorts to build a sophisticated and adjusted final cohort.

Another consequence of the 'physical' sub-set approach and moving the data induced by lack of performance or lack of other infrastructure related capabilities is the use of many different stand-alone analytical engines on different servers. The costs of the replication of infrastructure and may be the increase of licenses because of this, should be evaluated carefully as it can as well be quite significant.

Approaches to launch the system on the cloud are becoming more and more the common model. But what has been mentioned before, still holds true. We have seen the cohorts going to be built on the cloud, and then the data gets copied and used either on the cloud or the subset is even downloaded on premise and gets analyzed on an analytical platform which is available there. In this case of course some of the advantages of a cloud-based approach are eaten up by additional costs because of redundancy, or keeping other local infrastructure going-

Frequently we are hearing, that the cloud has the advantage to be scalable and burst-approaches in situations of high need and load could handle e.g. the performance related challenge mentioned above. Interestingly even this works in many cases out to work, the costs related to the cloud-based technology chosen, it might lead to an unexpected cost burst as well, as the technology demands for quite important infrastructure to be released to leverage the according performance. It might be of help to perform some initial pilot settings to avoid a surprise.

THE INTEGRATION / MODELING

More and more companies are moving from propriety (provider dependent) formats to a common mode to be applied on the RWE-data. The OMOP / OHDHSI initiative is gaining more visibility. The potential to apply the same models and algorithms on several data-sources increases the analytical efficiency and allows in an easier way to validate results across different data-sources.

As pointed out in the governance of the data, in many cases the data is residing in separated data bases. But there are still many organizations, which have not yet performed this step. Among several reasons one we heard about is the fact, that in applying the transformation of the initial OMOP approach, records which are not 'compliant' with the mapping rules will be suspended. In traditional and well design data-integration approaches (you can call it data-warehouses) these suspended records are still available to the user if needed.

Other hurdles are the in the meanwhile acquired of some expertise and skills, as well programs based on the proprietary formats. But the common model clearly adds more value and efficiency overall, despite there might be a short-term dip in migrating the code and performing change management in the usage of the data.

Another important trend and natural trend, is the intend of data-integration beyond the RWE-data itself. A goal is the combination of RWE-data and the clinical data.

If there would be a common modeling approach across this data-sets, the usage might become even more interesting. Approaches to run 'virtual' control arms on RWE-data and the treatment arm on a real trial might would be made much easier if the RWE data would come e.g. in CDISC like formats.

Another case is as well to combine hospital related data (HL7/FIHR) with CDISC data (as part of the trial setting) and as well the RWE-data licensed in.

The easier the use of such settings, the higher the likelihood, that such concepts and approaches are going to be realized.

We had many discussions, meetings, workshops around the right modeling approach. Our usual approach taken in many other industries is the creation of models, which put the nature of the data into the center of the logical modeling. If such an approach gets applied, as a sudden the data coming from clinical trials, RWE-data providers, or Hospital system are modelled the same way. It is patient related information and there is from a 'logical' perspective no difference of a blood-pressure measured as part of a trial-related visit or a hospital visit because of a control or disease related event. The only difference is the why the data was taken and of course this should be incorporated in the modeling approach.

During those meetings and workshops everyone tends to agree such an approach and the significant efficiency gains related to it, up to now we are not aware of any organization which is using such a concept. Interestingly enough, CDISC has built a model (CDISC-BRIDG) which is somehow founded on such an approach and even it is publicly available, not used in many organization according to our knowledge.

THE THIRD LEVEL OF INTEGRATION: IOT, FINANCE, GENOMES, BIOM, …

In the future the RWE-data will include more and more other data of different nature. This means, the 'integration' and scalability becomes more challenging from an infrastructure perspective. Genome-related data (e.g. provided by Flatiron and other providers), device or digital biomarker related data, financial data within the own organization, other publicly available reference databases (e.g. FDA, NIH, bioinformatic institutes, research publications), demographic and epidemiological data published by statistical authorities and health authorities etc.

We expect an increasing need of such data-sources being part of the ecosystem. The need of tailored, on demand solutions of integration can be easily foreseen and one of the challenges will be the way a solution can scale in the speed of the demand not only in the context of execution time, but as well simply in the context of adding and integrating the data in a meaningful way. We all do know, that in case an infrastructure or solution cannot leverage in time, there will be quickly a new analytical platform available. The challenge in the future will be, how far the environment can leverage good data and a performing, flexible analytical workbench.

There might be justified concerns, that such a challenge can never be handled, and we need to live with a fragmented diverse solution.

From my personal perspective, innovative ideas and existing concepts developed could help to leverage it. Crowd sourcing of data-integration, AI-driven approaches, and other, new concepts might help to handle this. Like other aspects, the volume of valuable and easy to use data, will guarantee the existence and growth of an analytical platform.

THE ANALYTICS ECOSYSTEM AND IT'S FLEXIBILITY

What we do see as well is an increasing diversity of analytical approaches, tools and methods. If the analytical platform cannot leverage all this, and new upcoming approaches, the interest in using the underlying data will quickly decrease. From an architectural/infrastructure perspective, a lot of the companies are trying to cope with this situation. This holds true for turn-key solutions provided on the cloud, as well on-premise solutions. To handle such a demand and diversity and keeping the governance of the data under control (how has access etc.) clearly demands some principles. The access control needs to be moved outside of applications and simply linked to the data. This means, as a user, I can use any analytical approach, the access and usage of the data is enabled and leveraged.

What we see as well is a mixed community where researchers, biostatisticians, bioinformatician and BI-users or visualization tools are all accessing the same data-sources are adding load and challenges. Workload management might become a need to handle priorities or a sophisticated scale up and down approach.

Analytical servers and engines are going to be plugged in and used (e.g. tensor-flow and other AI/ML based software packages).

The question, how far those needs, and the related extensions are going to leveraged on the cloud, on premise, by a service provider or the internal organization is not as relevant. Where it becomes relevant is how far the solution is still manageable in a cost-effective way. If the system isn't scalable because of its complexity, technology or related extension processes, new analytical platforms are going to be created and the effort of multi-system data-integration is eating up a lot of resources and will increase the overall costs dramatically.

THE LEGAL ASPECTS

Finally, I wanted to add a few words around the RWE-data access in the EU. In many countries of the EU, data-privacy is highly ranked and the use of RWE-data in a de-identified or anonymized form is simply not possible because of the regulations.

However, I recently got a phone-call from a member of the EU-commission in charge of assuring equal economic conditions among competitors. The reason was, that they were looking after two cases, where two-big global privately-owned companies, were involved of using RWE-data from university hospitals for testing AI approaches to be applied and leveraged. One of the cases was situated in England the other in Italy.

The question which was brought up was, how far any company not in the size of such a big organization can have access today to the same data to leverage their own solution or commercial product related to the data.

Being citizen of a non-EU-country, the question was delicate and could not fully answer directly the question. But what I was aware off, is the situation in Germany. At specific event, there was a minister of a 'Bundesland' who was concerned of the innovation driven out of data-science. They realized that start-up companies who wanted to leverage services out of RWE-data did move to the USA, as they had no money to license those large data sets nor had they any opportunity to get access to data from university hospitals because of the privacy concerns.

During the discussion the EU-representative brought to my attention, that generally, data collected as part of a process financed by the government, and many of the university hospitals are financed by the tax payers, must be made accessible for free to any organization requesting for it.

Fully aware, that such a statement completely contradicts to the existing experience and reality, I will leave it to a further discussion. But if this would be true, the use of RWE-data in Europe might become much more easier and I'm curious, who might have more insights related to it and who is going to stretch those boundaries.

REFERENCES

References go at the end of your paper. This section is not required.

CONTACT INFORMATION

Peter Grolimund PhD
Senior Industry Consultant Life Sciences

Peter.Grolimund@teradata.com
mobile: +41 79 619 37 20
Phone: +41 44 839 33 63

twitter.com/P_Grolimund
linkedin.com/in/peter-grolimund-2588964
Teradata (Schweiz GmbH)
Alte Winterthurstrasse 99
8304 Wallisellen (Switzerland)