Why data services should be invisible. A coordinated approach across universities.

Thomas Eifert[1], Denise Dittrich[2], Julia Opgen-Rhein[3]

[1,2,3]IT Center, RWTH Aachen University, Seffenter Weg 23, 52074 Aachen,
[eifert|dittrich|opgen-rhein]@itc.rwth.aachen.de

Keywords

data services, digitalization, reliability, collaboration

1. SUMMARY

"Digitalisation" is a ubiquitous theme at every university. Since by now literally all aspects of university work and life depend on IT and, in particular, on data, the availability and persistence of data and data services is essential. As trivial as this appears, scientific users need to rely on the availability of their data and data stores, either for using previously stored data or to run experiments that generate new data. This includes all aspects, from the availability to persistency.

Since in many cases access to the data happens in a tiered IT setup, there is no way of "direct" access to the data but a suitable - often server-based - software stack has to be up and running as well. Plus, it's not "a" software stack but a vast multitude. Therefore, the mentioned requirements go beyond bitstream preservation but include the infrastructure to access the information.

In addition, the justified expectation of all users is an (almost) 24x7-availability of services, and, in particular, an assured persistency of once-stored data. So, data services have to be invisible in the sense that they are expected to be ubiquitously and continuously available. Any "visibility" can be regarded as a disturbance of this expectation.

To cope with this combination of amounts of data, level of expectation, and technical multitude and a concurrent steady growth in all of these dimensions, a consortium of universities in North-Rhine Westphalia currently runs a project to develop structures for joint, university-spanning services for safe data storage and high storage and server availability.

1. The Challenge

Along the ongoing course of digitalization of either in teaching and learning, in research and in administration, almost all processes rely on a stable IT infrastructure with permanent availability. Even more, also in the event of an outage, the expectation is that the state immediately before the outage occurred can be restored.

From the users' perspective, "data storage" also includes the software layer they are interacting with, like a database, a versioning system like gitlab[1] and many others. It also refers to everyday actions like email communication or storing data in a shared directory. Therefore, the mentioned requirements go beyond bitstream preservation but include the infrastructure to access the information.

In all these cases, if the proposed – or implemented - service level is (or is felt to be) below the scientists' expectations, local circumventions are built either by the scientists themselves or, in a better case, by the IT staff. These solutions appear to solve a particular problem including a concrete infrastructure or software layer, but effectively hinder structured approaches. Thus, they are in total,

[1] https://about.gitlab.com/

from the organization's perspective more expensive than central services. Furthermore, specialized local solutions make the exchange of data and hence the collaboration between scientists more difficult as they can create the need to convert data to another format and store it in another system.

When talking to scientific users "service level" means much more than assured parameters like uptimes, restore parameters and so on. From the users' perspective, it also means functionality, performance, capacity, and price.

"Capacity" apparently is the simplest one, achieved by just enough amount of storage. We will recur to that apparent simplicity later.

"Performance" means several things. For automated processes, either in the lab or elsewhere, it must be simply fast enough to process a given amount of data in a given timeframe. It is much more complicated in the interactive realm where it comes to felt responsiveness of an IT system. This aspect gets even more complicated under the extended meaning of "data storage" we discussed earlier where one has to take the complete stack of software into account.

"Functionality" applies mainly to the higher-level software layers. Taking gitlab as an example, the core feature – version control – as well as aspects of usability are crucial for acceptance by users. For systems dedicated to research data handling, issues like metadata handling or access control are important.

Given we have such powerful services, systems, and solutions, users do not hesitate to use them heavily and, as said, rely on them. Which matches exactly the intention of these services.

2. The Downside

The widespread use of large numbers of users makes these services and the underlying infrastructure critical in the sense, that an outage will have huge impact on virtually all processes of a university. In contrast to former days when everyone worked on their isolated infrastructure with almost no impact on others, an outage of any kind immediately affects many users nowadays. In consequence, we are in the responsibility to make sure that the ongoing integration and consolidation builds and runs a reliable platform.

In this context, "reliability" has recently got an additional meaning. Recent events show that the IT-infrastructure of universities is under attack by malevolent people, ransomware etc. and there are voices saying that it is only a matter of "when" an institution gets attacked, rather than "if". Consequently, reliability nowadays means also resiliency against these forms of attacks.

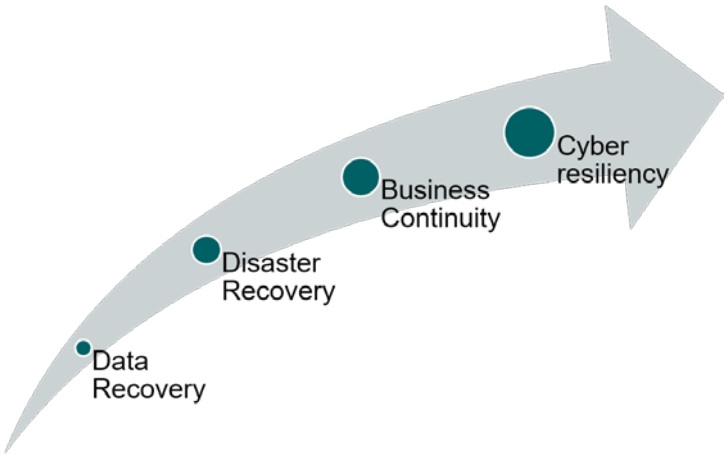

Fig. 1: business value of data protection

So, the combination of the primary storage platform and protection mechanisms have to ensure the reliability to the users.

3. The Challenge, Part II

Expressing the demand for a reliable platform covers the fact, that there is no such thing as "the" platform. On the contrary, the various flavours of storage-like services lead to an equal number of systems, with differing technical characteristics. So, protecting these in an appropriate manner has to consider all of these varieties. The hitherto widespread practice copes with this by simply having every "higher level" storage service dump its data to disk. Subsequently, this dump file can be collected on the file level. While this staged method works until now, it omits all sophisticated methods of saving only changed data and – at least when the dump files are removed after having them copied – leaves the "higher level" service without a mean to restore eventually lost data.

This approach is, besides its structural weakness, less and less feasible nowadays primarily due to the rapidly growing amounts of data: In more and more cases it is becoming a challenge to dump a service's data to disk in a reasonable timeframe.

The appropriate way to cope with this is to build data protection specifically to every type of system. Some things can be solved in a prototypic manner, but essentially, we have to tailor each of the above-mentioned different platforms – and those still to come -, systems, and software stacks to interact with the data protection layer.

4. The Approach

The scenario is quite the same for probably every university so it is somehow evident to join forces. Consequently, with support from the State ministry of Culture and Science of the Federal state of North Rhine-Westphalia, we started a project to re-organize the protection of data in a co-operative manner with a division of work between the involved IT departments. One part of the idea is that concentrating the infrastructure at a moderate number of sites lowers the overall operational effort. Furthermore, it allows techniques that fully unfold only at large scale, like an object storage system distributed over 6 or more local server rooms and erasure code protection. Such a setup simply is not affordable when talking about a few hundreds of terabytes. The other, evenly important part of the idea, is that the expertise about the operation and protection of the mentioned variety of systems does not necessarily have to be at the same site as the storage infrastructure. These ideas are the basis of the work-sharing basic concept that we regard as the key ingredient to cope with the depicted challenges.

This approach is thereby a unique chance to question the way data protection is done by now in order to create a better service.

In previous discussions we experienced that there are many differences in the perception of the state, functionality, and service level the current data protection services have at each university. These differences grow instantaneously when asking for the "ideal" state, but these differences occur not so much between universities but between the roles of the person one is talking to.

To systematically collect this information, we designed an online survey which is distributed to the central IT as well as the departments and institutes of all involved universities. This survey is conducted using the tool SoSci[2]. The interviewee can give answers by single and multiple-choice selection, by numerical values, and by free text. To assure the quality of the answers and to make sure that we can evaluate them, several core questions are mandatory.

To capture the different perspectives of the interviewees, we differentiate the roles "manager in central IT", "staff in central IT", "manager in department", "staff (non-IT) in department" and "staff (local IT) in department". Previous discussions showed that these are the distinguishable roles with respect to data protection.

[2] https://www.soscisurvey.de/

We ask the people involved with the current data protection service several quantitative questions like amount of data, number of systems and so on.

The next section of questions deals with the integration of data protection in processes: which role can trigger which action in the context. And, equally important, how are the bearers of these roles maintained? Furthermore, service level parameters and the organization of local support is asked. We collect all these parameters once how it currently is and once how it should be from the interviewee's perspective.

By this approach we are confident to re-align data protection with the expectations of our users and create a solid base for the res.

Regrettably, at the moment of submitting this contribution the survey is still running, so it is too early to give results here.

5. Next Steps

With the results of the survey we expect to have a quite precise picture of the quantitative demands and of the expectations concerning technical and process integration. This expectation is to be met by the concept how to realize the depicted co-operative approach.

6. REFERENCES

Eifert, Th., Stanek, D. (2012). *Maßnahmen für verlässliche und schnelle Datenwiederherstellung*. PIK - Praxis der Informationsverarbeitung und Kommunikation. Band 35, Heft 3, Seiten 195–198, ISSN (Online) 1865-8342, ISSN (Print) 0930-5157, DOI: 10.1515/pik-2012-0032

Eifert, Th., Schilling, U., Bauer, H., Krämer, F., Lopez, A. (2017). *Infrastructure for Research Data Management as a Cross-University Project*. S. Yamamoto (Ed.): HIMI 2017, Part II, LNCS 10274, pp. 493-502. [DOI: 10.1007/978-3-319-58524-6_39]

Bunsen, G. , Eifert, Th. (2013). *Grundlagen und Entwicklung von Identity Management an der RWTH*. PIK Band 36, Heft 2, Seiten 109–116, ISSN (Online) 1865-8342, ISSN (Print) 0930-5157, DOI: 10.1515/pik-2012-0053

Bischof, C., Bunsen, G., Eifert, T. (2007). *The Resource Cooperative North Rhine-Westphalia (RV-NRW). Lessons learned from the Organization of a State-wide IT Resource Pool*. PIK Band 30, Heft 2, Seiten 88–92, ISSN (Print) 0930-5157, DOI: 10.1515/PIKO.2007.88

7. AUTHORS' BIOGRAPHIES

Denise Dittrich, M.Sc is working at the RWTH Aachen University's IT Center since 2005. She received her Master Degree in Artificial Intelligence from Maastricht University in 2009. From 2010-2012 she was deputy head of the IT-ServiceDesk, thereafter responsible for IT Process support and Identity and Role Management. Since 2016, she is deputy head of the department for Systems & Operation with her focus on providing large-scale central services like Groupware, Identity Management and Collaboration platforms.

https://www.linkedin.com/in/denise-dittrich-8b2169195/

Julia Opgen-Rhein, M.Sc. is a research associate at the IT Center of RWTH Aachen University since 2019. She received her B. Sc. in Scientific Programming from FH Aachen in 2017 and a M.Sc. in Data Science for Decision Making at Maastricht University in 2019.

https://www.linkedin.com/in/julia-opgen-rhein-73a946159/

Dr. rer. nat. Thomas Eifert is Chief Technology Officer of the IT Center of RWTH Aachen University and as such responsible for the strategy for technological development and the corresponding third-party funding of the IT Center. The focus in this function includes concepts for research-oriented storage infrastructures. He is also a lecturer for Calculus in the study program "Applied Mathematics and Computer Science" at the FH Aachen University of Applied Sciences, where he is involved in the processing and digitalization of traditional teaching content.

https://www.linkedin.com/in/thomas-eifert-4338b389/