Aalto University
School of Science

Evaluation of the Isaacus project's data lake solutions in research use

Richard Darst, Mikko Hakala, and Kimmo Kaski

Department of Computer Science
Aalto University School of Science

Goal of evaluation

In this report, Aalto University evaluates the Data-Lake solutions of Helsinki Uusimaa Hospital District (HUS) and the Hospital District of Southwest Finland (VSSHP). The primary concern of this evaluation has been to see how suitable these solutions are for research purposes. The evaluation has been done by interviewing the parties involved in developing these platforms, namely HUS with Tieto Ltd. and VSSHP during March-April 2017. It should be noted that these solutions are still under development, since the wider integration between different Isaacus parties has not yet been implemented. It should also be noted that we do not go into the details of enterprise data management, business intelligence services, or internal medical uses of data.

Organizations

Sitra is leading the Isaacus project both in terms of funding and supervision of the data lake solution developers. HUS has an in-house team managing the project, with the implementation realised by Tieto Ltd. and VSSHP has a team, “Kliininen tietopalvelu” (KTP, Center for Clinical Informatics), which is currently developing the data analysis support and is the logical predecessor of this project as well as involved in the design of this platform. The VSSHP implementation of the platform is done by Medbit, which is a public company.

In addition, we were informed by Sitra as well as HUS & Tieto that as part of Isaacus project – the North Savo Hospital District (PSSHP) in Kuopio is having their own data lake by contracting with Tieto - like HUS has done - to build their data lake. In the light of this information, the PSSHP’s data lake is expected to be identical to the HUS data lake, only the administrative policies may be slightly different.

Current status

Overall the solutions are still in very early phase and no researchers have yet used the system for actual research purpose or testing. We have considered these plans and any existing solutions in this evaluation. We have found that the planned technical solutions are generally well suited for research and data storage with competent staff involved. VSSHP has a well-developed team and currently actively provides small amount of data to internal researchers, including data cleaning and processing services. They have an initial working system, which can provide access to larger datasets with more computational power. This system is an effective prototype, though it may not scale to larger number of users. HUS, on the other hand, is mostly waiting for the final platform - being realised by Tieto - to be available. This is currently planned for mid to late 2017.

Main points

When considering the entire research process, we have identified the following high-level points yet to be missing satisfactory answers:

1. **Discoverability** of existing dataset in the data lakes is very important when formulating research questions. If data cannot be found, or if the very existence of any data is unknown, then no relevant research could be done. Yet there does not exist any catalogue or promotional materials that contains both general level information of the available data, material for later usage, and the detailed metadata of individual datasets. There are plans to implement this feature as a part of Isaacus project, but the question is will it be on time and provide a sufficiently useful view to researchers? How structured will the metadata be, and will it be sufficiently complete and useful for researchers when designing research projects? The amount of promotion required should not be underestimated, since researchers have many different things they could do, and if Isaacus wants their attention, competition is needed.
2. **Data access** presents also some possible challenges. This is so because in all cases, research permits are granted by a third-party administrative group (research ethics committee or THL for example). If research permits are only granted for projects which match certain goals of their own, the use of the infrastructure will be limited. If those evaluating research permits do not share the vision of an analytics environment taking risks and creating new opportunities, then new research will not be done, not to speak about obtaining novel and useful results. Furthermore, if the perception is that research permits are not easily granted, then the number of applications will remain low and focused on insiders that once again raises the question about the possibility of novel results.
3. **Funding model.** The regions using the cloud for the analytics environment can scale to almost any required resources, but the question is who will pay for the research usage? Researchers do not usually have extra funds for computation, and will prefer projects for which they can use their existing resources. If there is not some free basic tier, then the use of the infrastructure will be limited.
4. **Support gap.** For both organizations, the main tool for heavy-duty research is virtual machines interfacing with advanced data tools. Many researchers have a relatively narrow skillset. In order to make an effective use of this data, the question is who will provide support the researchers in using the modern computational or methodological tools? In the modern economic environment, mid-level support staff are always cut first. Who will bridge the gap between the researchers and the particulars of the infrastructure? Within the VSSHP hospital this is somewhat covered by KTP, who provides analysis as a service for doctors. At HUS this is yet undecided. Even at VSSHP, will this scale to the number of expected users?
5. **Long-term** prospects, end of life, and data integrity. Because these solutions are being managed in-house, it should be possible to migrate data when needed. However, questions of long-term funding and data integrity are not yet answered. How long is funding currently scheduled to last? Will data be migrated and kept long-term? Is data duplicated and scrubbed to present a high enough reliability guarantee (number of 9s of reliability).
6. **High-level decisions.** We are evaluating the realisations from the technical perspective. Many unanswered questions will have answers that are decided politically and socio-economically, not by either the designers or support staff we have been interviewing.

Currently, Sitra is providing leadership, but once Sitra takes exit, one can ask what happens then? The issue of follow-up and its coordination is the key for securing successful outcome of Isaacus project.

General findings and points for further considerations

- **Usage.** The platforms are intended both to support simple “small data” use-cases via tools such as Microsoft Excel and “big data” use-cases in virtual machines that can run arbitrary software for intensive computational development.
- Both data lake solutions will use **Hadoop** as the backend for raw data storage. This is suitable. VSSHP (Medbit) will build their data lake domestically and in Finland, while HUS (Tieto) builds theirs in the Azure public cloud, located in Northern EU. *If data should reside only within Finland, which is an issue that should be carefully considered and discussed.* If development is done well, the high-level software will be interchangeable.
- For **heavy computing**, VSSHP currently has locally hosted virtual machines with researcher access. These virtual machines run standard Linux and can have any software installed. HUS will have cloud-hosted virtual machines with similar features. Cloud-hosted virtual machines can scale to almost any workload - however note the critical point about payment.
- For **light data analysis** (Microsoft Excel analysis), VSSHP has a remote desktop environment so that data can be viewed while it stays under the full control of the central organization. This is suitable, secure, and flexible. *However, HUS plans on this type of small (pseudo-anonymous) data being removed from the data lake to the researcher’s local computers for analysis*. This is quite surprising and because we had assumed one of the fundamental principles being “data does not leave.” Once data does leave, it is almost impossible to ensure it is managed properly. This must be evaluated and discussed within HUS.
- **Data will flow** from all possible hospital data systems into the data lake. It will pass through an anonymization and standardization layer. After access is granted, the specifically allowed data is copied in a processed form to a researcher-specific location where it can be accessed.
- **Data standardization** is important for portability and linking. The organizations are engaged in this activity and actively coordinating, however we have limited information on the current status. An explicit final goal is to be able to link data across regions.
- **Permission for data access** is outside the scope of this review. Permission for data access is granted by external groups, and technical staff simply make that access possible. *If these groups adopt a restrictive view of allowed research, there will be few users.* In particular, access must be sufficiently simple and consistent enough so that researchers will build projects around this data.
- **Documentation** is absolutely critical in order for a system to be used. The documentation must also be kept up to date, which is an issue that is currently considered. VSSHP has internal documentation including several user guides, and HUS will produce a researcher guide.

- **Access and permission management** will eventually be linked to a central portal. However, this portal does not exist yet and the level of automation is unclear. It seems that some amount of manual interaction is needed to set up each new researcher. The amount of effort needed for full automation can be quite high, so the currently proposed realization seems to be a reasonable first step. *If the goal is full automation of data access once a research permit is granted, this will require more decision making.* We believe that things should be reasonably automated, but a small amount of manual intervention is OK and improves the robustness and security.
- Currently, no region expects to link data together in its own environment. This will be handled by a new environment from Statistics Finland. We have not evaluated these plans, but it should attempt to be as suitable for both small and big data as the regional analytics environments.
- There is need for the actual users to (agilely) give feedback of the functionality, and doing this process iteratively would make it even better.
- The **attitude of support personnel** is good, however their mandate for end user support is limited by time and funding. Should there be too many users, the time limitations may become a bottleneck.
- The **security model** is considered in general to be good. At the same time, if someone takes the data out, it stays out, so it should be as convenient to use the data in the system than using it outside. Good security design is taken into account during the design phase of the project. There are only two surprising issues mentioned above: HUS data will be stored outside Finland (but within the EU) in the Microsoft Azure cloud. Also, currently HUS plans on “small data” leaving the analytics environment to researcher local computers. We strongly recommend that all parties involved agree on a model of when data can and cannot leave Finland and the analytics environments.
- **Measurability** is a broader issue and it does not directly affect research. However, without measurement of impact, long-term funding cannot be assured. *How will the success of the project be measured, and what are the criteria for continued funding?*

Summary

To summarize we are pleasantly surprised by the current progress and design of the data lake analytics environment. The people working on this project are qualified and highly skilled. The platforms have good design principles and should be suitable for both simple and demanding use cases. The platforms are based on an open design, which provides for long-term maintainability. There is an awareness of most of the critical issues at some level.

However, as in any large project, there are many different parts with imperfect communication and management. In particular, while there is a plan for the national level, this has not materialized yet. Sitra has so far been doing a good job at coordinating, but Sitra is exiting from this role at the end of this year. After that, the coordination must be kept up by some party or all the actors must have very closely aligned goals and values so that self-organization is possible. *This is clearly an issue to be considered and subsequently realized very carefully.*

We are confident that internal users will find these services useful. External (academic) users should also find these services useful, but the developers must make sure that it stays an attractive proposition in terms of usability and cost model.

To fully evaluate a complex system like a data lake without seeing the final construction is challenging, as there are many different interdependent pieces and unexpected differences in one piece can result in something less useful. Therefore, the tests and actual use by the researchers are critical in order to determine how functional and usable the infrastructure turns out to be. We encourage Sitra to monitor the research use of the infrastructure and to get a proper testing and access on the way as soon as it possible. In addition, there should be continued resources for implementing feedback from this testing over the next few years, in order to come up with the most useful and versatile system or product.

Appendixes

Organizations evaluated

HUS - Helsinki

HUS has an internal team of a few people – formed for this project – for implementing the Isaacus project and coordinating the data lake and analytics environment design within HUS. Consulting and actual building of production systems is done by Tieto Ltd. Tieto is using open-source software and aims to make a supported open-source product that can then be used by others as well. We have interviewed staff from both HUS and Tieto.

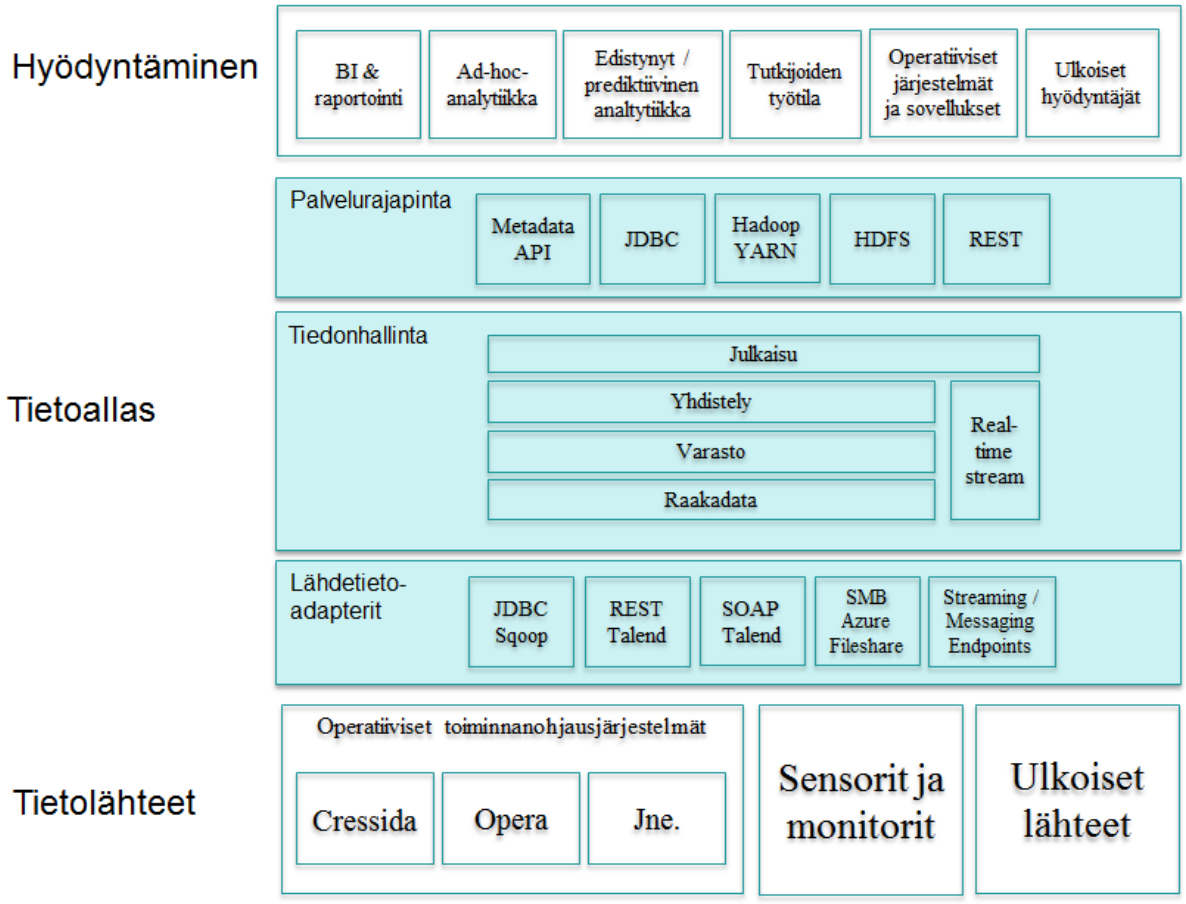

Figure 1. *HUS Data Lake schematic. This figure focuses on the flow of data from sensors, to the data lake, to users of the lake. Researcher access is one of the users.*

VSSHP - Turku

VSSHP has as part of it “Kliininen tietopalvelu” (KTP, “Center for Clinical Informatics”) which is the primary agency involved in this project. There are a group of six people who serve as an interface to data for clinicians. There the existing tasks involve receiving a request from clinician-researcher, finding the data from different clinical devices and data sources, merging them, anonymizing, and providing a secure access. KTP has provided its first-generation analytics environments, one suite for small data and another one for large-scale analytics.

Medbit is a public company, owned by VSSHP and other public health actors, which provides the IT services to VSSHP. VSSHP does not have any in-house IT staff besides KTP. KTP is serving as the data specialists and initial users of the system. Medbit is producing the actual data lake for bulk data storage.

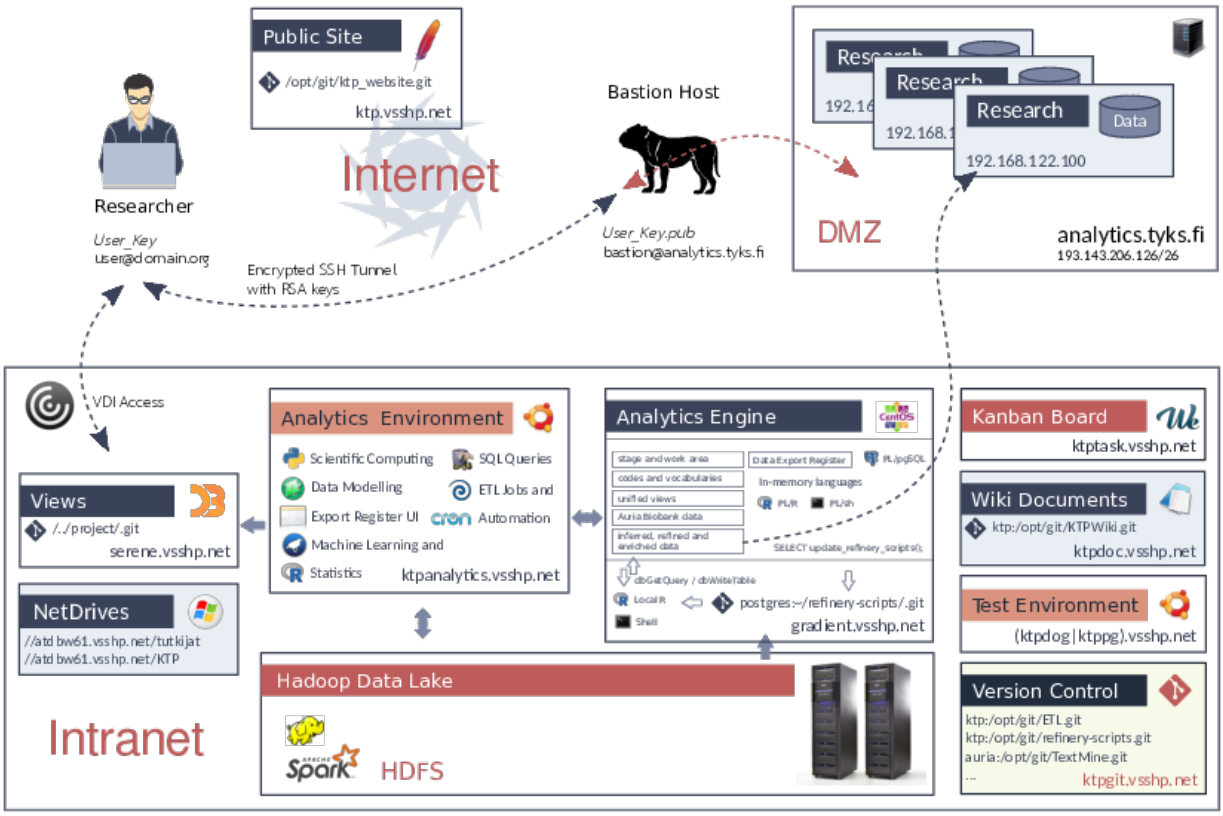

Figure 2. *Schematic of the VSSHP analytics environment solution, focused on researcher access to data (top). Note that this figure demonstrates a different part of the infrastructure stack compare to Figure 1.*

PSSHP - Kuopio

According to the information obtained from Sitra as well as HUS & Tieto, the PSSHP - Kuopio is contracting with Tieto Ltd. for the construction of their data lake and analytics environment. Then it is expected that the product PSSHP - Kuopio will get should be directly derived from the

HUS solution, even being constructed and maintained by the same staff as part of Tieto's new project. As such, there is currently limited information specific to PSSHP - Kuopio in this document. Any differences are likely to be most apparent in the administrative policies.

Part 1: Analytics environment

The analytics environment provides a place for researchers to use data. It must combine researcher access, data availability, computational power, and security. It is distinct from the data lake and it holds data pre-analysis. For research purposes, the analytics environment is the main interface and concern.

Intended audience

The needs of researchers looking at some small spreadsheets using Microsoft Excel are much different from the researches developing machine learning algorithms for big data analysis. In addition, users may have all sorts of skill levels.

HUS: The intended audience includes in-house business intelligence staff, doctors/medical thesis workers, and computational researchers. There is the possibility to extend to companies and external partners. The current limit on the audience comes from the number of available support personnel.

VSSHP: Their largest audience is in-house doctors/researchers, for example doctors who get Excel sheets and look at them. They do not have very demanding needs, and a network drive with the data and remote desktop are sufficient. They also support computationally intensive researchers via a different platform. Their solution is advertised internally (ktp.vsshp.fi), though there is not yet a concerted external marketing campaign.

Evaluation: Both districts have fairly similar goals, something which is usable by both small-data researchers and big data oriented computational researchers. In both cases, the limiting factor is availability of support staff and variable skills of researchers.

Software

All data analysis is done with software, and the availability of software will set the limits of what can be accomplished. In this and following sections, we evaluate this for two audiences: 1) "small data" researchers, who use tools such as Excel or stand-alone software packages. 2) computational scientists, who are more likely than not bringing in their own code and doing methodological development.

HUS: The environment is in the Azure cloud. 1) For small data, external connectors such as ODBC are provided, so that data can be extracted from the data lake and analyzed on their own computers. This provides simple connectivity for programs such as Microsoft Excel. Currently

no facilities within the lake are provided. 2) Researchers can get virtual machines (VMs) running Linux. Because these are in Azure, there is closer integration to the data lake itself, thus offering more possibilities for using big data tools directly. These come pre-installed with the necessary data tools, and any other necessary software can be installed.

VSSHP: 1) A Windows remote desktop server for clinicians allows them to run Microsoft Excel to examine small data. This is a simple and secure solution for their needs. Other standard software could be installed here, should it be requested. 2) A virtual-server based solution provides a raw Linux virtual machine image. In the work so far, data has been imported to a Postgres database. In the future, these VMs could be more tightly integrated to the data lake once it exists.

Evaluation: Both the remote desktop of VSSHP and the cloud VMs of HUS are ideal for their purposes. 1) The VSSHP remote desktop server provides a way for data to be analyzed within the environment, which is important. This should be contrasted with the HUS environment where small data is removed from the data lake. 2) For advanced use, both HUS and VSSHP provide virtual machines and necessary flexibility for practical usage. In principle, the HUS virtual machines with access to the actual data lake can be the most powerful option. Also see discussion on costs in the next section.

Hardware

While basic statistical research is relatively easy to perform, advanced research methods such as machine learning or deep data analysis requires sufficient computational resources (disk space, memory, CPU power, or even specialized hardware such as graphics processing units or GPUs for short).

HUS: The HUS analytics environment will be in the Azure cloud (which thus means that the data resides outside Finland, though in the EU). This means that acquiring hardware resources is in principle very easy: one simply pays for what is needed, and it can be immediately provided. Certain advanced hardware facilities, such as GPUs, must be available in the Azure region in order to be used. In particular, GPUs are only now coming to Azure and are not available in all regions. This may mean that deep learning type research is not yet possible. Cloud computing also raises the question of payment. Who will pay for the resources which are consumed? The normal cloud models have resources being turned on and off as needed to reduce price, but this may not be practical in the current setup without additional development. However, the HUS platform has also been designed flexibly, so that it can be deployed independent of the Azure cloud if that is needed.

VSSHP: Their initial solution uses physical hardware in a Medbit data center. This ensures that data stays under the control of Finnish organizations and in principle provides more flexibility as well. However, this does mean that one has greater management overhead, and the ability to quickly scale to larger problems is limited. It should be emphasized that this is a proof of concept and made to allow immediate use of resources, and a cloud solution could come later.

Evaluation: In short, the planned HUS system vs current VSSHP solution constitutes a trade-off between long term planning vs immediate usability. If the cloud has sufficient resources and matches the organization's security policy, it is likely better. However, it takes more time and effort to set up, thus slowing down the initial usage. Should the end usage be low, the simpler solution can provide a useful flexibility. We expect the VSSHP solution to adapt as development progresses. The cloud model creates a certain marginal cost for each research project. With local resources (current VSSHP), there is an initial investment and the desire is to get as many users as possible. With the cloud model, there must be continued funding which increases for each research project. These two models create different incentives which should at least be known.

Data flows

Data comes from some source devices, flows through the data lake, through some processing steps, and eventually to the analytics environment. The speed and ease of this process affects its usefulness.

HUS: Currently, the device service providers will do integrations to the data lake, which handles the initial data imports. All imported data is stored raw, and then there is a second processing step where data is anonymized and standardized. From here, data can be taken out to different users. HUS will also have metadata on the structure of imported data, which will allow a degree of automated processing of the raw data into secondary data.

VSSHP: Their data lake is still in the design phase, and thus less is presently known here. Currently, the KTP team is responsible for fetching data on request from doctors, processing it, and making it available to the users in the preferred form. Because the data lake does not yet exist, this is currently manual and data goes straight from source systems, to KTP's internal storage where it is cleaned, to the analytics environment for the user.

Evaluation: The data flows are mostly behind the scenes task. The HUS strategy is ideal and VSSHP has experience in accomplishing this.

Data linking

Data comes from a wide variety of sources, with different incompatible formats. In order to get the most value from the data, different sources must be linked.

HUS: The standardization process involves hashing identifiers using a consistent secret key. This allows all data to be linked. Regions are coordinating to have a consistent hashing strategy, so that in the long term all data can be linked.

VSSHP: The specialty of KTP is linking data from different sources, though their process is not automatic yet. Their skills will be very valuable once it comes time to automating data scraping. They are in communication with HUS.

Evaluation: Both organizations consider data linking. A standardized hashing scheme will allow data linking if implemented properly. The most important remaining consideration is approval for doing studies which link data, and this is being considered at a higher level. Currently, data is planned to be linked in a central platform managed by Statistics Finland. This platform is not evaluated here.

Access

In order for a technical solution to be useful, people must be granted access to it. This is primarily an administrative decision, but we attempt to evaluate it here. We also evaluate the technical system of access control.

HUS: Access is granted when researchers get research permits for accessing data. This involves evaluation by the ethics committee. Eventually, this may involve evaluation by THL for register research and data linking studies. When connecting, either hospital authentication or Suomi.fi provides strong authentication for outsiders. Other system can also be integrated. Currently there is not a plan for authentication of people outside of Finland, but they are considering this.

VSSHP: As with HUS, access is granted in accordance with hospital research permits. Currently the remote desktop solution uses hospital authentication, and access to researcher's virtual machines uses a two-layered ssh-key + password solution, with the initial setup done with some secure method. This also allows outsiders to get access.

Evaluation: The access is primarily an administrative decision, and there should be both an easy method for internal users and flexible method for external users. The planned strategies are suitable, but the exact implementation matters. A decision should be made about who the audience is (people within Finland, or anyone in the world) and consider how this fits in.

Permissions management

Once a researcher has permission to access data, this must be tracked and made active in the software. Ideally, this would be automatic, but in practice some manual work provides useful flexibility and a second check.

HUS: There is a HUS gateway where researcher permissions are registered. The gateway tracks users, permissions granted by research permits, and provides access to data lakes. Currently, this is a manual process, and prospects for automation depend on how structured the process becomes. Permissions can be assigned by user or group.

VSSHP: Currently, this is also a manual process pending a more structured method being developed. Permissions can be flexibly assigned, to specific users, group, etc.

Evaluation: A fully automated method for assigning permissions would be ideal, but would require structuring the process. Without knowing about the national permissions portal, it is difficult to make a full evaluation. For a relatively small amount of usage, a manual process could provide more overall flexibility and faster start-up time.

Documentation

In order for a system to be useful in the long-term, it must be documented.

HUS: They are currently making a user guide for researchers, however this is not yet finalized or available.

VSSHP: There is an internal wiki used for administrative documentation. For the analytics environment, researcher documentation is extracted from this and sent to users when needed. This documentation is in structured form, so that it can be compiled to a book for easy distribution.

Evaluation: Documentation is without doubt in a critical role to secure success and usefulness of any system. It should be as simple to maintain as possible, so that it will stay useful. We hope that documentation can be standardized and shared. Hopefully, user documentation can be made public, since that will provide users with direct information on using the system.

Automation

For long-term sustainability and freshness of data, ideally all steps of data processing or permissions management will be automated.

HUS: The final system will be capable of automatically converting the raw data to processed, standardized, and anonymized secondary data.

VSSHP: Currently, KTP's primary task is structuring the data. They are involved in different data standardization initiatives, thus are well placed for automation once the infrastructure is in place for this.

Evaluation: Both organizations have a goal of automatically processing data, and are involved in that process. However, fully automated data processing is a large goal, which requires much work and coordination. While goals are good, implementation remains to be seen.

Security

Any system containing medical data must have the strictest of security standards.

HUS: Currently, the small data (Excel-like) *is* planned to leave the environment. Researchers would connect their program to the lake and extract data to their local computer for processing. Larger data is examined on their own virtual machines without leaving the environment.

VSSHP: The primary model is that data never leaves the data lake. Small data is examined on their systems, through the remote desktop environment. Larger data is examined using the virtual machines.

Evaluation: Both organizations use good security measures. However, *the fact that HUS would allow small data to leave the data lake is quite unexpected*. If this should not be the case, it should be made clear that administratively this is not desired, and they can set up another solution such as that at VSSHP.

Standardization

Procedures and data formats should be standardized across districts, both for ease of use and for future combining of data.

HUS: Tieto aims to make the HUS solution a standard product that can be used across regions. In fact, the PSSHP - Kuopio is working to directly use the Tieto solution as well.

VSSHP: KTP is involved in different Finnish and Nordic data standardization processes. Their current solution is different from the HUS solution, but that is because Medbit has not yet constructed the data lake yet.

Evaluation: Both organizations do care about the standardization and are in communication with each other. However, we have so far heard about goals but the implementation remains to be seen.

Part 2: Data Lake

The “Data lake” is what actually holds all the raw data and makes it available for analysis purposes. Researchers do not have access to this: when they need data, the relevant pieces are extracted and made available through another interface, primarily for security reasons. The virtual machine based analytics environments may run the same software as the data lake, but are logically kept separate.

Technicalities

HUS: The data lake is produced with HDInsight on the Azure cloud within the EU but not in Finland. They intend to support more Hadoop-related software, for example Spark. Tieto is specifically making sure that their platform is portable to other deployment systems.

VSSHP: The data lake will also use the Hadoop stack, but from Cloudera. The stack is deployed within Finland in Medbit data centers.

Evaluation: Both organizations use the same platform. They use different distributions of Hadoop, however with proper use of APIs this should not be a major issue.

Data sources

For the platform to be usable, data must be integrated. This is a quite difficult task involving a lot of initial and ongoing work. In the end, this should be automated.

HUS: Data integrations are done by the data system providers under contract. They will integrate as many data sources as possible.

VSSHP: KTP has expertise in processing data sources, however Medbit will be the ones managing the final integrations. They report that each data source must be checked and valued, and then integration can be performed. We interpret this to mean that the customer (VSSHP/KTP) must request each data source. KTP has provided a list of initial important data sources.

Evaluation: Both organizations have a plan for data integration. HUS is optimistic about the variety of data sources. Medbit (VSSHP) reports that their integrations are done on request. We do not know who is making decisions about how many sources to integrate. Either way, data source integration is a continuous process and must be continually updated and improved.

Documentation

Documentation is needed, both of the system itself and the data within it. Metadata about the data can be used for automated processing of the data.

HUS: They will also collect structured metadata, which can be used for automated processing of the data. Documentation of the data lake itself (as opposed to researcher and user documentation) was not discussed.

VSSHP: This is currently under discussion.

Evaluation: Documentation is often the last thing considered in large projects. Long-term goals should be carefully considered at the highest levels, and it should be ensured that resources are available to make these goals possible. There is not currently enough progress to evaluate.

Portability

The solution of each region is ideally adaptable to other regions, and data and skills can be shared across regions.

HUS: Tieto intends for their platform to be portable and deployed to other customers, including other hospital districts. It is based on open source technologies. They are currently using Microsoft Azure, but the intention is to make the stack deployable on other infrastructure as well so that they are not bound to a single provider.

VSSHP: Their platform could also be used by other customers, should other customers choose Medbit.

Evaluation: Both organizations aim to make something re-usable. Whether it is reused is an administrative decision by other customers. There seems to be an understanding of the importance of portability and standardized interfaces, so we are confident in a good solution in the end.

Computation

Is computation supposed to be done inside of the data lake itself, or in the separate analytics environment? In general, it should be done outside of the lake, but for the largest tasks, it could be done inside of the lake. This requires a much higher threshold of security precautions in the code that is executed, and thus in all cases it would be decided on a per-case basis.

HUS: Data processing should be done outside of the lake, but if necessary could be done inside.

VSSHP: Data processing should be done outside of the lake, but if necessary could be done inside.

Evaluation: The situation is about as expected. With enough need and resources, analysis could be moved into the data lake.

Long-term plan

The environment is currently being developed, but what happens once the platform must be changed? How long is data expected to be kept for, and will it remain unchanged for that long?

HUS: Tieto intends to make this a long-term project with support. As the service provider, they will be responsible for migrating to the next solution when it comes out.

VSSHP: Medbit will do what the customers ask.

Evaluation: Since platforms are managed close to the customers, they should be able to do whatever is needed. It may be useful to prompt an administrative discussion on long-term plans, since it is primarily a non-technical decision. We did not see consideration of long-term data integrity (data scrubbing, 9s of reliability). Hadoop should have the necessary functionality, but it must be checked.