Governing the Amsterdam Innovation ArenA Data Lake

Finding the balance between innovation and control

S.E.J. Jeurissen MSc
is a consultant at KPMG IT Advisory and a member of the Enterprise Data Management team.
jeurissen.simone@kpmg.nl

N.L. Martijn MSc
is a senior consultant at KPMG IT Advisory and a member of the Enterprise Data Management team.
martijn.nick@kpmg.nl

Simone Jeurissen MSc and Nick Martijn MSc

Enterprises feel the need to create more value out of the data they are collecting, as well as the data that is openly available. Traditional data warehouses cannot support analyses of multi-format data, giving rise to the popularity of data lakes. However data lakes require controls to be effective, data governance is of utmost importance to data lake management. Amsterdam ArenA is an example of an enterprise that joined this movement, paving the way to the creation of a smart city.

Data-driven innovation

The collection of data is greater than ever seen before. Over 2.5 quintillion bytes of data are created every day ([IBM16]) and this number is rapidly increasing. In fact, the last two years alone have seen more data created than the entire history of the human race. In line with the rapid increase of data collection, data analytics has become an increasingly popular topic. Data analytics refers to the business intelligence and analytical technologies grounded in statistical analysis and data mining. Although increasingly popular, less than 0.5% of all data has ever been used and analyzed ([Marr15]) demonstrating that much of the potential is still untapped.

Organizations are not letting the potential value slip away, 75% of organizations either have already or are currently implementing data-driven initiatives ([GART16]) ([IDG16]). The aim of these initiatives is to increase operational efficiency, improve customer relationships and make the business more data-focused. However, this is only part of the potential, as business analytics generally only considers structured data that companies collect about their operations. Innovation is about thinking outside the box, ideally it would include more than structured internal data. The possibilities when combining different datasets of different formats are endless.

One such data-driven innovation application is the development of data-driven Smart Cities. Traditional cities are extremely inefficient in terms of waste. Smart Cities aim to better control the production and distribution of resources such as food, energy, mobility and water. This can be achieved through the means of data collection and analytics. For instance, real-time data about traffic can be used to suggest alternate routes for drivers or supply levels can be altered to better meet demand based on historical purchasing data. These are just a few examples of how data can increase a city's efficiency. Amsterdam ArenA, is an example of an organization that decided to join this movement. They have started to make use of the data that they and their partners have been collecting for years in new ways and switched their focus from the optimization of individual systems to the creation of effective network systems. This will lay the foundation for the creation of a Smart Stadium and eventually, a Smart City.

ArenA launched an initiative called the Amsterdam Innovation ArenA (AIA) that provides a safe, competition-free, open innovation platform where companies, governments

Organizations are not letting the potential value slip away, 75% of organizations either have already or are currently implementing data-driven initiatives

and research institutions can work together to make quick advancements and test smart applications and solutions. The stadium and its surrounding area serve as a living laboratory, a hotspot where innovations are tested in a live environment. Amsterdam ArenA has a data lake which stores a large array of data that is collected internally, this ranges from Wi-Fi location data, solar panel data, video camera data, and much more. They have also installed a data analytics platform. The platform allows projects to be carried out in data labs. Data sources are gathered from the lake and combined in these analytical environments.

Unfortunately, integrating a number of different datasets is more complex than it sounds, think about combining video camera data (unstructured data) with a table in an Excel sheet (structured data) for example. It poses risks to the ArenA as an organization, in terms of compliance to data privacy legislation, but also the misuse of data for purposes or analyses it was not intended for. Therefore, it is important to control the use of the platform, but with-

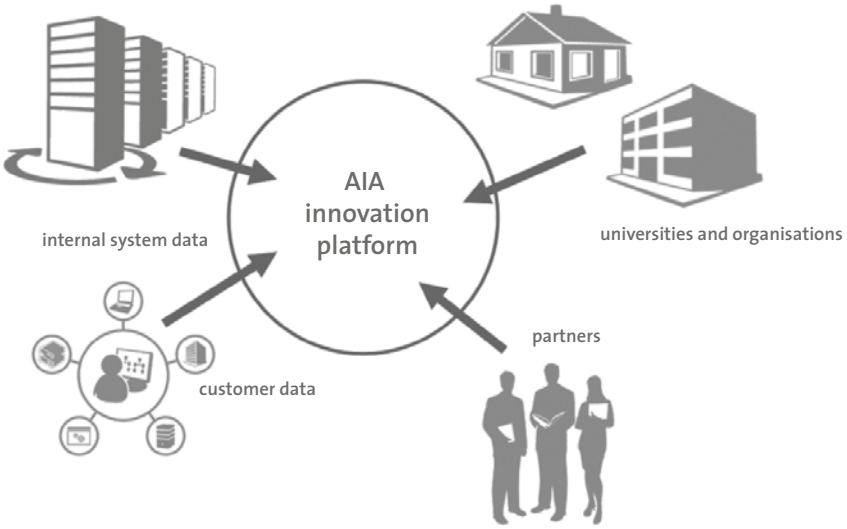

Figure 1. **Platform Scope.**

Data Warehouse, Data Lake: what is the difference?

The vast majority of collected data is of an unstructured nature. There are four main types of data; structured (formal scheme and data model), unstructured (no predefined data model), semi-structured (no structured data model) and mixed (various types together). Currently, only about 20% of all data is structured ([GART16]). Yet, traditional data warehouses only support structured data, meaning that the vast majority of collected data cannot be stored for analytical purposes. To resolve this, enterprises have begun using data lakes. A data lake is a storage repository that holds a vast amount of data in its native format. Neither the structure of the data nor its requirements are defined until needed. Unlike a traditional data warehouses, data lakes also support the storage of unstructured data types. In traditional data warehouses data is cleaned before it is stored, not having to do this when storing data in data lakes saves both time and money. Instead of having to clean all the data, analysts only have to clean the data that is relevant for their analysis. The costs of data storage are significantly smaller in data lakes as the architecture of the platform is designed for low-cost and scalable storage. However, there are two drawbacks to the use of data lakes, as they are still a relatively new topic the security standards of data lakes are not as high as those of data warehouses. Moreover, using mixed data formats requires experienced and skilled data scientists, who are often not present in the average organization ([Dull17]).

Topic	Data Warehouse	Data Lake
Data	Structured	Structured/semi-structured/mixed/unstructured
Handling	Processed	Raw
Storage	Expensive for large data volumes	Designed for low-cost storage
Agility	Fixed configuration	Highly agile
Security	Mature	Maturing
Users	Business professional	Data scientists

Table 1. **Data Warehouse vs. Data Lake: key differences.**

Ideally, anything would be allowed when analyzing the data in the data lake, however in practice this is impossible

out lowering the innovative value of the platform, the 'data playground'.

The trade-off between innovation and control

Ideally, anything would be allowed when analyzing the data in the data lake, however in practice this is impossible. On the one hand, enterprises should strive to store as much data in the lake as possible and let users have full freedom to innovate. Imagine combining real-time social media data, sales data and personalized promotions for example. The combination of this data would allow a firm to offer satisfied or dissatisfied customers (based on social media behavior) special promotions when sales are down. On the other hand, both data storage and user activity need to be controlled. There are legal mandates about the maximum storage time of specifics types of data. Video camera data may only be stored for a maximum of 4 weeks, and often even as little as 48 hours ([AUTP17]). When working with external suppliers, the data lake provider should also inspire confidence and reflect that they have control over the data lake. Data suppliers will not share their data on a platform where users handle data without restrictions. So how does one find the balance between innovation and control?

ArenA also faced this challenge when implementing their data lake and data labs. Implementing the data lake on an organizational scale, and allowing not only internal but also external users to make use of the data lake, poses risks to ArenA. Users should be given full freedom to stimulate innovation yet ArenA should maintain control over user activity to ensure data is used appropriately. ArenA also faced challenges considering privacy regulations, as part of the collected data is customer-related and saving it in its raw format infringes privacy regulations. To overcome these challenges, we developed *data governance* around

Overcoming related challenges: Amsterdam ArenA

Besides the trade-off between innovation and control, two of the most common challenges enterprises face when implementing a data lake is maintaining control of what is saved and finding the right people to carry out analyses. If everything is blindly saved in the data lake, data is simply being stored and never looked at again. Actually getting value from the data is the responsibility of the end user, increasing the risk that the data becomes a collection of disconnected data pools or information silos. This phenomena is also referred to as the creation of a data swamp rather than a data lake ([Bodk15]). Try to make use of it and you will drown. Data lakes require clear guidelines on what will be saved, data definitions and quality rules. Furthermore, carrying out analyses on a wide range of different data sources requires highly skilled analysts. An assumption which is often made is that data lakes can be marketed as an enterprise tool. It is said that if a data lake is created, employees will be able to make use of it, assuming that all employees have the skills to do so. The average company has a limited number of analysts or data scientists on their payroll.

The ArenA overcame these challenges when implementing their data lake. Firstly, the recruitment of skilled staff with knowledge of existing analytics methods and applications. Amsterdam ArenA overcame this issue by creating an open-sourced analytics platform. AIA does not rely on employees alone, as they have made the platform accessible to everyone. Due to the large variety of data and the innovative nature of the platform AIA does not need to make use of generalized data quality rules. Data preparation is the responsibility of the end user. In order to avoid creating a data swamp, Amsterdam Arena stores data in the data lake using metadata (such as date, content and event information), making it easy to find specific datasets quickly and combine datasets on an event-basis. Furthermore, only unique and interesting moments of video camera data are stored, scrapping the large amounts of valueless data (e.g. video footage of an empty stadium). The data analytics platform is built on top of the data lake, and data is only loaded into a project analytical environment is a project is initiated.

the data lake. This enables all (external) parties to become data-suppliers in a safe and reliable manner.

The need for data governance: finding the balance

Data governance is an overarching concept which defines the roles and responsibilities of individuals throughout data creation, reading, updating and deletion. Since data lakes employ a large array of data sources, clear rules must be laid down to control operations and to comply with legal regulations. With the establishment of increasingly strict laws in the privacy domain, companies must be especially mindful that their big data operations may not be compliant. Not only must companies be compliant, they need to protect themselves from possible future developments both within their firms as well as in the market. If something goes wrong, who is held accountable? Is sensitive personal information being analyzed? Who has ownership of the data? Who decides what data may and may not be used for? What controls are in place to ensure that data is used for the right purposes? Who deletes the data once it is no longer needed? Many questions arise when considering effective big data management. These questions can be answered by implementing data governance within an organization.

Implementing data governance

The first and fundamental element of governance is a virtual organization that defines the roles and responsibilities with regards to the handling of data. Depending on the size of the organization, a number of data related roles are defined. These are distributed over three levels; strategic, tactical and operational. Generally, all accountability and data strategy decisions are made at a strategic level and day-to-day decision making takes place at a tactical level. Daily operations such as data analysis and authorization management take place at an operational level. The strategic vision serves as a guideline for what new data is added to the data lake and which projects are carried out in the data labs.

Based on the defined roles, the data lifecycle process is defined step-by-step for every activity that takes place during the process. This ensures that all individuals with data-related tasks know what their responsibilities are and which activities they need to carry out. The making of such a process flow also clarifies where the potential risks in the process lie. This allows organizations to hedge against potential risks and implement the necessary controls to mitigate them.

At a tactical level, clear agreements must be made about the data and its use. This is important in order to create stable partnerships and inspire trust from external parties. Hence, data delivery agreements are made between data suppliers and data receivers. In this agreement the responsibilities of both parties are explained and agreed upon. The agreement also specifically defines both the expected content (attributes) and permitted uses of the data in question, which offers insight into the potential applications of the data and allows the organization to keep track of data and data attributes in the data lake preventing the creation of a data swamp.

At an operational level, the use of data must be controlled. Is data usage in compliance with the terms of the relevant data delivery agreement? For this reason, data usage agreements are signed by all users of the platform. To assess whether all users truly act according to the terms laid down in the data usage agreement, one of the roles in the data governance model is that of a controller, who is responsible for the continuous monitoring of user activ-

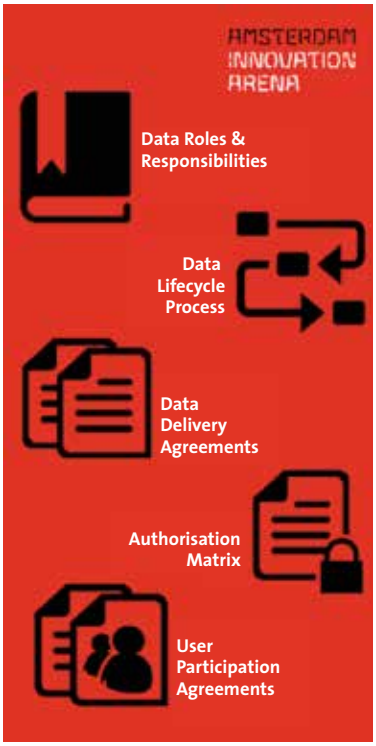

Figure 2. **Governance Documentation.**

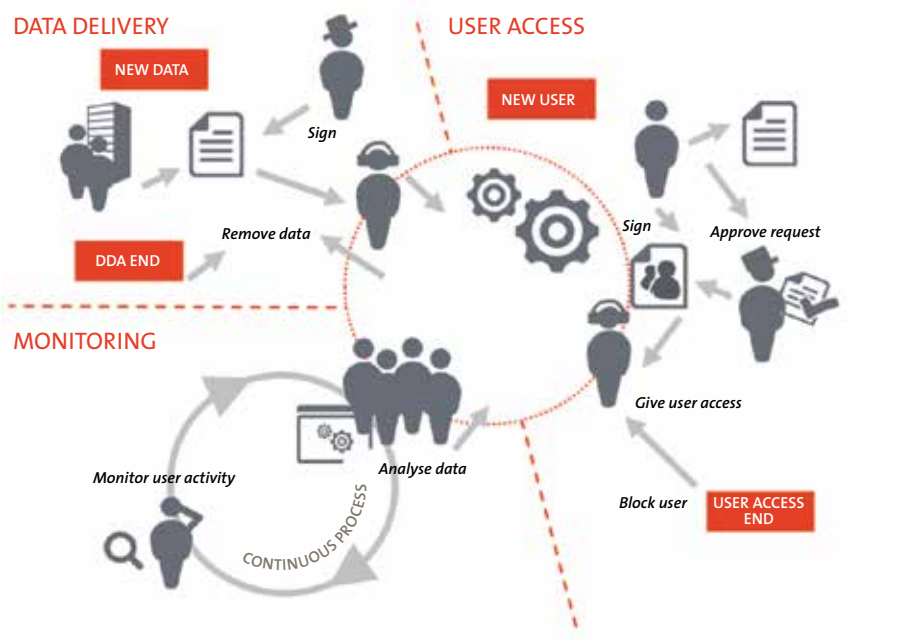

Figure 3. **Data Lifecycle Process.**

ity. An ideal analytical platform also logs all user activity making it easy to identify any wrongdoings.

Finally, there is one last control mechanism that is part of a data governance implementation; the authorization matrix. This document gives an overview of all available data and who has authorizations over the data. At any moment in time, the authorization matrix can be used to assess whether all necessary formal documentation reflects the current state of the platform.

The innovative governed data lake

The Amsterdam ArenA now has a running platform on which a wide variety of data is saved. From the data lake, data is periodically migrated to either the innovation platform or specific data labs for projects. These data labs are set up on a project basis when requests are made by external parties. The entire process is strictly governed and accurately documented through the means of data governance. Based on existing documentation, it is easy to assess in which stage of the process the various projects lie. During the implementation it became clear that both the data type and data transaction may influence what documentation is required. Data leaving the ArenA platform for example requires more documentation. Overall, partners feel confident in sharing their data with ArenA and many new partnerships are excepted in the near future. Similarly, there has been great interest in the platform from the user side, such as students who have been using the platform for university projects. A use case of one such project was that of students who analyzed purchasing data. They used the large food and beverage dataset to

Enterprises are creating more value from data, whether this is structured, semi-structured, unstructured or mixed

discover patterns in purchasing behavior at Amsterdam ArenA in relation to external factors such as weather, event type and customer demographics. ArenA's analytics platform is a safe innovation playground that will play a role in developing the data scientists of the future.

Concluding

Enterprises are creating more value from data, whether this is structured, semi-structured, unstructured or mixed. Combining all available information in a data lake creates innovative ideas and new insights that will add value to the business as well as society. With data comes knowledge, and with knowledge comes power. To avoid this power being abused data lakes require data governance. Implementing data governance allows enterprises to stimulate innovation within their firms without risking loss of control over user activity.

References

[AUTP15] Autoriteit persoonsgegevens, 2015, *Cameratoezicht op de werkplek* [online] Autoriteitpersoonsgegevens.nl, available at: https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/camera-toezicht/cameratoezicht-op-de-werkplek [accessed 13 Jan. 2017].

[Bodk15] R. Bodkin, 2015, *4 Ways to Avoid a Data Swamp* [online] Data Informed, available at: http://data-informed.com/4-ways-to-avoid-a-data-swamp/ [accessed 30 Nov. 2016].

[Dull17] T. Dull, 2017, *Data Lake vs Data Warehouse: Key Differences* [online] Kdnuggets.com, available at: http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html [accessed 4 Jan. 2017].

[GART16] Gartner, 2016, *Gartner Survey Reveals That 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years* [online] Gartner.com, available at: http://www.gartner.com/newsroom/id/2848718 [accessed 30 Nov. 2016].

[IBM16] IBM, 2016, *What is big data?* [online] IBM.com, available at: https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html [accessed 18 Dec. 2016].

[IDG16] IDG Enterprise, 2016, *2016 Data & Analytics Research - IDG Enterprise* [online] available at: http://www.idgenterprise.com/resource/research/tech-2016-data-analytics-research/ [accessed 30 Nov. 2016].

[Marr15] B. Marr, 2015, *Big Data: 20 Mind-Boggling Facts Everyone Must Read* [online] Forbes, available at: http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#624074d46c1d [accessed 13 Dec. 2016].

About the authors

S.E.J. Jeurissen MSc is a consultant at KPMG IT Advisory and a member of the Enterprise Data Management team. She advises enterprises in a variety of sectors in the development and implementation of data governance and information management (www.linkedin.com/in/simonejeurissen).

N.L. Martijn MSc is a senior consultant at KPMG IT Advisory and a member of the Enterprise Data Management team. He advises companies in various sectors in the development and implementation of data governance, document and records management and other data management measures (http://nl.linkedin.com/in/nickmartijn).