

Capitalizing on external data is not only an outside-in concept

Remi Verhoeven MSc



R.S. Verhoeven MSc is a senior consultant at KPMG IT Advisory, Enterprise Data Management. verhoeven.remi@kpmg.nl

Throughout most large and medium sized corporations data is generated and consumed within core processes. This has caused an increase in data on products, consumers and companies and has unfolded in an increasing availability of data generated by external sources. This availability has facilitated the rapid evolution of 'business as usual', but has also enabled service differentiation in the periphery. Whether external data is applied within the corporation (outside-in) or generated to be made available externally (inside-out), companies face unique data management challenges that often require tailored solutions.

Introduction

Corporations are integrating more and more data generated outside their company in their internal processes to increase efficiency, understand customers or gain new insights. This external data is becoming increasingly important when more and different data becomes available. However, very few organizations have standardized procedures to deal with external data. Even fewer organizations utilize the externally available data to its full potential.

C-level executives understand that concepts like analytics and process automation are essential for a successful business. Employees are already using data generated within the company in their daily work and their dependency on data will keep increasing in the years to come.

Due to upcoming technologies such as predictive analytics, machine learning and robotics, the corporate taste for applying data will further reduce human involvement in numerous business processes.

But this insight is not new. What is new, is that the value of data created externally is becoming more important than the data that you acquire from internal sources ([Wood12]). Data that is not captured internally can provide new

insights into customer behavior and preferences not previously accessible. This enables companies to further tailor their services, sales and marketing to those customers that are most likely to buy their products.

This is not the only method to capitalize on external data however. Companies can:

1. increase organizational or operational efficiencies;
2. use internal and external data in new business models;
3. directly monetize from making internal data externally available.

Increase Organizational Efficiencies

The most straightforward method to capitalize on external data is by integrating it within internal processes.

A case for external data integration that is becoming more common is using customer sentiment to improve the efficiency of marketing campaigns ([Fran16]). This typically starts passively by looking at the number of shares and likes a campaign generates on social media, but can go as far as real time interaction with customers and active participation in discussions to identify potential sales leads.

Corporations are integrating more and more data generated externally in their internal processes to increase efficiency, understand customers or gain new insights

A more internally focused example is product information. Retailers within the food industry integrate data on product labeling, product packaging and ingredients from their suppliers ([Swar17]). This not only reduces time at the moment of data entry, but also when the data is consumed further downstream, given the reliability and quality of this data is correct.

Besides this outside-in concept to capitalize on external data by increasing the bottom-line, recent developments change corporate perspective and turn data monetization inside-out.

New Business Models

If your company generates data or possesses proprietary information, you can consider how the data and informa-

tion you already possess can be reapplied. That staying close to your current business is beneficial is confirmed in an HBR study. The authors have found that top performing companies, across 62 industries, consistently expand their business on the periphery of their existing operations ([Lew16]).

There are three ways one can go about building new business models, by creating value added services:

1. from only internal data;
2. from only creating additional insights and reselling external data;
3. by combining internal and external data sources.

This is also shown in figure 1.

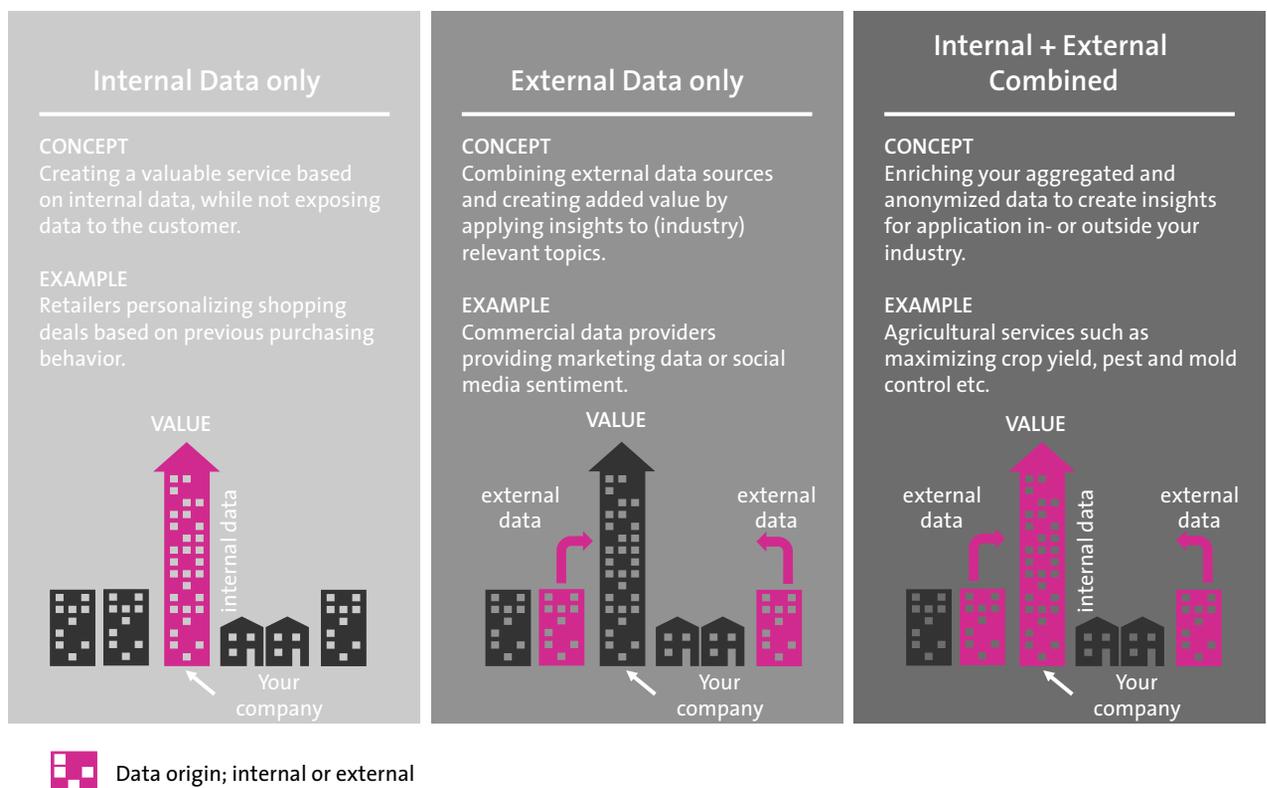


Figure 1. New business models can be constructed in one of three ways.

The most straightforward method to capitalize on external data is by integrating it within internal processes

Create value for customers only from internal data

A familiar example of generating value from internal data is Google. Google has always had a data driven business and has originally provided value-added data services from their search business. Since 2005, the company has been using their proprietary audience and machine learning data in their Google Analytics offering. Since March 2016, this offering has been expanded with an enterprise-class solution: Analytics 360, which adds functionality that enables personalized user experienced testing and much more.

Creating value for customers only from external data

The added value of businesses that capitalize only on external data (generally advertised as data providers) is the addition of structure through analytics. Transforming data to information is complex and not every company is able or willing to do this on its own. Companies like Experian, Acxiom and GNIP collect and aggregate data from social media, financial sources and open, publically available data and turn this into insight.

Through these value-added services, these data providers provide companies with the necessary information to increase the efficiency of their business. Examples include the evaluation of the borrowing capacity of clients, the likelihood estimation of a customer having any interest in a specific service, and the evaluation of customer sentiment on social media.

Combining internal and external data

Several of the large food and agriculture multinationals have acquired analytics and data driven business in the last few years. Presently they are competing with services that help farmers optimize crop yields. These services integrate proprietary data on their products' growth properties and externally available data such as historical weather patterns, soil properties and satellite imaging. Cargill's NextField DataRX, Dupont's Encirca Services and Monsanto's FieldView help farmers grow crops better and more efficiently and navigate weather shifts ([Bung14]). Moreover, Dupont's Pioneer company has developed mobile applications that help making farming easier and more reliable. These apps provide farmers with information on crop yield estimates, weed and pest control and even help farmers estimate the growth stage of their crops.

Direct Monetization

A specific new business model to generate value with your data is selling it directly to other companies. This sounds counterintuitive. Many companies' core business is collecting data on their customers, assets or users. And many suspect that selling this data could undermine their position. The key is to sell your data to companies that aren't your competitors, or aggregate the data sufficiently to ensure that true insights into your core business don't shine through ([Lew16]). Although not every company generates data that is of interest to other parties, we hope you get some inspiration from the following examples of companies which do.

UnitedHealth, an American health insurance provider, has created a business of selling aggregated claims and clinical information that it receives and generates from about 150 million individuals. UnitedHealth sells this data to pharmaceutical companies to give insights on how a product is being used, the product's effectiveness and its competitiveness. It provides data attributes such as, but not limited to, educational background, histology results of clinical exams, hours of self-reported sleep, tobacco or alcohol use and gender, ethnicity and age. This creates a model that can be easily adopted by similar companies, but hasn't found widespread application.

In January 2016, Telefonica announced they were launching a joint-venture in China to sell mobile consumer data. Besides their existing consumer base in several European and South American countries, Telefonica will now generate and sell anonymized and aggregated mobile network data on 287 million additional China Unicom users. The data is enriched with aspects such as social demographics, home and work location, modes of transport and other attributes, allowing sophisticated profiling. It is being used to find optimal locations for new store placement but also for safety initiatives such as crowd control.

Challenges dealing with external data

Although these examples of successful business models that incorporate or generate external data are inspiring, companies that adopt these models face many challenges.

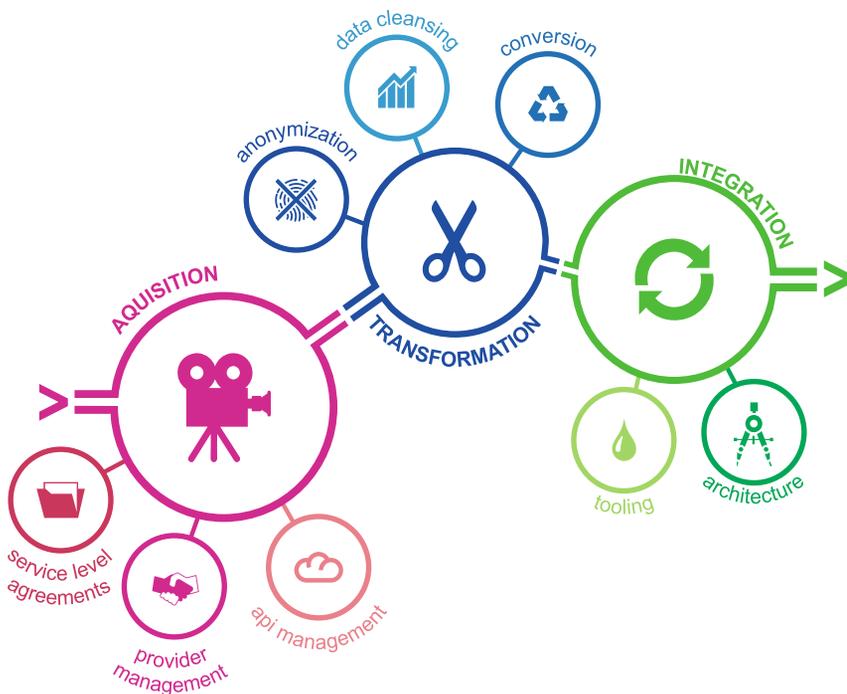


Figure 2. Different aspects contribute to the success of external data management.

These challenges are sometimes specific to external data, but are often generic inefficiencies that are resolved by good data management.

We can distinguish between these challenges in the phase of the data lifecycle they occur; 1) Acquisition, 2) Transformation and 3) Integration (ATI). This is similar to the Extract, Transform, Load (ETL) taxon, but broadens the scope slightly. Figure 2 contains a non-exhaustive overview of aspects involved in this process. This is similar to the Extract, Transform, Load (ETL) taxon, but broadens the scope slightly. Figure 2 contains a non-exhaustive overview of aspects involved in this process.

Acquisition

During the acquisition an organization should manage data acquisition concepts such as procurement, quality control and service level agreements (SLAs), but also interfaces with the data providers and authorizations within your organization. In corporations, external data is often incrementally introduced in different places in the organization. Local initiatives find solutions for the challenges they face. But these local solutions don't get implemented organization wide.

Not aligning these initiatives can be costly and are often not transparent to top level management because of their fragmentation throughout the organization. It is therefore important for top level management to lead a unified data management agenda to achieve alignment and leverage across the organization ([Logar16]). This not only covers data acquisition, but includes data governance, data quality and other organization-wide data management aspects.

Multiple people at your company have acquired the same data set

If one or more external data sets are relevant to a certain department in your organization, chances are high that the same data set is used in other departments for the same reason. Disregarding whether or not these data sets can be freely acquired, having to store, manage and update often large amounts of data multiple times is a costly affair. In order to combat this, a central data repository should keep track of all external data sets, interfaces, updates and authorized users within the company. Moreover, provider and API management should be implemented.

Dealing with different versions of a single data set

Not only duplicate data sets are a risk for an organization, storing and maintaining different versions of the same data set introduces similar complications. There are good reasons to keep different versions of a data set, such as to keep track of historic patterns. In this instance detailed version control is important to ensure traceability (or: data lineage) of your data through your companies processes. This is particularly important for regulatory compliance, but also for reporting entities to ensure consistency and prevent complications with aggregated data based on different versions.

Transformation

The transformational aspects of external data management should include data conversion efforts to unify data formats such as dates, numbers, addresses etc. with your internal data model. It should also include (automated) data cleansing for external data and anonymization of internal data when you are exposing data to the outside world.

Applying data set updates

Wilder-James estimates the amount of time spent on conversion, cleansing and previously mentioned acquisition activities make up 80% of the total time spent on data operations ([Wild16]). When dealing with a fixed portfolio

Successful integration depends on available architecture, technologies and desired data sources, but also heavily on strong data governance

of data sets, tooling can automate much of these tasks for each update of the data. This will significantly reduce effort on these currently manual tasks.

Automating alone does not solve this completely. Version control is also very important to ensure automation does not have adverse effects within the organization. In some cases, automated updating is not wanted. For example, when changes to the data are applied annually (vendor naming, model types etc.), a more frequent updating regime could bring unnecessary complexity and error.

Issues with data privacy

Data privacy is a hot topic ([A&O16]). With the newly adopted European General Data Protection Regulation (GDPR) in place by 2018, data subject consent and rights are aspects to apply to your data management processes. Selling aggregated and anonymized data outside the country it was collected in is becoming more complicated. More requirements are placed on asking subject consent and corporations have to ensure contractual documentation is adequate for GDPR standards.

Integration

Within the integration activities, the organization is faced with many technical challenges. Deploying a data lake or applying the data through direct (on demand) integration are complicated affairs. Successful integration depends on available architecture, technologies and desired data sources, but also heavily on strong data governance ([Jeur17]).

Reliability of external data

One aspect of data that is often overlooked is reliability. The reliability of data refers to a multitude of aspects of data that we deem important. For external data integration relevant reliability measures are a combination of the data's accuracy and appropriateness.

Accuracy is defined as the closeness from the data to its actual value. This is especially important for quantitative measures and should be taken into account when further calculations rely on this data. Sometimes, when dealing with external data sources, the data's accuracy is mentioned and can be taken into account during analysis.

Appropriateness is more difficult to quantify but can be a major factor in the reliability of the data as a whole. For

example, it plays an important role within the Solvency II regulatory framework for insurance and re-insurance companies. Within the context of external data, appropriate data sources should ensure that data corresponds to the goal it is intended for, it requires the data to be consistent with underlying statistical assumptions, and it requires the data to be acquired, converted and integrated in a transparent and structured manner.

The reliability of external data is difficult to pin-point accurately. However, during integration and application into analyses, keeping track of the data sources reliability can prevent issues later on. Imagine using a data set from a data vendor you have worked with for 5 years versus using a newly published data set from a not well-known university. The quality of these data sets might be the same, but their reliability could differ. Making c-level decisions based on the latter set might upset shareholders when the data set turns out to be unreliable.

We propose to keep track of data set accuracy and appropriateness throughout the data lifecycle. These aspects for each (critical) attribute can be documented in a data directory or similarly structured documentation. However, documentation alone does not achieve reliable results. End-users of data within your company should reflect on their analyses with such a data directory in order to ensure that their analyses are in their turn reliable.

Conclusion

Whether you are integrating external data in your processes, you are starting a new business by using external data, or you are selling the data you own to third parties, there is great potential in the application of external data.

However, throughout the life cycle of external data, either created by or applied within your organization, many challenges need to be tackled. Even if company-wide data management maturity is high, there might still be aspects unique to external data integration or creation that should be organized. Topics mentioned such as API and Provider management, automated data cleansing and conversion, and sufficiently anonymized data for publication, largely contribute to the success of an external data-driven business model.

References

- [A&O16] Allen & Overy, *The EU general data protection regulation*, Allen & Overy, 2016.
- [Bungr14] Jacob Bunge, *Cargill releases data-analysis service for farmers*, WSJ, 2014.
- [Fran16] Andrew Frank and Martin Kihn, *Marketing Data and Analytics Primer for 2016*, Gartner, 2016.
- [Jeur17] Simone Jeurissen and Nick Martijn, *Governing the Amsterdam Arena Data Lake*, Compact, 2017/1.
- [Lewi16] Alan Lewis and Dan McKone, *To get more value from your data, sell it*, HBR, 2016.
- [Loga16] Valerie A. Logan, Jamie Popkin and Mario Faria, *Gartner CDO survey reveals that chief data officers drive both data management and analytics for maximum impact*, Gartner, 2016.
- [Rijs17] Arno van Rijswijk, Sander Swartjes and Ruurd van der Ham, *Data Quality GS1*, Compact, 2017/1.
- [Wild16] Edd Wilder-James, *Breaking down data silos*, HBR, 2016.
- [Wood12] Dan Woods, *Do you suffer from the Data Not Invented Here Syndrome?*, Forbes, 2012.

About the author

R.S. Verhoeven MSc is a senior consultant at KPMG IT Advisory, Enterprise Data Management. He is an expert in the fields of data monetization, (asset) data management and data quality. He advises large and medium sized companies within several different industries on tactical and operational data management topics.

