Mikhail Lysyakov

Extracting value from Digital Data Streams: A case study in developing a DDS platform ecosystem

Abstract

The paper demonstrates the design of a value-generating platform ecosystem for harvesting and exploiting Digital Data Streams (DDS). It provides a general discussion of value creation through DDS and demonstrates some of the concepts with a case study. The case focuses on the use of DDS pertaining to public companies earnings calls. The contributions are: a general platform ecosystem archetype that could be used by researchers to implement DDS harvesting and unlock the hidden potential of data.

Introduction

The literature defines digital data stream (DDS) as, "the continuous flow of digital representations of events" (Piccoli & Pigni, 2015, p. 5). Examples include tweets, sensors data, and GPS location data. With increasing computer-mediation of human activities and the sensorization of physical objects (Yoo, 2010, p. 216), DDSs have become pervasive. By some accounts, humans create zettabytes of data every [day | month | year]. The genesis of these data is often in digital form in real-time as events occur (Piccoli and Watson, 2008). The Internet of Things represents perhaps the most visible recent example of DDS generation. "Two hundred billion things are expected to be connected to the Internet in five years, including about 150 million cars, 300 million utility meters, 100 million light bulbs, and 6 billion consumer devices" (Agarwal & Tiwana, 2015). The volume of real-time data has a high potential for value discovery.

DDS and proposed platform ecosystem

The literature identifies five DDS value creation archetypes (Pigni, Piccoli, & Watson, 2015). Each archetype represents a class of business models that modern organizations use to extract value from DDS. Firms adopting the *generation* archetype originate one or more data streams that other organizations may find useful. Twitter represents perhaps the best example. Organizations employing the *aggregation* archetype collect, accumulate, and/or repurpose existing DDS. Socrata represents an example. This company exploits Open Data and data.gov initiatives, aggregates data streams generated by government agencies, and makes them available to the public. Firms exploit the *service* archetype to provide new services or to improve existing ones. They do so by merging and manipulating DDSs. MyTaxi provides a new service by linking a passenger's transportation needs with a taxi cab's location. The *efficiency* archetype is used to optimize internal operations or track business performance. A telling example is UPS, which has optimization algorithms for efficient parcel delivery. And, finally, the *analytics* archetype's usage involves enhancing decision-making by producing superior insights. Dashboards, data mining and visualization tools could be used for that purpose. Semantria is an example a company related to that archetype. The firm performs sentiment analysis on DDSs that is integrated into the customers' decisions making processes.

In this paper we advance the design of a DDS platform ecosystem for research. The design follows the aggregate archetype and utilizes the DDS of public companies earnings calls. While this is a case study intent in showing the design of the platform, we

hope that researchers in both the academia and organizations will find value in the data we have aggregated and will contribute to the development of the ecosystem.

In the remainder of this paper we discuss the literature on platform ecosystems and apply this research to the specific case of the DDS of earnings calls.

"A software-based platform is the extensible codebase of a software-based system that provides core functionality shared by the modules that interoperate with it and the interfaces through which they interoperate." [A module is] "an add-on software subsystem that connects to the platform to add functionality to it (e.g., iPhone apps and Firefox extensions)" (Tiwana, et al., 2010, p. 675). A platform ecosystem is a "collection of the platform and apps that interoperate with it" (Tiwana, 2014, p. 10).

An example of an ecosystem is a Co-Created Digital Ecosystem (Brohman & Negi, 2015, p. 7) in which value is generated through a collaborative co-creation process between the firm and customers. The platform here acts as the operand (the resource to be acted upon) while the knowledge, experience, and skillset of the researchers is the operant (a resource that can act on other resources). The potential value in the system is only created once the operant operates on the operand (Vargo & Lusch, 2004, p. 2). The operand has minimal utility and value without the action of the operant on it. A Co-Created Digital Ecosystem encourages an active interaction, between customers and the firm whereby the customer plays a pivotal role in product innovation (Brohman & Negi, 2015).

DDS aggregation: the case of earnings calls as a platform ecosystem

In the context of digital data streaming, we propose that a platform ecosystem has elements of a co-created digital ecosystem (Figure 1). The researchers in an organization

research team (or researchers in multiple cooperating organizations or universities) play the role of the customers. A platform in our case is a core database that could be expanded. The database (operand) on its own has minimal value, as it has a limited functionality. The additional value for the database is generated through the creation of dynamic applications serving multiple requirements. The platform also includes a core filtering layer, a core harvesting layer, and a core monitoring layer, as well as a quality control block.

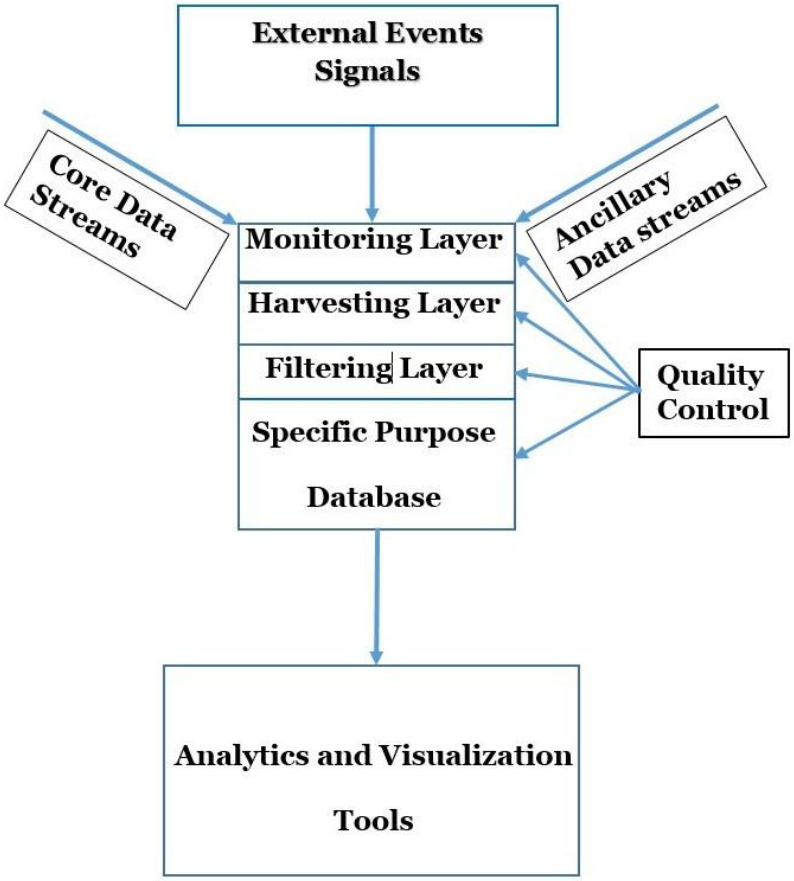

Figure 1. Proposed Value-Generating Platform Ecosystem

In our ecosystem the apps are the analytics/visualization tools that are interacting with the platform to generate a value. The interfaces are access channels or protocols by

which different software programs (RStudio, MySQL Workbench etc.) connect to the database.

First, we describe the process of a core platform creation followed by suggestions for future collaborations. High synergy of human – platform interaction is achieved when work is performed by a person in response to the relevant and interesting questions. So, the outcomes of the interaction depend on the specific business/research needs. After determining the business/research need(s), a research team should start harvesting relevant data streams (core data streams as well as ancillary data streams), taking into consideration the external signals from the relevant events. A monitoring layer manages and controls those signals. Collecting internal data streams (data that a company owns) is a much easier procedure as compared to harvesting external data streams. Because the latter process requires following specific signals from the external sources of data. Those signals are, for instance, subscription emails that are sent whenever new updated content is posted on the website, changes in the periods of regular updates of the websites with needed numerical (or textual) information, changes in the format of web pages etc. The changes in an external source might interrupt harvesting, so the platform should exploit a control system monitoring such signals (changes). Another possible way of collecting external data streams is through cooperation with other organizations that offer the service of data harvesting. Or the data should be obtained through some mutually beneficial data sharing partnerships with relevant organizations. So, the process of data harvesting will depend on the decision of an organization/research team leadership to be the full owner of the data, or have ownership partnerships.

After being harvested, the data streams should come through the process of filtering (cleaning) to make sure good quality data (with no duplicates, no incorrect values etc.)

enter the specific purpose database. Although an organization might have all internal data in its local system, the specific purpose database is a crucial element that combines all relevant internal and external data and makes them readily available for data mining, analysis and visualization. Additionally, the quality control mechanism is an important element of the platform ecosystem. Such control is performed at each stage.

At the monitoring stage control involves ensuring that monitoring is done correctly, and that it evolves as external events change. At the harvesting stage quality control procedures should guarantee that what should be harvested is harvested properly (correct format, correct number of data items etc.). At the filtering stage the filtering output should be checked for possible deviations from expected good quality data. At the database level the control should involve checking the proper data types' assignment, null values absence etc. At the first stage of platform development quality control will most likely be manual, but later the special procedures and codes could make it a semi- and fully-automated process.

The research team employing the proposed framework will be able to create a platform ecosystem for value-generating activities. Anticipation of some degree of a core platform expansion should be considered at a platform design stage. The specific purpose database will likely be inflexible in terms of expanding beyond its core capabilities, so the research question/business need should cover several areas to make sure that the scope of a platform ecosystem's value-generating activities is broad enough to provide valuable insights for a long period. Additionally, the project(s) may attract many researchers as shown in the case below. Moreover, the current research might give rise to new research ideas related to the platform ecosystem. In that case the platform

ecosystem is to be expanded by adding specific analytics/visualization tools to investigate those new areas.

Also, the researchers will likely develop transferable skills that will allow them to participate in the future projects. For that reason the proposed platform ecosystem could also be considered as a skills incubator and an ideas exchange system.

Analytics and visualization tools set expansion involves incorporating a new tool(s) into the existing platform ecosystem. For example, those tools are the visualization software programs, applications and websites designed to allow external users to be consumers of a value(s) generated by a platform ecosystem.

The case below shows the practical implementation and usage of the proposed platform ecosystem.

Platform database

In a platform ecosystem designed to support the aggregate archetype it is critical to organize the data for later use by apps connecting to it. Thus, we paid considerable attention to the design of the database. Figure 2 shows the proposed data model for storing Earnings Calls documents and other appropriate contextual information.

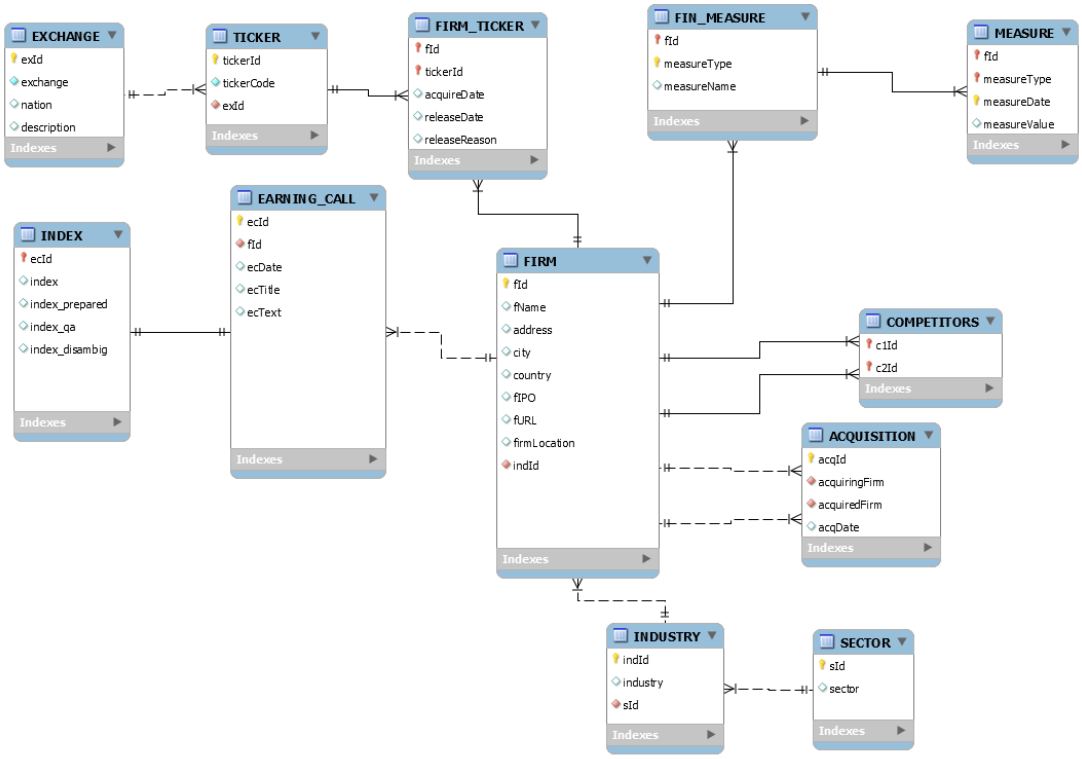

Figure 2. Proposed Data Model

Each entity and its corresponding attributes are captured inside rectangular boxes with gold key symbols indicating the primary key for each table (multiple keys indicate a composite primary key and red key symbols indicate primary keys that are also foreign keys). Solid lines indicate identifying relationships while dashed lines indicate non-identifying relationships between entities. Red diamonds denote foreign keys and blue diamonds simply indicate not-null designations (while white diamonds indicate attributes that accept null values). The database is a part of the infrastructure (Piccoli, Rodriguez, & Watson, 2015) that allows automatic harvesting of Earnings Calls from the Internet, pre-processing and cleaning of textual information to make it suitable for uploading into the database, adding ancillary data (like sector and industry data, competitors' data, acquisitions' data, financial metrics etc.), and then performing

analytics and data mining activities, including, but not limited to, sentiment index calculation, generating macro- and micro-level forecasts, topic modelling, data visualization etc. The research need driving the creation of such a database is the following – unlocking hidden values of textual information contained in companies' Earnings Calls. However, our model is extensible to any other form of corporate communication (e.g., letters to shareholders, press releases). Anticipated insights and first results of system exploitation are provided below.

Earnings Calls (ECs) have a number of valuable qualities: they are issued regularly, they are formal in nature, they have stable structure, and their content is detailed enough to provide a lot of interesting insights. Moreover, executives might not be fully aware that the language of Earnings Calls transcripts possesses much more hidden information than what they intended to share.

The database for the project allows storing harvested DDSs from multiple publicly available sources. Stored data allow performing advanced analytics to obtain valuable insights into the hidden value of formal textual information combined with other relevant data/metrics.

Mapping the relationship between the proposed platform ecosystem and the earnings calls database, the following important points should be noted:

- The external events signals are subscription emails sent daily by a source(s) of earnings calls.
- Once an email is received, the monitoring algorithm triggers the harvesting process. Core data streams in our case are earnings calls, while ancillary data streams are contextual data that are either static (like company address, location

etc.), or dynamic (like regularly updated financial metrics harvested from another source).

- The text of the earnings calls is filtered by pre-processing, which involves the following steps – extracting text, as well as ticker and date information from an html page, checking that the correct data format has been extracted, and uploading the data into the database.
- The quality control procedures check, for example, that the number of new earnings calls posted on an external source(s) corresponds to the number of earnings calls uploaded into the database. Otherwise, in case of differing numbers, the log is generated for further investigation by a researcher.
- Next, the analytics and visualization applications extract valuable information from the text of the earnings calls. That procedure, combined with contextual data, generates interesting insights shown below.

Valuable insights

One of the examples of current research is the creation of the P&W index (Piccoli, Rodriguez, & Watson, 2015) that has a potential to reflect economic sentiment of the USA (but does not have to be limited to a single country). The index combines macro- and micropotential, enabling interpretation and prediction of a microsentiment of a single company as well as aggregated macrosentiments for an industry, sector, country etc. Also, our team has separated Earnings Calls into two sections – a prepared statements section and a question & answer section. We hypothesize that the latter section is more truthful with less exaggeration. Current research will show if there is a

significant difference between two sections and which section is more accurately reflecting companies' sentiment and predicting major economic events.

Figure 3 shows the Quarterly Index for a 2 years period. Currently the P&W index is at the stage of additional validation. To validate it, researchers need to confirm its correlation with major economic indicators (like GDP), and other established economic indicators (like Michigan Consumer Confidence Index). Previous research showed promising results. Currently, the work is directed towards using a much larger sample of US companies (the number will go up to 6,000) as well as towards increasing the time period for Earnings Calls. We are working towards having a database with 15 years of Earnings Calls data.

Figure 3. Quarterly P&W index.

Source: https://pwindex.shinyapps.io/dashboard/

Another direction of current research involves developing a second index based on disambiguated word meanings. The P&W index uses the "bag of words" method - by matching specific key financial terms (positive and negative) from a financial dictionary

with the words in an Earnings Call. An example of that usage is the word “challenging”, which has a negative meaning and thus gets the value of -1. Positive words like “encouraging” receive a value of 1. The P&W index is just a reflection of the number of positive words versus the number of negative words. But you may see some disadvantages of current approach. For example, the word “capital” has multiple meanings depending on the context. If used as “something related to money”, the meaning is positive. But “capital gain” and “capital loss” are completely different combinations of words. Moreover, if the word refers to “a capital of some country”, then the meaning is neutral. But if the word phrase is “capital punishment”, the meaning is negative. The example above does not guarantee that the “capital punishment” might ever appear in an Earnings Call, but, nevertheless, the first two meanings are quite possible. In that case the “bag of words” approach will blindly consider the word “capital” as positive, neutral or negative (depending on the dictionary used) without relating it to the context. For that reason our research team is developing an index that will incorporate word sense disambiguation into the calculation algorithm and will also include decimal values assigned to word meanings. For instance, strongly positive words will get a value of 1, while “somewhat positive” words will receive a value of 0.5, 0.3 etc. The same approach will be used with negative words. The new index might be used separately, or might strengthen the P&W index if both indexes are used in combination.

One more example is the usage of the sentiment index to predict bankruptcy (work in progress). The micro-level bankruptcy index shows some promising preliminary results. The algorithm is based on the P&W index. The validation of the bankruptcy index involves several steps. The process starts with benchmarking of a company index with other companies matched by market capitalization, earnings and other financial

metrics. Additional validation involves industry benchmarking. The index for a company who will file for bankruptcy is hypothesized to show significant deviations from the industry average index and/or from the index of the companies similar to the bankrupt one but with healthy prospects. The difference should be especially pronounced as a company is approaching the quarter Zero, which is the quarter of bankruptcy filing. After validation the bankruptcy index could be used in combination with other financial metrics to show an early “red flag” some quarters before a company decides to file for bankruptcy. An interesting case will happen when official financial metrics show some enduring healthy performance, but the sentiment index starts to decline. Enron could be one of the companies to test that proposition. Our future research might address that case.

One of the other possible directions of our research is topic modelling and data visualization. Imagine that some companies have prevailing topics of discussion (or discussion of a product) for a specific week and a specific sector or an industry. An analyst may not notice that discussion is held among several companies. But text mining techniques applied to a big number of earnings calls have a potential to capture that relation. Our future website might provide that information to the user. There are many ways to visualize the topics/products discussions as well as most frequent words for the week, month, quarter, year etc. Interactive graphs will likely use pop-up additions and referral links to the appropriate content tailored based on the user input.

Overall, current research is expanding in many directions. The project has great potential to last for many years leading to many opportunities for contribution to knowledge.

References

Agarwal, R., & Tiwana, A. (2015). Editorial-Evolvable Systems: Through the Looking Glass of IS. *Information Systems Research*, *26*(3), 473–479. http://doi.org/10.1287/isre.2015.0595

Brohman, K., & Negi, B. (2015). Co-Creation of Value in Digital Ecosystems: A Conceptual Framework. *Twenty-first Americas Conference on Information Systems.* Puerto Rico.

Piccoli, G., & Pigni, F. (2015). Selecting Digital Data Stream Winners. *SIM Advanced Practices Council*.

Piccoli, G., Rodriguez, J., & Watson, R. T. (2015). Leveraging Digital Data Streams: The Development and Validation of a Business Confidence Index. *2015 48th Hawaii International Conference on System Sciences (HICSS)* (pp. 928-937). IEEE.

Piccoli, G., & Watson, R. (2008). Profit from Customer Data by Identifying Strategic Opportunities and Adopting the Born Digital Approach. *MIS Quarterly Executive*, *7*(3).

Pigni, F., Piccoli, G., & Watson, R. (2015). Digital Data Streams: Creating value from the real-time flow of big data. *California Management Review*, Forthcoming.

Tiwana, A. (2014). *Platform Ecosystems: Aligning Architecture, Governance, and Strategy.* Waltham, MA, USA: Elsevier Inc.

Tiwana, A., Konsynski, B., & Bush, A. A. (2010, December). Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics. *Information Systems Research, 21*(4), 675-687.

Vargo, S. L., & Lusch, R. F. (2004, January). Evolving to a New Dominant Logic for Marketing. *Journal of Marketing, 68*(1), 1-17.

Yoo, Y. (2010). Computing in Everyday Life: A Call for Research on Experiential Computing. *MIS Quarterly, 34*(2), 213-231.