ISSN:2231-0711

Claudia Loebbecke et al | IJCSET |May 2013 | Vol 3, Issue 5, 152-156

A Parallel Platform for Big Data Analytics : A Design Science Approach

Claudia Loebbecke[1], Joerg Bienert[2], Ali Sunyaev[3]

[1]*Department of Media and Technology Management, University of Cologne*
Pohligstr. 1, 50969 Cologne, Germany
claudia.loebbecke@uni-koeln.de
[2]*ParStream GmbH*
Große Sandkaul 2, 50667 Cologne, Germany
joerg.bienert@parstream.com
[3]*Departmentof Information Systems and Information Systems Quality, University of Cologne*
Pohligstr. 1, 50969 Cologne, Germany
sunyaev@wiso.uni-koeln.de

***Abstract*— Following a Design Science approach, at the core of this paper we propose a technically innovative parallel platform for Big Data analytics. The design of the proposed platform allows for analyzing and filtering billions of records, querying data structures with 1,000s of columns, getting answers in milliseconds without cubes, continuously importing data with low latency, and executing 1,000s of concurrent queries. Deploying the platform has empowered organizations across many industries to capture new business opportunities from better analytic quality of very large, close to real-time data. With our single platform design project, we hope to provide an interim attempt at theorizing [1] about achieving data quality and business opportunities from Big Data analytics.**

***Keywords*— Big Data Analytics, Parallel Platform Design, Design Science.**

I. Data Quality from Big Data Analytics: Challenges and Status

Big Data is a trend that is about to change enterprises fundamental, so that those players who fail to tackle the subject will go out of business. Digital data resources have exploded in the past decade (e.g. [2], [3]).For example, from 2008 to 2009, the digital universe has grown by 62% or up to 800,000 petabytes or 0.8 zettabyte; until 2020, the amount of data is expected to be 44 times as big as it was in 2009, reaching approximately 35 zettabytes. The organizational benefits of Big Data analytics undoubtedly depend on the resulting quality and suitability, i.e., the organization value, of the analyzed data (for approaches to assessing data quality see, for instance, [4] or [5]].

Today's data analytics applications require large and expensive IT and database infrastructures which consume vast amounts of energy and resources (e.g. [6]). In the future, increasingly more terabyte-scale datasets will be used for research, analysis and diagnosis. The demand for applications based on searching and analyzing large datasets such as price comparisons, call data records, or financial transactions will significantly grow in many industries including eCommerce, telecommunications, energy, social networks, and finance (e.g. [7] or [8]). Example applications in those industries include web analytics, SEO, affiliate nets in eCommerce; profiling, targeting, billing in telecommunications; smart metering, smart grids, wind parks in energy; ad serving, profiling, targeting in social networks; and automated trading, trend analysis, fraud detection in finance (e.g. [9], [10]). The growing demand for such applications based on Big Data analytics should cause remarkable market growth; so do [11] and [12] expect Big Data to be worth billions in increased sales and productivity across industries.

Conventional database technology is not engineered to meet today's analytics demand ([13]-[15]) and hence not able to fully exploit the potential of analyzing data volumes of up to 100 million records in real time (e.g. [16]). Architectures are often 20-30 year old; their data and index structures mostly employ sequential algorithms. CPUs are close to reaching their physical speed limits with respect to cycle time. Over the last decades, the performance of processors with a calculating engine has doubled each year in accordance with Moore's law. Nowadays, they are achieving only lower percentage increases.

Algorithmic procedures for indexing large amounts of data have seen relatively few innovations in the past years and decades. A commonly used product like MapReduce is not suitable for near real-time calculations. The much needed innovative parallel algorithmic procedures for indexing large amounts of data are only slowly becoming available [17].

Further performance enhancements will not only come from innovative algorithmic procedures. Another, perhaps even more important, driver will be the concurrent use of multiple processors on a single chip. Modern graphic cards with hundreds of processor cores on one chip – paired with ultra-fast memory – are the cutting edge of this technology. For instance, the leader in manufacturing graphic processors, NVIDIA, launched its 'Fermi' product line in 2009. The memory error-correcting codes built-into the product line can correct single bit errors automatically. This significantly increases the use efficiency of any multiple processor technology. As high-performance computers based on multiple parallel processors are increasingly deployed for 3D graphics applications and scientific simulations and analyses[a], the embedded specialized graphic processors have become more universal, even though not fully exploited yet – neither by commercial database suppliers nor in value creating user scenarios.

[a] Big Data analytics using high-performance computers are also increasingly deployed in humanistic and social science disciplines where they support respective claims as quantitative sciences and objective methods and help providing for quantifiable results in many social spaces [5].

II. A Design Science Approach for Platform Development

The objective of this research is to design and implement a framework for a technically innovative parallel platform for Big Data analytics. For designing our framework and the proposed platform (i.e., the artifacts that address the problem of Big Data analytics based on truly parallel processing), we adopt a design science approach following the seven guidelines promoted by [18]: Design as an Artifact, Problem Relevance, Design Evaluation, Research Contribution, Research Rigor, Design as a Search Process, and Communication of Research. For a discussion of the actual design of the artificial (see also Simon 1996 [19], Niederman & March 2012 [20]).

Our proposed parallel platform is an artifact which supports truly Big Data analytics (Design as an Artifact). It allows for searching and analyzing very large datasets in real time – a need which has become prevalent in many industries and is expected to reach new heights in the not so distant future (Problem Relevance). Several performance measures during real-life deployments serve as proof-of-concept to demonstrate the implementability and the value of the designed platform (Design Evaluation). The main contribution of this work lies in the innovative design of several artifacts which are molded together to a new parallel platform for analyzing Big Data in real-time. Further, this research points to innovative business opportunities based on 'fit-for-purpose data quality resulting real time Big Data analytics (Research Contribution). This research relies on rigorously gained insights from multiple academic fields. Both, the construction and evaluation of the artifact, are based on the knowledge base of computer science, data and information retrieval, business intelligence, and web mining (Research Rigor). During all stages of the platform design and development, critical feedback from potential and actual users led to many iterations and adaptations (Design as a Search Process). Sharing the resulting design with both practitioner audiences and the academic community has triggered additional design iterations – often related to specific real-life deployments (Communication of Research).

III. Platform Design

A. Overview

The proposed platform for high-performance Big Data analytics is based on using GPUs – in addition to CPUs – and a new index structure to execute queries. Figure 1 provides an overview of the platform design.

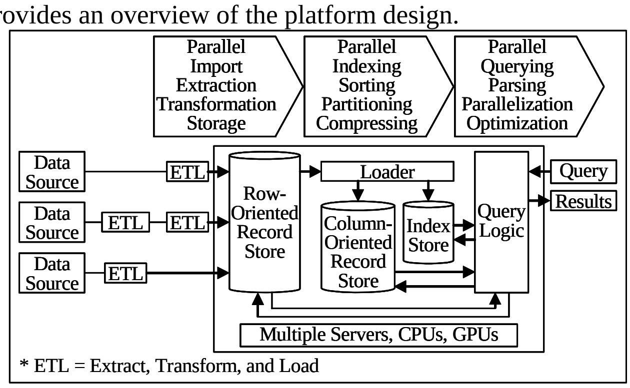

Fig. 1 Platform design overview, after: anonymous for review

The 'Extract, Transform and Load (ETL)' process supports continuous data import even with billions of records via three interface levels. The supplied data is read and processed in such a way that it can be forwarded to the actual loader. The platform then generates and stores the necessary index structures and data representations which allow for using GPUs and CPUs to index and execute queries. Input data can be stored in row and / or column-oriented data stores. This offers configurable optimizations regarding sorting, compression, and partitioning, which can be adjusted to specific analytic needs. Once data is loaded into the platform server, the user can pose queries over a standard interface (SQL) to the database engine. A parser interprets the queries and a query executor executes them in parallel – thus delivering close-to-real-time query processing. From the declarative request, the optimizer generates the initial query plan, which is used to start processing. The query plan can be altered dynamically during runtime; for instance, the level of parallelism can be changed. The query executor manages the use of the available infrastructure as it executes parts of the query plan on the GPU and other parts on the CPU. It thereby balances GPU and CPU usage at runtime. The query process exploits bitmap information wherever possible, so that logical filter conditions and aggregations are calculated and forwarded to the client via bitmap operations.

B. *Platform Elements*[b]

The *index structure technology* (see Figure 2), a column-oriented bitmap index using a unique data structure, enables efficient parallel processing in compressed form. It eliminates the need for index decompression as in other database and indexing systems. Thus the highly efficient query engine offers faster index operations, shorter response times, substantially less CPU-load, efficient parallel processing of search and analysis, and the capability of adding new data to the index during query execution. In the end the new index structure supports close-to-real-time importing and querying of data.

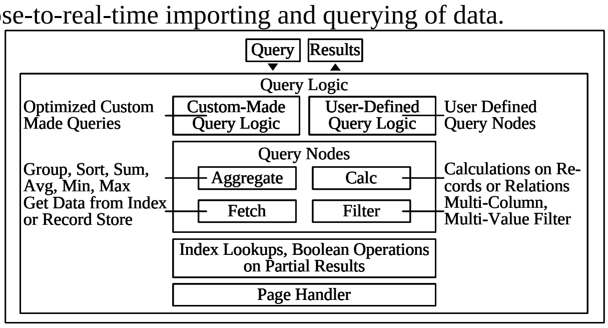

Fig. 2 Query processing design, after: anonymous for review

The *query processing* is based on the concept of query nodes. Each query node represents a specific operation, e.g., filter or aggregation nodes. The entire flow of data through the processing engine is represented by a number of query nodes that form a tree. The fetch nodes are found at the bottom of the tree. On the fetch level, the platform provides several users all available information to minimize the I/O load of the actual data fetching. It thereby makes use of filter conditions and index structures. Above the fetch level,

[b] After anonymous for review.

various operations such as aggregation, filtering, calculations, and sorting are performed on the data by the query nodes. A special query node transports data over network connections so that single queries can be distributed over multiple instances.

The query processing engine automatically analyzes data dependencies and parallelizes relevant parts of the query tree, so that the execution engine can execute all query nodes at maximum throughput and minimum latency. The support of modern graphic cards decreases the query response time. Combining GPU and CPU use in hybrid scenarios, i.e., using GPU boards for best performance on most used index data and CPU processors for less frequently used data, balances performance and cost.

The entire query process is completely transparent to both developers and users. Further, developers can write their own customer-specific query nodes and plug them into the execution engine. These customized nodes support the provision of customer-specific solutions, which can be implemented with minimum effort.

The *data import* of CSV files[c] runs at unprecedented speed due to the index structure and the parallel processing. It is possible to separate the import's CPU and I/O-load and move them to different machines. If preferred, the platform can operate on a CSV record store instead of using the more conventional column stores. Operating on a CSV record store accelerates the data import as only the indexes need to be written. Since one usually wants to save the original CSV files, no additional hard drive memory is wasted.

Three main *interface levels* enhance query options: At a high level, a JDBC driver[d] enables the implementation of a cross-platform front-end, so that the platform can be used with any application capable of using a standard JDBC interface. A Java applet can be built to query the server from within any web browser. At mid-level, queries can be submitted as SQL code or other descriptive query implementation (e.g., JSON format[e]). In the latter case, the results are sent back to the client as CSV text or in binary format. At a low level, a C++ API and base classes can be used to write user defined query nodes that are stored in dynamic libraries. Users can integrate their own query node into a tree description and register it dynamically into the platform server. Such user-defined queries can be executed via a TCP/IP connection and are integrated into the parallel execution framework.

IV. SECURING PLATFORM RELIABILITY AND SCALABILITY

A sustainable roll-out of new business opportunities built on the high quality data from close to real-time Big Data analytics requires continuous platform reliability and flexible platform scalability in the context of ever increasing data volumes and throughput needs.

Using an arbitrary number of servers, which all can be configured to replicate the entire or only parts of the data store, sustains *platform reliability*. Several load-balancing algorithms either pass a complete query to one of the servers or they pass parts of one query to several, even redundant servers fostering a load-balancing.

Finally, *platform scalability* up to petabytes is provided in several directions – upward, outgoing, or resource-oriented. A single node can contain anything up to four CPU sockets or eight GPU cards. All TCP/IP capable node interconnections and mixed node configurations are feasible. Depending on performance needs, reliability and ease of administration, data can be stored and distributed on local disc arrays or on a centralized storage system. While multiple nodes perform the index processing completely independently, only the minimal possible result sets are transferred for further processing. Data load is balanced over the nodes at run-time. Additional nodes can be added during run-time to meet new performance requirements. Multiple data copies can be stored on different nodes to increase the availability. Inside a single processing node or inside single data partitions, query processing can be parallelized by using all CPUs, GPUs, IO-Channels.

V. PERFORMANCE MEASURES GAINED FROM REAL-LIFE DEPLOYMENTS

The demonstrated performance of the platform has already attracted several 'Big Data' customers from different industries:. The proposed platform has repeatedly outperformed PostgreSQL, supposedly "the world's most advanced open source database" (www.postgresql.org), by a factor of 1,000 delivering results in sub-seconds on large data volumes.

For instance, an online search engine for travel offerings requested to deliver 100 travel offers out of one billion data records based on 25 independent (optional) filter criteria. Compared to the client's previous conventional DBMS X solution, the newly designed platform offered a 80 times faster response time: it required only about five megabytes compared to five gigabyte per query in the previously used DBMS X. Whereas the solution with the conventional database needed about 400 servers, the proposed platform ran on twenty servers, thus allowing for a significant decrease in infrastructure cost and power consumption compared to the conventional DBMS X solution.

As part of a new SEO initiative, deploying the platform brought data mining efforts to a new level: It allowed for joining together one billion records in less than one second. New market research potential arose from flexibly executing 5,000 queries with filtering and grouping more than 1,000 columns per online-analysis.

Harvesting business opportunities in web analytics, the platform supported the concurrent calculation of one billion user records in fifteen milliseconds – down from three minutes. Thereby it was up to 12,000 times faster than the established MySQL-cluster.

Increased analytic back-end performance was shown when the proposed platform supported global mining company in its in geo-spatial analytics as it supported filtering and geo-clustering three billion records in 100 milliseconds – in total scaling up to three petabytes.

---

[c] CSV (Comma-Separated Values) files are a common, relatively simple file format widely used by business and the scientific community.

[d] A JDBC (Java Database Connectivity API) driver enables a Java application to interact with a database. It connects with the database and implements the protocol for transferring the query and result between client and database.

[e] JSON (JavaScript Object Notation) is a data-interchange text format, that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages. It is rather easy to read and write for humans and to parse and generate for machines.

VI. DISCUSSION

Scientific research and the press agree that many enterprises will go out of business if they will not take advantage of Big Data analytics for bringing data quality to a new level. In this context, we proposed the design of a platform for Big Data analytics. The proposed, cost-efficient platform design offers an innovative index structure technology which supports parallel processing on parallel architectures. It uses GPUs in addition to CPUs for indexing and executing queries and obtains close to real-time analytic results through the simultaneous importing, indexing and querying of data.

Certainly, the data quality of the analytic results gained is difficult to measure ([21]-[25]). The value of the generated data depends on users' quality perceptions and on the use context, or – in other words – whether it is 'fit for use' or 'fit for purpose' ([26], [27]; on a broader scale see also [28]). In the case of the proposed Big Data analytics platform, the 'fit for purpose' lies where query filters have to use many columns in various combinations (ad-hoc queries) and truly complex queries are to be performed frequently. The commercial potential comes from balancing performance in the sense of new heights of commercially valuable data quality with infrastructure and operating costs (for extensive discussions of the value resulting from data quality see [29]).

Methodologically, with this work we followed a pragmatic Design Science approach as it has been extensively covered in the academic IS literature (e.g. [30]-[32]). In particular, we adhered to the seven guidelines for Design Science studies as suggested by Hevner[19]. We provided the architecture and various technical elements of a parallel platform for truly Big Data analytics *(Design as an Artifact)*. We underlined the problem relevance with the growing need and business potential of Big data analytics paired with needed technical solutions not yet being available in the market *(Problem Relevance)*. With early real-life deployments of the proposed platform, we demonstrated the technical performance of the proposed platform in meeting client needs and outperforming traditional database solutions and pointed new business opportunities resulting from better analytic quality of big data sets and *(Design Evaluation)*. Accordingly, with out study, we offer two contribution. First and foremost, we presented the design of a truly parallel, innovative platform for Big Data analytics. Secondly, we see outline business opportunities to be gained from generating state-of-the-art data quality from truly complex, close to real-time Big Data analytics *(Research Contribution)*. We reached our contribution by harking back to insights from multiple scientific fields, such as data base technology, computer science, engineering, and web mining on the rather technical side and business model design on the management side *(Research Rigor)*. Our design efforts have gone through numerous cycles and interactions in the labs and in the context of direct interactions with actual and potential users and clients *(Design as a Search Process)*. Finally, we undertake ongoing efforts to communicate the proposed design and related, newly gained insights both in academic meetings and in presentations for practice *(Communication of Research)*.

VII. IMPLICATIONS AND FUTURE RESEARCH

We believe that our study is timely and has several implications for research and practice. Our platform allows enterprises to exploit Big Data opportunities and move parallel Big Data analytics platforms into mainstream usage. Business managers and CIOs can deploy the platform to close to real-time Big Data analytics in a wide range of companies, business models, and contexts – aiming at ensuring data quality in complex query settings while balancing platform performance and infrastructure costs. While a growing body of literature recognizes the need and the potential of Big Data analytics (e.g. [33]), it has barely provided the design of a viable technical artifact for achieving data quality and harvesting the business potential of Big data analytics. Here we hope that our work offers a foundation for future work – both from a technical perspective and on the business side.

The focus of this work lay on proposing a newly designed artifact, here a truly parallel platform for close to real-time Big Data analytics. We have not developed any design theory, neither a rather general one (e.g., [33]-[38]). Instead, we adhere to [1], who stresses that single projects like ours can be of scientific value, but at best be interim attempts at theorizing and hopefully encourage further theory-building research efforts.

Future research may want to evaluate our platform design and compare its performance to alternative platforms for Big Data analytics. This can be done both in the labs and – wherever possible – in real-life implementations and applications. Such future research will not only help to further develop and improve the proposed design, it will also contribute to the generalizability of our design results and thus also serve as theoretical foundation for achieving data quality from Big Data analytics.

REFERENCES

[1] K. Weick, *Sensemaking in Organizations, Foundations for Organizational Science*, London, UK: Sage Publications, 1995.

[2] M. Chau, and J. Xu, “Business Intelligence in Blogs: Understanding Consumer Interactions and Communities,” *MIS Quarterly*, vol. 36, pp. 1189-1216, Dec. 2012.

[3] A. McAfee, and E. Brynjolfsson, “Big Data: The Management Revolution,” *Harvard Business Review*, vol. 90, pp. 60-68, Oct. 2012.

[4] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for Data Quality Assessment and Improvement,” *ACM Computing Surveys*, vol. 41, pp. 1-52, Jul. 2009.

[5] D. Boyd, and K. Crawford, “Critical Questions for Big Data,” *Information, Communication & Society*, vol. 15, pp. 662-679, Jun. 2012.

[6] A. Cuzzocrea, I. Song, and K. Davis, “Analytics over Large-Scale Multidimensional Data: The Big Data Revolution!,” *ACM International Workshop on Data Warehousing and OLAP*, Glasgow, UK, Oct. 2011.

[7] P. Hemp, “Death by Information Overload,” *Harvard Business Review*, vol. 87, pp. 82-89, Sep. 2009.

[8] C. Bizer, P. Boncz, M. Brodie, and O. Erling, “The Meaningful Use of Big Data: Four Perspectives – Four Challenges,” *ACM SIGMOD Record*, vol. 40, pp. 56-60, Dec. 2011.

[9] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, “The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature,” *Decision Support Systems*, vol. 50, pp. 559-569, Feb. 2011.

[10] B. Otto, Y. Lee, and I. Carballero, “Information and Data Quality in Business Networking: A Key Concept for Enterprises in its Early Stages of Development,” *Electronic Markets*, vol. 21, pp. 83-97, Jun. 2011.

[11] B. Brown, M. Chui, and J. Manyika, “Are you Ready for the Era of 'Big Data',” *McKinsey Quarterly*, pp. 24-35, Oct. 2011.

[12] J Bughin, J. Livingston, and S. Marawha, “Seizing the Potential of 'Big Data',” *McKinsey Quarterly*, pp. 103-109, Oct. 2011.

[13] K. Asanovic, R. Bodik, J. Demmel, T. Keaveny, K. Keutzer, J. Kubiatowicz, N. Morgan, D. Patterson, K. Sen, J. Wawrzynek, D. Wessel, and K. Yelick, “A View of the Parallel Computing Landscape,,” *Communications of the ACM*, vol. 52, pp. 56-67, Oct. 2009.

[14] A. Jacobs, “The pathologies of Big Data,” *Communications of the ACM*, vol. 52, pp. 36-44, Aug. 2009.

[15] R. Casonato, A. Lapkin, M. Beyer, and Y. Genovese, *Information Management in the 21st Century*, Stamford, CT: Gartner, 2011.

[16] N. Geri, and Y. Geri, “The Information Age Measurement Paradox: Collecting Too Much Data,” *Informing Science*, vol. 14, pp. 47-59, 2011.

[17] C. Boja, A. Pocovnicu, and L. Batagan, “Distributed Parallel Architecture for 'Big Data',” *Informatica Economica*, vol. 16, pp. 116-127, Jun. 2012.

[18] A. Hevner, March, S., Park, J., and Ram, S., “Design Science in Information Systems Research,” *MIS Quarterly*, vol. 28, pp. 75-105, Mar. 2004.

[19] H. Simon, *The Sciences of the Artificial*, 3rd ed., Cambridge, MA, US: MIT Press, 1996.

[20] F. Niederman, and S. March, “Design Science and the Accumulation of Knowledge in the Information Systems Discipline,” *ACM Transactions on MIS*, vol. 3, pp. 1-15, Apr. 2012.

[21] A. Aladwani, and P. Palvia, “Developing and Validating an Instrument for Measuring User-Perceived Web Quality,” *Information & Management*, vol. 39, pp. 467-476, May 2002.

[22] D. Ballou, and H. Pazer, “Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems,” *Management Science*, vol. 31, pp. 150-162, Feb. 1985.

[23] D. Ballou, R. Wang, H. Pazer, and G. Tayi, “Modeling Information Manufacturing Systems to Determine Information Product Quality,” *Management Science*, vol. 44, pp. 462-484, Apr. 1998.

[24] S. Knight, and J. Burn, “Developing a Framework for Assessing Information Quality on the World Wide Web,” *Informing Science*, vol. 8, pp. 159-172, 2005.

[25] Y. Lee, D. Strong, B. Kahn, and R. Wang, “AIMQ: a Methodology for Information Quality Assessment,” *Information Management*, vol. 40, pp. 133-460, Dec. 2002.

[26] D. Strong, Y. Lee, and R. Wang, “Data Quality in Context,” *Communications of the ACM*, vol. 40, 103-110, May 1997.

[27] R. Wang, and D. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *J. of Management Information Systems*, vol. 12, pp. 5-34, Mar. 1996.

[28] T. McLaren, M. Head, Y. Yuan, and Y. Chan, “A Multilevel Model for Measuring Fit between a Firm's Competitive Strategies and Information Systems Capabilities,” *MIS Quarterly*, vol. 35, pp. 909-929, Dec. 2011.

[29] S. LaValle, E. Lesser, R. Shockley, M. Hopkins, and N. Kruschwitz, “Big Data, Analytics and the Path from Insights to Value,” *MIT Sloan Management Review*,” vol. 52, pp. 21-32, Jan. 2011.

[30] A. Hevner, “A Three Cycle View of Design Science Research,” *Scandinavian J. of Information Systems*, vol. 19, pp. 87-92, 2007.

[31] S. March, and G. Smith, “Design and Natural Science Research on Information Technology,” *Decision Support Systems*, vol. 15, pp. 251-266, Dec. 1995.

[32] M. Sein, O. Henfredsson, S. Purao, M. Rossi, and R. Lindgren, “Action Design Research,” *MIS Quarterly*, vol. 35, pp. 37-56, Mar. 2011.

[33] H. Chen, R. Chiang, and V. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact,” *MIS Quarterly*, vol. 36, pp. 1165-1188, Dec. 2012.

[34] S. Gregor, “The Nature of Theory in Information Systems,” *MIS Quarterly*, vol. 30, pp. 611-642, Sep. 2006.

[35] S. March, and V. Storey, “Design Science in the Information Systems Research: an Introduction to the Special issue on Design Science Research,” *MIS Quarterly*, vol 32, pp. 725-730, Dec. 2008.

[36] J. Pries-Heje, and R. Baskerville, “The Design Theory Nexus,” *MIS Quarterly*, vol. 32, pp. 731-756, Dec. 2008.

[37] M. Markus, A. Majchrzak, and L. Gasser, “A Design Theory for Systems that Support Emergent Knowledge Processes,” *MIS Quarterly*, vol. 26, pp. 179-212, Sep. 2002.

[38] F. Mueller-Wienbergen, O. Mueller, S. Seidel, and J. Becker, “Leaving the Beaten Tracks in Creative Work – a Design Theory for Systems that Support Convergent and Divergent Thinking,” *J. of the Association for Information Systems*, vol. 12, pp. 714-740, Nov. 2011.