# A concept of enterprise Big Data and BI workflow driven platform.

Maciej Pondel
Wroclaw University of Economics
ul. Komandorska 118/120, 53-345
Wrocław, Poland
Email: Maciej.Pondel@ue.wroc.pl

*Abstract*—**This paper describes the author's concept for universal enterprise platform that allows to benefit from modern approaches to data analysis based on Business Intelligence and Big Data technologies. This idea is based on SOA architecture that enables workflow to coordinate the business processes and communicate domain applications with bespoke Business Intelligence and Big Data Solutions. Whole platform is designed the way that can be provided in a cloud environment.**

## I. INTRODUCTION

IN current times business requires scalable IT platforms supporting whole business processes from the beginning till the successful finish. Due to a huge complexity of the processes they require very often several domain applications like ERP, CRM, DMS and many more. In most cases those processes are handled manually and need human interaction between various IT Systems. Such approach leads to mistakes or inconsistencies in processes. Business Intelligence systems and Big Data solutions in current world provide essential information and knowledge for managers on the strategic or tactical level. The most important challenge defined in this paper is to take advantage of them also on the operational level of a business.

## II. BIG DATA AND BUSINESS INTELLIGENCE

There are several definitions of Big Data and Business Intelligence solutions. Some of them emphasize that Big Data is a modern concept replacing older approach defined as Business Intelligence [1], [2].

We can also find the approach stating that Big Data resolves new problems, brings new opportunities and meets new challenges that are different or complementary to the known Business Intelligence and Data Warehouses areas of usage [3], [4].

Author of this paper claims that both Big Data and Business Intelligence solutions meets similar problems when we look at it from the very general perspective. In details they have slightly different foundations and some problems could be better satisfied with multidimensional approach based on data warehouse and Business Intelligence solutions (OLAP, reporting, querying) and some problems are better satisfied with Big Data approach (NoSQL data structures, map reduce parallel processing).

We have to remember that Business Intelligence is no only a technology or a set of IT tools. It is considered a management strategy used to create a more structured and effective approach to decision making [5]. Of course, to fulfill this strategy, we need IT tools providing an access to data and capabilities allowing analysis and presentation of information. We can consider traditional BI based on ETL Process, data warehouses, data marts, OLAP, dashboards, scorecards and analytics. The Business Intelligence approach is still evolving. We can mention Business Intelligence 2.0 that includes following Features [5]:

- Proactive alerts and notifications.
- Event driven/ real time/ instant access to information.
- Advanced analytics.
- Enterprise Integration.
- Mashups and portal integration.
- Mobile/ Ubiquitous access.
- Improved visualization, Rich Interfaces (RIA).
- BI as a service (SOA and SaaS)
- In-memory analytics
- Open Source BI.

We can also meet (mainly on the BI Vendors websites) the term Business Intelligence 3.0. Author of this paper couldn't find a consistent scientific definition of the term. Different sources define Business Intelligence 3.0 as another evolution of BI systems that focus on the user interface, that is so intuitive to the regular user, that little or no training is necessary to explore data. We can list the following terms that appear when we consider this new approach: Data Discovery, Advanced Visualization, Visual Analytics, Business Discovery, Self-Service Business Intelligence. Example tools delivered by the top vendors are [7]:

- Visual Insight by MicroStrategy,

- Visual Intelligence by SAP,
- Visual Analytics by SAS,
- PowerPivot, Power View Power BI by Microsoft,
- Cognos Insight by IBM,
- Endeca by Oracle.

Also, it's important to highlight that there are two kinds of self-serve BI user:

- Analytics Power Users who create visual apps from multiple data sources – both internal and external.
- Regular Users that can fully explore the visual apps created by power users or IT.

Literature also mentions that BI 3.0 focuses on mobile devices. Mobile devices are here treated as the interface of data visualization but also as a source of the data together with other sensor-based Internet-enabled devices equipped with RFID, barcodes, and radio tags (the "Internet of Things") [8].

According to the quoted key factors of Business Intelligence 3.0 we can state that it is not a completely new concept but more development of Business Intelligence 2.0 with emphasized selected features. In some cases BI 3.0 concentrates on the same data sources as Big Data approach.

From the author's point of view the main difference between Business Intelligence and Big Data concept regards the following:

- Business area: Business Intelligence tools and Warehouses are concentrated on business transactions (sales, costs, money transfers) [13] and Big Data stores the records describing all the activities leading to the transaction or leading to the decision about transaction (tracking the customer path in on-line or off-line shop, social activity)
- Data quality – in most cases data in warehouses come from transactional databases that are consistent, stable and there is no room for any errors. In case of Big Data systems that are loaded very often with social content, by devices (Internet of Things) or server logs, the quality of data can be lower. In any case the lower data quality doesn't contradict the value of Big Data solutions. If there is an error in the data describing transaction in data warehouse it can result with the serious problems in financial / tax reports. The outcome of the missing or incorrect items the Big Data system usually do not affect the business so much (of course it the wrong data do not represent the majority of data).
- Data structure – the structure of data warehouses is based on relational structures and usually are denormalized multidimensional star or snowflake structure. Basing on such structure we can easily

build the OLAP cubes. Big Data system store data in non-relational databases.

### III. WORKFLOW BASED PLATFORM

Workflow systems are considered mainly as tools supporting business processes. A workflow application implements a business process model. The model describes the process steps to be performed to achieve a specific business goal, business rules for coordination of those steps and responsibilities of process participants[9]. The steps include tasks that should be performed by agents that can be human, computer systems or combination of both [10]. Workflow systems, with the benefits of efficient and flexible process modelling and process automation, have been widely used for managing business processes [11]. The main advantages of using workflow platforms are:

- Ability to model and deploy processes by a business user without engagement of software developers and deployment procedures.
- Ability to measure the performance of the processes and the whole business area related to the process.
- Full control over processes. Every step must be undertaken during the process and whole process (there is no possibility to forget some activity) has to be performed according to the business rules (no risky shortcuts allowed). Workflow makes processes less dependent on users mistakes and faults.
- Workflow can automate both business critical processes and also those back-office and automation makes processes to consume less time and resources.

Workflow activities execute programs that consume and produce data (parameters values and files). An output data produced by an activity can be consumed as input data to another activity, establishing a dependency relation between those activities [12].

### IV. CONCEPT OF BI AND BIG DATA PLATFORM THAT IS WORKFLOW DRIVEN

Being aware that steps in workflow produce some data it seems natural that they could be connected directly to the data storage such us data warehouse or big data repositories. Such approach can change the current well known and stable ETL approach (Extract Transform and Load).

Regarding to the consumption data by workflow activities we can assume that beside the data produced on the earlier stages of the process, the activity can be also fed with the data or information coming from data warehouses or even knowledge produced by a knowledge discovery modules of both BI and Big Data Systems.
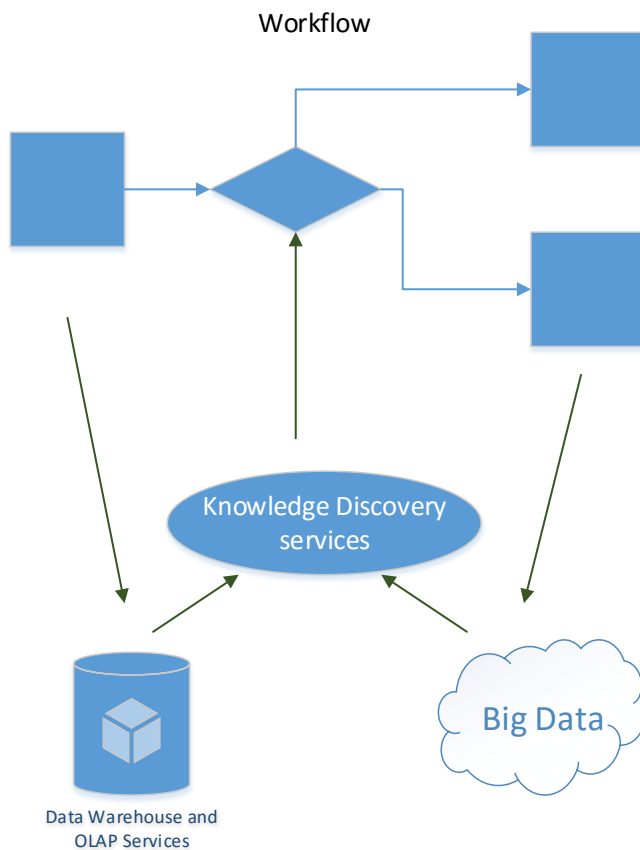
Workflow



**Figure 1. Concept of workflow and Big Data and Warehouse communication**

Some key items of the system's architecture are visualized on the Figure 1. There we can business process supported by workflow tool. Some steps (decision blocks) can use the Knowledge acquired in Knowledge Discovery module. Other steps (actions) can load the data into:

- Data Warehouse
- Big Data repository

The compound layer called Knowledge Discover services is working on the collected data and in this layer the Data Mining models are prepared. We can use well the known methods of data discovery like:

- Artificial neural networks
- Machine learning
- Naïve Bayes networks
- Collaborative filtering
- Clustering
- Association rules

and any other that can produce the valuable conclusions in the business environment.

The most important assumption regarding this layer is accessibility to its actions through SOA protocols like SOAP and REST. Regarding the actions it is meant mostly executing the models on the real object to gather the classification or assign the object to the cluster. Author assumes that through the service model we could also launch the training, testing or validation processes.

In SOA architecture we can efficiently send data for the technology different:

- Workflow platform
- Data loading services
- Knowledge discovery services

Example of a workflow driven combination of such activities can be the sales process in an example company. During the process, the customer goes the path (live or virtual) among various products and stops to watch some of them. The information about user focusing on the product and making decision regarding putting it to the basket or not is a valuable item to store in big data repository. After the user makes some particular activity, we can ask the conclusion model (bulid on previous customers behaviors) if it is efficient to present the customer directed proposal (or any other marketing message). If the recommendation of conclusion model is positive we can communicate with customer by:

- sending him message on his mobile (using sms, mms or push notification in the bespoke mobile app),
- presenting him the message on the screen visible in the market,
- make the shop assistant to approach the customer and ask some more information and present the dedicated offer.

We can imagine many more ways the store is communicating the customer to make the offer more suitable for him and optimize the income and turnover from the customer service.

Of course the areas of usage of such platform are not limited only to sales and marketing processes, although it is probably the most efficient and most obvious to use data mining based on big data and Business Intelligence modules there.

The important remark regarding the proposal of workflow using Big data and BI states that such workflow can automate nearly every process in the company (both Front Offices and Back Office as well). Workflow is able to communicate not only with Knowledge Discovery services but also with any application working in the enterprise environment that is build using Service Oriented Architecture approach.

## V. PROJECT OF PLATFORM IMPLEMENTATION IN ENTERPRISE ENVIRONMENT

Due to various business processes and the logic of existing IT environment in enterprises we cannot assume that the implementation of proposed platform may be done in a standard way and be processed easily. Of course whole platform will contain a set of coupled services, ready to cooperate, but they all will need to be integrated with enterprise data sources and configured to fulfill the business requirements.

The project of the platform implementation will contain the following stages:

1. Identification of business goals. The processes or business areas of efficiency improvement must be defined. We can imagine a whole variety of platform usage like:
   a. Sales processes
   b. Customer relation and loyalty management
   c. Logistic processes
   d. Back office processes
   And many more
2. Information and data sources identification. We should investigate the possible sources of data required to acquire sufficient knowledge useful in chosen processes improvement. In current approach the data may come from:
   a. The classic electronic sources like:
      i. IT systems databases.
      ii. Server logs describing usage of the IT tools.
      iii. Search engines' queries corresponding to the knowledge flows and users' needs.
      iv. Social media activities of internal users and external individuals (customers, suppliers, fans, competitors). Social systems are now full of interesting information. We can distinguish 2 ways of social systems:
         1. Open – like Facebook, Twitter, Youtube, LinkedIn and many more
         2. Internal – available only for internal users or intentionally invited externals – like Yammer, HighQ, eXo Platform and many more.
   b. The electronic sources emerging especially to feed the database of implemented platform. They can register activities of individuals that have not been registered yet eg.:
      i. Sensors identifying paths of customers movement around market or employees in a warehouse.
      ii. Barcode scanners, RFID tags / readers.
      iii. Mobile devices (tablet/smartphones) equipped with applications to register some facts that have not been registered previously (Customer assistant may register the conclusion of conversation with customer eg.

Customer has just bought a house and soon he will need pieces of furnishings or the logistics employee may calculate the time every activity takes and register that)
      iv. Stationary machines allowing customers to give a feedback or provide a satisfaction survey in a simple way.
3. Verification of individuals acceptance for registering the data describing individuals and their behavior. If customers feel surveilled they can avoid the services or the company that implemented platform. Internal employees who are aware of continuous tracking may feel strange and not work fully efficient. Another important aspect is validation of law and regulatory compliance. Personal data is protected by the law and we should verify if the data we gather can be processed without breaking the law.
4. Data load processes design. We can distinguish in here:
   a. ETL (Extract, Transform and Load) processes – mainly for data coming from traditional electronic sources.
   b. Services design (SOAP/REST) – mainly for data describing events that were when event occurred. In such approach the database is filled immediately

5. Data structures design that will include:
   a. Data warehouse structure in the form of multidimensional structure (star or snowflake). Basing on the structure we can build OLAP cubes structure or we can process a data mining models.
   b. Non – relational database that can be much more efficient while storing the behavioral data. We can distinguish in here:
      i. Column database like Cassandra or HBase
      ii. Document database like MongoDB or HyperDex
      iii. Key-value like FoundationDB or CouchDB
   NoSQLdatabases are more efficient for huge data sets processing and the performance of data loading is much higher.
6. Validation of the data designed structure especially it's adaptation to the business goals that were discovered on the earlier stage. Verification of the data quality must also be made.
7. To build the knowledge discovery services it is crucial to choose the correct data exploration

methods that fits best to business problem solutions. The huge topic is extraction of the algorithms parameters and their adjustment for models to generate the best possible results.

8. Since the main concept is based on the fact, that business processes using knowledge services and feeding them with data should be supported by workflow tool – it is important on the design stage of the project to model the business processes and automate them in the workflow software.

## VI.  PLATFORM STRUCTURE ASSUMPTIONS

In the current state the platform is technology independent. The necessary components like:

- Data warehouse
- NoSQL databases
- Services (SOAP/REST) protocols
- ETL tools
- Workflow Engines
- Data exploration algorithms

have implementation in various operating systems and software development platforms, that is why from today's perspective it is not crucial to select the final technology.

Figure  2 . presents  the example of the concept platform architecture basing on real life examples of use cases.
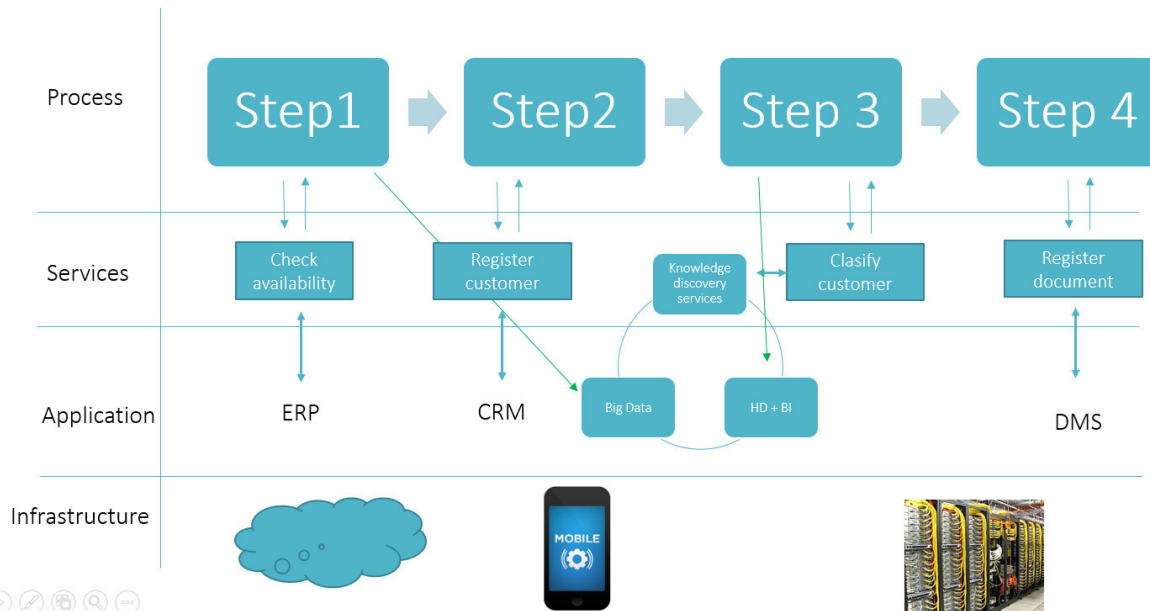


**Figure 2. Example of concept platform**

## VII. Challenges

Without a doubt, implementation of presented platform is a risky project and includes a lot of challenges. Among them we can distinguish:

- Performance – the knowledge discovery services and intended to respond on-line for process requests. If there is any significant delay – it may harm the efficiency of core business processes.
- Data mining models' relevancy – it is crucial to acquire valuable knowledge form data and provide it to the business scenarios supported by presented platform.
- Business efficiency – implementation of the platform is high cost affair. If we choose wrong processes or wrong methods of knowledge discovery and usage - the return of investment bay be not rewarding
- Stability of the platform - some of the technologies / solutions intended to be used in the platform are modern and in some cases may be unstable. It may influence the stability of the whole platform.

## VIII. Conclusions

The article presents modern approach for building enterprise platform, that could be used in majority of enterprises. The advantages of such approach are the expected high performance and most efficient usage of analytical platform (BI and Big Data) not only on a strategic level as management support systems but also on an operational level where business processes supported by knowledge can be most efficient from a business perspective. The future researches will concentrate on choosing example technologies and implementation of pilot atomic use cases to prove the correctness of proposed approach.

## References

[1] Tabakow M., Korczak J., Franczyk B., "Big Data – Definicje, Wyzwania I Technologie Informatyczne" In Informatyka Ekonomiczna Business Informatics 1 (31) 2014, Publishing House of Wroclaw University of Economics, Wrocław 2014

[2] D. Che, M. Safran, and Z. Peng, "From Big Data to Big Data Mining: challenges, issues, and opportunities," in Database Systems for Advanced Applications, pp. 1–15, Springer, Berlin, Germany, 2013

[3] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[4] J. Manyika, C. Michael, B. Brown et al., "Big data: The next frontier for innovation, competition, and productivity," Tech. Rep., Mc Kinsey, May 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[5] Nelson G., Business Intelligence 2.0: Are we there yet? SAS Global Forum 2010 (http://support.sas.com/resources/papers/proceedings10/040-2010.pdf)

[6] Pondel M.: Business Intelligence as a service in a cloud environment, in: Proceedings of the 2013 Federated Conference on Computer Science and Information Systems / Ganzha Maria, Maciaszek Leszek, Paprzycki Marcin ( red. ), 2013, IEEE, ISBN 978-1-4673-4471-5, ss. 1269-1271

[7] Cabrio B. What Is Business Intelligence 3.0?, Strategic Analytics Blog, http://blog.strat-wise.com/2015/03/what-is-business-intelligence-30_2.html , 2015

[8] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4 (2012): 1165-1188.

[9] Schmidt, Marc-Thomas. "Building workflow business objects." Business Object Design and Implementation II. Springer London, 1998. 64-76.

[10] Demeyer, Romain, et al. "Declarative workflows to efficiently manage flexible and advanced business processes." Proceedings of the 12th international ACM SIGPLAN symposium on Principles and practice of declarative programming. ACM, 2010.

[11] Liu, Xiao, et al. "Managing large numbers of business processes with cloud workflow systems." Proceedings of the Tenth Australasian Symposium on Parallel and Distributed Computing-Volume 127. Australian Computer Society, Inc., 2012.

[12] Chirigati, Fernando, et al. "Evaluating parameter sweep workflows in high performance computing." Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies. ACM, 2012.

[13] Alsqour, M., Owoc, M. L., & Ahmed, A. S. (2014, September). Data warehouse as a source of knowledge acquisition. An empirical study. In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on (pp. 1421-1430). IEEE.