Position Papers of the Federated Conference on Computer Science and Information Systems pp. 233–238
DOI: 10.15439/2016F584
ACSIS, Vol. 9. ISSN 2300-5963

Big Data solutions in cloud environment

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

Jolanta Pondel
University of Business in Wroclaw
Ostrowskiego 22
53-238 Wroclaw, Poland
Email: Jolanta.pondel@handlowa.eu

***Abstract*—Current business faces new challenges that require modern and adjusted IT models. Authors of this paper try to identify and indicate selected challenges that are addressed by by cloud computing concept and Big Data solutions. Authors of this paper concentrate od Microsoft Azure cloud offering mainly in area of Big Data and they want to prove that development of Big Data solutions in cloud environment can be efficient from financial and functional perspective.**

I. Introduction

Nowadays organizations are facing new challenges due to emerging business models. They operate on a market where competitors provide customers with more sophisticated and advanced services. Enterprises need to benefit from modern technologies if they want to reach or even overtake competitors. There is number of examples of companies that invented new service or product and seriously changed the way the market works. Examples of Uber, Netflix, Spotify, Airbnb or WhatsApp show that basing on communication technologies and efficient data processing and usage one can build business worth much more than the one working in traditional model. Of course market leaders are forced to continuously streamline their business basing on 2 main pillars: delivery of unique value to their clients and customers and controlling efficiency of business processes – mainly based on cost-cutting. Companies like Microsoft, Google, Amazon, Facebook are in a process of continuous improvement of business strategies, models and operational processes to keep their current position or expand the business.

II. Business and IT Challenges

There are several IT challenges that companies need to consider in current business. Among them we can distinguish:

- Innovation agility [1],[2] – understood as flexibility of IT architectures and business approaches to quickly deliver value, evaluate efficiency and market potential and if necessary scale up the innovation or withdraw with reasonable incurred expenses
- Interoperability and microservices architectures [3], [4] – which delivers IT architectures that are from IT perspective understandable, maintainable, scalable. Loosely coupled components are reusable. They collectively provide the complete functionality of a large software application. The services cooperate by exchanging data and information with other services without any human interaction. The services should be black boxes with precisely defined input parameters and output results. From business perspective such approach allows to efficiently create innovative services for clients that are composition of carefully selected microservices delivered by various vendors that we consider to be most efficient, proficient and capable. Various software applications are capable to cooperate because of interoperability between different programming languages, systems that allow integration of services. Varous types of resources are federated, what allows transparently mapping multiple autonomous resources to be treated by users as one federated resource.
- Mobile interfaces [5],[6] – modern requirements regard ability to complete interaction with software application using all possible devices. Users or customers spend more and more time in Internet using not only smartphones or tablets but also wearable devices providing discrete interfaces that allow control of mobile devices through subtle gestures in order to gain social acceptance. If company provides customer/user with the capabilities empowering use various possible devices and ways of communication with their service, it will be considered to be more

attractive but also impressive and efficient than others.

- Consumerization [6] of business activities expressed mainly in BYOD approach. It addresses the adoption of consumer devices and applications in the workforce. Employees bring computer tablets and smartphones into the workplace and harness social media applications and special purpose apps for their work lives. Such behavior if not properly managed may impact enterprise security but also can increase productivity of individuals and teams. The effects of consumerization are considered to be a major driver that redefines the relationship between employees (in terms of consumers of enterprise IT) and the IT organization [7]. Examples of consumerization we can observe in a phenomenon of existence iPhone and iPad in business scenarios, the usage of Facebook in companies and popularity of such application like Yammer or Hipchat that are based on Facebook. Another example are Google-like enterprise search indexing internal business content and documents in company that seem to be efficient in knowledge management scenarios.
- Internet of things measurability. The advancements and convergence of micro-electro-mechanical systems (MEMS) technology, wireless communications, and digital electronics has resulted in the development of miniature devices having the ability to sense, compute, and communicate wirelessly in short distances. These miniature devices called nodes interconnect to form a wireless sensor networks (WSN) and find wide ranging applications in environmental monitoring, infrastructure monitoring, traffic monitoring, retail, etc. [8]. IoT approach is not narrowed only to connecting things such as devices, machines etc. to the Internet, but it also allows them to interact with each other, exchange specific data, and complete some tasks without human interaction. Machine-to-machine (M2M) communication is nothing new, but this way of using sensors and wireless connection is revolutionary. Basic example of M2M operation is when sensors gather data, send it to a network either wirelessly or via cable connection, where its directed to a central server. We are able to create a wide range of IoT based business scenarios that give a real benefits to end users. We also have to be aware that all those signals and messages exchanged by IoT devices occur to be a valuable data that can be gathered and used, for example for decision making or enhancing operations. The amount of data generated by IoT devices is grooving rapidly. The figure 1 presents forecasts regarding number of connected devices in use globally. Assuming that part of them produce data that is worth storing and analyzing we can conclude that it is impossible to build efficient analytical system basing on relational databases and data warehousing technologies. Due to anticipated efficiency issues dig data technologies have to be involved in the processing of Internet of Things originated data.

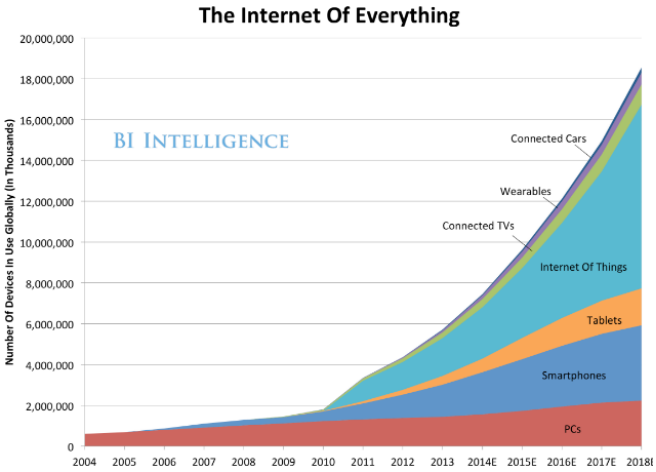

Figure 1 Forecasts for the entire Internet-connected ecosystem. Source: BI Intelligence [9]

Mentioned challenges are addressed mainly by two emerging IT concepts that are significant nowadays and affect business models. They are Big Data and Cloud Computing. Formally separable concepts in some areas are strongly related and their association can deliver real profits for users. Authors are aiming to present both concepts basing on an example of Microsoft Azure offering that is one of the most advanced platforms addressing mentioned challenges.

II. CLOUD COMPUTING OFFERING

Of course Microsoft platform is not the only existing. There are competitive platform provided by Amazon (Amazon Web Services), Google Cloud Platform provided by Google. Authors decided to mention those 3 because according to Gartner's report called Magic Quadrant they are worldwide leaders. Magic quadrants evaluates solution in 2 dimensions [10]:

- Ability to Execute - Gartner analysts evaluate technology vendors on the quality and efficacy of the processes, systems, methods or procedures that enable IT providers' performance to be competitive, efficient and effective, and to positively affect revenue, retention and reputation. The following criteria were evaluated: Product/Service, Overall Viability, Sales Execution/Pricing, Market

Responsiveness/Record, Marketing Execution, Customer Experience, Operations.

- Completeness of Vision - Gartner analysts evaluate technology vendors on their ability to articulate logical statements convincingly about current and future market direction, innovation, customer needs and competitive forces, as well as how they map to Gartner's position. The following criteria were taken into consideration: Market Understanding, Marketing Strategy, Sales Strategy, Offering (Product) Strategy, Business Model, Vertical/Industry Strategy, Innovation, Geographic Strategy.

Gartner prepared separate quadrants in the following categories:

- Infrastructure as a Service (IaaS) is a type of cloud computing service; it parallels the infrastructure and data center initiatives of IT. Cloud compute IaaS constitutes the largest segment of this market (the broader IaaS market also includes cloud storage and cloud printing)[10]. Figure 2 presents IaaS Quadrant.
- Platform as a Service (PaaS) is a cloud computing model that delivers applications over the Internet. In a PaaS model, a cloud provider delivers hardware and software tools -- usually those needed for application development -- to its users as a service. A PaaS provider hosts the hardware and software on its own infrastructure.[12]

Figure 2.Magic Quadrant for Cloud Infrastructure as a Service, Worldwide
Source [10]

Figure 3. Magic Quadrant for Cloud Platform as a Service, Worldwide
source [11]

Regarding IaaS offering Amazon Web Services is considered to be worldwide leader. Microsoft Azure is located on the 2nd place and Google on 3rd. In terms of PaaS offering Salesforce is acknowledged to be a leader. PaaS model is only delivered by Salesforce that is why the company was not mentioned by authors as cloud provider. Microsoft and Google have 2nd and 3rd place.

The aim of authors is not to evaluate the proficiency of every cloud provider. Microsoft Azure was chosen because of its existence in the cloud providers rankings and because of other research works conducted by authors in that platform.

The main advantages of using cloud services boil down to:

- Savings in terms of initial expenses on infrastructure and platform creation. Cloud services require payments for the resources that company really utilizes. In most cases there are no upfront costs assigned to hardware and software licenses purchase.
- Scalability - If application needs a large computing power for relatively short time the cloud can provide resources dynamically what is financially efficient. It is also available to allocate a large amount of resources in a relatively short time what is impossible in regards to on-premises environment.
- The operational costs are reduced because they are spread over a number of clients. In most cases it is more efficient to subscribe cloud services then to provide independently such constituents like energy, air conditioning, renting premises, security, maintenance, networking, management and many others.
- Wide range of well standardized services that can enrich capabilities of planned solution. In MS Azure the number of new services increased by 500 during

one year. Of course some of them are available in a preview edition (what means they have not been fully stable yet) but majority of them is available to use and they are relatively easy to integrate with other solutions.

There are also some boundaries related with cloud services that company needs to accept before starting usage of cloud services:

- Standardization – public cloud providers have to standardize their services in order to be able to maintain their quality. If clients defines individual requirements because of some special resources demands or he needs individual services or contracts – cloud vendor most probably will not be able to fulfill client's requirements.
- Limited SLA – in most cases cloud providers offer reasonable SLA (MS Azure offers from 99,9% to 99,95% SLA depending on service and chosen architecture). If client requires higher value – in most cases it is impossible to assure such quality.
- Distance between data and their user. If we store data in Local Area Network we have a high speed access. If data are stored in cloud we reach them through Internet connection that in most cases is slower than LAN. If we have a problem with Internet reliability we can be isolated from our data or applications.

Authors do not mention the security as a boundary of cloud offering because all leading vendors addressed the issue and are able to assure security on much higher level than it is possible in on premises infrastructure.

Indicated advantages in most cases address the challenges mentioned earlier. Cloud eliminates entrance barriers that supports innovation agility. A number of available, well standardized services support the benefits of interoperability and microservices architectures.
Scalability is curtail in every Internet of Things solution. We cannot predict precisely the final number of connected devices and amount of produced data that is why such solutions require a flexible resource allocation what is beneficial from performance perspective.

III. BIG DATA ECOSYSTEM OVERVIEW

As it was mentioned current business generate and possess a huge amount of data created during their operational activity. The data is stored in various IT systems in miscellaneous formats and very often in different locations. Continuous development of Computer Sciences and Information Technologies is increasing the level of knowledge concerning methods of designing and creating databases and processing the data that give us new possibilities and opportunities in acquiring information and knowledge increasing the business performance.

Big data concept is not aimed to deliver a business critical transactional systems. It is more about performing analytics.

Recently big data and big data analytics have been used to describe the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data [13] . The challenge of big data is to manage a large volume of data with optimal processing time [14].

Describing Big Data architecture we have to start from MapReduce concept is one of the most important techniques, and currently the preferred choice of cloud providers, for providing cloud-based data analysis services [15]. This platform invented by Google allows distribute processing on large number of computers. It is basing on 2 steps. Mapping is a step splitting the complex task into subtasks and reduce is about aggregation of results to combine them into one answer. Map function processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function merges all intermediate values associated with the same intermediate key [16],[17].

Basing on a MapReduce concept Apache™ Hadoop® platform emerged. The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules[18]:

- Hadoop Common: The common utilities that support the other Hadoop modules.
- Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

Basing on Apache Hadoop interesting projects are developed like:

- Cassandra™: A scalable multi-master database with no single points of failure.
- HBase™: A scalable, distributed database that supports structured data storage for large tables.
- Hive™: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- Mahout™: A Scalable machine learning and data mining library.

- Pig™: A high-level data-flow language and execution framework for parallel computation.
- Spark™: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.

IV. BIG DATA IN MICROSOFT AZURE

Microsoft is an active contributor to the Apache Software Foundation development effort. Their engineers works, committing code and driving innovation in partnership with the open source community across a range of Hadoop projects. Microsoft understand benefits of contributing Apache Hadoop development because they include most important Hadoop Components into MS Azure scope.

Most important Azure service called Azure HDInsight is a 100% Apache Hadoop-based service in the Azure cloud. It offers all the advantages of Hadoop, plus the ability to integrate with Excel, your on-premises Hadoop clusters and the Microsoft ecosystem of business software and services [19]. Azure HDInsight includes the most important modules and components like:

- MapReduce
- Pig
- Hive
- HBase
- Storm
- Spark
- RServer.

Azure also includes 3rd party solutions basing on Hadoop concept. They are:

- Cloudera Enterprise Data Hub
- Hortonworks Sandbox on Azure

Why to implement Hadoop in the cloud? Deploying Hadoop on-premises requires hardware and skilled Hadoop experts to set up, tune and maintain them. Cloud service possess preconfigured platforms that we can launch in minutes without up-front costs. Most big data tasks require individual tasks of data processing and after it is finished we can shut down the platform and stop incurring expenses.

If we want to initiate Hadoop cluster in Microsoft Azure platform we have to specify:

1. Type o a cluster. We have to choose from: Hadoop, Storm, HBase, Spark and R Server (2 latest existing in a preview edition). We also need to define Operationg system (Linux or Windows) and verision of chosen software. Confoguration of a cluster type is presented on figure 4.

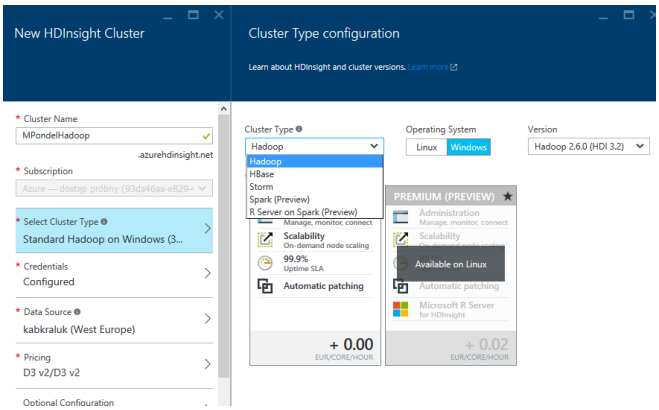

Figure 4. Configuration of cluster type
Source: own work on Azure Portal

2. In a next step we have to provide credential – admin password and an name and password of a remote desktop user
3. We have to indicate the storage account for the cluster
4. We have to choose the pricing tier. We specify the number of nodes (computers in cluster) and their parameters. This step is presented on a figure 5.
5. We can also define Optional configuration regarding:
    a. Virtual network
    b. External Metastores
    c. Script Actions
    d. Linked Storage Accounts
6. We should also allocate our cluster to resource group what determines the localization of servers.

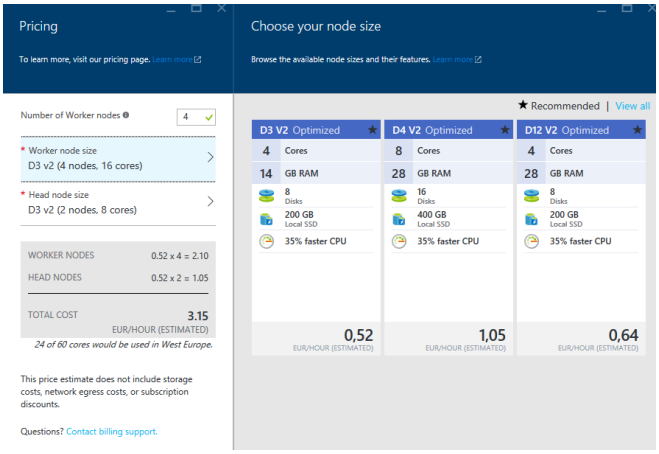

Figure 5. Definition of a pricing tier.
Source: own work on Azure Portal

After initiation of a process crating platform we had to wait about 20 minutes for Hadoop cluster to be created. We can connect using remote desktop and perform operations.

CONCLUSIONS

New business challenges require modern and adjusted IT models. Some of challenges are addressed by cloud computing concept. In modern business we need analytical tools that are able to perform analytical tasks efficiently. Big

Data solutions can support those business needs. Authors of this paper wanted to prove that building Big Data solutions in cloud environment in some cases can be more efficient and give additional value supporting enterprise business.

REFERENCES

[[1] Boyer, M. J., & Mili, H. (2011). Agile business rule development (pp. 49-71). Springer Berlin Heidelberg
[2] Wilson, K., & Doz, Y. L. (2011). Agile innovation: A footprint balancing distance and immersion. California Management Review, 53(2), 6-26.
[3] Maciaszek, L. A. (2008). Building Quality into Web Information Systems. In WEBIST (1).
[4] Pondel, M. (2013, September). Business Intelligence as a service in a cloud environment. In Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on (pp. 1281-1283). IEEE.
[5] Rico, J., & Brewster, S. (2010, April). Usable gestures for mobile interfaces: evaluating social acceptability. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 887-896). ACM.
[6] Harris, J., Ives, B., & Junglas, I. (2012). IT Consumerization: When Gadgets Turn Into Enterprise IT Tools. MIS Quarterly Executive, 11(3).
[7] Niehaves, B., Köffer, S., & Ortbach, K. (2012). IT consumerization–a theory and practice review
[8] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645-1660.
[9] BI Intelligence, Business Insider Research http://www.businessinsider.com/the-internet-of-everything-2014-slide-deck-sai-2014-2
[10] Leong, L., Toombs, D., & Gill, B. (2015). Magic Quadrant for Cloud Infrastructure as a Service, Worldwide. Analyst (s), 501, G00265139.
[11] https://azure.microsoft.com/pl-pl/blog/microsoft-the-only-vendor-named-a-leader-in-gartner-magic-quadrants-for-iaas-application-paas-cloud-storage-and-hybrid/
[12] http://searchcloudcomputing.techtarget.com/definition/Platform-as-a-Service-PaaS
[13] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.
[14] Business Opportunity Detection in the Big Data
[15] Ranjan, R., Georgakopoulos, D., & Wang, L. (2016). A note on software tools and technologies for delivering smart media-optimized big data applications in the cloud. Computing, 98(1-2), 1-5.
[16] Schutt, R., & O'Neil, C. (2013). Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.".
[17] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
[18] http://hadoop.apache.org/
[19] https://azure.microsoft.com/en-gb/solutions/hadoop/