BIG DATA FROM THE CLOUD COMPUTING AND SOCIAL MEDIA: A NEW FRONTIER

Research-in-Progress

Hak J. Kim
Hofstra University
134 Hofstra University, Hempstead, NY 11549
hak.j.kim@hofstra.edu

Abstract

With the development of cloud computing and smartphones, social media today is popularized as a new communication platform and becoming a new way of life to the people. Social media have created the tidal wave of data, which is so-called Big Data. This paper discusses Big Data issues in the cloud computing and social media, and introduces IBM Cognos as one of typical Big Data Analytics tool. We also demonstrate a dashboard of the US financial companies to monitor social media activities, including Facebook, Twitter, and YouTube. The future research will analyze Big Data (unstructured data) in financial sector collected from social media (i.e., Twitter and Facebook) using the IBM Cognos system, and then compared to traditional financial data (structured data).

Keywords: Big Data, Cloud Computing, Social Media, Business Analytics

Introduction

For several decades, Internet has been explosively growing and generating tremendous benefits for our world. As pointed out by Werbach (1997), Internet was fundamentally different from other communications technologies (i.e., the traditional telephone network). Since Internet has open and flexible architecture, it could provide the endless spiral of connectivity; that is, any form of network could connect to and share data with other networks through the Internet. As a result, the services provided through the Internet are separated from the underlying infrastructure to a much greater extent than with other media.

Cloud Computing (Harmer et al., 2009; Hayes, 2008; Milojicic, 2008; Weiss, 2007) is emerged as new generation of business infrastructure environment. Different from the traditional wired and client/server-based system architecture, this platform consists of wireless and cloud-based system environment. It supports new business models, such as user-driven purchase and click install on any device. It also creates new service deployment models by enabling lower total own cost (TCO), scalability and short time-to-usage.

People can communicate and interact with anyone, anytime, and anywhere using smartphones. Smartphones (i.e., iPhone and Galaxy) with their rich application support are one of the fastest growing fields in mobile industry. Unlike traditional cellular phones as a communication tool, today's smartphones are used for sharing information (i.e., social networking and geographic location services) and enjoying entertainment (i.e., games and sports), which is called '*Infotainment*'. The wide adoption of smartphones has opened new opportunities to business organizations, driving innovation in business. As a consequence of these initiatives, the business firms will experience better productivity and increased efficiency.

Another area of distinctive growth in IT ecosystem is *Social Media* (Cusumano, 2011; Häsel, 2011; Violino, 2011; Foster et al., 2010; Beckman, 2010; Hathi, 2009). It is becoming a new way of life to the people, such as multi real-time access behavior. With this mobile cloud computing and social media tools, the world is changing and becoming more intelligent and interconnected. These phenomena have become a revolutionary driving force for the development of new digital era, which is called *Big Data*. People are becoming to enjoy *Intelligent Digital Life*.

Revolution of Computing Platform

Modern enterprises are heavily relying on information systems (Sturdevant, 2011). They introduced a mainframe system in 1960s to the 1970s which is a timesharing system to serve many connected terminals with large and powerful data processing systems. In 1980s, personal computers (and workstations) were connected each other, which is called '*networked PCs'*, but still they were communicated within the company using private networking software. In 1990s, Internet-based enterprise information systems were introduced. The employee could use their enterprise information systems through the Internet regardless of geographical distance. Rather, web-based standards and protocols were embedded in the enterprise systems.

Recently, we have seen the emergence of new enterprise information system platform which is called '*Cloud Computing Platform*' (Cusumano, 2011). This platform uses the concept of *Grid* (Kurdi et al., 2008; Abramson et al., 2002) which is to build virtually a supercomputer to connect many networked computers and then to aggregate resources (i.e., CPUs, storage, power supplies, network interfaces, etc) for utilizing them collectively. Cloud computing has been made possible by the shift to Internet technologies that are built on Web-based standards and protocols. Figure 1 shows the comparison of architecture among mainframe, client/server, and cloud computing platforms.

For last decades, the client/server architecture (Abdul-Fatah, 2002) has been the main architecture of the Internet. This client/server is built on the distributed environment. Recently virtualization technologies are introduced in server systems to create a virtual form of operating systems, storage devices and network resources. There are different levels of virtualization, such as users, applications, processors, storages, and networks. Virtualization allows multiple accesses to different devices by users. It is like one computer controlling other machines by consolidating information to improve efficiency. The model is to build a virtual environment in which different partition resides on one hard drive and this in turn manages the virtual machine.

Cloud Computing (Mell and Grance, 2011) is a form of virtualization which involves data outsourcing with no up-front cost and provides just-in-time services. Cloud Computing is a model for enabling ubiquitous, convenient, on-

demand network access to a shared pool of computing resources that can be rapidly provisioned with minimal management effort or service provider interaction. It provides resources over the internet on demand and eliminates the cost for in house infrastructure. The key drivers for cloud computing are bandwidth increase in networks, cost reduction in storage systems, and advances in database.

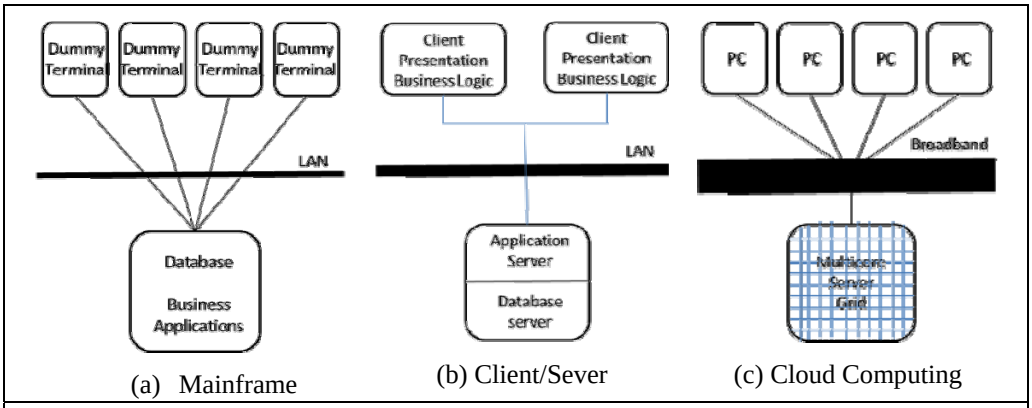

Figure 1. Comparison of Computing Platforms

As shown in Figure 2, Cloud Computing has three typical types of business models (Sotomayor et al., 2009); Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). In SaaS, customers can use applications, but cannot control operating system, hardware or network infrastructure which are running. In PaaS, customers can use hosting environment (i.e., servers) as well as their applications, but still cannot control operating system, other hardware and network infrastructure. In IaaS, customers can use the fundamental computing resources, such as processing power, storage, network components. And also they can control operating system, storage, deployed applications and possibly networking.

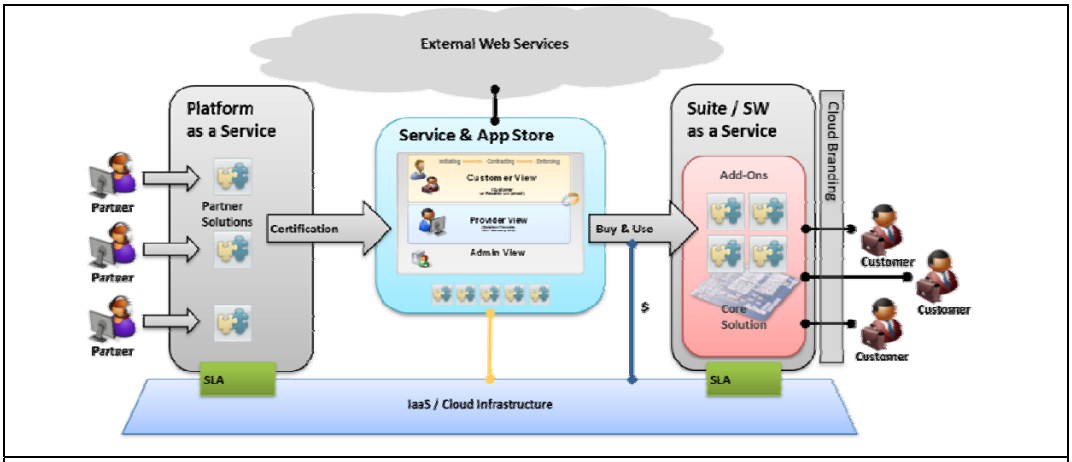

Figure 2. Cloud Computing Platform (*Source: SAP, 2011*)

Big Data as a New Tidal Wave

Big Data is newly spotlighted with the popular use of social media (Lacho and Marinello, 2010), such as Twitter (Weng et al, 2010; Jansen et al., 2009), Facebook (Chiu et al., 2008), and Flickr (Cha et al., 2009). Social media are becoming more prevalent and emerged as a new way of life to the people (Hathi, 2009). According to IBM (2011), more than two billion Internet users and 4.6 billon mobile phones are in the world. Facebook (Foster et al., 2010)

has more than 500 million users and created 30 billion pieces of content every month. And about 340 millions of data every day in Twitter are exchanged. As a result, we are living the *Age of Big Data*.

Nowadays *Big Data* terminology is popularly used in the business world. Then, what is *Big Data*? There is no single definition of *Big Data* until now, but broadly speaking it is the tidal wave of data, not only volume but also velocity and variety, from the cloud computing and social media. Narrowly, it can be defined as datasets whose size is beyond the ability of typical database software tools to capture, store, manage, analyze, and visualize. However, the size of database that qualifies as *Big Data* is changed, so its definition varies by industry sectors. *Big Data* today ranges from few dozen tera bytes (TB) to multiple peta bytes (PB).

There are two types of data; structured and unstructured. Structured data refers to data with high degree of organization in a structure so that it is identifiable, such as data in database. While unstructured data is the opposite. It is simply the lack of structure. The typical types of unstructured data include video clips, weblogs, social media feeds, etc. For example, e-mail is a type of unstructured data because it does not generally write about precisely one subject and even the format. Data in spreadsheets, on the other hand, is an example of structured data because it can be arranged in a database system. In reality, about 80% of the world's data in the business world is unstructured. It may be data we've been aggregating before, but could not process with current data mining tools.

The characteristics of *Big Data* are big volume, high velocity, and wide variety. Data volume is expanding due to the increase of social media, online data collection and location data, to name a few. Volume is accelerating with additional online activity and usage of sensor-enabled devices. The pace of business activity and competitive pressure increases as companies begin to use data occurring on a more frequent basis, including streaming data.

Big Data Analytics Tool: IBM Cognos

In the past, there was a lot of work to analyze on the flow of society (unstructured data), but that relies on the analysis of specific areas of expert with the sensibilities of the average people and also is difficult to provide a quantitative basis. But the *Big Data* analytic techniques are possible to take advantage of the more objectifying data. For example, to the question of "how much more people can be represented as a numerical?" We can answer "This year is 2.5 times hotter in summer than last year" instead of "this summer is very hotter than last year." It is the so-called *Big Data Analytics*.

Big Data Analytics is newly spotlighted the field can be a clue to solve the economic and social issues. Positioning system functions in the path or destination to move to the smartphone, Internet search history or search pattern can be analyzed and used to investigate the history of the credit card, it is possible to analyze individual consumption patterns. Computational capacity and a wide range of smart devices and the Internet penetration is increasing, ranging from the activities of the individual to society as a whole, the data became available, collect data analysis methodology development coupled with a rapidly growing trend. Such as Google, Amazon, Facebook and IBM, the IT companies are entering competitively.

The IBM Cognos is a web-based architecture which is separated into three tiers; Web Server, Applications, and Data. It is a Windows®-based system that provides business intelligence to the needs of different users. It leverages existing corporate IT resources such as web servers, application providers, and application servers, and also supports multiple languages. The components of IBM Cognos include reporting, analysis, scorecarding, dashboarding, business event management, and data integration from a wide array of data sources.

The *Business Insight* module is to create sophisticated interactive dashboards using IBM Cognos content as well as external data sources according to a user's specific information needs. The user can view and open favorite dashboards and reports, manipulate the content in the dashboards, and email your dashboards. The *Report Studio* is to create, edit, and distribute a wide range of professional reports. It can also define corporate-standard report templates. The *Query Studio* module is to design, create, and save reports to meet reporting needs that are not covered by the standard, professional reports created in the *Report Studio*. The *Analysis Studio* module is to explore and analyze data from different dimensions of their business. The user can also compare data to spot trends or anomalies in performance. The *Event Studio* module is to set up agents to monitor your data and perform tasks when business events or exceptional conditions occur in your data. When an event occurs, people are alerted to take action. Agents can publish details to the portal, deliver alerts by email, run and distribute reports based on events, and monitor the status of events. The *Metric Studio* module is to create and deliver a customized score carding environment for monitoring and analyzing metrics throughout your organization. The user can monitor, analyze, and

report on time-critical information by using scorecards based on cross-functional metrics. The *Administration* mdule is a central management interface that contains the administrative tasks for IBM Cognos. It provides easy access to the overall management of the IBM Cognos environment and is accessible through IBM Cognos Connection. The *Framework Manager* module is the IBM Cognos modeling tool for creating and managing business related metadata for use in IBM Cognos analysis and reporting. Metadata is published for use by reporting tools as a package, providing a single, integrated business view of any number of heterogeneous data sources.

Once the IBM Cognos is installed, the user types 'user name' and 'password'. The first screen is shown in Figure 6 by default. The user can choose several activities to perform, such as to query, analyze data, or perform administrative tasks.

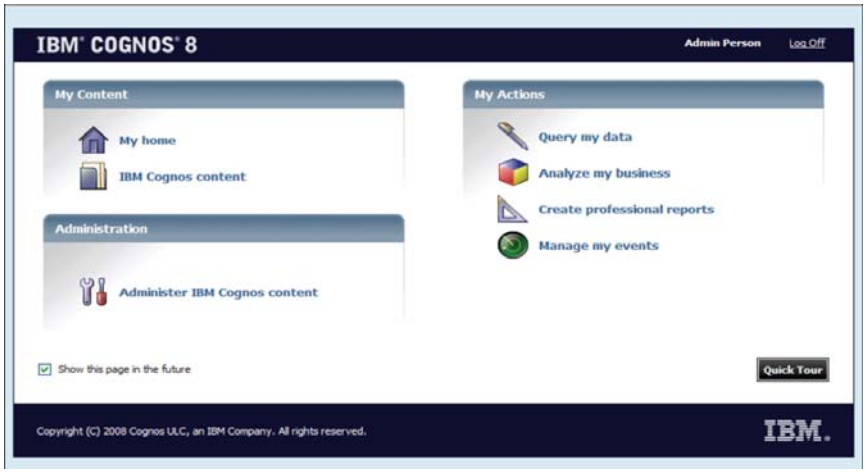

Figure 3. The First Screen of IBM Cognos

A Case Study of Big Data Analytics

In this section, we attempt to build a dashboard of financial companies in the USA including Bank of America, JP Morgan Chase Bank, and CITI bank. This dashboard will develop into business intelligence system in the future. Figure 4 shows its sample. We monitored the 19 financial companies in Fortune 500 companies in US. We daily are monitoring and capturing social media data from Facebook, Twitter, and YouTube.

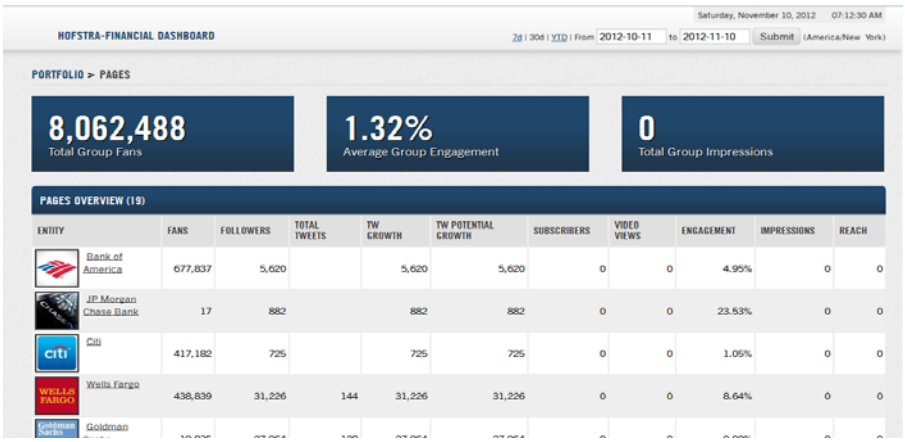

Figure 4. Dashboard of Financial Sector

Figure 5 shows the analysis of social media data using data mining tool (i.e., Rapid Miners). In this Figure 7, we can see the high-level information, such as fans, likes, and engagement. We will drill down and analyze these data to get the impact of social media to company's performance.

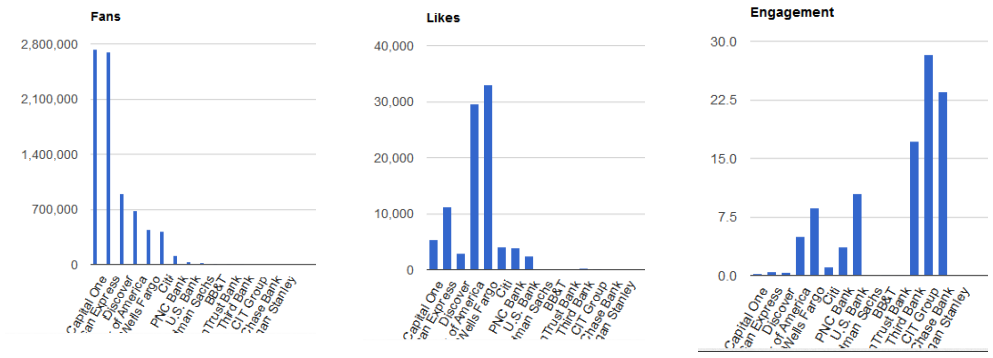

Figure 5. Analysis of Social Media Data (Frequency)

Figure 6 shows the comparison of engagement rate in the 19 financial companies.

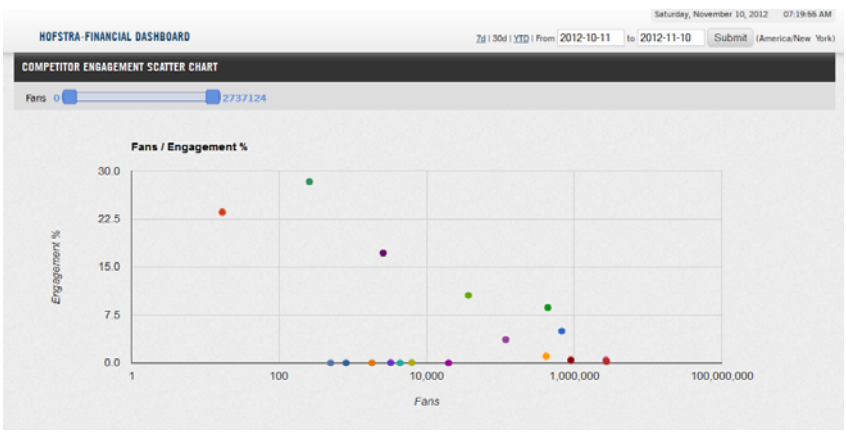

Figure 4. Analysis of Social Media Data (Engaement)

Concluding Remarks

Until now, we discuss new IT environment including mobile cloud computing, social media, and *Big Data*. And we also present briefly about IBM Cognos as an example of *Big Data Analytics* tools. The Internet is the backbone of our society, while mobile cloud computing is a central source of social change. Today social media has created *Big Data* which is beyond the ability of typical database software tools to capture, store, manage, analyze, and visualize. Today businesses firms are challenged by *Big Data* because it grows so large that they become awkward to work with using on-hand database management tools. However, *Big Data* has big potential that it can generate significant value across sectors, such as healthcare, retail, manufacturing, public sector, etc. The future research will analyze *Big Data* (unstructured data) in financial sector collected from social media (i.e., Twitter and Facebook) using the IBM Cognos system, and then compared to traditional financial data (structured data).

References

Abdul-Fatah, I. 2002. "nPerformance of CORBA-based client-server architecturesSource," *IEEE Transactions On Parallel And Distributed Systems* (13:2), pp. 111- 127.

Abramson, D., Buyya, R., and Giddy, J. 2002. "A computational economy for grid computing and its implementation in the Nimrod-G resource broker," *Future Generation Computer Systems* (18:8), pp. 1061–1074.

Beckman, M. 2010. "Enterprise Security vs. Social Media," *System iNEWS*, SystemiNetwork.com, pp.21-27.

Cha, M., Mislove, A., and Gummadi, K. 2009. "A measurement-driven analysis of information propagation in the Flickr social network," *In Proceedings of the 18th international conference on World Wide Web*.

Chiu, P., Cheung, C., and Lee, M. 2008. "Online social networks: why do we use Facebook?," *Communications in Computer and Information Science* (19), pp.67-74.

Cusumano, M.A. 2011. "Technology Strategy and Management: Platform Wars Come to Social Media," *Communications of the ACM* (54:4), pp. 31-33.

Foster, M. K., Francescucci, A., and West, B.C. 2010. "Why Users Participate in Online Social Networks," *International Journal of e-Business Management* (4:1), pp.3-19.

Harmer, T., Wright, P., Cunningham, C., and Perrott, R. 2009. "Provider-Independent Use of the Cloud," Proceedings on *The 15th International European Conference on Parallel and Distributed Computing.*

Häsel, M. 2011. "OpenSocial: An Enabler for Social Applications on the Web," *Communications of the ACM* (54: 1), pp. 139-144.

Hathi, S. 2009. "How Social Networking Increases Collaboration at IBM," *Strategic Communication Management*, (14:1), pp. 32-35.

Hayes, B. 2008. "Cloud computing," *Communications of the ACM* (51:7), pp. 9–11.

IBM. 2011. "Better Business Outcomes with Business Analytics," *White Paper*, IBM Software Group.

Jansen, B.J., Zhang, M., Sobel, K. & Chowdury, A. 2009. "Twitter power-tweets as electronic word-of-mouth," *Journal of the American Society for Information Science andTechnology* (60:11), pp.2169-2188.

Kurdi, H., Li, M., and Al-Raweshidy, H. 2008. "A classification of emerging and traditional grid systems," *Distributed Systems Online* (9:3), pp. 1-13.

Lacho, K. J. & Marinello, C. 2010. " How Small Business Owners Can Use Social Networking to Promote Their Business," *Entrepreneurial Executive* (15), pp. 127-133.

Mell, P. and Grance, T. 2011. "The NIST definition of cloud computing," *Special Publication 800-145 Retrieved from* http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

Milojicic, D. 2008. "Cloud computing: Interview with russ daniels and franco travostino," *IEEE Internet Computing* (12:5), pp. 7–9.

SAP. 2011. "Presentation slides," retrieved from http://www.sdn.sap.com/irj/scn

Sotomayor, B., Montero, R., Llorente, I., and Foster, I. 2009. "Virtual Infrastructure Management in Private and Hybrid Clouds," *IEEE Internet Computing* (13:5), pp. 14-22.

Sturdevant, C. 2011. "Socializing the Enterprise," *eWeek*, (28:1), p34-34.

Violino, B. 2011. "Social Media Trends," *Communications of the ACM* (54:2), pp. 17-17.

Weiss, A. 2007. "Computing in the clouds," *netWorker* (4), pp. 16–25.

Weng, J., Lim, E., Jiang, J., and He, Q. 2010. "Twitterrank: finding topic-sensitive influential twitterers," *In Proceedings of the third ACM international conference on Websearch and data mining*, ACM.

Werbach, K. 1997. "Digital Tornado: The Internet and Telecommunications Policy," *Working Paper* , FCC Office of Plans and Policy (29).