International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 08, August 2016
www.ijirse.com
ISSN: 2454-9665

DATA ANALYTICS AS A SERVICE (DAaaS) CLOUD BASED ANALYTICAL MODEL

Harpreet Kaur[1], Harminder Singh[2]

[1,2]Research Scholar Punjabi University, Patiala

ABSTRACT

Cloud computing is an evolving paradigm that has redefined the way information technology based services can be offered. Storing big data is only part of the picture. Special techniques are needed to analyze big data in cloud computing. Executives need to become familiar with the big data methodologies, adopt the technology appropriate for their business, and ensure that employees develop skill with the technology. Data storage techniques differ depending on whether the data are unstructured or structured. In this paper DAaaS is introduced as another approach of aggregating data into clouds and analyze such data with the powerful computational capacity of clouds. For handling big account of data like in tens of terabytes (TBs) or tens of petabytes (PBs) , a new approach is required such as distributed data store and complex event processing which are basic technologies for big data processing in cloud environments

Index Terms: Big Data, Daaas, Cloud Computing, ICT, RTAE. App Store.

I. INTRODUCTION

Data Analytics as a Service (DAaaS) represents the approach to an extensible platform that can provide cloud-based analytical capabilities over a variety of industries and use cases. From a functional perspective, the platform covers the end-to-end capabilities of an analytical solution, from data acquisition to end-user visualization, reporting and interaction. Beyond this traditional functionality, it extends the usual approach with innovative concepts, like Analytical Apps and a related Analytical Appstore.

Architecturally, and due to the intrinsic complexities of analytical processes, the implementation of DAaaS represents an important set of challenges, as it is more similar to a flexible Platform as a Service (PaaS) solution than a more fixed Software as a Service (SaaS) application. Aspects like the PaaS internal architecture, the distinction between real-time vs. non real-time processing, the specific characteristics of the Analytic Services, the needs for data storage and modeling, the delivery over hybrid cloud models and several others, make its design a complex challenge.

II. BIG DATA PROCESSING IN CLOUD

Information and communication equipment and networks prices shows downfall in recent years .As the number of ICT equipment increased, the amount of data is also increasing at rapid rate which is being accumulated in cloud data centers. These massive amounts of data are called big data.

For example, information from position sensors of the Global Positioning System (GPS) mounted on mobile phone handsets or automobiles for tracking and transaction records from big departmental stores cash registers

are stored along with the location and time of their generation, and transferred via networks to data centers, where they are accumulated.

This big department store's data should be analyzed for varies type of information regarding the behavior of purchasers like which is the product having most sales with respect to time ,month and the various changes in demands for different product can be predicted which will be valuable information for the owner of the store for maintaining the inventory in the store. According to a trial calculation, the amount of event data generated in the U.S.is estimated to be 7 million pieces per second, which adds up to a few tens to hundreds of PBs per month if accumulated as they are without compression. This value is not equal to the amount of data actually transferred to data centers and processed. However, the future availability of such detailed data is raising hopes for the acquisition of valuable information for enterprises and purchaser , such as estimates of what a person will purchase, and where. Thus data analysis is important from many prospective of profits for enterprises with big retails stores, ecommerce websites for planning online offers for websites.

III. DATA ANALYTICS AS A SERVICE (DAaaS)

Data Analytics as a Service (DAaaS) is an extensible analytical platform provided using a cloud-based delivery model, where various tools for data analytics are available and can be configured by the user to efficiently process and analyze huge quantities of heterogeneous data. Customers will feed their enterprise data into the platform, and get back concrete and more useful analytic insights. These analytic insights are generated by Analytical Apps, which orchestrate concrete data analytic workflows. These workflows are built using an extensible collection of services that implement analytical algorithms; many of them based on Machine Learning concepts. The data provided by the user can be enhanced by external, 'curated' data sources. The DAaaS platform is designed to be extensible, in order to handle various potential use cases. One concrete case of this is the collection of Analytical Services, but it is not the only one. For example, the system can support the integration of very different external data sources. To enable DAaaS to be extensibility and easily configured, the platform includes a series of tools to support the complete lifecycle of its analytics capabilities.

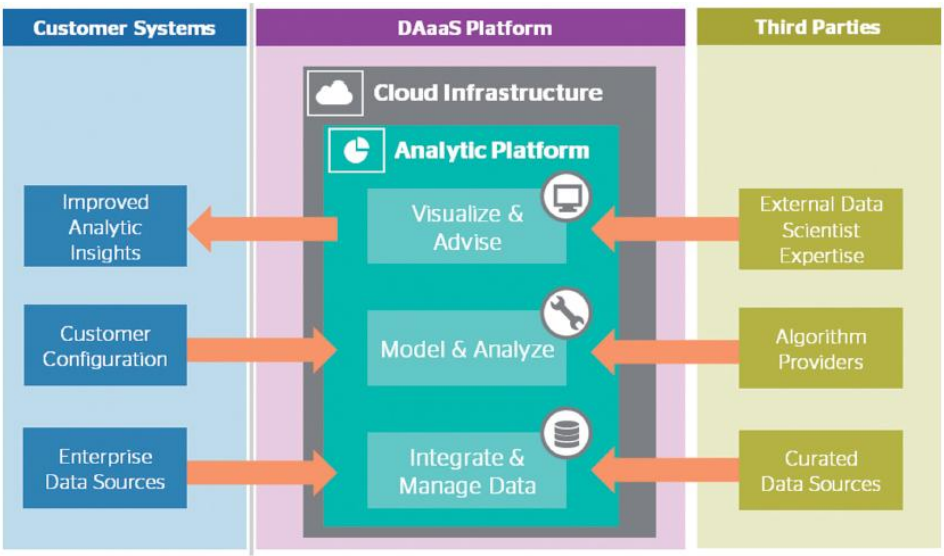

Fig1: DAaaS Concept

IV. BIG DATA AND CLOUD CONVERGENCE

It's really difficult to miss Big Data in the media. Major news outlets like the New York Times or the Guardian have made everyone aware of the exponential growth of the information that businesses need to cope with. The importance of Big Data is however not the volume of data, neither its heterogeneity, but instead it is about the insights that businesses can get from understanding this entire "data deluge" using the appropriate analytical methods.

"Doing" Big Data in a proper way is not easy. In addition to the complexity of tools and infrastructures that are required to manage huge volumes of data, we need to identify and resolve the scarcity of talent that can properly take advantage of these tools to make data "talk". Some of the techniques associated with Big Data are not really new (statistics and machine learning have a long history) but these techniques need people with a deep understanding; in a new role that now is often called the 'Data Scientist'. So, Big Data can be a tough proposition for many companies as conventional tools and on-premise techniques could be the wrong approach.

To ease many of these "pain points" another transformative trend has entered the market place - Cloud Computing and the continuous movement towards a Utility ICT model. So instead of deploying complex solutions in-house, companies can take advantage of services provided by third parties, using economic models that offer them flexibility and adaptability to the changing needs of their environment.

It's easy to understand that the combination of Big Data with Cloud can ease the adoption of advanced analytic capabilities over the bigger and more heterogeneous data sources that companies need to handle, letting companies benefit of the insights derived from it. Value is in the data itself, and not in the technology that is used to process it. What companies need is not the deployment of a complex Big Data infrastructure and the associated capital investment, but the access to the services that provide advanced Data Analytics on their data now and in the future. Providing this service in a flexible and scalable format is the main purpose of Data Analytics as a Service.

V. DAaaS FUNCTIONAL ELEMENTS

In order to deliver all the capabilities of a DAaaS solution, a complete platform needs to be implemented, as shown in the following figure:

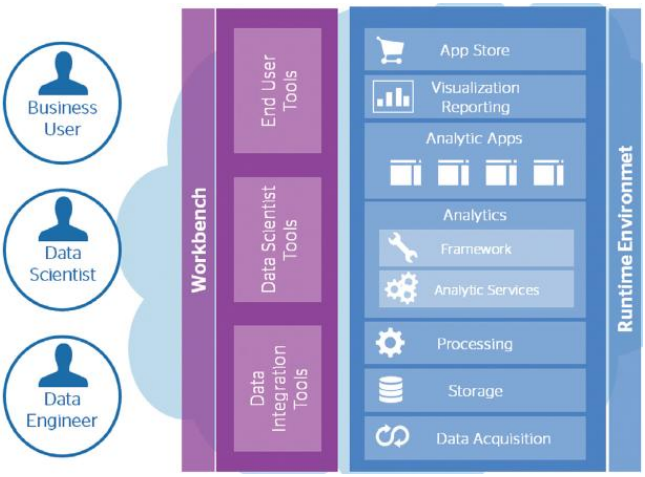

Fig2: DAaaS Functional Elements

ISSN: 2454-9665

First, we need to differentiate the two groups of elements:

- Those related with the runtime aspects of the solution, that is, the platform that processes the data. We've identified this as Runtime Environment.
- Those that control the interaction with the user, mainly for the configuration of the system, using a set of tools that we have called Workbench Environment.

To the last point, we use here the concept of a user quite broadly, including not only the end business users, but also all other individuals that interact with the platform in more "technical" roles, like programmers in Data Modeling and Integration roles, or Data Scientists that configure Analytical Services and Data Flows.

5.1 Runtime Environment

The Runtime Environment is the execution platform of the DAaaS solution. We describe the various components from the bottom to the top as shown in the figure, following the logical data flow from ingested data (input) to generated insights (output).

Data Acquisition: it provides a web services / messaging interface to the external world, for the acquisition of data, both from the end customer but also for the alternative 'curated' data sources. To provide the flexibility to cope with very diverse data sources and protocols, the solution needs to be modular with components that implement widely recognized EIP (Enterprise Integration Patterns).

Storage: the core data repository is used to store the information in the system, both from the customer but also other data. The solution should be able to cope with up to petabyte-size quantities of data, but also needs to be flexible enough so very different data models can be implemented and supporting strict multi-tenant capabilities. There are several potential NoSQL databases that allow these kinds of capabilities, each having strong and weak points for different scenarios.

Processing: In order to be able to process huge volumes of data, it is essential that some kind of distributed processing capability exists, so that different processing algorithms can be implemented and executed in parallel. This processing layer acts as an interface between the Data Storage and the analytical services. Nowadays the more popular solution for this kind of Distributed Processes is the Hadoop Map-Reduce solution. It is widely supported, both by different NoSQL data sources, but also by programming languages and analytic tools in upper layers of the stack. However other emerging alternatives like Spark, or Storm, are appearing with some specific advantages, for example in real-time scenarios.

Analytics: in a sense, this is the most crucial part of the platform. This is where all the analytic processes reside. We can deconstruct it further in two main elements:

a) The Analytic Services, which are well-defined components that implement concrete data analysis algorithms. These can be quite varied and could be implemented using different base technologies. But they have a defined, concrete scope related to a specific Data Analysis technique over specific dataset classes. Many of them will be based in Machine Learning techniques, both using Supervised and Unsupervised approaches.

b) An Analytic Framework, that acts like a glue that brings together different Analytic Services in order to achieve a concrete business outcome. So a specialized programmer could use this framework to implement a complete analytic functionality.

ISSN: 2454-9665

Analytic Apps: as we've just seen we can combine different Analytic Services using the capabilities of the Analytic Framework. We call these bundles of complete analytic functionality an Analytic App. These Analytic Apps are the elements that expose the final, business-oriented capability to the end-user.

Visualization / Reporting: although a great part of the functionality of the DAaaS platform could be accessed using a web services interface, a complete solution will integrate some form of end-user oriented visualization / reporting to simplify access to information. An ample set of tools that provide these kind of capabilities exist, some commercial (Tableau, QlikView) and some others open source (Pentaho, Jasper Reports).

App Store: The solution will provide high-level functional bundles or Apps to end-users. An enterprise App-Store front-end will provide the mechanisms to control the lifecycle of an app in an organization, from acquisition to retirement.

5.2 Workbench Environment

In addition to the runtime environment, some specific tooling is needed to customize the solutions (the Analytic Apps) to the specific needs of the end-user. We call this set of tools the Workbench Environment. It includes tooling for different kind of roles:

Data Integrator. This role takes responsibility for interfacing the existing internal enterprise systems with the DAaaS system. Specific tooling for the Data Acquisition subsystem (and, to a lesser degree, for the data storage and processing layers) is needed, that enables ETL (extraction, transformation and load) of this information, so it is properly modeled and incorporated into the DAaaS environment.

Once the data is in the DAaaS Storage subsystem, the next role takes care of a correct implementation of the analytical capabilities required by the enterprise.

Data Scientist. A substantial part of the work of a Data Scientistis to correctly model, test and check analytic workflows exposed in Analytic Apps, for validity for the specific datasets used. Going even further, we can assume that specific Analytic Apps (or even, Analytic Services) could be built by them. Data Scientists can be employees of the company, but they could also work for a specialized service provider.

The final step is by the Business Users. One important point to bear in mind is that from the DAaaS perspective, there exist different kinds of business users. Obviously, we have the non-technical, business end-user, who only needs to have the results from the tool, either integrated in their existing enterprise apps or using simple to use visualization tools.

VI. CHALLENGES OF ANALYTICS IN THE CLOUD

Analytic solutions that need to support Big Data services present additional challenges. This is even more the case if these services are intended to be delivered through a cloud environment.

- Information Lifecycle Management: the complete analytical workflow can get very complex with lots of important steps: data acquisition, data modeling, data mining and visualization. In contrast with transactional solutions, which are more fixed in nature, Analytics need a flexible approach to adapt to all this potential variability.
- Data model diversity: a diversity of potential types of data models exists for specific business needs - and these data models are tightly coupled to specific types of analytics. For example, time series data is modeled quite differently from social network data and also the potential algorithms to be used are distinct.

ISSN: 2454-9665

- Analytic knowledge: although not really new, many of the advanced techniques related to advanced analytics (like Machine Learning) are quite complex and demand people with very specific knowledge, such as the previously mentioned Data Scientists.
- Data volume: even when technology exists for processing huge volumes of data, it is not easy. Moving big volumes of data to a cloud solution can be difficult, and sometimes, it is much easier to bring computation to where the data is.
- Real-time analytics: more and more, the value of analytics demands quicker insights, progressing towards the concept of real-time analytics. Even if we are talking here of soft real-time, this has a definite impact on how a cloud solution is designed to handle these loads.
- Security: like in any other cloud solution, security is a very complex issue. Some companies, due to the data criticality or regulatory constraints, may be reluctant to move data to the cloud, but could benefit of the analytical capabilities offered in a private cloud.
- Privacy: for some specific types of data, privacy considerations may impact the potential of cloud analytics - not only due to the data in itself, but also due to the potential that data will not remain anonymous after analysis.

In addition, we should note that in addition to issues that are specific to Analytics, there are others that are of a more technical nature, related to how existing Analytics and Big Data platforms need to adapt to a cloud infrastructure, as many of them started out in more classical 'on-premises' infrastructure.

VII. REAL-TIME VS NON REAL-TIME ANALYTICS

Another important issue regarding DAaaS is related to its potential real-time capabilities. Note that here we are talking about a soft real-time; response times are in the order of seconds, not the hard real-time that is in use in embedded systems for example.

Usually, traditional analytics have been in the realm of non real-time. However, technological advances and business pressures are demanding shorter response times for analytical processes at the same time that data volumes grow exponentially.

So, the architecture must provide mechanisms for this kind of near-real-time response.In the case of the DAaaS platform, there is an additional problem that needs to be taken into account. Some technologies in the DAaaS stack are not very well aligned with the needs of this kind of real-time response. Perhaps the most significant one of those 'troublesome' system components is the Hadoop Platform.

It is the most popular Big Data platform but at the same time it is not really real-time oriented. For example, the Map-Reduce framework is based on batch execution modes of operation. This imposes an important burden onto the system if real-time modes of operation are required.

ISSN: 2454-9665

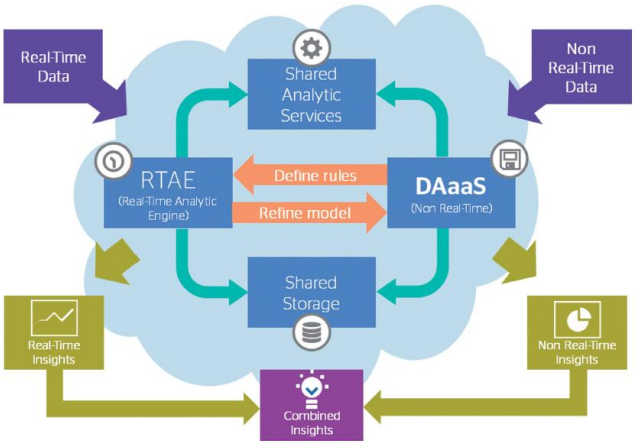

Fig3: DAaaS and Real-Time Analytics

We can say that we have two systems running in parallel:

a) The DAaaS platform.

b) A Real-time Analytics Engine (RTAE).

The latter takes care of the real-time analytic capabilities, processing streaming data and looking for patterns on it. These patterns may be of different degrees of complexity and may be expressed using different mechanisms. Lower levels of complexity may be represented by a "business rules language", like the one included in Complex Event Processing (CEP) solutions.

For a higher level of complexity, complete analytic services could apply that may share common elements with those included in DAaaS.

For example, a potential approach to the RTAE part is an extension of the Atos' Context Broker Platform (CBP)5. The latest version of this solution will use a distributed messaging framework based in the 'actor model', called Akka implemented in the Scala Language. These models will define a 'distributed complex event processing (CEP)' platform and on top of it, the real time analytical capabilities can be implemented.

As shown in the figure, both DAaaS and RTAE can be interrelated, so that they complement each other:

- They can share some specific analytical services.
- They can share persistent storage.
- The rules at the RTAE can be configured by the outcomes of the DAaaS Platform. So, if a pattern emerges in a DAaaS App the RTAE can be configured to detect this pattern.

VIII. DAaaS IN HYBRID CLOUD ENVIRONMENT

A pure public cloud approach to DAaaS may have to address some important challenges for some user scenarios:

- When the volume of data is huge or rapidly growing, it may be not practical or efficient to move data from customer's premises to a DAaaS provider infrastructure due to size or communication constraints.
- In addition to this size issue there may be other limitations such as security or privacy policies, that demand on-premise or private cloud approaches.

ISSN: 2454-9665

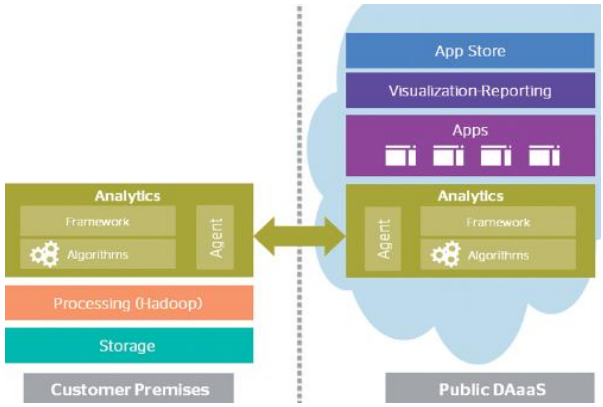

Fig4: DAaaS working in Hybrid Cloud model

Taking into account these restrictions is there a model that enables the benefits of the DAaaS model while still preserving data in the locations of the customer?

The answer may be an approach towards a Hybrid Cloud which would bring part of the DAaaS approach to customer premises, in effect bringing computation to the data instead of the other way round.

The approach is based in the work being done internally by Atos Scientific Community in the concept of Cloud Services Brokerage6 and can be summarized in the following figure:

In this model, the architecture is split with a part residing on customer premises, which comprises of the Data Storage and Processing Layers. Part (or even all) of the processing for the Analytic Services then takes place on customer's premises as Hadoop Map/Reduce processes for example, coordinated by the higher level elements of DAaaS available as external public cloud elements and using the Analytic Services present in the Public DAaaS solution. This coordination is achieved by using a series of Agents which take charge of distributing jobs between both environments.

So this Hybrid PaaS provides important advantages regarding flexibility of deployment, according to customer restrictions on data availability and minimizes the needs of data transfers as well as integrating DAaaS in a higher view of cloud services coordinated by some kinds of Cloud Orchestration services.

8.1 Benefits of DAaaS

The main benefit of the DAaaS is to lower the barrier of entry to advanced analytical capabilities, without demanding that the user commits to large internal infrastructures and human resources to the project. Instead of a complex custom project the customer follows simpler steps:

- Data Scientists working for the organization explore the AppStore for an Analytical App that fits the problem.
- They rent the Analytical App for a specific time or quantity of data.
- They configure the Analytical App to its needs including, for example, the usage of external data sources provided by the DAaaS.
- Then the data is fed from the internal systems to the Analytical App.
- The SMEs in the company validate the results and even enhance them with some customization.
- Outcomes are available for all other uses.

Certainly using DAaaS may not be as direct as using other kinds of SaaS software. Any analytical process demands certain preparatory work: explore initial data, define analytical processes, implement and validate

ISSN: 2454-9665

results using test data and optimize it as new data comes. But even so, effort is diminished. And that is without taking into account the benefits of a Cloud delivery model: no upfront costs in infrastructure and a pay per usage model allow experimentation or even temporal usage scenarios.

Technically we've seen the complex issues that the implementation of a functionally complete Big Data Analytics solution needs to overcome, if developed internally by an organization. So a DAaaS solution minimizes this technical complexity even more if it is properly designed to manage a hybrid cloud model, for those cases where information needs to be on-premises. Also there is the issue of expertise scarcity: Data Science is hard and expert resources are not easily available. DAaaS doesn't eliminate the need for Data Scientists but alleviates some of the problems as some pre-packaged applications are provided for specific use cases. In addition to providing a DAaaS platform, analytic services companies can offer to their customers' access to Data Scientists on demand. This way combining a growing collection of Analytical Services and Data Scientist expertise, the richness of the platform and the value for customers grows.

IX. CONCLUSION

However hyped it may be currently Big Data is certainly a business changing trend, as the facts are evident: the data explosion is real and some companies have shown clear competitive advantage by creating and implementing new analytic capabilities over previously unused data. But getting this kind of capability may be not easy for some companies. Here the flexibility that Cloud delivery models bring can simplify adoption for some companies and even those that could have the resources to implement it internally can obtain significant cost advantages with DAaaS. Data Analytics as a Service, the model we propose in this paper, can be applied to multiple use cases and industries even as the analytic approaches to different scenarios may vary considerably. Beyond that DAaaS puts analytics as a first-level element component in a new vision of Enterprise Computing which makes extensive usage of the advantages of Cloud technologies.

REFERENCES

[1] Kiron, David, Rebecca Shockley, Nina Kruschwitz, Glenn Finch and Dr. Michael Haydock, "Analytics: The widening divide: How companies are achieving competitive advantage through analytics" IBM Institute for Business Value in collaboration with MIT Sloan Management Review. October 2011.

[2] LaValle, Steve, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitz. "Analytics: The new path to value: How the smartest organizations are embedding analytics to transform insights into action." IBM Institute for Business Value in collaboration with MIT Sloan Management Review. October 2010

[3] Dr. Marc and Dr. Michael Haycock. "Customer analytics pay off: Driving top-line growth by bringing science to the art of marketing." IBM Institute for Business Value. September 2011

[4] D.J. Abadi, Data management in the cloud: Limitations and opportunities, IEEE Data Engineering Bulletin 32 (1) (2009) 3–12.

[5] Amazon redshift, http://aws.amazon.com/redshift/.

[6] Amazon data pipeline, http://aws.amazon.com/datapipeline/. [7] Amazon Elastic MapReduce (EMR), http://aws.amazon.com/elasticmapreduce/.

[7] Amazon Kinesis, http://aws.amazon.com/kinesis/developer-resources/.

[8] R. Ananthanarayanan, K. Gupta, P. Pandey, H. Pucha, P. Sarkar, M. Shah, R. Tewari, Cloud Analytics: Do We Really Need to Reinvent the Storage Stack? in: Proceedings of the Conference on Hot Topics in Cloud Computing (HotCloud 2009), USENIX Association, Berkeley, USA, 2009.

[9] G. Andrienko, N. Andrienko, S. Wrobel, Visual analytics tools for analysis of movement data, SIGKDD Explor. Newsl. 9 (2) (2007) 38–46.

[10] Announcing Suro: Backbone of Netflix's Data Pipeline, http://techblog. netflix.com/2013/12/announcing-suro-backbone-of-netflixs.html.

[11] Apache S4: distributed stream computing platform, http://incubator.apache. org/s4/Attention, shoppers: Store is tracking your cell, New York Times. URL http://www.nytimes.com/2013/07/15/business/attention-shopper-storesare-tracking-your-cell.html.

[12] A. Balmin, K. Beyer, V. Ercegovac, J.M.F. Ozcan, H. Pirahesh, E. Shekita, Y. Sismanis, S. Tata, Y. Tian, A platform for eXtreme Analytics, IBM J. Res. Dev. 57 (3–4) (2013) 4:1–4:11.

[13] R.S. Barga, J. Ekanayake, W. Lu, Project Daytona: Data Analytics as a Cloud Service, in: A. Kementsietsidis, M. A. V. Salles (Eds.), Proceedings of the International Conference of Data Engineering (ICDE 2012), IEEE Computer Society, 2012, pp. 1317–1320.

[14] G. Bell, T. Hey, A. Szalay, Beyond the Data Deluge, Science 323 (5919) (2009) 1297–1298.

[15] I. Bhattacharya, S. Godbole, A. Gupta, A. Verma, J. Achtermann, K. English, Enabling Analysts in Managed Services for CRM Analytics, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009), ACM, New York, USA, 2009, pp. 1077–1086.

[16] bigdata@csail, http://bigdata.csail.mit.edu/.

[17] 'Big Data' has Big Potential to Improve Americans' Lives, Increase Economic Opportunities, Committee on Science, Space and Technology (April 2013). URL http://science.house.gov/press-release.

[18] Birst Inc., http://www.birst.com.

[19] R. Bonney, J.L. Shirk, T.B. Phillips, A. Wiggins, H.L. Ballard, A.J. Miller-Rushing, J.K. Parrish, Next steps for citizen science, Science 343 (2014) 1436–1437.

[20] D. Borthakur, J. Gray, J.S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash, R. Schmidt, A. Aiyer, Apache Hadoop Goes Realtime at Facebook, in: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2011), ACM, New York, USA, 2011, pp. 1071–1080.

[21] C. Bunch, N. Chohan, C. Krintz, J. Chohan, J. Kupferman, P. Lakhina, Y. Li, Y. Nomura, An Evaluation of Distributed Datastores Using the AppScale Cloud Platform, in: Proceedings of the 3rd IEEE International Conference on Cloud Computing (Cloud 2010), IEEE Computer Society, Washington, USA, 2010, pp. 305–312.

[22] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Gener. Comput. Syst. 25 (6) (2009) 599–616.