

A Reference Architecture for Big Data Systems

Go Muan Sang, Lai Xu, Paul de Vrieze

Faculty of Science and Technology, Bournemouth University

Poole, Dorset, BH12 5BB, UK

{gsang, lxu, pdvrieze}@bournemouth.ac.uk

Abstract

Over dozens of years, applying new IT technologies into organizations has always been a big concern for business. Big data certainly is a new concept exciting business. To be able to access more data and empower to analysis big data requires new big data platforms. However, there still remains limited reference architecture for big data systems. In this paper, based on existing reference architecture of big data systems, we propose new high level abstract reference architecture and related reference architecture notations, that better express the overall architecture. The new reference architecture is verified using one existing case and an additional new use case.

Keywords—*Big Data; Reference Architecture; Data Analytics;*

I. INTRODUCTION

The scale of data collection and data availability becomes increasingly high due to cheaper storage and evolution of Internet of things technology and their applications (digital data collection devices such as mobile, sensors, etc.) [7, 9, 38]. The current digital landscape enables more people interact with data and more data is shared ever before. Besides, the challenge created by fast growing amount of data, there are opportunities for organizations as the world becomes more and more digital allowing context-specific aggregation and analysis of data [1, 7, 8].

Data and information have become primary assets for organizations as they recognized that the data they own, and how they can use it can make them different than others [1, 2, 3]. As a result, businesses collect vast amounts of data from their daily activities [4, 5, 6].

Contemporary reports and decision support systems cannot adequately handle big data. As such, a new generation of methods and tools is needed. Current systems handling big data consist of inflexible and complex platforms, tools and information. This needs to be restructured into much more centralized but flexible analytical infrastructure. The problem of how to do this presents a challenge to businesses and the research community.

The need of enterprise agile analytics is growing rapidly as traditional systems are not able to satisfy the demands driven by big data, and increasingly complex business analysis and analytics [1, 7, 8]. Also, the current departmental implementations such as data mart are unable to meet the demands of increasingly complex KPIs, metrics and dashboards [1, 7, 8].

Data-driven discovery such as analytics is becoming a mainstream process for companies [1, 7, 8]. The term “analytics” is often described in various ways from web analytics to predictive analytics in related area [39]. [40] defined analytics as “techniques used to analyze and acquire intelligence from big data”. In other word, it is the overall process of producing insight (value) from big data. We refer analytics as “the collection, processing, analysis (machine learned models, statistics, etc.) and visualization of big data to produce insight”.

Analytics implies a requirement for providing efficient and effective decision-making processes with the right data that is transformed to be meaningful information [1, 3, 7]. The nature of complex business demands businesses to find innovative ways to differentiate among competitors by becoming more collaborative, virtual, accurate, synchronous, adaptive and agile [2, 7, 8]. To survive in such a market, it is essential to respond to market needs and changes. Analytics enables business to identify these needs leading to more efficient business [1, 3, 7].

Current big data based analytics systems are expensive and complex. This is particularly limiting for smaller organizations without the technical knowledge and financial resources to use these systems. In addition, there still remains a lack of reference architectures as well as a coherent architecture of big data systems. This paper explores the simplification of the realization of a big data system, and subsequently proposes a new high level abstract reference architecture. Key aspects and considerations such as architecture design of the system, utilization of underlying processes, technologies and services are important for the realization of a big data system.

When a big data design system is realized, key considerations (architecture of the system, utilization of underlying processes, technologies and services) are to be recognized [10]. The contributions of this work are: a) to provide high level independent reference architecture for design a big data platform based on existing work [17], b) to verify the proposed reference architecture using existing different big data systems, c) using the new reference architecture to present Amazon big data services.

The structure of the paper is as follows: related work and literature study are presented in Section 2. Production and mapping of the reference architecture with big data use cases, and mapping the reference architecture with Amazon Web Services are presented in Section 3. An analysis is discussed in Section 5, and the future work and conclusion are provided in Section 6 and 7.

II. RELATED WORK

A. Big Data Use Case

In recent years, several big data use cases have been published from Social network domain such as Facebook, Twitter and LinkedIn to Entertainment Video-streaming such as Netflix. There has also been a huge interest and opportunity of big data in the health industry. Facebook and LinkedIn collect from both traditional database and streaming data from users whereas Twitter mostly deals with streaming data [26, 27, 28, 29, 32, 33]. The collected data are then handled on a batch or streaming processing with each own defined processing functionalities. Data analysis such as Deep Analytics, trained models and specified jobs, algorithm service with Hadoop HDFS are predominately executed in clustering and distributed computations. Similarly, Netflix collects users' events, which are then processed in Online, Nearline and Offline computations [34, 35]. Recommendations processed via the Online data analysis are available to users.

B. Reference Architecture of Big Data Systems

Various reference architectures have been published. For example, Angelov, Grefen and Greefhorst [11] presented a framework for design and analysis of software reference architectures which forms of a multi-dimensional classification space, and five types of reference architectures. They [11] argue that the proposed reference architectures facilitate better architecture design, and hence will lead to better success. Galster and Avgeriou [16] also carried out an empirical study and proposed software reference architecture. The approach is a step-wise process which involves deciding a type for reference architecture, selection of design strategy, empirical acquisition of data, construction of reference architecture, enabling of variability, and evaluation.

Service-oriented architecture and its principles facilitate software design, architecture and implementation in the enterprise software domain [18, 19]. For the context of big data, limited architectures have been suggested. Paakkonen and Pakkala [17] presented reference architecture and classification of technologies, products and services based on published big data use cases and associated commercial products. A high level description of big data lifecycle and infrastructure for a big data architecture framework was presented by [20]. Schmidt and Möhring [21] suggested a service and deployment model for implementing big data pipeline in the cloud domain. [22] proposed reference architectures for the integration of SQL and NewSQL databases in order to support different growth patterns in enterprise data traffic. [23] proposed and validated in-memory storage system and distributed task execution analysis in big data architecture. [24] proposed tiered architecture (SOLID) for separating big data management from data generation and semantic consumption. A generalized software architecture was suggested for predictive analytics of historical and real-time temporally structured data [25].

Literature review shows that there are limited reference architectures in the big data context as well as lack of concrete

or coherent reference architecture. This demonstrates a need for further research in reference architecture for big data systems.

III. REFERENCE ARCHITECTURE FOR BIG DATA PLATFORMS

Reference architecture can help in facilitating the creation of concrete architectures and increasing understanding of the overall picture by containing typical functionality and data flows in a big data system [11]. A reference architecture is also useful for analyzing existing big data systems, providing the base of classification of data analysis processes and technologies. Categorizing the processes, technologies and services into groups (components) further facilitate decision making regarding the realization of system processes and functionalities.









In the comparison of big data use cases as shown in Table 1, we have divided the key elements of big data uses into five components: Data source; Data Collection, Processing and Loading (CPL); Data Analysis and Aggregation; Interface and Visualization; and Job and Model Specification.

- *Data Source*: it refers to the original source of data to be collected. This can be traditional data such as relational data or streaming data. Data can be structured, unstructured, semi-structured or streaming.
- *Data Collection, Processing and Loading*: Data Collection represents getting data from the (normally multiple) sources for storage or analysis. There are several techniques for this such as Snapshot. Data Processing covers the executions required for processing the source data before moving into the Loading stage. Data Processing represents functionalities such as data cleaning, replication, filtering, algorithm service, etc. Data Loading can then be executed, meaning the collected and processed data are now loaded into a data storage such as Hadoop HDFS.
- We group three items (Data Collection, Processing and Loading) in one component because each item closely links to each other. For example, data collection requires data processing for cleaning or formatting as well as loading the data into storage. At high level architecture, we believe that this improves logical linkage and flow between items closely related, and hence provides clearer and effective overall architecture.
- *Data Analysis and Aggregation*: Data analysis refers to the related data analysis tasks and processes whereas the Aggregation refers to the data storage (including multi-dimensional) which stores the results of the analysis.
- *Interface and Visualization*: they represent the end users as well as applications such as dashboards.
- *Job and Model Specification*: This covers models trained, specifications and scheduling of jobs with their storages.

Table 1 Big Data Use Cases Comparison

Big Data Use Case	Data Source (Traditional and Streaming data)	Data CPL (Collect, Process, Load)	Data Analysis and Aggregation	Job and Model	Interface and Visualization
Facebook	MySQL (structured) and Web Servers Logs (unstructured, streaming)	Dump database copy and Scribe. Hadoop Raw Data with replication.	Cube Transformation, Deep analytics (Hive jobs)	Jobs are stored a database	Microstrategy Visualization and User Applications.
LinkedIn	Oracle (structured) and Users Activity Data (streaming)	Snapshot of Oracle and Kafka producer of streaming. Hadoop data stores (Raw, Prepare, Sandbox, and Enterprise). Data cleaning, de-duplication and replication.	Deep Analytics (Hive, Pig, MR jobs), Voldermort with avatara transformation process.	Azkaban Framework for batch processing and jobs.	Visualization applications for enterprise users as well as LinkedIn users.
Twitter	Firehose (tweets), Updater, Queries	Tokenization and annotation of Firehose(tweet), Stats collector of Queries, Filter and Personalization of Firehose (tweet), Updater and Queries, Ranking Algorithm	Hadoop HDFS for Ranking algorithm and analysis results. Cache database for Twitter's users.	Ranking algorithm is mainly used.	Twitter users.
Netflix	Netflix users (streaming) and online data service (streaming)	Chukwa agent and Stream signals. Stream Processing Manhattan Framework	Deep Analytics: Offline, Nearline, Online. Algorithm Service	Machine Learning, Pig for batch processing jobs.	Netflix Users, Visualization applications for enterprise.

Table 2 Reference Architecture Notations

No	Notation	Description
1		The notation of a data store or database which represents structured, unstructured, semi-structured or real-time (streaming) data. We often indicate a short text or description of the notation.
2		This represents the Collection Process of data from a source to a destination. We often indicate a short text or description of the notation.
3		This notation represents Data Processing and Computation. We often indicate a short text or description of the notation.
4		This indicates the Interface and Visual applications. We include a short text.
5		This represents the Job Schedule for batch processing. A short text is included.
6		This represents the Job Specification for batch processing. A short text is included.
7		This indicates the flow between two notations. The left arrow of the line indicates the flow of the connection occurs from left to right notation.
8		This line indicates the flow between two notations. The line does not have any arrow so we can say that the flow can occur both ways.

For the purpose of simplifying a big data system and effective understanding of its overall architecture, we have introduced our own mapping notations (based on data flow, processing and analysis) for modelling the big data uses. In addition, from our study of literatures and use cases, we found that there is a gap for improvements in mapping notations. Hence, we create our mapping notations which are presented in Table 2.

Notations 1, 2 and 3 present in ‘data source’, ‘data collection, processing and loading’, ‘data analysis and aggregation’, ‘job model specification’. Notation 4 is used in ‘interface and visualization’. Notations 5 and 6 are used in ‘job and model specification’. Finally, Notations 7 and 8 are links between elements of Notation 1-6.

IV. USAGE OF REFERENCE ARCHITECTURE NOTATION

To demonstrate the use of the notation and its broad applicability, we use a number of cases. In this section, we present the production of the mapping reference architecture.

Each use case is presented and subsequently each use case is mapped to the reference architecture. For most accurate information on implementation architectures, the reader is referred to the original citations.

Due to the page limitation, we only present our works on Facebook and Amazon Web Services for Analytics.

1) Facebook

Facebook collects data from two sources: a federated MySQL tier containing user data; and web servers for event log data. The structured data from the Federated MySQL is copied, compressed and stored in the Production Hive-Hadoop cluster. The Scribe servers aggregate event logs and process the data in Hadoop Distributed File System (HDFS). HDFS data is compressed periodically, and moved to Production Hive-Hadoop clusters for further processing.

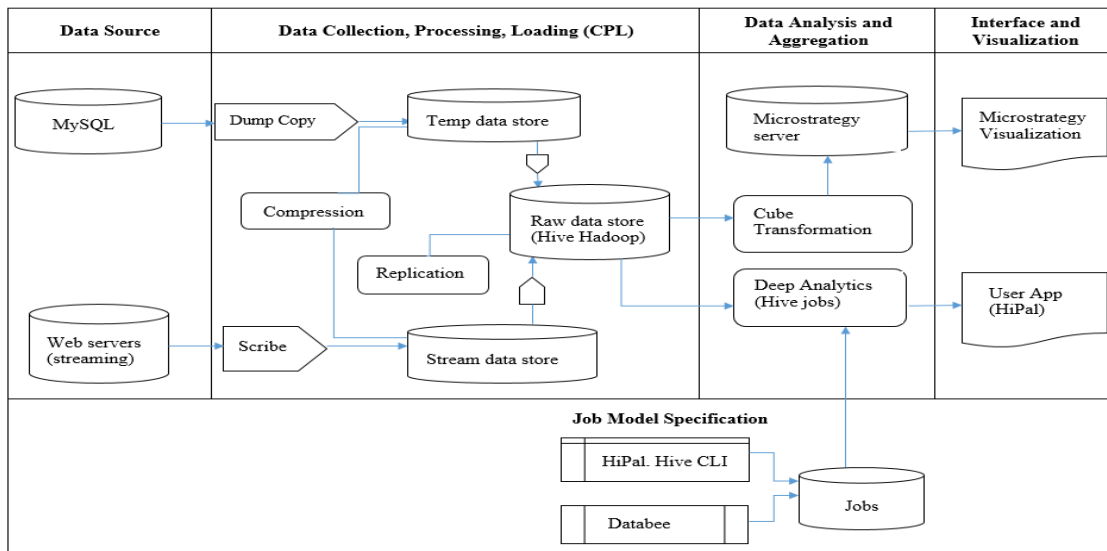


Fig. 1. Mapping Facebook Use Case Reference Architecture

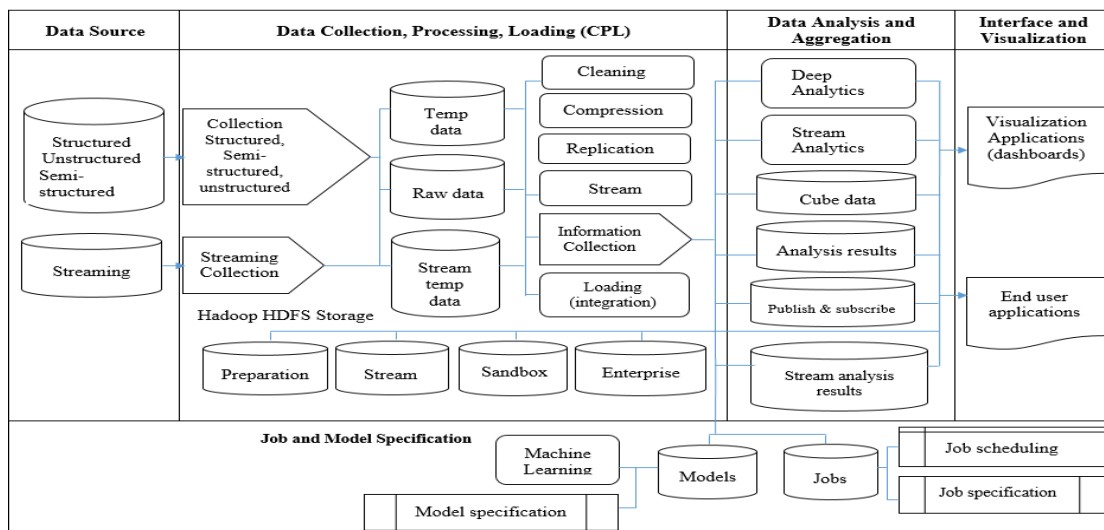


Fig. 2. Big Data Use Case Reference Architecture

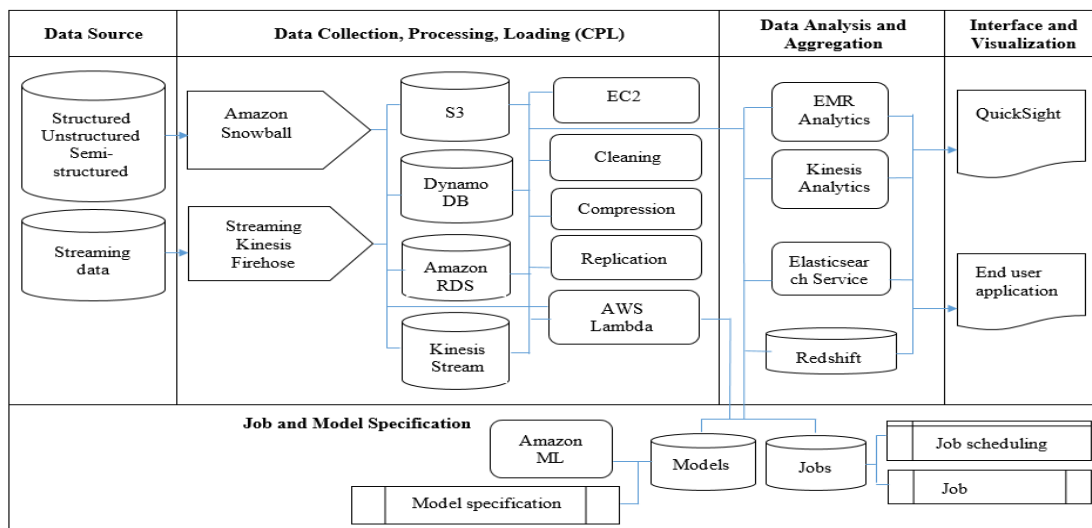


Fig. 3. Amazon Analytics Services Reference Architecture

Facebook differentiates low or high priority clusters for data analysis. High priority jobs are executed in the Production Hive-Hadoop cluster whereas lower priority jobs and ad hoc analysis jobs are executed in the Ad hoc Hive-Hadoop cluster. In addition, data is replicated from the Production cluster to the Ad hoc cluster. The results of data analysis are stored in the Hive-Hadoop cluster or MySQL tier for Facebook users.

HiPal, a graphical user interface with Hive CLI (a Hive command-line interface) is used to accommodate ad hoc analysis. Databee (a python framework) is used for execution and scheduling of periodic batch jobs in the production cluster. Microstrategy Business Intelligence (BI) tools are used for cube dimensional analysis.

The mapping Facebook use case reference architecture is presented in Fig. 1.

2) Reference Architecture

From the mapping with each big data use case, we compiled a comparison table in order to produce the reference architecture for Big Data system. The comparison table is presented in Table 1 as well as the reference architecture at Fig. 2.

3) Amazon Web Services with Reference Architecture

We also mapped our reference architecture with Amazon Web Services [37]. At this stage, we do not target for a specific requirement, business needs or use case, but rather focus on evaluating our produced architecture with a commercial service, Amazon. Fig. 3 shows the corresponding mapping of the reference architecture.

V. ANALYSIS

We have explored and compared big data use cases at Facebook, LinkedIn, Twitter and Netflix, and subsequently we formed a reference architecture by mapping big data use case. Our comparison table can be found at Table 1 and our reference architecture is presented in Fig. 2. Furthermore, we mapped and produced a reference architecture based on the commercial technology and services of Amazon. Based on this mapping it is clear that the reference architecture can cover this broad range of cases.

Our reference architecture for big data systems is composed of different components in which different functionalities and processes are described in graphical notations with short descriptions. It also provides specific notations such as data stores, data collection, and data processing. Furthermore, we grouped closely linked items such as data collection, processing, loading into one component to represent logical flow and linkage of closely related items as well as to provide effective overall architecture at high level abstract.

Our work provides alternative approach to Paakkonen et al [17]. That approach includes eight blocked items and similar notations of different processes or functionalities. Furthermore, the multiple connections between data flows of items and several occurrences of similar notations existing in the approach can sometimes be difficult to follow.

In this paper, we only looked at one existing big data use case and subsequently produced our reference architecture due

to the page limitation. In addition, we produced mapping reference architecture with Amazon services. At this stage, our reference architecture can still be generic, and different use cases of big data and technologies are emerging at speed. In the future, we plan to carry out further big data use cases from different domains as well as evaluate the reference architecture in business environments, and subsequently make improvements to the reference architecture.

We did not concentrate on big data methods or technologies but our focus was rather on reference architecture at high level abstract. Considering big data key topics such as stream and batch processing, virtualization, machine learning, data mining, data stores, business intelligence and visualizations, cloud computing and effect of CAP (Consistency, Availability and Partition Tolerance) theorem will also enable us to make effective improvements to the reference architecture as well as to further work on lower level details (implementation).

VI. FUTURE WORK

All use cases explored in this paper are high-tech internet related businesses (Facebook, LinkedIn, Twitter, Netflix and Amazon), hence further research on other industries with their use cases will provide further input and validation. Even though, this will require extensive work and research, it enables to discover key elements, challenges, techniques and methods, hence would allow us making better reference architecture.

In addition to exploring further cases, we also plan to work on data analytics related technology and services including open sources and commercial offers, and cloud computing. We see that a well-designed big data system needs the realization of key and emerging technology and services.

Finally, industrial application and evaluation of the architecture will provide additional understanding and validity.

VII. CONCLUSION

The paper looked at reference architecture for big data systems. The overall goal was to simplify the realization of a big data system and gain clearer understanding of its overall architecture. First, we produced a high level reference architecture with own architecture notations for big data platform based on existing work [17]. We presented the reference architecture which was formed of five components upon logical linkage and flow, and modelled with specific notation of items such as data collection, processing, job model specification and visualization. Second, we validated the work by mapping the reference architecture with one existing big data use cases and presented the mapping use case architectures. Finally, we analyzed, produced and presented Amazon analytics services reference architecture by mapping the analytics services of Amazon with the reference architecture. It was proved to be producing a coherent result in all cases.

ACKNOWLEDGEMENT

This research has been partially sponsored by EU H2020 FIRST project, Grant No. 734599.

REFERENCES

- [1] D. Barton, D. Court, Making Advanced Analytics Work For You, Harvard business review, 90 (10) (2012), pp. 79–83
- [2] A. McAfee, E. Brynjolfsson, Big data: the management revolution, Harvard business review, 90 (10) (2012), pp. 60–68
- [3] B. H. Wixom, B. Yen, M. Relich, Maximizing Value from Business Analytics, MIS Quarterly Executive, June 2013(12:2)
- [4] J.M. Tien, Big data: Unleashing information, Journal of Systems Science and Systems Engineering, Springer
- [5] E. Turban, R. Sharda, D. Delen, D. King, 2011. Business intelligence, A Managerial Approach. 2nd ed. Boston: Prentice Hall
- [6] T. H. Davenport, Harvest Business Review, Competing on Analytics [online]. January 2006 Issue.
- [7] H. Demirkan, D. Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decision Support Systems 55 (2013) 412–421
- [8] F. Provost, T. Fawcett, Data Science and its relationship to Big Data and Data-Driven Decision Making, Big Data, Mary Ann Liebert, Inc. Vol.1 No.1 March 2013
- [9] M.D. Assuncao, R. N. Calheiros, S. C. Bianchi, M. A. S. Netto, R. Buyya, R., Big Data computing and clouds: Trends and future directions J. Parallel Distrib. Comput. 79–80 (2015) 3–15
- [10] C.L.P. Chen, C. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on Big Data, Inf. Sci. 275 (2014) 314–347.
- [11] A. Angelov, P. Grefen, D. Greefhorst, A framework for analysis and design of software reference architectures, Inf. Softw. Technol. 54 (2012) 417–431.
- [12] M. Chen, S. Mao, Y. Liu, Big data: a survey, Mob. Netw. Appl. 18 (2014).
- [13] C.L.P. Chen, C. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on Big Data, Inf. Sci. 275 (2014) 314–347.
- [14] E. Begoli, A short survey on the state of the art in architectures and platforms for large scale data analysis and knowledge discovery from data, in: The 10th Working IEEE/IFIP Conference on Software Architecture & 6th European Conference on Software Architecture (WICSA/ECSA), Helsinki, Finland, 20–24 August, 2012.
- [15] X. Wu, G. Wu, W. Ding, Data mining with big data, IEEE Trans. Knowl. Data Eng. 28 (2014) 97–106.
- [16] M. Galster, P. Avgeriou, Empirically-grounded reference architectures: a proposal, in: Joint ACM SIGSOFT Conference on Quality of Software Architectures and ACM SIGSOFT Symposium on Architecting Critical Systems, Boulder, Colorado, USA, June 20–24, 2011.
- [17] P. Paakkonen, D. Pakkaka, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems", Big Data Research 2 (2015) 166–186
- [18] A. Zimmermann, K. Sandkuhl, M. Pretz, M. Falkenthal, D. Jugel, M. Wissotzki, Towards and integrated service-oriented reference enterprise architecture, in: International Workshop on Ecosystem Architectures, Saint Petersburg, Russia, 19 August, 2013.
- [19] J. A. Zachman A framework for information systems architecture, IBM Systems Journal, Volume: 26, Issue: 3, 1987
- [20] Y. Demchenko, C. Ngo, P. Membrey, Architecture framework and components for the Big Data Ecosystem, SNE Technical Report, University of Amsterdam, September 12, 2013.
- [21] R. Schmidt, M. Möhring, Strategic alignment of cloud-based architectures for big data, in: 17th IEEE International Enterprise Distributed Object Computing Conference Workshops, Vancouver, Canada, 9–13 September, 2013.
- [22] K.A. Doshi, T. Zhong, Z. Lu, X. Tang, T. Lou, G. Deng, Blending SQL and NewSQL approaches reference architectures for enterprise big data challenges, in: The International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Beijing, China, 10–12 October, 2013.
- [23] T. Zhong, K.A. Doshi, X. Tang, T. Lou, Z. Lu, H.Li, On mixing high-speed updates and in-memory queries a big-data architecture for real-time analytics, in: IEEE International Conference on Big Data, Santa Clara, California, USA, 6–9 October, 2013.
- [24] C.E. Cuesta, M.A. Martinez-Prieto, J.D. Fernandez, Towards an architecture for managing big semantic data in real-time, in: 7th European Conference on Software Architecture, Montpellier, France, 1–5 July, 2013.
- [25] M. Westerlund, U. Hedlund, G. Pulkkis, K. Bjork., A generalized scalable software architecture for analyzing temporally structured big data in the cloud, New Perspect. Inform. Syst. Technol. 1 (2014) 559–569.
- [26] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. S. Sarma, R. Murthy, H. Liu, Data warehousing and analytics infrastructure at Facebook, in: 2010 ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, 6–11 June, 2010.
- [27] J. Kreps, N. Narkhede, J. Rao, Kafka: a distributed messaging system for log processing, in: The 6th International Workshop on Networking Meets Databases, Athens, Greece, 12 June, 2011.
- [28] L. Wu, R. Sumbaly, C. Riccomini, G. Koo, H.J. Kim, J. Kreps, S. Shah, Avatara: OLAP for web-scale analytics products, in: 38th International Conference on Very Large Databases, Istanbul, Turkey, 27–31 August, 2012.
- [29] R. Sumbaly, J. Kreps, S. Shah, The “Big Data” Ecosystem at LinkedIn, in: 2013 ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 22–27 June, 2013.
- [30] G. Mishne, Fast data in the era of big data: Twitter’s real-time related query suggestion architecture, in: The 2013 ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 22–27 June, 2013.
- [31] J. Lin, D. Ryaboy, Scaling big data mining infrastructure: the Twitter experience, ACM SIGKDD Explor. Newsl. 14 (2013) 6–19.
- [32] G.L. Lee, J. Lin, C. Liu, A. Lorek, D. Ryaboy, The unified logging infrastructure for data analytics at Twitter, in: The 38th International Conference on Very Large Databases, Istanbul, Turkey, 27–31 August, 2012.
- [33] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, J. Lin, EarlyBird: real-time search at Twitter, in: 2012 IEEE 28th International Conference on Data Engineering, Washington, DC, USA, 1–5 April, 2012.
- [34] X. Amatriain, Big & Personal: data and models behind Netflix recommendations, in: The 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Chicago, Illinois, USA, 11 August, 2013.
- [35] X. Amatriain, J. Basilico, System architectures for personalized recommendations, Available via Netflix, <http://techblog.netflix.com/2013/03/system-architectures-for.html>, accessed 05 August, 2016.
- [36] J. Boulon, et al., Chukwa: a large-scale monitoring system, in: Cloud Computing and its Applications, Chicago, Illinois, USA, 22–23 October, 2008.
- [37] Amazon, Big Data Analytics Options on AWS, January 2016. Available via http://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf, accessed 15 August, 2016.
- [38] Z. Sun, K. D. Strang, J. Yearwood, Analytics service oriented architecture for enterprise information systems, Proceeding iiWAS '14 Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services Pages 508-516
- [39] Gartner, Analytics, Gartner IT Glossary. Available via Gartner <http://www.gartner.com/it-glossary/analytics/>, accessed 01 September 2016.
- [40] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management 35 (2015) 137–144