Simplifying Big Data Analytics Systems with A Reference Architecture

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. The internet and pervasive technology like the Internet of Things (i.e. sensors and smart devices) have exponentially increased the scale of data collection and availability. This big data not only challenges the structure of existing enterprise analytics systems but also offer new opportunities to create new knowledge and competitive advantage. Businesses have been exploiting these opportunities by implementing and operating big data analytics capabilities. Social network companies such as Facebook, LinkedIn, Twitter and Video streaming company like Netflix have implemented big data analytics and subsequently published related literatures. However, these use cases did not provide a simplified and coherent big data analytics reference architecture as well as currently, there still remains limited reference architecture of big data analytics. This paper aims to simplify big data analytics by providing a reference architecture based on existing four use cases and subsequently verified the reference architecture with Amazon and Google analytics services.

Keywords: Big Data, Analytics, Reference Architecture.

1 Introduction

Cloud computing and the evolution of Internet of things technology with their applications (digital data collection devices such as mobile, sensors, etc.) have exponentially increased the scale of data collection and data availability [1, 2]. Digital technology (social network applications, etc.) facilitates people interact with data and more data is shared ever before [1]. Across every industry, businesses focus on putting data at the core of business transformation to better understand their customers, create product and service differentiation as well as to simply lower their costs and to increase efficiency [1, 2, 3]. As a result, data becomes one of the primary assets for organizations [2].

The need of enterprise agile analytics is growing rapidly to satisfy the demands driven by big data, and increasingly complex business analysis and analytics [1]. Thus, data-driven analytics become a mainstream process for businesses [2]. [4] defined analytics as "techniques used to analyse and acquire intelligence from big data". We refer analytics as "the collection, processing, analysis (machine learned models, statistics, etc.,) and visualization of big data to produce insight".

Contemporary data analytics systems cannot sufficiently handle big data as well as 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 [1]. The problem of how to do this presents a challenge to businesses and the research community. Reference architecture can help in the creation of concrete architectures [19], however there still remains a lack of reference architectures as well as a coherent architecture of big data analytics.

The aim of this paper is to explore the realization of simplifying big data analytics systems, 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 analytics system [5]. The contributions of this work are: a) to provide high level independent reference architecture for design a big data analytics platform based on existing publications b) to verify the proposed reference architecture using existing different big data analytics.

The structure of the paper is as follows: Related work and literature study are presented in Section 2. Production of the reference architecture with big data use cases as well as LinkedIn big data use case and its mapping the reference architecture are presented in Section 3. Due to page limitation, we do not present all other use cases. Finally, a short conclusion is provided in Section 4.

2   Related Work

In this section, we reviewed existing big data analytics use cases and related reference architectures based on past publications.

2.1   Big Data Analytics System

Over the last few years, several large-scale data driven analytics have been published from Social network domain such as Facebook, Twitter and LinkedIn to Entertainment Video-streaming such as Netflix. Also, other industry like the health industry has recognized and demonstrated their huge interest and opportunity of big data analytics. Data including streaming are collected by Facebook, LinkedIn and Twitter from their users [6, 9, 10, 11, 12]. Facebook and LinkedIn often deals with both traditional data such as web logs and streaming whereas Twitter mostly deals with streaming data [6, 7, 8, 10, 11, 12]. The collected data are then managed on a batch or streaming processing with own custom functionalities. Several data analysis and aggregations such as Deep Analytics, trained models and specified jobs, algorithm service with Hadoop HDFS are largely implemented in clustering and distributed computations. Furthermore, Netflix collects users' events, which are then processed in Online, Nearline and Offline computations [13, 14]. Recommendations processed via the Online data analysis are available to users.

2.2   Reference Architecture of Big Data Analytics System

Service-oriented architecture and its principles facilitate software design, architecture and implementation in the enterprise software domain [16, 17]. In addition, several reference architectures have been published. Galster and Avgeriou [18] proposed software reference architecture based on a step-wise process. It includes six steps; deciding a type for reference architecture, selection of design strategy, empirical acquisition of data, construction of reference architecture, enabling of variability, and evaluation. Angelov, Grefen and Greefhorst [19] presented software reference architectures and subsequently claimed that it facilitate better architecture design, and hence will lead to better success. Their framework is for design and analysis of software reference architectures which forms of a multi-dimensional classification space, and five types of reference architectures.

For the context of big data, limited architectures have been suggested. Paakkonen and Pakkala [5] 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]. A service and deployment model for implementing big data pipeline in the cloud domain was suggested by Schmidt and Möhring [21]. Reference architectures for the integration of SQL and NewSQL databases was proposed for supporting different growth patterns in enterprise data traffic [22]. In addition, big data architecture for in-memory storage system and distributed task execution analysis was suggested and validated [23]. Another tiered architecture (SOLID) for separating big data management from data generation and semantic consumption was also proposed [24]. For predictive and real-time analytics, a generalized software architecture was suggested. This was intended for structured historical and temporal data [25].

Literature review shows that there are limited reference architectures in the big data analytics context as well as lack of concrete or coherent reference architecture. This establishes a need for further research in reference architecture for big data analytics.

3   Reference Architecture for Big Data Analytics

Reference architecture can help in the creation of concrete architectures as well as the understanding of the overall architecture with typical related functionality and data flows in a big data analytics system [19]. A reference architecture is also useful for analysing 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 analytics use cases as shown in Table 1, we have divided the key elements of big data uses into five components: Data source; Data

Table 1. Big Data Analytics User Case Comparison

Big Data Analytics Use Case	Data Source	Data Integration	Data Analysis, Aggregation	Model, Job	Interface, Visualization
Facebook	MySQL, Web Servers Logs	Dump database copy, Scribe, Hadoop raw data, Replication.	Cube, Deep analytics, Hive jobs	Jobs stored Database	Microstrategy Visualization and User Applications
LinkedIn	Oracle, Users Activity Data	Oracle (Snapshot), Kafka, Hadoop Data cleaning, De-duplication, Replication.	Deep Analytics, Hive, Pig, MR, Voldemort, Avatara.	Azkaban Framework	Visualization applications (enterprise), LinkedIn users.
Twitter	Firehose (tweets), Updater, Queries	Tokenization, Stats Queries, Filter and Personalization of Firehose, Ranking	Hadoop HDFS (Ranking algorithm, analysis), Cache database	Ranking algorithm	Twitter users.
Netflix	Stream data (Users and online service)	Chukwa, Stream signals, Manhattan Framework	Deep Analytics: Off/Online Nearline. Algorithm	Machine Learning, Pig .	Netflix Users, Visualization applications (enterprise).

Integration; Data Analysis and Aggregation; Interface and Visualization; and Job and Model Specification.

a. *Data Source*: it refers to the original source of data to be collected. Data can be stored as traditional database such as relational data or NoSQL. Data can be structured, unstructured, semi-structured or streaming.

b. *Data Integration*: it involves the Collection, Processing and Loading of data. The collection deals with getting data from sources (multiple) 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 a 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.

c. *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.

   d. *Interface and Visualization*: they represent the end users (i.e. ad hoc analysis, queries, etc.) as well as visualization based applications such as dashboards.
   e. *Model and Job Specification*: This covers machine learning, models trained, specifications and scheduling of jobs with their storages.

Furthermore, we have introduced our own mapping notations (based on data flow, integration, processing, analysis and interface) for modelling the big data uses. Our primary intention was of simplifying a big data system and effective understanding of its overall architecture. Additionally, our study of literatures and use cases of big data analytics prompted a gap for improvements in mapping notations. Thus, we created our mapping notations which are presented in Table 2 and subsequently produced a reference architecture which is presented in Figure 1.

Table 2. Big Data Analytics Reference Architecture Notations

No	Notation	Description
1		The notation of a data store or database which can represent structured, unstructured or semi-structured data. We often indicate a short text or description of the notation.
2		This represents the Extraction 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 specifically indicates Visualization based applications. We include a short text.
5		This specifically indicates Ad Hoc or End Users based applications. We include a short text.
6		This represents the Job Schedule for batch processing. A short text is included.
7		This represents the Job Specification for batch processing. A short text is included.
8		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.
9		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.

Notations 1, 2 and 3 present in 'data source', 'data integration', 'data analysis and aggregation, 'job model specification'. Notation 4 and 5 are used in 'interface and visualization'. Notations 6 and 7 are used in 'job and model specification'. And Notations 8 and 9 are links between elements of Notation 1-7.

3.1 Mapping LinkedIn

At LinkedIn, users' activity data is collected and transferred with database snapshots. The activity data refer to the usage of LinkedIn' services which are treated as streaming events. Kafka, a distributed, publish/subscribe messaging system, is used for the collection of the streaming events [7].
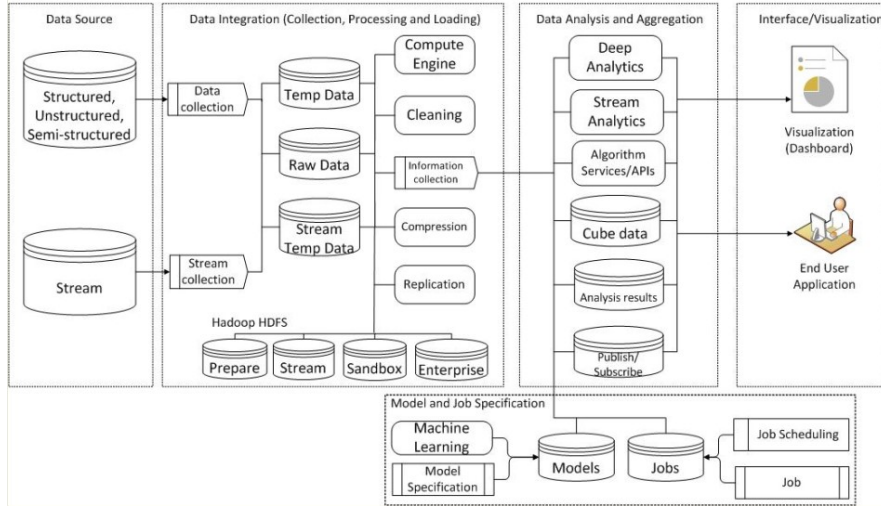
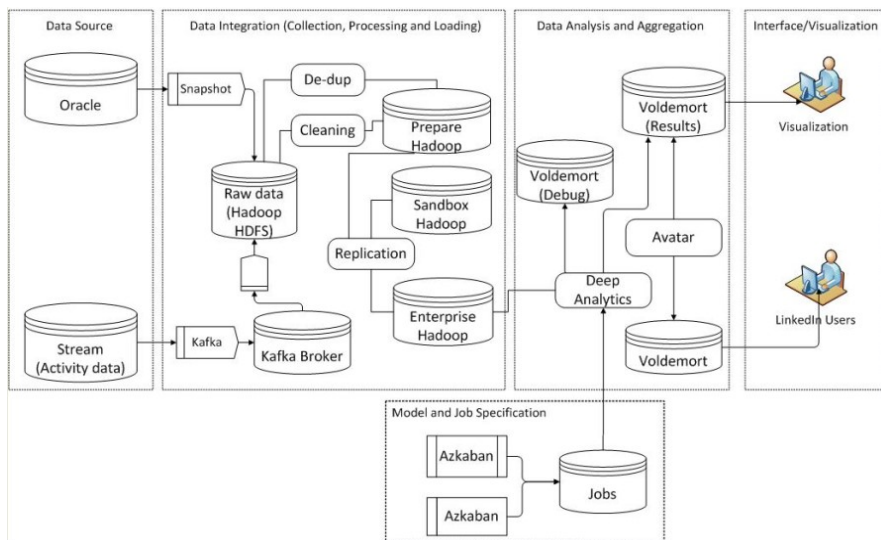**Fig. 1.** Big Data Analytics Reference Architecture

Fig. 2. Mapping LinkedIn Big Data Analytics Reference Architecture

With Kafka, producers report events to a broker, and consumers read data at their own space [7]. The event data of Kafka are moved to Hadoop ETL cluster for further processing tasks such as combining, de-duplication. Data from the Hadoop ETL cluster is copied into production and development clusters.

Workloads are typically performed in this way; initial experiments are carried out in the Development cluster, and then are transferred to the production cluster upon successful review and testing. Analysis results from the production cluster are stored

in both offline (debugging) and online database. Moreover, these results may also be used again in the Kafka cluster [7].

Azkaban is used for scheduling workloads due to its support of a diverse set of jobs. Azkaban is a batch workflow job scheduler created at LinkedIn to execute Hadoop jobs. An instance of Azkaban is run in each Hadoop environment. Scheduled Azkaban workloads can be MapReduce, Pig, shell script, or Hive jobs. Avatara, a scalable and fast OLAP serving system created at LinkedIn is used for analysing and processing data which then are stored in Voldemort databases [8].

We mapped LinkedIn big data analytics with our produced reference architecture which is presented in Figure 2.

4 Conclusion

Our aim was to simplify big data analytics systems, hence we explored and compared big data use cases at Facebook, LinkedIn, Twitter and Netflix and subsequently formed a reference architecture with own mapping notations for simplifying a big data analytics system and effective understanding of its overall architecture. Our comparison table can be found at Table 1 and our reference architecture is presented in Figure 1. Furthermore, we mapped and produced LinkedIn Use Case Reference Architecture which is presented in Figure 2. Due to the page limitation, we did not present our work of the other use cases. Based on the mappings we produced, it is clear that the reference architecture can cover this broad range of cases and produced a coherent result. Industrial application and evaluation of the architecture will provide additional understanding and validity.

References

1. Demirkan, H., Delen, D.: Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decision Support Systems 55, 412–421 (2013)
2. Barton, D., Court, D.: Making Advanced Analytics Work For You, Harvard business review, 90 (10), pp. 79–83 (2012)
3. McAfee, A., Brynjolfsson, E.: Big data: the management revolution, Harvard business review, 90 (10), pp. 60–68 (2012)
4. X. Wu, G. Wu, W. Ding, Data mining with big data, IEEE Trans. Knowl. Data Eng. 28 (2014) 97–106.
5. Paakkonen, P., Pakkaka, D.: "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems", Big DataResearch2, 166–186 (2015)
6. 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, Indi-anapolis, Indiana, USA, 6–11 June (2010)

7. Kreps, J., Narkhede, N., Rao, J.: Kafka: a distributed messaging system for log pro-cessing, in: The 6th International Workshop on Networking Meets Databases, Athens, Greece, 12 June (2011)

8. Wu, L., Sumbaly, R., Riccomini, C., Koo, G., Kim, H.J., Kreps, J., Shah, S.: Avatara: OLAP for web-scale analytics products, in: 38th Inter-national Conference on Very Large Databases, Istanbul, Turkey, 27–31 August (2012)

9. Mishne, G.: 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)

10. Lin, J., Ryaboy, D.: Scaling big data mining infrastructure: the Twitter experience, ACM SIGKDD Explor. Newsl. 14, 6–19 (2013)

11. Lee, G.L., Lin, J., Liu, C., Lorek, A., Ryaboy, D.: The unified logging infrastructure for data analytics at Twitter, in: The 38th International Conference on Very Large Databases, Istanbul, Turkey, 27–31 August (2012)

12. Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S., Lin, J.: EarlyBird: real-time search at Twitter, in: 2012 IEEE 28th In-ternational Conference on Data Engineering, Washington, DC, USA, 1–5 April (2012)

13. Amatriain, X.: Big & Personal: data and models behind Netflix recommendations, in: The 2nd International Workshop on Big Data, Streams and Heteroge-neous Source Mining: Algorithms, Systems, Programming Models and Applica-tions, Chicago, Illinois, USA, 11 August (2013)

14. Amatriain, X., Basilico, J.: System architectures for personalized recommendations, Avail-able via Netflix, http://techblog.netflix.com/2013/03/system-architectures-for.html, accessed 05 August, 2016.

15. Boulon, J., Konwinski, A., Qi, R.: Chukwa: a large-scale monitoring system, in: Cloud Computing and its Applications, Chicago, Illinois, USA, 22–23 October (2008)

16. A. Zimmermann, K. Sandkuhl, M. Pretz, M. Falkenthal, D. Jugel, M. Wissotzki, Towards and integrated service-oriented reference en-terprise architecture, in: International Workshop on Ecosystem Architectures, Saint Petersburg, Russia, 19 August, 2013.

17. J. A. Zachman A framework for information systems architecture, IBM Systems Journal, Volume: 26, Issue: 3, 1987

18. M. Galster, P. Avgeriou, Empirically-grounded reference architectures: a pro-posal, in: Joint ACM SIGSOFT Conference on Quality of Software Architectures and ACM SIGSOFT Conference on Quality of Software Architectures and ACM SIGSOFT Symposium on Architecting Critical Systems, Boulder, Colorado, USA, June 20–24, 2011.

19. A. Angelov, P. Grefen, D. Greefhorst, A framework for analysis and design of software reference architectures, Inf. Softw. Technol. 54 (2012) 417–431.

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 architec-tures 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 Soft-ware Architecture, Montpellier, France, 1–5 July, 2013.

25. M. Westerlund, U. Hedlund, G. Pulkkis, K. Bjork,, A generalized scalable software architecture for analyz-ing temporally structured big data in the cloud, New Perspect. Inform. Syst. Technol. 1 (2014) 559–569.