BIG DATA ANALYSIS IN SOCIAL MEDIA

JelenaLjubenović*, OgnjenPantelić, Ana Pajić Simović[1]
[1]University of Belgrade, Faculty of organizational sciences, Serbia
*Corresponding author, e-mail: jelena.ljubenovic@fon.bg.ac.rs

Abstract: *The Internet community has changed a lot in the past few years. Social media has changed the way of communication and connected millions of people. With the increase of information that is being generated in these communication channels, there is a new problem: storing all the data. Relational databases are no longer most suitable storage for social media data. Two technical entities have come together. First is big data for massive amounts of data. Second, there's advanced analytics, which is a collection of different tool types for getting meaningful information from the data. Together they make big data analytics, a new practice in BI today.*
This paper presents big data, social media, and social media data infrastructure. It also shows tools for analyzingbig data. The contribution of this paper is to provide an overview of tools and techniques for analyzing social media data either for research or business purposes.

Keywords*: big data, big data analysis, social media, social media data, social media data infrastructure*

1. INTRODUCTION

Information overload and the limits of human cognition in dealing with information are much older than new technologies and Internet network. As hardware technology has improved, software technology has improved as well. It's never been that easy to communicate and stay connected with other people. But, the information and data overload have also increased. A brand new way of communicationand new technologies have brought massive amounts of various data and a new problem: How to store it? The answer is big data. It is not regular, structured data and can be explained through three dimensions: volumes, variety, and velocity.

Companies are recognizing the potential value of this data and they are putting great efforts in order to extract some useful and meaningful information from it. The only way to derive value from big data is the use of analytics. It must be analyzed and the results should be used by decision makers and organizational processes in order to generate value. The type of information extracted from social media is constantly in motion. Analyzing social media data can contribute to better understanding people's behavior, their needs and habits. New data sources bring new challenges and new challenging questions.

2. BIG DATA

Big data is a very popular buzzword used by IT companies to increase their sales. However, it is more than just a word. It is real and extremely important technology trend with huge business potential. Big data can often be associated with social media. Rather, it is a combination of data-management technologies that have evolved over time. Big data enables organizations to store, manage and manipulate vast amounts of data. Over time, size and type of data that needs to be stored have changed and new ways of storing data appeared. Big data does not represent regular, structured data. It can be explained through three dimensions: volumes, velocity, and variety, so-called three Vs (Xiang, Du, Ma, and Fan, 2017).

Volume is also known as the big part. It represents the amount of data that is being generated in everyday use. As we generate more and more data, this point of big data will keep growing. Some experts consider that the size of memory that is used for storing data is measured in Petabytes.

Velocity represents the speed of generating new data or the frequency of delivering data. It shows how fast that data is processed. Big data improves real-time data analyzing and enables fast access to a large amount of data.

Variety is about the different type of data and file types available in everyday use. For example there are tweets (500 million per day), video files on YouTube (3.25 billion hours of watching video each month), posts on Facebook (510,000 per second), profiles on LinkedIn (over 5 millions) and then there are also all the different log files and other data produced by any computer system we use. So the explanation of big data could be that it is dealing withhigh-velocity and high-volume data streams while data types are highly diverse.

Big data requires methods and tools that can be used for analyzing and extracting patterns from large-scale data. The large Volume of data poses a challenge to conventional computing environments. It requires

scalable storage and a strategy to data querying and analysis. (Choi, Chan, and Yue, 2017) The Volume of data is also a major advantage of big data. Dealing with the Variety among different data is a unique challenge for Big Data, which requires preprocessing of unstructured data in order to extract structured information. The importance of dealing with Velocity in big data is the quickness of the feedback, translating data input into usable information (Bello-Orgaz, Jung, and Camacho, 2016).

3. SOCIAL MEDIA

Social media was originally intended for people who want to interact with friends and family but later many companies started using it as a very powerful tool for communication with customers. Now we can define it as a technology based on computers and Internet that is meant for facilitating the sharing of information and ideas. A huge advantage that social media has is the ability to connect with anyone and share information as long as they also use social media. Some of the most popular social media services are Facebook, Twitter, LinkedIn, Instagram, Pinterest, YouTube etc.

Twitter is a real-time service that allows users to post short messages of 140 or fewer characters. Messages posted on Twitter are called tweets. The network infrastructure of “friends“ and “followers“ on Twitter is asymmetric. User’s friends are the accounts that user is following and user’s followers arethe accounts that are following the user. Very important dimension of Twitter is the ability to analyze and track user activity. Retweet function is a form of data sharing. It is a way for Twitter users to participate in discussion without starting it. Twitter offers Apps for various mobile phones and tablets. Interactions or integrations with Twitter are done using the Twitter APIs (Newman, 2017).

Facebook is social networking site. Originally it was designed by Mark Zuckerberg for college students at Harward University. Today, Facebook is the largest social network with more than 1 billion users all over the world. Facebook allows users to send messages, post pictures, videos, links, share information and like content posted by other users.

LinkedIn is a social network for business people. LinkedIn started in 2003 and had only 2,700 members the first week. Today, it has more than 350 million members. It provides a way to connect with other professionals. LinkedIn is used for exchanging knowledge, ideas, as well as finding a job.

4. SOCIAL MEDIA DATA INFRASTRUCTURE

Next three chapters show data analytics infrastructure at Facebook, LinkedIn, and Twitter.

4.1. DATA INFRASTRUCTURE AT FACEBOOK

There are two sources that Facebook collects data from. These are federated MySQL tier and web servers tier.Web servers generate event-based log data and the data is collected to Scribe servers which are executed in Hadoop clusters. The Scribe servers aggregate log data, which is written to Hadoop Distributed File System (HDFS) (KamarSahu, 2015). Data from HDFS is compressed and transferred to Production Hive-Hadoop clusters for processing. Federated MySQL tier contains user data.

There are two different clusters for data analysis. Tasks that have a higher priority are executed in the Production Hive-Hadoop cluster. Tasks with lower priority and ad hoc analysis tasks are executed in Ad-hoc Hive-Hadoop cluster. Data from Production cluster is replicated to the Ad hoc cluster. Data analysis results are saved in Hive-Hadoop cluster or to MySQL tier. The graphical user interface (HiPal) or Hive command-line interface (Hive CLI) are used for specifying queries for Ad hoc analysis.

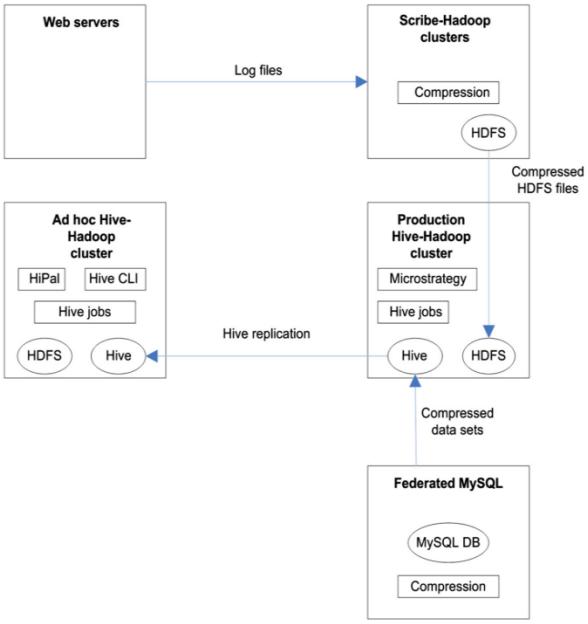

Figure 1. Data analytics infrastructure at Facebook
(SourcePääkkönen, P., &Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems)

4.2. DATA INFRASTRUCTURE AT LINKEDIN

LinkedIn started as a single monolithic application called Leo. As the site began to grow, it was necessary to "Kill Leo" and break it up into many services. As the amount of data that needed to be collected also grew, LinkedIn developed many custom data pipelines for streaming and querying data. The site needed to scale and each pipeline needed to scale. The result was the development of Kafka, distributed messaging platform which is used for collection of streaming events.

Data is collected from two sources: activity data that includes streaming events based on usage of LinkedIn's services and database snapshots. Kafka producers report events to topics at a Kafka broker, and Kafka consumers read data at their own pace. Kafka's event data is transferred to Hadoop ETL cluster for further processing. Data from the ETL cluster is copied into production cluster and development cluster.

Azkaban is a workload scheduler for supporting various types of tasks. It is realized as MapReduce, Pig, Hive jobs or shell script. Workloads are tested in the development cluster and transferred to production after testing. Analysis results are moved to an offline debugging database or to an online database.It can also be back to Kafka cluster.

Avatara is used for the preparation of OLAP data. Analysed data is read from the Voldemort database, pre-processed, and aggregated for OLAP, and saved to another Voldemort read-only database (Clemm J. 2015).

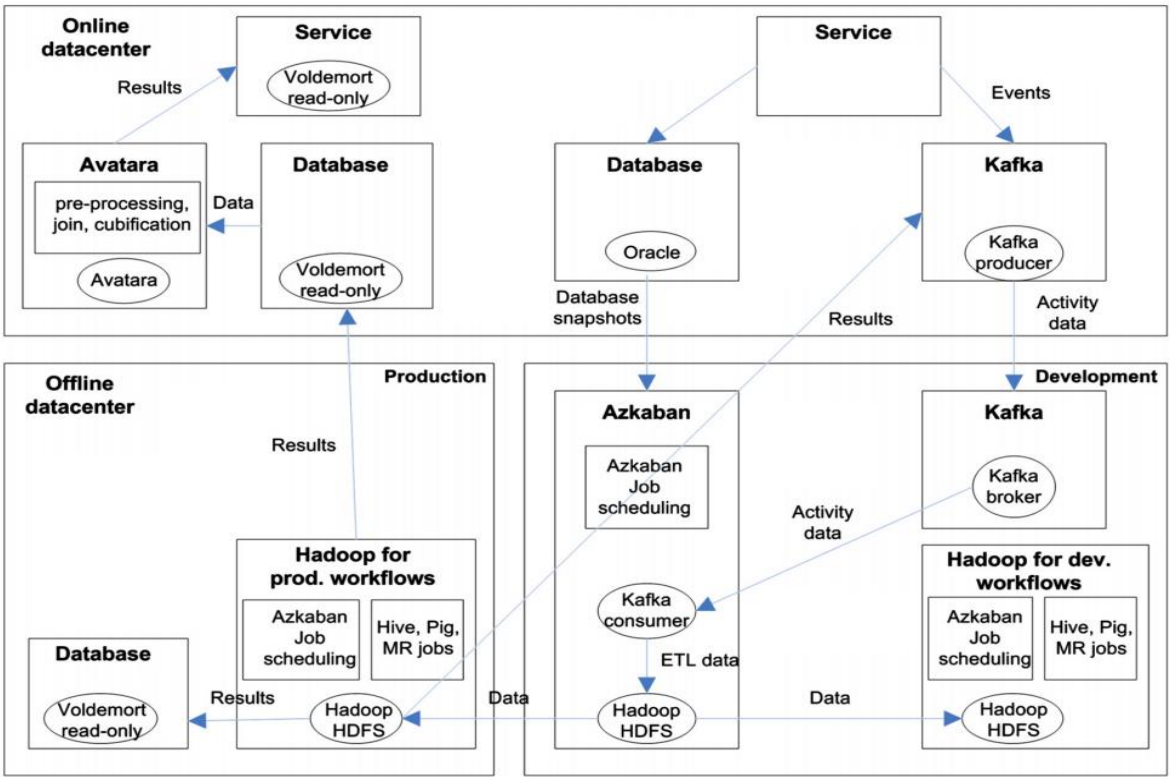

Figure 2. Data analytics infrastructure on LinkedIn
(SourcePääkkönen, P., &Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems)

4.3. DATA INFRASTRUCTURE AT TWITTER

Twitter has published two different implementation architectures of their infrastructure: one based on Hadoop batch processing which processes streaming data and the other custom-made solution that meets real-time requirements.

In infrastructure for real-time services, all requests (searching for tweets or user accounts via a QueryHose service) coming from Twitter are managed by Blender. The EarlyBird is a real-time engine designed for providing low latency and high throughput for search queries. FireHose service uses tweets as an input. After receiving tweets, EarlyBird servers filter data, personalize and index data. The EarlyBird servers also serve incoming requests from the QueryHose/Blender.
When tweet or a query is served, search assistance engine saves statistics into one of three in-memory stores. Session store saves user sessions, Query statistics store saves individual queries information, and Query co-occurrence store saves information about pairs of co-occurring queries. The results of the analysis are saved into Hadoop HDFS. In the end, Front-end cache takes results of analysis from the HDFS and presents them to Twitter users.

Twitter has three streaming sources from which data is extracted: tweets, Updater, and queries. REST API transforms tweets and queries in JSON format. Data format from Updater is not known. Blender and ingestion pipeline work as Stream temp data stores. The data is being processed. EarlyBird servers contain processed stream-based data (Stream data store). Hadoop HDFS storing the analysis results is modeled as a Stream analysis data store. Front-end cache (Serving data store) provides data to a Twitter app.

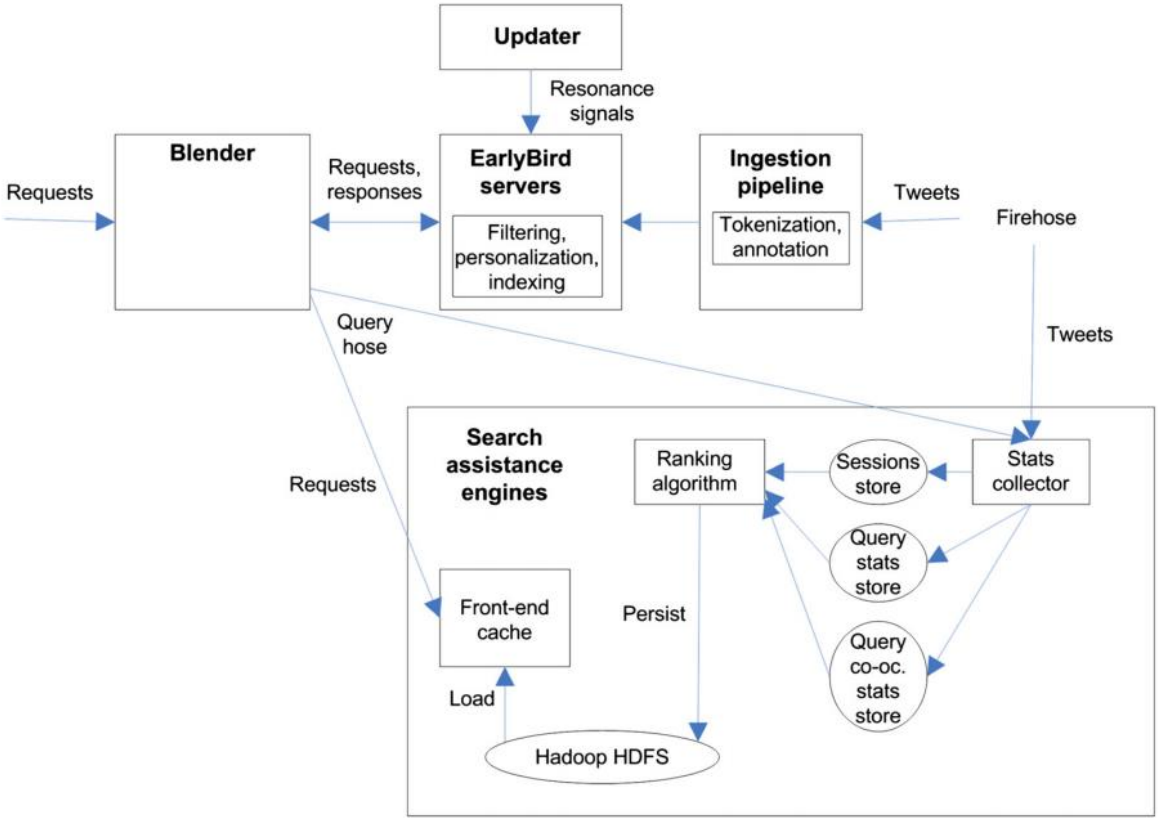

Figure 3. Data analytics infrastructure on Twitter
(SourcePääkkönen, P., &Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems)

5. ANALYSING BIG DATA IN SOCIAL MEDIA

Social media store a great amount of data. Underneath the surface of a Facebook profile, Facebook page, Instagram business profile or just tweets on Twitter, there is a huge information potential that is still just mostly unstructured data. Social media analysis is the analysis of structured and unstructured data from its channels.

At the moment, data in social media is available either via simple routines or require analysis in some research programs that use programming languages such as MATLAB, Java or Python. Researchers require:
• **Analytics dashboards** — non-programming interfaces for 'deep' access to 'raw' data.
• **Holistic data analysis** — tools required for combining multiple social media and other data sets.
• **Data visualization**—researchers also require visualization tools so information can be visualized in some schematic form.

There is a number of tools for big data text analyzing that can be used for analyzing social media data.Attensity's software is one of them.Attensity is one of the original text analytics companies. It uses a Hadoop framework (MapReduce, HDFS, and HBase) to store data. Another text analytics vendor is Calarbridge. Clarabridge CX Social is its tool specialized for exploring social media data.IBM offers several solutions such as Watson, SPSS, IBM Content Analytics with Enterprise Search (ICAES). IBM BigInsights for Apache Hadoop is an industry standard Hadoop. It is a data management platform. SAS is also solving big data problems. The SAS High Performance Analytics Server is an in-memory solution that allows a user to analyze complete data. It uses Hadoop Distributed File System (HDFS).

There are also custom applications for big data analysis. The purpose of these applications is to increase the speed of decision making or action taking. For example, the "R" environment is based on the "S" statistics and analysis language. It is an integrated suite of software tools and technologies designed to create applications that are used to ease data manipulation and analysis. It supports effective data-handling, operations for arrays and other ordered data types and tools to a wide variety of data analyses.

The Google Prediction API is an example of an emerging class of big data analysis application tools. It functions by looking for patterns and matching proscriptive, prescriptive or the other existing patterns. While doing this matching, it also learns. The source can be postings from Facebook, Twitter, Amazon, LinkedIn and the goal finding specific patterns of behavior. That can be very useful in business for a consumer products company. Based on the gathered information, the company might launch a new product or upgrade the old one. It is implemented as a RESTful API and supports .NET, Java, PHP, JavaScript, Python, Ruby, and many other programming languages.It also provides scripts for accessing the API as well as a client library for R.

Hadoop is an open source framework which is designed to solve problems associated with distributed data storage, analysis, and retrieval of big data. A distributed file system (HDFS) stores data and replicates it so it is always available. MapReduce is a distributed processing system for parallelizable problems. The idea is to design map functions that are used for generating a set of key/value pairs after which the reduce function will merge all the values associated with the same key. In the first step (Map), a problem is divided into many small problems and sent to servers for processing. In the second step (Reduce) the results of the previous step are combined to create the final results of the problem. Reduce can't begin until all the mapping is done, and it isn't finished until all instances are complete. The output of mapping and reduce are key-value pairs. Hive is a data warehouse system in which the user can specify instructions convert them to MapReduce tasks. Pig is another member of Hadoop family. It has similar functions like Hive, but it uses a programming language called Pig Latin, which is more oriented to data flows. HBase is another component of the Hadoop ecosystem, which implements Google's BigTable data store. Bigtable is a multidimensional sorted map. Elements in the map are an array of bytes. They are indexed by a row key, a column key, and a timestamp. The official Hadoop project is Apache. Hadoop solutions are also offered by companies such as Cloudera and Hortonworks. There is also a company MapR. It offers a commercial implementation of Hadoop (Zadrozny and Kodali, 2013).

6. COMPARISON OF SOCIAL MEDIA DATA INFRASTRUCTURE

With large incoming data and the fact that more and more data needs to be properly stored in order to allow data analysis, many social media platforms use Hadoop tools. Hive-Hadoop is an ad hoc cluster. It is a data warehousing framework built on Hadoop. It was created by Facebook and then given to Hadoop as a subproject of Hadoop ecosystem. Hive allows users to query and analyze data using SQL which makes data processing easier. In LinkedIn event data from Kafka cluster is transferred to Hadoop ELT (Extract – Transform - Load). ELT on Hadoop is a data integration process that provides flexibility in data processing environment. Also, Azkaban, a workload scheduler for supporting various types of tasks, is realized as MapReduce, Pig, Hive jobs or shell script. All the three social media networks, Facebook, Twitter and LinkedIn use Hadoop HDFS typically for storing structured data or storing the results of the analysis.

LinkedIn data is collected from two sources: activity data and database snapshots. Facebook also collects data from two sources. From federated MySQL tier it collects structured and stream-based user data. From web servers tier it generates event–based log data. The data is applied for batch-based data analysis. Every time user sends a Facebook message, add a contact on LinkedIn or tweets on Twitter digital data is being generated. These platforms are able to gather real-time data. All these events are captured, streamed and processed in digital form as they occur (Pigni, Piccoli, and Watson, 2016). Twitter uses a new stream processing system called Heron. It provides significant performance improvements and other advantages such as debugging-ability, manageability, and scalability. If something goes wrong, the design of Heron makes it transparent as to which part is failing. Each Heron Instance is executing a single task, so it is easier to debug that instance (Kulkarni, Bhagat, Fu, Kedigehalli, Kellogg, Mittal, and Taneja, 2015).

Relational databases are still applied for storing important user data as MySQL for Facebook and Oracle for LinkedIn. For storing data analysis results NoSQL databases or in-memory stores are used (LinkedIn's Voldemort). Log facilities are used for storing stream-based data. For Facebook it is Scribe, For LinkedIn Kafka. Technologies for data processing can be classified as a batch and stream processing. Real-time jobs require special technologies and algorithms (Twitter's ranking algorithm and the EarlyBird architecture). Batch processing is used for jobs with less strict timing requirements (Facebook's and LinkedIn's MapReduce, Hive and Pig scripts). Jobs for batch processing in LinkedIn are scheduled with Azkaban, in Facebook's with Databee. Processed data can also be presented with commercial Business Intelligence tools. Facebook uses MicroStrategy (Pääkkönen and Pakkala, 2015).

7. CONCLUSION

Big data analysis is a very powerful way to get right insights, information, and meaningful data from a huge volume of various data on social media platforms. It is now easier to collect data than it has ever been before. But extracting and utilizing useful information is not easy at all. From the business viewpoint, online reviews, photos, and the reviewer's personal information are very important in order to understand and influence user behavior. Social media isn't just used by marketers to find out about the position of their brand on the market or about customers opinion. It is used by healthcare institutions to discover health threats all around the world. The government is using it to predict terrorist attacks. Business intelligence can benefit from big data analytics. Social media data hides great informational potential. With the right tools and the right techniques, data can be transformed into valuable information.

As big data evolves, software vendors are competing to make the best tool for all sorts of analysis. There are numbers of programs designed to meet the needs of data analyzing. It's up to analyzer to choose the one that is most suitable for the given problem.

In this paper, various definitions of big data and social networks have been presented. A primary focus has been on big data analytics tools and social media data infrastructure for three different social media platforms: Facebook, LinkedIn, and Twitter. The paper can be a good starting point for further social media data exploring.

REFERENCES

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, 45-59.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51-65.

Newman, T. P. (2017). Tracking the release of IPCC AR5 on Twitter: Users, comments, and sources following the release of the Working Group I Summary for Policymakers. *Public Understanding of Science*, *26*(7), 815-825.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., &Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Zadrozny, P., &Kodali, R. (2013). *Big data Analytics Using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources*. Apress.

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*." O'Reilly Media, Inc.".

Clemm, J. (2015), A Brief History of Scaling LinkedIn, retrieved from https://engineering.linkedin.com/architecture/brief-history-scaling-linkedin

Kumar Sahu, B. (2015), Big data Analytics Reference Architectures - Big data on Facebook, LinkedIn and Twitter, retrieved from https://www.linkedin.com/pulse/big-data-analytics-reference-architectures-facebook-sahu.

Thusoo, A., Shao, Z., Anthony, S., Borthakur, D., Jain, N., SenSarma, J., ...& Liu, H. (2010, June). Data warehousing and analytics infrastructure at facebook.In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1013-1020).ACM.

Pääkkönen, P., &Pakkala, D. (2015).Reference architecture and classification of technologies, products and services for big data systems. *Big data Research*, *2*(4), 166-186.

Gandomi, A., &Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Batrinca, B., &Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, *30*(1), 89-116.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*(4), 1-34.

Borthakur, D., Gray, J., Sarma, J. S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., ...& Schmidt, R. (2011, June). Apache Hadoop goes realtime at Facebook. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (pp. 1071-1080).ACM.

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *CAIS*, *34*, 65.

Das, T. K., & Kumar, P. M. (2013). Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, *5*(1), 153.

Choi, T. M., Chan, H. K., &Yue, X. (2017).Recent development in big data analytics for business operations and risk management. *IEEE transactions on cybernetics*, *47*(1), 81-92.

Kulkarni, S., Bhagat, N., Fu, M., Kedigehalli, V., Kellogg, C., Mittal, S., ...&Taneja, S. (2015, May). Twitter heron: Stream processing at scale. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 239-250).ACM.

Pigni, F., Piccoli, G., & Watson, R. (2016). Digital Data Streams: Creating value from the real-time flow of big data. *California Management Review*, *58*(3), 5-25.