Analysis of issues and trends in Big Data Platforms

Muzamil Mehboob, Maruf Pasha, S. M. Waqas Shah and Urooj Pasha
(Bahauddin Zakariya University, Multan, Pakistan)

Abstract:
In this paper, we discuss the Big Data issues and related solutions in market. We are living in an era where information.is of utmost importance. Due to emergence of new technologies in recent years, data production is also increased for example IOT technology. Internet of things technology created a data production opportunity for every device in the world. We also see that different organizations are also using their raw data for statistical purposes as it is helping in predicting customer's behavior and market trends. Use of social network sites is also increased in past few years. Social sites are producing a large amount of big data that is not useful every time so to make sense of this useless data is a big challenge. In business perspective it is also important to know whether storing, managing and describing this raw data is useful in terms of business output or not. Different techniques and tools are being used to manage and understanding the big data. Map Reduce and Hadoop are well known in BigData market. We also presented big data production trends since last few years and new challenges to big data. We also lined up some big data solutions in market with some features that can be helpful for big data implementation for a business organization. We also conducted a test using different frameworks and presented the results. We also presented some good practices for big data implementation to achieve high business goals.

Keywords:
Big Data, Data Mining, Map Reduce, Hadoop, Database, GridGain, HazelCast, DAC

1. Introduction

The use of social network is increased as people, friends, family and colleagues are being connected 24/7. Twitter, Facebook, LinkedIn, WhatsApp and Skype are well known participants in connecting people together and producing huge amount of data every day. From 2014 Facebook is producing 600+ terabytes of data every day that is increasing day by day [Pamela et al. 14].

Game is not over here weather forecasting systems, flight operation, telescopic data, scientific and academic research are also producing terabytes, Exabyte to petabytes of data in each day. Near future challenge is to deal with IOT data as every single connected device in IOT system on earth will produce large amount of data every day. Big data includes structure, semi structure and unstructured data. Actual challenge is to manage this meaningless unstructured data, as traditional technologies are not able to make this unstructured data meaningful.

Map Reduce; an effective programming model that plays an important role in big data domain. Map Reduce works in scalable programming paradigm with two core functionalities; Map function and second Reduce function. We will give details of this technology in further sections. Big data can be defined with following properties:

Velocity: defines the speed of data, which is being produced from various sources. This is not only the speed at which data is producing. This speed includes the continuous data flow from source to destination. For example, in a patient monitoring system a sensor is continuously sending the data from body of a patient to the server.

Variety: we know that data is being produced from different sources which may be raw, structured, unstructured or semi structured data. For example, data from web pages, web logs, social sites and email is different than available in traditional databases. Now, IOT trend also increasing the variety of data for example sensor data collected from patient body, sensor data collected from environmental monitoring systems and so on. So, this data is totally different from traditional data.

Volume: the big data term defines itself that data in huge amount i.e. petabytes to zettabytes. Many social sites and IOT devices are producing large amount of data every day. Not only IOT devices and social media are producing large amount of data many other sources are also contributing in big data production.

Complexity: Data is not only from different sources or in huge amount but is also with increased complexity. There is need to deal with a systematic way. Need to link, make relationships, connect, correlate and make hierarchies.

Value: Data is being produced from different sources and can be effective for business analytics so need to extract valuable data, which will help in business analysis. But problem is for IT professionals they have to deal with data in two perspectives:

i. They have to design efficient systems and mechanism that should be able to manage big data from different sources.
ii. IT professionals need to deal with data in such a way that will add value to the business.

Data production is more than resources available to handle it. Computing resources are needed to be available widely if we just handle the data in a traditional way. There should be some solutions that will deal the data at two level; hardware level and software level. Hardware level focuses on the hardware technology development according to the big data requirements and software level should emphasis on software

Manuscript received July 5, 2017
Manuscript revised July 20, 2017

Further sections will present the big data production trends in last few years then different issues and challenges to the big data, associated technologies with big data in market and will conclude the paper with good practices in implementing big data for meeting business requirements.

2. Recent Thirst for Big Data

Big data production varies depending upon sources. Now a day new technological trends increase the big data production for example e-commerce where selling, purchasing and even payment is made in e-currency, Use of IOT technology is increased now, computerized academic researches has been extended worldwide, Space sciences, astrology, astrophysics and many related fields also producing big data.

In warehouses, data needs to be cleansed and secured which is different from big data. Data should also be compatible with basic warehouse structure. In Big data term, data has to be managed either it is fit with the warehouse structure or not. Following technology trends enhances the big data importance.

2.1 IoT data

Internet of things technology makes an extensive use of sensors to accumulate the data from various sources for example in smart health system data is collected from patient body with sensors. These sensors are accumulating and forwarding the data either continuously or after a certain time. This data is more important as it has to be analyzed either in rest state. So, dealing with this large amount of data is challenge for big data [Bashir et al., 16]. IOT data obtained from sensors can be categorized into two types:

i. *Static data* which is stored and need to be stored for long time.
ii. *Motional data* which is continuously produced without long term storage requirement (e.g. patient data received from the body is just need to compare with the already stored data).

So, both types of data required to be analyzed and managed safely, profitably and efficiently.

2.2 Social media data

We can see social media usage everywhere Twitter, Facebook, LinkedIn, Skype and many other social sites are producing big data which can be used to find out the customer trends toward products, customer feedback about products [Mansour, 16]. So, companies can enhance business outcomes using social media data. Social media data can be categorized into two types:

i. *Social data* which is only relate to the personal or social activities of user (e.g. user pictures, friend's information).
ii. *Economic data* which just relates to user comments feedback, products buying and etc.

So, challenge is to extract and utilize only useful data which is required for business analysis from big data of social media.

2.3 Data in traditional IT industry

It is a traditional practice to generate logs for problem identification in an organization but problem occurs rarely and logs are being generated continuously. So, dealing with these logs is a challenge as logs are generated differently with software or hardware updates. Big data solutions not only extract logs related to current problems but also predict future happenings [Heavin et al., 14]. Telecommunication networks also make use of these logs for anomaly detection to detect unusual behavior of networks.

Financial institutes need to analyze risks continuously. Institutes model the data to analyze the risk. Data model continuously calculates risk so that business remains within an acceptable range. A large amount of data that is being produced in an organization needs to be mapped with data model to accurately measure the financial risks. Whenever there is a large amount of data, which is being produced and stored since long time become more important as business solutions can predict risks and trends more accurately. But storing and managing big data is difficult tasks so this is a big challenge for big data. We can also see that various big data projects have been launched by governments and private organizations to enhance the importance of big data.

3. Big data challenges

No doubt, big data is increasing business profits but technically an IT expert or service provider bear all challenges in big data management. In this section we are presenting big data challenges which can be categorized into two types:

i. Basic challenges (deployment challenges)
ii. Advance challenges (continuous monitoring more technical challenges)

Basic deployment challenges are security, privacy, access, sharing and processing of big data. Second type of challenges are more technical challenges like fault tolerance, heterogeneity, scalability and data quality.

3.1 Basic challenges

Various challenges and issues are required to be addressed in big data. Basic challenges that should be dealt at first footsteps may be following:

3.1.1 Security issues

We are extracting useful information from data mines and third parties are involved in doing so. Mostly third party analytical tools are being used to extract, analyze and manage the data. So, this can be security and privacy risk [Bertino, 16]. Personal identification of any person can be misused it can be used to predict something about that person. Predicted information might be very sensitive for that person. Law and enforcement agencies are also making use of big data that can create serious consequences to the irrelevant people without knowing. Personal information of anyone can also be used to add value to the business by seeing in personal lives.

Companies are not able to store and process the big data. Organizations are using cloud solutions to store and process the data which require guaranteed security at cloud.

3.1.2 Data Access issues

Data should be available to access it anytime so that predictive analysis can be done [Carter at al., 16]. Data should be available in accurate format to access it accurately. There may be a need to share the data with law and enforcement agencies and other organizations so data should be available for access.

3.1.3 Data Storage issues

We know that big data is in great volume and need tremendous storage capacity. For example, data produced from social media sites and IOT sensors. Data that is being produce is too much great in volume than available storage capacity. There are different solutions to store the big data for example cloud storage. Cloud storage can offer great storage capacity at remote end but there is need to upload terabytes of data which is time consuming especially it is important when data is producing continuously and when there is a need to store the real time data. We also know that cloud offers distributed storage that is also problematic.

3.1.4 Data Processing issues

Big data is not facing only storage issues it is also facing processing problems. There may be need to process the data at another place than storage place. Data processing and storage may be at same place and only results are processed to the other place. To maintain the Integrity of data during process either at storage place or at other place is a big challenge. Processing large amount of data also consumes time so only relevant data sets need to extract and process to save the time.

3.2 Advance challenges

Advance challenges require continuous monitoring and advance technical skills to deal with it. Following are some challenges big data is facing.

3.2.1 Scalability

Scalability issues related to performance issue over the cloud. Cloud uses distributed technology and share the resources that are not a cheap task. Multiple jobs are required to be done cost effectively. Large clusters in clouds are required to be managed effectively in case of any failure. Storage technology is also being developed from time to time so choosing best technology on a large scale is a big challenge.

3.2.2 Fault management

It is not possible to have 100% fully fault tolerated system but it is possible to maintain a threshold level. Cloud computing tends toward mitigate the effect of any failure in any part of whole system. Big data tasks are divided into sub tasks that are being processed at different nodes if any failure occurs only that part would be affected and need to be restarted. But the issue is to deal with the task that is recursive in nature in which one part needs to be completed first and its output will be input of next part. In this case, failure can lead to whole task failure. This can be prevented by introducing check points at various stages of task in case of any failure restart will be from last check point only.

3.2.3 Heterogeneity

We also know that data is being produced from different data sources some data is structured and some is unstructured. Structured data is well organized and linked one and can easily be managed, processed and stored but big challenge is to deal with unstructured data which is being produced widely [Jirkovsky et al.,17]. It is not possible to link and make structure of an unstructured data. All unstructured data cannot be converted to structured data. So, heterogeneous nature of big data is also a big challenge to deal while playing with big data technology.

3.2.4 Data quality

Big data tries to deal with quality data rather than irrelevant data that is not required [Rao et al., 15]. Data storage and processing is available at cost that's why only relevant and quality data is utilized. Quality data can be

useful for business analytics, future predictions, problem identifications, behavior or pattern recognition and many other purposes. But the issue is how to identify and extract the useful data.

So, to deal with the big data issues different technical and non-technical skills are required. Research and analytical approach is also required to deal with big data. Specialized training programs at professional level and in academics are required to enhance the skills for big data technology.

Big data is also increasing the earning benefits due to its vast application in different fields of life. So technological skills of big data technology for a professional can grant earning benefits and IT companies can invest in providing big data solution to get more profits.

4. Big data Platforms

Below table 1 is showing the different big data platforms and analytic solutions in market.

Table1: Big Data Platforms and analytics software solution

Platforms	Solutions	Features
IBM Big Data Analytics [ibm, 17]	i. InfoSphere Stream ii. InfoSphere Big Insights iii. IBM Watson Explorer iv. IBM Smart Analytics System v. DB2 with BLU Acceleration vi. InfoSphere Information Server vii. IBM Pure Data Powered by Netezza Technology.	Distributed Storage, Processing of large data (structured and unstructured data) Real time data processing, Analytics, Harness data stream including IOT,
HP Big Data [hpe, 17]	i. HPE Vertica Advanced analytics ii. HPE IDOL iii. HPE Heaven OnDemand HP Big Data OEM	Heaven on demand provides cloud based BIG Data Platforms, Secure data lake, Integrated data Governance and MapR support,
ORACLE Big Data Analytics [oracle,17]	i. ORACLE Big Data Appliance ii. ORACLE Exadata database Machine iii. ORACLE Exalytics in Memory Machine iv. Oracle Database v. Oracle NoSQL database vi. Oracle Coherence In database Analytics	Engineered system pre-integrated to reduce the cost and complexity of IT infrastructure, Work with all data types and technologies, Integrates the BigData with existing data, applications and reports data integrity and security
HPCC Systems Big Data [hpccsystems,17]	i. Thor ii. Roxie	Open source platform, Thor performs ingesting, data profiling, and data linking. Roxie provides high concurrency

5. Big Data Frameworks

Following are the big data frameworks that we utilized to process the large datasets.

i. Hadoop
ii. GridGain
iii. Hazelcast
iv. DAC

Given platforms are explained below in further sections.

5.1 Hadoop for Big Data

Hadoop is an open source project by Apache software foundation [Hadoop, 17]. Hadoop manages the distributed processing of big data sets over computer clusters. Hadoop provides scalability in local storage and computation. Hadoop consist of following basic components:

i. Hadoop distributed file system (HDFS)
ii. Map Reduce

Hadoop also supports subprojects that provide additional capabilities to the Hadoop project. For example, Zookeeper provides high performance concurrency for distributed applications. Tez is a flexible engine that

provides exceptional capabilities to process data for batch and interactive cases. Spark is a computational engine to support wide range of applications like machine learning, graph and stream processing. Pig supports for parallel computing and Hive provide warehouse infrastructure for ad hoc querying and data summarizations. Cassandra supports multiple master databases to avoid from single point of failure and Ambri is a web based tool for Hadoop to manage the Hadoop cluster with supports to work on all Hadoop sub projects.

5.1.1 Hadoop Distributed File system (HDFS)

HDFS is a distributed file system to support storage of large size of files and runs on clusters of commodity hardware. HDFS supports 64MB of default block size for file system thus it reduces the number of disk seeks. HDFS consists of name nodes and data nodes. Name node manages the tree of directories, metadata for files and file system namespaces. Client instructs to the data node to store the file then retrieve a block of data and data node also reports to the name node with the block information of stored data. Name node is always required to access the files and thus it becomes the single point of failure.

5.1.2 MapReduce

MapReduce functionality can be inferred from its name it performs two basic functions:

i. Map function

ii. Reduce function

Map tasks are performed on the input value of file system. Map task produces sequence of key values. These values are collected by master controller and sorted by keys. These keys are distributed among reduce tasks. Sorting mechanism ensures the same key values with reduce tasks. Map and Reduce tasks are created by the master controller and assigned to the worker nodes. Master node keeps track of the status of map and reduce task. When worker nodes complete a task they report to the master node then Master node reassigns another task to the worker nodes. Master node continuously pings the worker node after a certain time whenever any worker node fails master node knows about it. Whenever a worker node fails master node restarts the task that a fail node was performing even task was completed successfully.

We can conclude that Apache Hadoop system provides distributed storage and processing of the large data sets based on simple map reduce program by using commodity hardware which reduces infrastructure cost. It is highly scalable easy to handle machine failure and cheaper in cost. It supports java language and contains open source apache license but disadvantage of the Hadoop system is its bounded programming model. But there are many other big data analysis tools which are based on Hadoop distributed file system for example GridGain which is based on HDFS and seems an alternative to Hadoop and provide real time in memory data processing for fast analysis so, it gives fast performance.

5.2 GridGain for Big Data

GridGain is a JVM-based middleware programming that is used for big data systems and uses the in memory computing to increase the throughput for data and reduces the latency. Applications created with GridGain can scale up on any base - from a solitary Android gadget to an expansive cloud.

There are two areas of functionality provided by GridGain:

i. Compute Grids

ii. Data Grids in-memory

On top of that it provides the compatibility of surrounding technologies many of which are frequently used by big data clients.

5.2.1 GridGain Features

Following features can be provided by GridGain [Ivanov, 10]:

i. Application can work in a zero-deployment mode.
ii. Application can be scaled up and down according to the demand.
iii. Application can be cached in data grid.
iv. Against cached data we can run sql queries.
v. Tasks can be speed up using MapReduce.
vi. Work can be divided on the grid.
vii. All resources on the grid can be discovered automatically.
viii. It provides advance data enquiring and provision for SQL, TEXT, and FULL SCAN queries.
ix. Complete provision for Document-style facts arrangements such as JSON.
x. Asynchronous & synchronous data preloading.

These GridGain features are prominent but there are also many other features related to GridGain platform.

5.3 HazelCast for Big Data

Open source in memory data grid framework for big data. HazelCast is based on Java. HazelCast can work in cloud, on virtual machines and in Dockers containers. HazelCast can support many programming languages like Java, scala, .Net framework, C++ and python.

5.3.1 HazelCast Features

Following features can be implemented by HazelCast [hazelcast, 17]:

i. Provides support for membership events & cluster information.
ii. Support for dynamic clustering.

iii. Support for dynamic HTTP session clustering.
iv. Scaling to hundreds of servers.
v. Provides support for dynamic fail-over.
vi. Very fast; supports thousands of operations/sec.
vii. HazelCast is very efficient and super nice for CPU & RAM utilization.

Hazelcast features are not limited to the above.

5.4 DAC for Big Data

Dynamic agent computation is an effective java based cloud/ matrix processing environment which is intended for executing multi frameworks [dacframe, 17].

5.4.1 DAC Features

Following are the features provided by the dynamic agent computation framework:

i. Its use is extremely easy.
ii. DAC also Provides high performance.
iii. Also provides multiple support for Transports i.e. Cajo, JMS & Oracle Advanced Queuing.
iv. Resource management, dynamic code execution and provides extensible architecture.
v. It also provides session support with user privileges (for security).
vi. DAC also provides task prioritization and accountability features in its framework.

6. Testing Big Data Frameworks

We are going to test big data frameworks which are discussed in previous section on the basis of following parameters:

i. Task distribution time
ii. CPU usage
iii. Network usage

6.1 Testing Environment

This test is carried out using 5-intel machines named as intel1 -intel5. Each machine having twofold Quad-Core Xeon E5410 2.33GHz processor and 4GB on-board RAM. Test environment is also shown in figure 1.

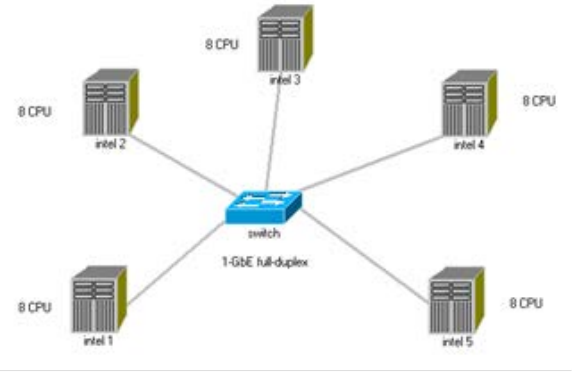

Figure 1: Test environment for big data frameworks

6.2 Test Results for Task Distribution Time

We processed 3 different assignments on each platform by dividing into multiple tasks. First we processed the 341 task then 2705 and then 37500 tasks on each platform. Following table 2 is presenting the results in milliseconds. Results are clearly showing that different platforms take different time to process the task.

Table 2. Results overview in milliseconds

Library name	Number of Task:341	Number of Task:2705	Number of Task:37500
DAC 0.7.11	305 834.701	299 507.701	304 971.931
GridGain 2.1.11	372 279.701	338 310.401	350 744.001
HazelCast 1.81	348 716.701	321 922.701	335 363.301
DAC 0.9.11	306 076.201	299 815.701	305 303.301
Hadoop 0.20.11	467 042.701	384 331.601	365.0.401

6.2 Test Results for CPU usage

CPU usage on different platforms is shown in figure 2 below.

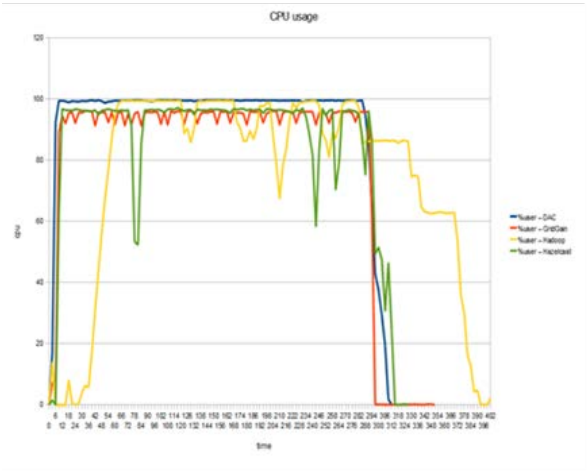

Figure 2: CPU usage results

In above figure, blue lines representing the DAC for intel machine usage. Red line representing the GridGain platform use over the intel machine. Similarly, yellow and green line are representing Hadoop and HazelCast usage over the intel machine and graph showing the cpu usage of intel machine.

6.3 Test Results for Network usage

Bytes received in per second on intel machine and transferred from intel machine are also calculated and shown in figure 3 and 4 below.

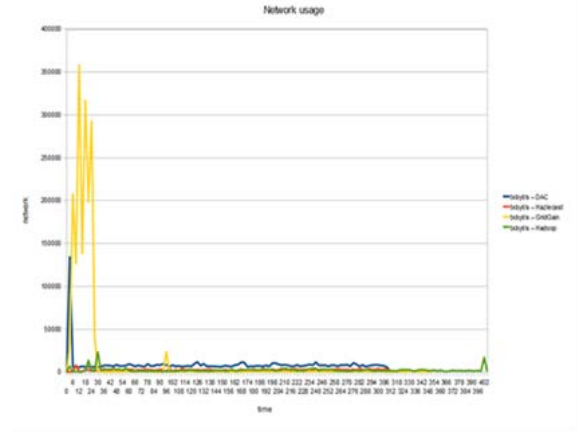

Figure 3: Transferred bytes/sec on intel machine

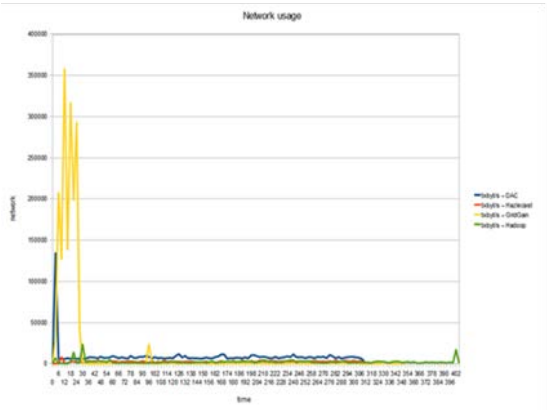

Figure 4: Received bytes/sec on intel machine

It can be concluded that GridGain and HazelCast are the good choice for CPU-concentrated tasks. HazelCast utilizes low bandwidth of network and CPU. GridGain utilizes small amount of memory. For processing of big data Hadoop is the best choice and DAC is not suitable in case of any node failure. DAC is not effectively pre-configured to handle node failures.

7. Conclusion and Future work

It can also be concluded that technologies that are becoming the source for big data production are growing day by day and new big data challenges are also increasing day by day. Right now, IOT and social sites are producing and will continue increasing big data. Different IOT projects have been started throughout the world thus upcoming challenges to IOT data requires in advance measurements. Social sites are becoming more popular than that of few years back. Social media data production is too large thus extracting useful data from social sites securely is also a big challenge in big data challenges. So, new enhanced technologies are required to manage them.
Our future research work will include real time IOT data management in future perspective using currently available big data solutions. This work will open the new opportunities in instantly handling real time IOT data using existing solutions.

References

[1] [Pamela, 14] Pamela. V. & Kevin. W. (2014, April 11). Scaling the Facebook data warehouse to 300 PB [Web log post]. Retrieved from https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/.

[2] [Bashir, 16] Bashir, M. R., & Gill, A. Q. (2016). Towards an IoT Big Data Analytics Framework: Smart Buildings Systems. 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd

International Conference on Data Science and Systems (HPCC/SmartCity/DSS). doi:10.1109/hpcc-smartcity-dss.2016.0188
[3] [Mansour, 16] Mansour, R. F. (2016). Understanding how big data leads to social networking vulnerability. Computers in Human Behavior, 57, 348-351. doi: 10.1016/j.chb.2015.12.055
[4] [Heavin, 14] Heavin, C., Daly, M., & Adam, F. (2014). Small Data to Big Data - The Information Systems (IS) Continuum. Proceedings of the International Conference on Knowledge Management and Information Sharing. doi:10.5220/0005133802890297
[5] [Bertino, 16] Bertino, E. (2016). Big data security and privacy. 2016 IEEE International Conference on Big Data (Big Data). doi:10.1109/bigdata.2016.7840581
[6] [Carter, 16] Carter, E. L., & Lee, L. T. (2016). Information Access and Control in an Age of Big Data. Journalism & Mass Communication Quarterly, 93(2), 269-272. doi:10.1177/1077699016646790
[7] [Jirkovsky, 17] Jirkovsky, V., Obitko, M., & Marik, V. (2017). Understanding Data Heterogeneity in the Context of Cyber-Physical Systems Integration. IEEE Transactions on Industrial Informatics, 13(2), 660-667. doi:10.1109/tii.2016.2596101
[8] [Rao, 15] Rao, D., Gudivada, V. N., & Raghavan, V. V. (2015). Data quality issues in big data. 2015 IEEE International Conference on Big Data (Big Data). doi:10.1109/bigdata.2015.7364065
[9] [ibm, 17] Big data at the speed of business. (2017, March 17). https://www-01.ibm.com/software/data/bigdata/
[10] [hpe, 17] Big Data Software. https://saas.hpe.com/en-us/software/big-data-analytics-software
[11] [oracle, 17] Oracle Big Data Products. https://www.oracle.com/big-data/products.html
[12] [hpccsystems, 17] HPCC Systems. https://hpccsystems.com/
[13] [Hadoop, 17] Welcome to Apache™ Hadoop®! http://hadoop.apache.org/
[14] [Ivanov, 10] Ivanov N, Real Time Big Data Processing with GridGain, Gridgain Project, 2010.
[15] [hazelcast, 17] Documentation - Hazelcast. https://hazelcast.org/documentation/
[16] [dacframe, 17] http://www.dacframe.org

Author List:

Muzamil Mehboob received his MS Degree in Information Technology in 2014 from Baha-ud-Din Zakariya University, Multan, Pakistan. He obtained his BS in Information Technology in 2012 from Baha-ud-Din Zakariya University, Multan, Pakistan. He is currently working as Lecturer in Dept. of Computer Science & I.T at Pakistan University of Gujrat Sialkot Campus. His research encompasses Big Data and IoT Systems.

Maruf Pasha received his PhD from University of South Brittany, France ad has obtained his MS (IT) from NUST, Pakistan and is currently serving as head in department of Information Technology at Bahauddin Zakariya University, Multan. His research encompasses semantic web, big data and IoT systems. He has number of publications in International Journals and conferences.

Syed Muhammad Waqas Shah received his BS (Telecommunication System) from Bahauddin Zakariya University, Multan, Pakistan. Mr. Waqas research interests are telecommunication and distributed systems. This research is carried out during his MS (Information Technology) degree in Department of Information Technology, Bahauddin Zakariya University, Multan, Pakistan.

Urooj Pasha received her PhD from Molde University College in 2014 and holds MSc degree in Computer Science from Bahauddin Zakariya University Multan in 2000, and a MSc degree in Logistics from Molde University College in 2008.Currently she is serving as Assistant Professor in Institute of Management Sciences, Bahauddin Zakariya University Multan. Her research interests are logistics, vehicle routing, big data and computational complexity.