sciendo RIGA TECHNICAL UNIVERSITY
Applied Computer Systems
ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
December 2018, vol. 23, no. 2, pp. 90–100
doi: 10.2478/acss-2018-0012
https://content.sciendo.com

Spatiotemporal Aspects of Big Data

Saadia Karim[1*], Tariq Rahim Soomro[2], S. M. Aqil Burney[3]
[1–3] *CCSIS, Institute of Business Management, Karachi, Sindh, Pakistan*

***Abstract* – Data has evolved into a large-scale data as big data in the recent era. The analysis of big data involves determined attempts on previous data. As new era of data has spatiotemporal facts that involve the time and space factors, which make them distinct from traditional data. The big data with spatiotemporal aspects helps achieve more efficient results and, therefore, many different types of frameworks have been introduced in cooperate world. In the present research, a qualitative approach is used to present the framework classification in two categories: architecture and features. Frameworks have been compared on the basis of architectural characteristics and feature attributes as well. These two categories project a significant effect on the execution of spatiotemporal data in big data. Frameworks are able to solve the real-time problems in less time of cycle. This study presents spatiotemporal aspects in big data with reference to several dissimilar environments and frameworks.**

***Keywords* – Apache Hadoop, big data analytics, spatiotemporal, Samza, Storm, Spark, Flink.**

I. Introduction

Big data is used for collecting datasets, which are high in volume, complex in variety, high in analysing the speed of data (velocity), and ambiguous with respect to expansion at a steady and frequently rapid rate (veracity). To store both unstructured and structured data, it is necessary to have new techniques with new tools, which are capable of analysing, making decision, working efficiently and optimising the processes. The complexity and volume of big data are always a big challenge for analysing the data [1], [2] as shown in Table I below.

The data is a valuable asset of an organisation and big data is a question to the following needs of an organisation [1]:

- Availability of massive data in the form of clear and accurate information.
- Development in terms of big data for efficiency and performance is a key feature for an organisation.
- Discovery of the knowledge base, which is extracted from big data, will be the primary information for an organisation to form new business standards with new business policies that adjust new service offerings to employees and customer. This is business analytics, which is performed by third parties.
- Division of massive data to related department, customers for offering business and marketing flow, which are changing day by day.
- Executed information is used as the power of taking decision for an organisation.

In many studies, the big data potential value has been discussed in different applications, such as big data in the US health care, European public sector administration, global personal location data, U.S. retail, and global manufacturing; here value is over $1 trillion U.S. dollars per year [3].

TABLE I
BIG DATA 4VS BY PWC [1]

	Traditional techniques and issues	Big data differentiators
Veracity	Does not account for biases, noise and abnormality in data	Data is stored, and **mined meaningful to the problem being analysed** **Keep data clean and keep "dirty data" from accumulating in your systems**
Velocity	No real time analysis	In real-time: **Dynamically analyse data** Consistently integrate new information Automatically delete unwanted to ensure **optimal storage**
Variety	**Compatibility Issues** Advance analytics **struggle with non-numerical data**	Framework **accommodates varying data types and data models** **Insightful analysis with very few parameters**
Volume	Analysis is limited to **small data sets** Analysing large data sets = **High Costs & High Memory**	Scalable for **huge amounts of multi-sourced data** Facilitation of **massively parallel processing** **Low-cost data storage**

* Corresponding author's e-mail: saadia.karim@iobm.edu.pk

The value of big data in the areas of customer intelligence, supply chain intelligence, performance improvements, fraud detection, and quality and risk management is $41 billion per year in the UK alone [4], [5]. The reusability of data for analysing and making strong decision in a critical situation causes to think about the security and privacy measures of data, and it is an important issue to be taken into consideration. The worth of big data depends on their use and reuse. Normally, cost is unconfined when dissimilar datasets are joined together to answer big questions. The answer of big data received after analysing is more valuable than the amount of data [6]. The big data has a high risk factor, because the analysis of big data helps predict the future output. If the analysis is not done properly, it causes the prediction or decision to be wrong. That is why the quality of data is a significant point to be taken into consideration for any organisation. The big data nowadays is highly useful in many organisations, such as the banking sector, healthcare, industrial, learning and transportation, which helps obtain the analytics to get accurate results [7].

The study of big data extends the thoughtful knowledge to the area, where big datasets are divided into major aspects involving storage, execution, method, and speed as spatial and temporal data execution. Spatial data characterises location, shape and size of an object, such as a mountain, lake, building, or township. It represents the information by including attributes that will deliver more information about the object. The GIS is mostly widely used to access, manipulate, visualise, and analyse the geospatial dataset. In its turn, temporal data represents an attribute in the form of time instances and period [8]. It describes different data types and stores relevant information to the past, present, and future, e.g., the land-use patterns of Karachi in 1990, or a heavy rainfall in Karachi, Pakistan on Thursday, 31 August 2017. The time datasets are collected to analyse the dependent and independent variables in data [9], [10]. The data has been gathered from many sources like manual data entries, data from simulation models or data using observational sensors and stored in a temporal database not in a conventional one, as a conventional database captures individual current portrait of reality whereas a temporal database stores data in time and allows time-based reasoning [11]. A temporal data management system (TDBMS) is a system that delivers built-in provision for the time dimension, as well as special services for storage, querying, and updating data with respect to time. A temporal DBMS can distinguish between historical data, current data, and data that will be effective in future. A temporal DBMS provides a temporal version of SQL, including enhancements to the data definition language (DDL), constraint specifications and their enforcements, data types, data manipulation language (DML), and query language for temporal tables [12], [13]. The spatial and temporal datasets have many types of frameworks to achieve big data and design on the basis of analysis, storage, and access availability at any time [13]. The big data processing frameworks are responsible for calculating complete data in a database system. The framework is divided into three focal points of areas, as follows [14].

A. Batch-only Framework

The running of large programs without communicating to an end user is called batch jobs and due to frequent execution of these programs they are called batch processing. Based on batch concept, Apache Hadoop framework was developed to run large size of programs with nominal human interaction [14]–[16].

1. Apache Hadoop

Hadoop was the first open-source framework for big data processing with execution of batch processing technique. Hadoop re-applied the Google algorithm and modules for huge scale of batch processing to be available. The Hadoop modern version works on multiple layers, such as HDFS layer – stands for Hadoop Distributed File System that can storage and replicate data across the cluster nodes, YARN layer – stands for Yet another Resource Negotiator, which is a cluster coordinating component of the Hadoop Stack, and MapReduce layer that is a native batch processing engine. Hadoop methodology influences analysis, permanent storage, and generating of multiple tasks per time cycle that lead to slow processing. However, Hadoop – MapReduce can run huge datasets on less expensive components using Hadoop cluster functionality. Apache Hadoop – MapReduce processing method is a well-defined batch processing model, which can run where no significant time factor is involved. Compatibility and grouping with other frameworks mean that Hadoop can frequently assist as the foundation for numerous processing workloads [13], [14], [17], [18].

B. Stream-only Framework

Streaming is a process of infinite tuples of the procedure generated continuously in a time (a1, a2, a3 … a*n*, *t*). Stream framework provides processing before storing data as compared to Hadoop that stores data then processes it [19]. To make processing efficient, Apache Storm and Samza frameworks were introduced.

1. Apache Storm

Apache Storm is used for pure stream processing loads with very strict potential requirements. Storm has option for assignments that need nearly to be processed in real time. It can execute a huge amount of data and deliver outputs with minimal delay as used by other solutions. Storm is possibly the best resolution obtained for near real-time execution of data. Storm does not perform batch processing. However, Storm with additional software, like Trident, provides micro-batch processing capabilities, which give users a flexible environment to solve the problems. This makes Storm advantage over other processing methods. In terms of compatibility of Storm, it can integrate with Hadoop's YARN, to make increase in remaining Hadoop deployment. Storm has very extensive language support, enabling users to make many decisions for significant topologies [13], [14], [20], [21].

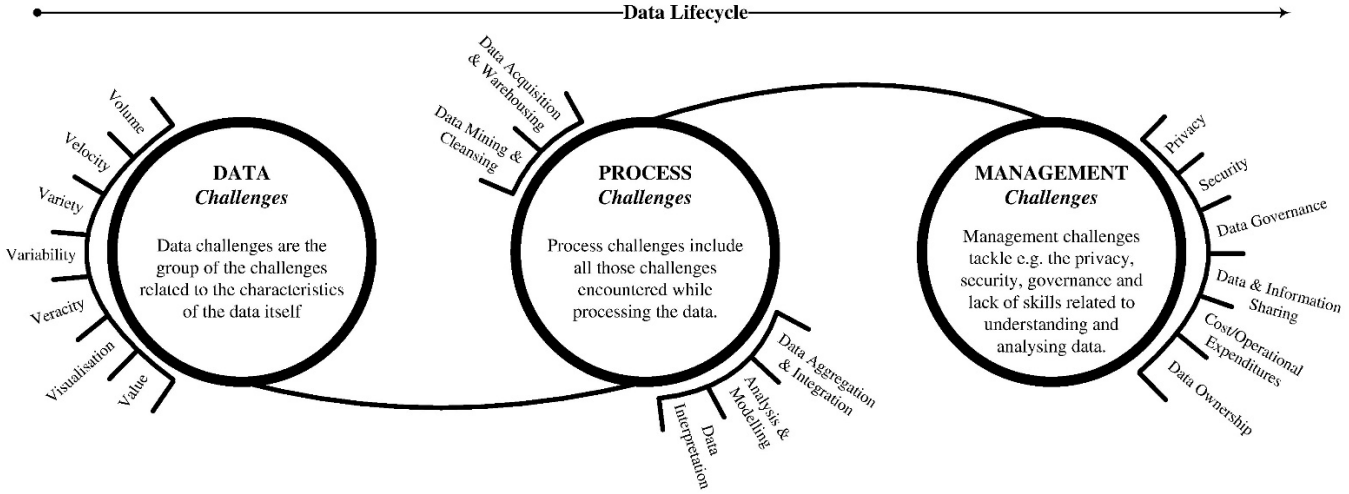

Fig. 1. Big data intellectual challenges [27].

2. Apache Samza

Apache Samza is another example of a streaming framework. Samza has an optimal solution for streaming workloads that strongly merge with Hadoop and Apache Kafka communication system. Samza is a design to take advantage of Kafka's architecture. It uses Kafka to provide buffering, state storage and fault tolerance. Kafka mirrors the way of MapReduce using HDFS, when batch processing calculation is performing, many problems arise, while stream processing provides solutions to many problems. Java Virtual Machine (JVM) language is the only language supported by Samza. Samza cannot provide low-cost performance and compatibility with current systems [13], [14], [22].

C. Hybrid Framework

Hybrid framework is a group of two or more frameworks that can be modified and retrieved by a user [23]. Based on this concept, both batch-only and stream-only frameworks are combined together to work more efficiently for users. Apache introduced Apache Spark and Flink frameworks based on a hybrid framework.

1. Apache Spark

Apache Spark is a next generation framework, which includes benefits of batch processing with stream processing abilities as a hybrid framework for big data analysis and optimization of workloads. It uses many of the Hadoop – MapReduce principles. Spark focuses on the speed issue of batch processing by implementing the complete in-memory computation with advanced directed acyclic graph (DAG) scheduling and fast optimization in processes. Spark deals with micro-batches concept, which is a design to execute streams of data as a sequence of very small procedures (batches) that can be controlled using the native semantic of batch engine. For this buffering technique, it is used as sub-second increments. Spark's main benefit is its adaptability as a standalone cluster or combined with an existing Hadoop cluster. Spark also has an environment of libraries that can be used for machine learning. Spark runs on RAM, which increases the speed of processing much faster [13], [14], [24], [25].

2. Apache Flink

Apache Flink is also a hybrid framework using stream processing and batch tasks. It considers data stream batches with finite boundaries as a process of executing batch processing as a subdivision of stream processing. Flink's stream processing approach follows the Kappa architecture or Lambda architecture processing method. Kappa architecture simplifies the model to handle all incoming data. Flink offers data streaming API to work on absolute streams of data. For storing state, Flink can work with a number of state back ends depending on varying levels of complexity and persistence. Flink's batch processing does not work on real-time streaming data; it supports bounded datasets. Therefore, Flink stores snapshots of data, which are removable from batch, but still recoverable. Flink is possibly best suitable for organisations, which have heavy stream execution requirements and some batch-oriented tasks. Storm and Hadoop programs are able to run on YARN-managed cluster [13], [14], [26].

The frameworks are implemented to solve the execution of big data with spatiotemporal datasets. In the real world, we now have data as bigger datasets, which are in the form of structured/unstructured, and storing of big datasets is possible (using cloud services, extensive memory space, etc.), but managing, creating data links/relationships, execution and finding more accurate results to take decision or prediction are the challenges of big data as shown in Fig. 1 [27], [28].

The following problems are solved by using spatiotemporal datasets:

- Live traffic monitoring [29].
- Real-time moving objects [30].
- Google Earth and Google Maps [31], [32].

The ArcGIS in version 10.6 release explores the spatiotemporal big data storage, which has many features plus real-time data and tracking Id concept [33].

Real-time execution of large-scale data was introduced in Apache frameworks to make a more productive environment for analysing the un-structured, semi-structured and structured big data, which have been produced by different mechanisms using GPS, GIS and others from social media. The frameworks have different approaches to solve spatiotemporal issues in big data, which help in the study to develop a thought that frameworks should be divided into architectural characteristics and feature attributes for finding the accuracy and performance adorable. The focus of the study is to represent work done in spatiotemporal aspects of big data with different frameworks and show attributes that can help in managing, storing, execution and processing of spatiotemporal data. The paper is organised as follows: Section II provides a review of literature; Section III describes material and methods used; the research results are provided in Section IV, and Section V discusses future work.

II. Literature Review

Authors in [34] focus on spatiotemporal system with both practical and theoretical areas. Some topics are discussed in respect to spatiotemporal data, such as visualization for natural, data mining, network cell modelling, parallel algorithms, veracity of data, complex simulation model, privacy and legal issues, etc. In [35], authors describe the challenges of datasets called spatial big data (SBD) problem. A new technique is used as a temporal detailed (TD) road map, which uses road sensors to produce complete information about roads and travel time with respect to speed that helps solve SBD challenges. According to authors in [36], finding potential patterns in datasets involves spatial and temporal data mining. Different types of algorithm approaches can be used here, such as regression model, Markov random model, and Gaussian process learning. It opens the way to develop different new approaches and methods for the compression and sampling of datasets.

Vatsavai and Bhaduri in [37] illustrate that the drastically inverse spatial and temporal big data has been observed by devices and social media sites. The analysis of huge data is very useful, but the execution of processes in a spatial big dataset makes them slow, costly and complex. To overcome these issues, strong algorithms are required, such as spatial auto regression (SAR) model, Markov random field classifiers (MRFC), and Gaussian process (GP). Some applications as discussed in [36] are high evolutional part of temporal datasets to execute the accurate geographic location of substance. They need efficient algorithms and support of spatiotemporal infrastructure. To support these challenges, a strong computational and I/O model is required to achieve full benefits of big spatiotemporal data. Authors in [38] focus on the problems of spatial datasets that involve space, storage, execution, analysis etc. and compare them to traditional data models. SBD has challenging tasks to compute, technology and methodology, which reflect as platform, analytics, and science. For these issues, some solutions are HDFS, OLAP, and SAR based ensemble model. For SBD platform, apache Hadoop, (MapReduce framework) is best suitable to execute data and deduct the node failure automatically.

According to [39], the big data grows from multiple sources and new technologies that make it cost effective, as well as big data has multi-dimensional nature, sometimes isolated source and sometimes multiple sources. The Hadoop GIS is used with top level of MapReduce technique for integration of spatial queries and lower level with different types of queries. Ming-H. Tsou in [40] provides a review on the big data application, techniques, and tools in reference to Geo informatics system in geography and GI Science and focuses on fast computing, cloud and grid computing, data mining, machine learning, 2D- 3D spatial databases, GPS, GIS and navigation. In [41], authors clarify that there are many spatial datasets introduced on the market, such as global positioning system (GPS), vehicle EM, social, etc. Spatial big data (SBD) refers to the value prepositional dataset that helps find a computational platform. SBD executes the point value, linear dataset, and merges with SBD raster dataset using GPS, VGI, and UAV, which involve the real-time data. The SBD involves infrastructures, such as MapReduce, Hadoop, Spark, etc. According to authors in [21], big data gives benefits to the business and improves the performance. Big data involves execution, storage and large datasets with space and time to achieve the accuracy and efficiency. Apache Storm is one of the frameworks used; it works based on real time processing and is best for fault tolerance issues. As compared to Apache Hadoop, Apache Hbase, Apache Spark, YahooS4 (Apache Incubator), Apache Storm is present to assist in real-time distribution of stream data. Apache Storm has five key features, which help execute twitter tweets. Authors in [25] describe how big data is analysed and helps to retrieve valuable information. An example of Twitter tweets is used with Apache Spark. Apache Spark is a cluster-computing tool and performs micro-batch real-time called Spark streaming. Spark processing proves that it works on real time streaming of datasets at low latency, flexibility, but Spark is not an easy tool to work and requires powerful data representation.

Authors in [42] present an idea of cartographer on big data and social media data, which has been created by different devices that can help transform global space to real-time space. The collective struggle on developing cartographers can make an efficient visualization approach improve the reliability of crowd-sourced big data and approximation of probability of generating social media messages. In [43], authors describe the large crowd during Hajj and discuss how to manage this crowd. This is a challenge, and the aim of a method is to collect multivariate records from a very large crowd equipped with multi-sensory smart mobile phones, automobiles, and community networks. According to authors in [44], the rapid growth of data from different mobile devices and applications contains GPS and GIS functionality in datasets. The datasets are divided into two types: ranging and migration. Some general additional solutions are used to handle the index layer, e.g., HBASE, CASSANDRA, and Mango DB. The three types of geospatial big data models are discussed, e.g., MPP-based relational databases, NoSQL (MapReduce) based data store,

and the hybrid solution SQL-on-Hadoop. The aim of the combined solution is to use the layer of MMP, SQL-on-Hadoop and Hadoop on a cloud data platform. In [45], authors explain the drastic increase of spatiotemporal big data along with the development of urban atmospheres. To solve the problem, a spatiotemporal data model is presented, which has the basic concept of a compressed linear reference (CLR) approach to transform three-dimensional (3D) system into a two-dimensional (2D) one that can be stored and managed in traditional spatial databases. According to [46], big data causes numerous aspects for GIS. The big data volume and velocity affect the spatiotemporal concepts of storing and analysing. Many challenges drawn by big data and solution of big data variety by land use/land cover change (LUCC) are based on geospatial cyber infrastructure (GCI), which optimises the big variety of data and identifies the changes in pixels, which develop an image to identify the relative number of changes The method introduces spatial temporal LUCC with RS datasets, which increase the variety of datasets used in big data. Temporal topology is a more flexible and efficient method used for solving problem. According to authors in [47], the increasing demand in spatial big data and high resolution applications of geographic information system (GIS) are causing issues for a traditional GIS system to manage and perform efficient computation. The present state of parallel GIS has evolved into two parallel GIS architectures on high performance computing (HPC) cluster and Hadoop cluster; furthermore, some GIS algorithms are implemented. Authors in [48] describe the data analytics as a self-importance feature of big data that contains the spatial processing impact as a historical data prototype. The analysis helps solve the critical tactics of data analytics. In [49], authors present big data challenges and a basic method for gathering big data dynamically. When the big data is fed into GIS, data has a source and transformation service with a dedicated path to information and getting result. Big data involvement in GIS will increase a hypothesis for collection of geographic dataset, which improves scientific solutions. According to authors in [50], the climate change and simulation modelling are generating large datasets with involvement of spatiotemporal attributes. The datasets provide the global challenges as climate change, natural disasters, and diseases. The indexing approach is introduced to handle critical data analysis efficiently. The spatiotemporal grid parallel approach, grid partitioning strategy, and more infestation are required with Spark. In [51], authors state that house pricing is an issue, because pricing increase or decrease causes many issues to evaluate the process by space-time dynamics with big data. The nearest neighbour analysis is applied to calculate the spatial distribution. The kernel density estimation is used on ArcGIS 10 for calculation of datasets. Then system is able to identify high and low frequency of prices. According to authors in [52], spatial-temporal data models are characterised by some differential equations, which identify the spatial-temporal features and comparison between the storage efficiency and management of data. The different types of spatial-temporal data models are used to compare lower storage capacity and space complexity, but can also solve limitations of spatial-temporal difference equations by compressing spatial-temporal data. According to authors in [53], the growth in new sensing devices help discover the spatial, time and range determination problems. For presenting the big data features suitably and automatically as compared to a classical approach, the texture descriptors from diverse margins of wavelet transformation are proposed. The descriptors contain statistical, directional (gradient histogram), periodical (autocorrelation), and a low-frequency statistical (Gaussian mixture model) models. The rationality of texture descriptors with confirming outcome from three different factors of wavelet transformation give better outcomes as compared to classical texture categories. Therefore, the texture descriptors are suitable for remote sensing of big data general prospects with simple design and spontaneous meaning. Kuien et al. in [54] describe the gap between the cloud storage and data warehouse storage, as the cloud based architecture uses the outside storage having unlimited planetary at high obtainability and low cost of preservation/repairs. The infrastructure of data warehouse is different from the cloud infrastructure as massive parallel processing (MPP). To solve the problem of MPP for a cloud, a middleware design has been built as an open-source model named GPCloud. According to [55], big data is a challenge, which is a difficult problem to be addressed in real world; a solution is anticipated based on the association analysis by storylines. The first method is storytelling and matching the dots by distance-based Bayesian inference, which includes spatial data for finding similar events; the second method is inference and forecasting of spatial association index and last method is a link analysis using spatio-logical inference for calculating the storylines in diverse positions, limited data volumes to forced districts, the concluding powerful event with filtering unrelated events and the means of events to aggregate big data. These methods help analyse, search, broadcast and predict the events.

In [56], authors describe the change in extreme weather and quality of infrastructure, which cause the high valuable devices to calculate the weather impacts using spatiotemporal technique in big data. The government introduced the solar system to predict the weather impact on the nature. To develop a robust system with accurate analysis of weather conditions, they need the spatiotemporal data as big data to make decision and avoid the influence of weather on human and nature. For implementing any approach to overcome the effect of weather on the electric power system, a wide-ranging examination should be executed with high volume of data from risk based insulation coordination, lightning hazard, and prediction of vulnerability with Gaussian conditional random fields (GCRF) regression. The solar generation prediction has been also proposed in the model that leads to the temporal and spatial correlations among different locations and improves the accuracy of prediction approach. Authors in [57] describe the analysis of massive datasets by using existing tools, such as Apache Hadoop, MapReduce, Spark, GeoSpark, and Spatial Spark. A new tool such as STARK is introduced among all these. In [58], authors discuss fast growth of spatial and temporal data by applications, such as mobile computing,

wireless messaging, and global navigation satellite systems (GNSS). The proposed model is a keystone on the way to efficient real-time working and tracking of objects through spatial and temporal stream grouping of data. According to [59], authors bring an idea of ST-Hadoop for the first time in the market of spatial and temporal data. The ST-Hadoop works on pre-defined processes of Hadoop, such as languages, data types and improved indexing approach of loads, and it divides data over Hadoop Distributed File System that makes computation efficient through the nodes and works on spatiotemporal queries of range and joins. The extensive approach allows enhancing the large-scale dataset over billions of spatiotemporal minutes.

According to authors in [60], the drastic advancement in data technologies is the cause of high volume of data that needs to be stored. The data is generated by IoT, GPS traffic data, and analysis of meteorological observations of physical world, which improve the area of wildfire management risk analysis. By using real-time meteorological spatial and temporal data, a system is proposed that integrates a disaster management system to address the challenges of wildfire. Lakshmi et al. in [32] present a concept of big data to analyse from different sources like Twitter, Facebook, satellites, etc. The method deals with heterogeneous data from satellite imagery patterns, remote sensing data, aircraft data and other devices. The classification method uses Google Map reduce C4.5 algorithm with Hadoop MapReduce framework, and the output obtained is more accessible, cost effective and efficient as compared to previous outputs. In future, the social media data is also going to be integrated with this classification method. According to [61], the authors state that geospatial information is increasing along with the increase of big data. Spark framework is developed for distributed computing framework in Hadoop ecosystems. To handle spatial queries at large-scale datasets, the GeoSpark SQL framework is proposed, which gives achievement on execution of efficient storage and high level-parallel computing. The experiment is performed for storage, spatial operator approach, and spatial query optimization that gives result as GeoSpark SQL, which is able to achieve the goal in real-time and perform more accurately spatial database processing. Furthermore, the GeoSpark SQL deals with kNN and spatial join queries.

III. Material & Methods

Spatiotemporal data involves the time and space factors, which make them distinct from traditional data. Nowadays the data is growing very drastically as big data is of high volume, velocity, veracity, and variety, which make it complex to manage. The big data with spatiotemporal aspects helps achieve more efficient results and, therefore, many different types of frameworks have been introduced in cooperate world. In this study, a qualitative approach is used to present the framework classification in two categories: architectures and features. Frameworks have been compared on the basis of architectural characteristics and feature attributes as well. These two categories project a significant effect on the execution of spatiotemporal data in big data. This study is based on qualitative research methods [62].

The frameworks based on architecture have the batch-only, stream-only, hybrid processing frameworks. The batch-only involves the Apache Hadoop methodology influence analysis, permanent storage, and generation of multiple tasks per time cycle that lead to slow processing, but Hadoop – MapReduce can run huge datasets on less expensive components using Hadoop cluster/ grid computing functionality. Hadoop uses the HDFS, YARN, and MapReduce layers to store the spatiotemporal data [13], [14], [18], [63]. The stream-only framework involves Apache Samza and Apache Storm processing systems [64]. Apache Samza has an optimal solution for streaming data and uses the KAFKA communication system and Samza has the storage layer of HDFS – MapReduce for batch processing with partitioned scalability. It executes data in the form of message using only JVM programming. Apache Storm works on real-time streaming framework for processing huge amount of data and delivering output with minimal delay. Storm uses the Trident to execute the micro-batch processing with processing of each tuple of data at once and YARN storage layer is used by storm with hash map function to execute in-memory problems. Storm works with any programming language [20], [22], [64].

The hybrid involves the combination of batch-only and stream-only framework and presents faster computing framework Apache Spark and Apache Flink [65]. Apache Spark is known as “Lightning-Fast Cluster Computing” framework. It uses Apache Hadoop Hbase, Cloudera, and Impala processing engine libraries for near real-time stream processing and micro-batching processing with layer of MapReduce on top and for complete in-memory computation, it performs advanced directed acyclic graph (DAG) scheduling and fast optimising using Hadoop cluster and is reliable to execute data in the form of streams. It works with any programming language like Java, Python, Scala, R, etc. [24], [66]. Apache Flink uses Kappa/ Lambda processing system with multiple layers for storing the data and performing in-memory data execution. The processing is scalable to around 1000s nodes and beyond with high throughput and low latency of performance power [26], [67].

IV. Research Results

The spatiotemporal big data aspect focuses on different frameworks and learning of them. The big data is a large scale of 4V’s of data, and data type is increasing in the form of unstructured data, which needs proper analysis to perform and optimization of work to make big data the structured data. As spatiotemporal big data is complex to manage, store and execute in normal database, it requires advanced architectural characteristics and feature attributes of databases that can assist in managing, storing, and decision making. Thus, spatiotemporal big data can create a strong framework that can help predict the future results based on the current knowledge. The framework makes an intelligent system of artificial intelligence, mobility, Internet of Everything (IoE) and Internet of Things (IoT) of big data from many devices, and social

TABLE II

COMPARISON OF ARCHITECTURAL CHARACTERISTICS OF HADOOP, SAMZA, STORM, SPARK, AND FLINK ON THE BASIS OF SPATIOTEMPORAL BIG DATA

	Apache Hadoop [13], [18], [57], [59], [68], [69], [70], [71]	**Apache Samza** [14], [22], [72]	**Apache Storm** [20], [22], [64]	**Apache Spark** [24], [30], [57], [61], [71], [73]	**Apache Flink** [26], [58], [67], [73]
Processing engine	Batch Processing	Streaming – Kafka	Streaming – Trident	Real Time & Micro Batch Streaming – Hbase, Cloudera, and Impala	Run time streaming Kappa or Lambda
Storage layer	HDFS, YARN, MapReduce	MapReduce, HDFS	YARN	MapReduce	MapReduce, Storm
In-memory processing	No	No	hash map (topology)	DAG	Explicit memory management
Processing power	Petabytes data	API Pipelines	Millions T/s/n*	100x faster	1000s nodes
Fault tolerance	Clusters	YARN	Auto Reset	High	exactly-once processing
Scalability	Distributed/ Grid	Partitioned	Parallel	Parallel/Cluster	Distributed
Reliability	Single batch	Message	Tuple	Objects/Data Stream	Objects/Data Stream
Programming language	Java	JVM only	Any	Any (Java/Python/R/Scala)	Any (Java/Scala)

*millions of Tuples per second per nodes (millions T/s/n)

TABLE III

FEATURE COMPARISON OF HADOOP, SPATIAL HADOOP, SAMZA, STORM, SPARK, GEOSPARK, SPATIAL SPARK, FLINK BASED ON SPATIOTEMPORAL ATTRIBUTES OF BIG DATA

	Apache Hadoop		**Apache Samza** [14], [22], [73]–[80]	**Apache Storm** [20], [22], [30], [64], [78], [81]–[83]	**Apache Spark**			**Apache Flink** [26], [58], [67], [73], [82], [84], [85]
	Hadoop [13], [18], [57], [59], [86], [87], [88], [89], [90]	**Spatial Hadoop** [57], [59], [81], [91], [92], [93]			**Spark** [24], [30], [94]–[97]	**Geo Spark** [30], [57], [58], [61], [98], [99]	**Spatial Spark** [30], [57], [58], [100], [101]	
Query language / Domain specific languages	Y	Y	Y	Y	Y	Y	N	Y
	Y	Y	N	Y	Y	Y	Y	Y
Spatiotemporal data	N	Y	N	Y	N	Y	Y	Y
Spatial partitioning	N	Y	N	N	N	Y	Y	Y
With indexing	Y	Y	N	Y	Y	Y	Y	Y
Without indexing	N	Y	N	Y	Y	N	Y	Y
Topological relationships:								
Contains	Y	Y	Y	Y	Y	Y	Y	Y
Contained by	N	Y	N	Y	Y	N	Y	Y
Cross	Y	Y	N	N	Y	N	Y	Y
Equal	Y	Y	Y	Y	Y	Y	Y	Y
Disjoints	Y	Y	N	N	Y	N	N	Y
Intersects	N	Y	Y	Y	Y	Y	Y	Y
Joins	Y	Y	Y	Y	Y	Y (limit)	Y	Y
Touch	N	N	N	N	N	N	N	Y
Within distance	N	Y	Y	N	Y	N	Y	Y
K-nearest neighbours	Y	Y	Y	Y	Y	Y	N	Y
Clustering	N	N	N	Y	Y	Y	Y	Y
Real-time spatiotemporal query tools	Suitable ESRI tool		Samza SQL	Tiny Storm SQL	Spark SQL	Geospatial SQL	Spatial SparkSQL	Flink streaming SQL

medium of data generation needs accuracy of spatial and temporal datasets. The growth of big data and contribution of Apache Foundation come with Apache Hadoop for big data processing. This innovation leads to a new knowledge ground that will assist in processing, managing, storing of big data in respect to spatial and temporal features in the database. Although many system tools are developed by other vendors in the market, they all use the basics of Apache Hadoop framework (as it is open source) and make one layer up on Hadoop as their own contribution of work to spatiotemporal big datasets. More and more advanced frameworks are developed by Apache and among them five different frameworks are discussed in the research. These five frameworks are arranged in the form of Table II and Table III. The focus of study is to portrait work done in the aspect of spatiotemporal big data on different frameworks. These two tables show the comparative study of five frameworks focused on the architectural characteristics and feature attributes of spatiotemporal big data. Table II shows the comparison of five frameworks based on eight architectural characteristics of spatiotemporal big dataApache Spark and Apache Flink are more effective frameworks to be used for executing the spatiotemporal big data. The frameworks use different programing languages as Apache Spark uses all big data programing language (Java, Python, Scala, and R), which makes it more effective to be used by more customers compared to other frameworks. Apache Spark is a fast cluster computing framework for big data with real-time analyses, and Apache Flink uses the run time streaming of data to execute more data faster.

Table III subdivides frameworks into stations of ten different attributes to solve the spatiotemporal big data problems. The study implies query language, domain specific language, spatiotemporal data, spatial partitioning, with/without indexing, topological relationships, k-nearest neighbour, and clustering features offered in these different frameworks. The working on these attributes is reflected as references to different authors that gives a proof of study in the form of yes (y) or no (n) and also additional study was required to perform the remaining features.

V. Discussion & Future Work

The spatiotemporal aspect in big data allows for additional projection of learning on different frameworks that can assist in performing more tasks during less time of cycles. The study is based on the comparison of architectural characteristics and feature attributes of frameworks. Some frameworks have been discussed in [57], where some features are available and unavailable in frameworks (Hadoop-GIS, Spatial Hadoop, GeoSpark, SpatialSpark, and Stark) with some ambiguity in [57]. According to the qualitative study on the spatiotemporal aspects in big data presented in Table II and Table III, Apache Spark and Apache Flink are more advanced and stable frameworks to work on spatiotemporal big data but some features still need to be trained and tested to find the strong proof of work done in the study. The comparison of architectures and features among many frameworks has been performed with many high priority attributes used to handle the big data with spatial and temporal data in them. Tables II and III give a complete overview of aspects involved in managing spatiotemporal aspects in big data. First, in Table II, architectural characteristics are processing of engine, storing capacity in layers, in-memory processing, processing power, fault tolerance, scalability, reliability, and programming language to execute programs/functions in the system. The architectural characteristics show that Apache Spark and Apache Flink handle spatiotemporal aspects in big data more accurately and efficiently. Apache Spark gives real-time data streaming with micro-batch streaming using HBase, Cloudera, and Impala structures and Apache Flink gives run-time streaming using Kappa or Lambda structures. Spark has MapReduce storage scheme with 100x faster processing because of DAG in-memory processing method and Flink has MapReduce but also Storm storage scheme with 1000s nodes processing because of explicit memory management. Spark gives parallel and cluster scalability, high tolerance, object/data streaming and Flink gives distributed scalability, exactly-once processing tolerance, objects/data streaming. Spark has been supported by Java/Python/R/Scala and Flink has Java/Scala programming languages. Second, in Table III, feature comparison of frameworks includes query language/domain specific languages, spatiotemporal data, spatial partitioning, with indexing, without indexing, topological relationships (contains, contained by, cross, equal, disjoints, intersects, joins, touch, within distance), k-nearest neighbours, clustering, and real-time spatiotemporal query tools. The feature comparison shows that Apache Flink is more accurate, efficient, and effective framework among all of them. Flink has all features available in it and has Flink StreamingSQL environment to work on spatiotemporal aspects in big data. These attributes show how the frameworks are similar/dissimilar to each other and which one is more powerful to handle big data with spatiotemporal features in it.

References

[1] PWC, “Big Data Analytics - UN Data Innovation Lab 4,” University of Nairobi, Nairobi, 2017.

[2] J. Kerber, “Demystifying Big Data: A Practical Guide To Transforming The Business of Government,” pp. 1–40, 2012.

[3] McKinsey & Company, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Glob. Inst., Report, p. 156, 2011.

[4] CEBR, “Data equity Unlocking the value of big data,” Report for SAS, pp. 1–44, April 2012.

[5] CEBR, “The Value of Big Data and the Internet of Things to the UK Economy,” Rep. SAS by Cent. Econ. reforms, 2016.

[6] B. NT, “10 key things to remember while dealing with big data,” Big Data Made Simple: A Crayon Data Resource, 2014. [Online]. Available: http://bigdata-madesimple.com/10-key-things-to-remember-while-dealing-with-big-data/. [Accessed: 25 Oct. 2017].

[7] “7 Big Data Examples – Application of Big Data in Real Life,” Intellipaat. [Online]. Available: https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/. [Accessed: 2 Nov. 2017].

[8] R. H. Güting, and M. Schneider, *Moving Objects Databases*, 1st ed. Morgan Kaufmann, 2005.

[9] S. Rathee, and A. Yadav, “Survey on Spatio-Temporal Database and Data Models with relevant Features,” *International Journal of Scientific and Research Publications*, vol. 3, no. 1, pp. 152–156, 2013.

[10] I. Ali, H. Samoon, and A. Khan, "23 killed as monsoon rains lash Karachi," Dawn News, 2017. [Online]. Available: https://www.dawn.com/news/1355132. [Accessed: 01-Nov-2017].
[11] "Temporal Database," Teradata Database, Tools and Utilities Release 16.00. [Online]. Available: https://www.info.teradata.com/HTMLPubs/DB_TTU_16_00/index.html#page/SQL_Reference%2FB035-1182-160K%2Fyxa1472240621730.html%23wwID0EX1BI. [Accessed: 03-Nov-2017].
[12] "Temporal Database Management System," Teradata Database, Tools and Utilities Release 16.00. [Online]. Available: https://www.info.teradata.com/HTMLPubs/DB_TTU_16_00/index.html#page/SQL_Reference%2FB035-1182-160K%2Fedi1472240621683.html%23. [Accessed: 03-Nov-2017].
[13] T. White, *Hadoop: The definitive guide*, 4th ed., United States of America: O'Reilly Media, Inc, 2015.
[14] J. Ellingwood, "Hadoop, Storm, Samza, Spark, and Flink: Big Data Frameworks Compared," *Digital Ocean*, 2016. [Online]. Available: https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared. [Accessed: 17-Oct-2017].
[15] "What is batch processing?," *IBM Knowledge Center* , 2010. [Online]. Available: https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc_whatisbatch.htm. [Accessed: 25-Nov-2017].
[16] W. Stallings, *Operating Systems: Internals and Design Principles*, 7th ed. Prentice Hall, 2012.
[17] V. Prajapati, *Big Data Analytics with R and Hadoop*. Birmingham: Packt Publishing Ltd., 2013.
[18] "Welcome to Apache™ Hadoop®!," Apache Software Foundation., 2014. [Online]. Available: http://hadoop.apache.org/. [Accessed: 05-Dec-2017].
[19] S. Kamburugamuve, and G. Fox, "Survey of Distributed Stream Processing," Indiana University, Bloomington, 2013.
[20] "Apache Storm," Apache Software Foundation, 2015. [Online]. Available: http://storm.apache.org/. [Accessed: 04-Dec-2017].
[21] M. H. Iqbal, and T. R. Soomro, "Big Data Analysis: Apache Storm Perspective," *Int. J. Comput. Trends Technol.*, vol. 19, no. 1, pp. 9– 14, 2015. https://doi.org/10.14445/22312803/IJCTT-V19P103
[22] "What is Samza?," Apache Software Foundation. [Online]. Available: http://samza.apache.org/. [Accessed: 04-Dec-2017].
[23] P. Sams, *Selenium Essentials*. Packt Publishing Limited, 2015.
[24] "Apache Spark™ - Unified Analytics Engine for Big Data," Apache Software Foundation. [Online]. Available: http://spark.apache.org/. [Accessed: 04-Dec-2017].
[25] A. G. Shoro, and S. & T. R. Soomro, "Big Data Analysis: Ap Spark Perspective," *Glob. J. Comput. Sci. Technol.*, vol. 15, no. 1, 2015.
[26] "Apache Flink: Stateful Computations over Data Streams," Apache Software Foundation, 2017. [Online]. Available: http://flink.apache.org/. [Accessed: 04-Dec-2017].
[27] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017. https://doi.org/10.1016/j.jbusres.2016.08.001
[28] D. Boyd, and K. Crawford, "Critical Questions for Big Data," *Information, Commun. Soc.*, vol. 15, no. 5, pp. 662–679, Jun. 2012. https://doi.org/10.1080/1369118X.2012.678878
[29] Y. Chen, M. Guizani, Y. Zhang, L. Wang, N. Crespi, and G. M. Lee, "When Traffic Flow Prediction Meets Wireless Big Data Analytics," CoRR abs/1709.08024, 2017.
[30] F. Zhang *et al.*, "Real-Time Spatial Queries for Moving Objects Using Storm Topology," *ISPRS Int. J. Geo-Information*, vol. 5, no. 10, p. 178, 2016. https://doi.org/10.3390/ijgi5100178
[31] R. Ravanelli *et al.*, "Monitoring the Impact of Land Cover Change on Surface Urban Heat Island through Google Earth Engine: Proposal of a Global Methodology, First Applications and Problems," *Remote Sens.*, vol. 10, no. 9, p. 1488, Sep. 2018. https://doi.org/10.3390/rs10091488
[32] C. R. Lakshmi, K. RammohanRao, and R. RajeswaraRao, "Exploring Big Data Analytics for Satellite Imagery Data Using Hadoop Technique," *Int. J. Eng. Res. Comput. Sci. Eng.*, vol. 4, no. 8, 2017.
[33] R. Kachelriess, "Managing spatiotemporal big data stores," ArcGIS Enterprise. [Online]. Available: http://enterprise.arcgis.com/en/geoevent/latest/administer/managing-big-data-stores.htm. [Accessed: 10-Nov-2018].
[34] J. F. Roddick, M. J. Egenhofer, E. Hoel, D. Papadias, and B. Salzberg, "Spatial, temporal and spatio-temporal databases - hot issues and directions for phd research," Newsletter ACM SIGMOD record, vol. 33, no. 2, 2014. https://doi.org/10.1145/1024694.1024724
[35] S. Shekhar, V. Gunturi, M. R. Evans, and K. Yang, "Spatial big-data challenges intersecting mobility and cloud computing," *Proc. Elev. ACM Int. Work. Data Eng. Wirel. Mob. Access - MobiDE '12*, New York, pp. 1–6, 2012. https://doi.org/10.1145/2258056.2258058
[36] R. R. Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, and S. Shekhar, "Spatiotemporal Data Mining in the Era of Big Spatial Data: Algorithms and Applications," in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, 2012. https://doi.org/10.1145/2447481.2447482
[37] R. R. Vatsavai and B. Bhaduri, "Geospatial Analytics for Big Spatiotemporal Data: Algorithms, Applications, and Challenges," *NSF Work. Big Data Extrem. Comput.*, 2013.
[38] D. Cugler, D. Oliver, and M. Evans, "Spatial Big Data: Platforms, Analytics, and Science," *Spatial.Cs.Umn.Edu*, 2013.
[39] X. Chen, H. Vo, A. Aji, and F. Wang, "High performance integrated spatial big data analytics," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data - BigSpatial '14*, Nov. 4, 2014. https://doi.org/10.1145/2676536.2676538
[40] M.-H. Tsou, "Big data: techniques and technologies in geoinformatics," *Ann. GIS*, vol. 20, no. 4, pp. 295–296, 2014.
[41] M. R. Evans, D. Oliver, K. Yang, X. Zhou, R.Y. Ali, and S. Shekhar, "Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities," GeoJournal Library, pp. 143–170, Jun. 2018. https://doi.org/10.1007/978-94-024-1531-5_8
[42] M.-H. Tsou, "Research challenges and opportunities in mapping social media and Big Data," *Cartogr. Geogr. Inf. Sci.*, vol. 42, no. sup.1, pp. 70–74, 2015. https://doi.org/10.1080/15230406.2015.1059251
[43] B. Sadiq *et al.*, "A spatio-temporal multimedia big data framework for a large crowd," in *Proc. 2015 IEEE International Conference on Big Data,* Nov. 2015. https://doi.org/10.1109/BigData.2015.7364075
[44] K. Liu, Y. Yao, and D. Guo, "On managing geospatial big-data in emergency management," in *Proc. 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management - EM-GIS '15*, 2015. https://doi.org/10.1145/2835596.2835614
[45] B. Y. Chen, H. Yuan, Q. Li, S.-L. Shaw, W. H. K. Lam, and X. Chen, "Spatiotemporal data model for network time geographic analysis in the era of big data," *International Journal of Geographical Information Science*, vol. 30, no. 6, pp. 1041–1071, Nov. 2015. https://doi.org/10.1080/13658816.2015.1104317
[46] J. Xing and R. E. Sieber, "A land use/land cover change geospatial cyberinfrastructure to integrate big data and temporal topology," *International Journal of Geographical Information Science*, vol. 30, no. 3, pp. 573–593, Nov. 2015. https://doi.org/10.1080/13658816.2015.1104534
[47] L. Zhao, L. Chen, R. Ranjan, K.-K. R. Choo, and J. He, "Geographical information system parallelization for spatial big data processing: a review," *Cluster Comput.*, vol. 19, no. 1, pp. 139–152, 2015. https://doi.org/10.1007/s10586-015-0512-2
[48] C. M. Dalton and J. Thatcher, "Inflated granularity: Spatial 'Big Data' and geodemographics," *Big Data Soc.*, 2015.
[49] M. Frank and S. Zander, "Smart web services for big spatio-temporal data in geographical information systems," in *CEUR Workshop Proceedings*, 2016.
[50] Z. Li, F. Hu, J. L. Schnase, D. Q. Duffy, T. Lee, M. K. Bowen, and C. Yang, "A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce," *International Journal of Geographical Information Science*, vol. 31, no. 1, pp. 17–35, Jan. 2016. https://doi.org/10.1080/13658816.2015.1131830
[51] S. Li, X. Ye, J. Lee, J. Gong, and C. Qin, "Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective," *Applied Spatial Analysis and Policy*, vol. 10, no. 3, pp. 421–433, Mar. 2016. https://doi.org/10.1007/s12061-016-9185-3
[52] D. Zhu, "Spatial-temporal difference equations and their application in spatial-temporal data model especially for big data," *Journal of Difference Equations and Applications*, vol. 23, no. 1–2, pp. 66–87, Apr. 2016. https://doi.org/10.1080/10236198.2016.1167890
[53] L. Wang, W. Song, and P. Liu, "Link the remote sensing big data to the image features via wavelet transformation," *Cluster Computing*, vol. 19, no. 2, pp. 793–810, May 2016. https://doi.org/10.1007/s10586-016-0569-6

[54] K. Liu, H. Wang, and Y. Yao, "On storing and retrieving geospatial big-data in cloud," in *Proceedings of the Second ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management - EM-GIS '16*, 2016. https://doi.org/10.1145/3017611.3017627

[55] R. F. Dos Santos, A. Boedihardjo, S. Shah, F. Chen, C. T. Lu, and N. Ramakrishnan, "The big data of violent events: algorithms for association analysis using spatio-temporal storytelling," *Geoinformatica*, vol. 20, no. 4, pp. 879–921, 2016. https://doi.org/10.1007/s10707-016-0247-0

[56] M. Kezunovic *et al.*, "Predicting Spatiotemporal Impacts of Weather on Power Systems Using Big Data Science," in W. Pedrycz, SM. Chen. Eds. Data Science and Big Data: An Environment of Computational Intelligence. Studies in Big Data, vol 24, Springer, 2017. https://doi.org/10.1007/978-3-319-53474-9_12

[57] S. Hagedorn, P. Götze, K.-U. Sattler, "Big Spatial Data Processing Frameworks: Feature and Performance Evaluation," in *Proc. 20th International Conference on Extending Database Technology (EDBT)*, March 21–24, 2017. https://doi.org/10.5441/002/edbt.2017.52

[58] Z. Galić, E. Mešković, and D. Osmanović, "Distributed processing of big mobility data as spatio-temporal data streams," *Geoinformatica*, vol. 21, no. 2, pp. 263–291, Apr. 2016. https://doi.org/10.1007/s10707-016-0264-z

[59] L. Alarabi, M. F. Mokbel, and M. Musleh, "ST-Hadoop: A MapReduce Framework for Spatio-Temporal Data," Lecture Notes in Computer Science, pp. 84–104, 2017. https://doi.org/10.1007/978-3-319-64367-0_5

[60] Z. Wang, *et. al*., 2017, "A large-scale spatio-temporal data analytics system for wildfire risk management," in *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data,* Chicago, Illinois, May 14–14, 2017. https://doi.org/10.1145/3080546.3080549

[61] Z. Huang, Y. Chen, L. Wan, and X. Peng, "GeoSpark SQL: An Effective Framework Enabling Spatial Queries on Spark," *ISPRS International Journal of Geo-Information*, vol. 6, no. 9, p. 285, Sep. 2017. https://doi.org/10.3390/ijgi6090285

[62] W. M. K. Trochim and J. P. Donnelly, "Qualitative Unobtrusive Measures," in *Research methods knowledge base*, 3rd ed., Mason, OH : Thomson Custom Pub., 2007, pp. 141–153.

[63] D. De Capite, "Techniques in Processing Data on Hadoop," *Pap. SAS033*, SAS Institute Inc., 2014.

[64] P. Zapletal, "Comparison of Apache Stream Processing Frameworks: Part 1," [Online]. Available: https://www.cakesolutions.net/teamblogs/comparison-of-apache-stream-processing-frameworks-part-1. [Accessed: 05-Dec-2017].

[65] I. Mushketyk, "Apache Flink vs. Apache Spark - DZone Big Data," 2017. [Online]. Available: https://dzone.com/articles/apache-flink-vs-apache-spark-brewing-codes. [Accessed: 11-Dec-2017].

[66] "Apache Spark," *GitHub Inc*, 2017. [Online]. Available: https://github.com/apache/spark. [Accessed: 11-Dec-2017].

[67] "Apache Flink," *GitHub, Inc*, 2017. [Online]. Available: https://github.com/apache/flink. [Accessed: 12-Dec-2017].

[68] "Hadoop & Big Data," *MapR Technologies, Inc*, 2016. [Online]. Available: https://mapr.com/products/apache-hadoop/. [Accessed: 13-Dec-2017].

[69] R. Paulls, "Apache Hadoop: A Big Data Solution in a Single Unit | Prowess Consulting," *Data Center*, 2014. [Online]. Available: http://www.prowesscorp.com/apache-hadoop-a-big-data-solution-in-a-single-unit/. [Accessed: 13-Dec-2017].

[70] S. P. Bappalige, "An introduction to Apache Hadoop | Opensource.com," *Red Hat, Inc*, 2014. [Online]. Available: https://opensource.com/life/14/8/intro-apache-hadoop-big-data. [Accessed: 13-Dec-2017].

[71] Vardhan, "Apache Spark vs Hadoop: Which is the Best Big Data Framework?," *Brain4ce Education Solutions Pvt*, 2015. [Online]. Available: https://www.edureka.co/blog/apache-spark-vs-hadoop-mapreduce. [Accessed: 14-Dec-2017].

[72] F. H. MD, "The Apache Software Foundation Announces Apache® Samza™ v0.13 : The Apache Software Foundation Blog," 2017. [Online]. Available: https://blogs.apache.org/foundation/entry/the-apache-software-foundation-announces11. [Accessed: 14-Dec-2017].

[73] D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "A comparison on scalability for batch big data processing on Apache Spark and Apache Flink," *Big Data Anal.*, vol. 2, no. 1, p. 1, Dec. 2017. https://doi.org/10.1186/s41044-016-0020-2

[74] "Samza - State Management," *The Apache System Foundation.Inc*, 2014. [Online]. Available: http://samza.apache.org/learn/documentation/0.8/container/state-management.html. [Accessed: 14-Dec-2017].

[75] M. Pathirage, *et. al.*, "SamzaSQL: Scalable fast data management with streaming SQL," IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), May 23–27, 2016. https://doi.org/10.1109/IPDPSW.2016.141

[76] "Samza - Concepts." [Online]. Available: https://samza.apache.org/learn/documentation/latest/introduction/concepts.html. [Accessed: 19-Dec-2017].

[77] "Announcing the release of Apache Samza 0.13.0," *Apache Software Foundation*, 2017. [Online]. Available: https://blogs.apache.org/samza/. [Accessed: 19-Dec-2017].

[78] Y. Jimu *et al.*, "SQLS: A Storm-Based Query Language System for Real-Time Stream Data Analysis," *Chinese J. Electron.*, vol. 25, no. 6, pp. 1025–1033, Nov. 2016. https://doi.org/10.1049/cje.2016.10.003

[79] G. Grover, T. Malaska, J. Seidman, and G. Shapira, *Hadoop Application Architectures: Designing Real-World Big Data Applications*, 1st ed. O'Reilly Media, Inc., 2015.

[80] "SamzaSQL: Fast Data Management with Streaming SQL and Apache Samza," *Online*, 2017. [Online]. Available: https://github.com/milinda/samza-sql. [Accessed: 10-Dec-2017].

[81] A. Eldawy, L. Alarabi, and M. F. Mokbel, "Spatial Partitioning Techniques in SpatialHadoop," *Pvldb*, vol. 8, no. 12, pp. 1602–1605, 2015. https://doi.org/10.14778/2824032.2824057

[82] F. Hueske, "Stream analytics with SQL on Apache Flink," in *Big data conference: Strata Data Conference*, 2017.

[83] Jekyll and J. Lee, "Tiny Storm SQL: A Real Time Stream Data Analysis Interface for Apache Storm · Json Lee." [Online]. Available: https://lijiansong.github.io/java/2017/06/05/tiny-storm-sql/. [Accessed: 18-Dec-2017].

[84] F. Hueske, "[FLINK-1538] GSoC project: Spatial Data Processing Library - ASF JIRA." [Online]. Available: https://issues.apache.org/jira/browse/FLINK-1538?jql=labels %3D spatial. [Accessed: 19-Dec-2017].

[85] F. Hueske, S. Wang, and X. Jiang, "Apache Flink: Continuous Queries on Dynamic Tables." [Online]. Available: https://flink.apache.org/news/2017/04/04/dynamic-tables.html. [Accessed: 20-Dec-2017].

[86] I.-H. Joo, "Spatial Big Data Query Processing System Supporting SQL-based Query Language in Hadoop," *J. Korea Inst. Information, Electron. Commun. Technol.*, vol. 10, no. 1, pp. 1–8, Feb. 2017. https://doi.org/10.17661/jkiiect.2017.10.1.1

[87] I. Portugal, P. Alencar, and D. Cowan, "A Preliminary Survey on Domain-Specific Languages for Machine Learning in Big Data," *2016 IEEE International Conference on Software Science, Technology and Engineering (SWSTE),* Jun. 2016. https://doi.org/10.1109/SWSTE.2016.23

[88] M. Jadhao, S. Bailmare, and K. Gaikwad, "Searching, Indexing And Sentimental Analysis On Big Data," *Int. J. Scientific Research & Development,* vol. 4, no. 2, 2016.

[89] "Apache/Hadoop - CheckingTheChanges #41," *GitHub, Inc*, 2015. [Online]. Available: https://github.com/Shubh91/hadoop/blob/c1957fef29b07fea70938e971b30532a1e131fd0/hadoop-yarn-project/hadoop-yarn/hadoop-yarn-common/src/main/java/org/apache/hadoop/yarn/nodelabels/CommonNodeLabelsManager.java. [Accessed: 22-Feb-2018].

[90] M. Bomewar, *et. al.,* "Searching And Indexing On Big Data," *Int. Journal of Research In Science & Engineering* , vol. 2, no. 3, pp. 20–23, 2016.

[91] E. Eldawy, "SpatialHadoop," *Proceedings of the 2014 SIGMOD PhD symposium on - SIGMOD'14 PhD Symposium,* 2014. https://doi.org/10.1145/2602622.2602625

[92] A. Eldawy and M. F. Mokbel, "SpatialHadoop: A MapReduce framework for spatial data," *2015 IEEE 31st International Conference on Data Engineering*, Apr. 2015. https://doi.org/10.1109/ICDE.2015.7113382

[93] M. Kramer, "Controlling the Processing of Smart City Data in the Cloud with Domain-Specific Languages," *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, Dec. 2014. https://doi.org/10.1109/UCC.2014.134

[94] "Spark SQL Programming Guide - Spark 1.2.0 Documentation." [Online]. Available: https://spark.apache.org/docs/1.2.0/sql-programming-guide.html. [Accessed: 14-Dec-2017].

[95] "Apache Spark Key Terms, Explained." [Online]. Available: https://www.kdnuggets.com/2016/06/spark-key-terms-explained.html. [Accessed: 17-Dec-2017].
[96] S. Hagedorn, P. Götze, K.-U. Sattler, "The STARK framework for spatio-temporal data analytics on spark," Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn, 2017.
[97] "Apache Spark: Introduction, Examples and Use Cases | Toptal." [Online]. Available: https://www.toptal.com/spark/introduction-to-apache-spark. [Accessed: 14-Dec-2017].
[98] "GeoSpark," *GitHub, Inc.*, 2017. [Online]. Available: https://github.com/DataSystemsLab/GeoSpark. [Accessed: 14-Dec-2017].
[99] J. Yu, J. Wu, and M. Sarwat, "GeoSpark," *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15*, 2015. https://doi.org/10.1145/2820783.2820860
[100] S. You, J. Zhang, and L. Gruenwald, "Large-scale spatial join query processing in Cloud," in *Proc. International Conference on Data Engineering Workshops*, pp. 34–41, 2015. https://doi.org/10.1109/icdew.2015.7129541
[101] D. Xie, F. Li, B. Yao, G. Li, L. Zhou, and M. Guo, "Simba: Efficient In-Memory Spatial Analytics," *SIGMOD Int. Conf. Manag. Data*, pp. 1071–1085, 2016. https://doi.org/10.1145/2882903.2915237

Saadia Karim graduated from Muhammad Ali Jinnah University and received MS degree in Computer Science (2015). She is pursuing her PhD in Computer Science (2017) at the College of Computer Science & Information Systems, Institute of Business Management (IoBM).
She has more than 7 years of experience as a Software Programmer and Team Lead. She has been a Web Analyst and Adjunct Faculty at the College of Computer Science & Information Systems, Institute of Business Management (IoBM) since 2015. Recent fields of interest are machine learning, artificial intelligence, big data, fuzzy set theory, spatiotemporal data, and information retrieval.
E-mail: saadia.karim@iobm.edu.pk
ORCID iD: https://orcid.org/0000-0002-3113-4365

Dr. Tariq Rahim Soomro has received BSc (Hons) and M.Sc degrees in Computer Science from the University of Sindh, Jamshoro, Pakistan and his Ph.D. in Computer Applications (1999) from Zhejiang University, Hangzhou, China.
He has more than 23 years of extensive and diverse experience as an Administrator, Computer Programmer, Researcher and Teacher. He has been a Professor and HoD of Computer Science at the College of Computer Science & Information Systems, Institute of Business Management since 2017. His research focuses on GIS, cloud computing, big data, IoT, databases.
He is a Member of the Editorial Board "Journal of Geosciences and Geomatics" and "Journal of Software Engineering"; Member of Advisory Committee "Journal of Information and Communication Technology" (HEC recognised Journal). He is Secretory IEEE Karachi section and a Senior Member of IEEE, IEEE Computer Society and IEEE Geosciences & RS Society since 2005 and IEEE Member since 1999.
E-mail: tariq.soomro@iobm.edu.pk
ORCID iD: https://orcid.org/0000-0002-7119-0644

Dr. S. M. Aqil Burney received M.Sc. (Statistics), M.Phil. (Risk Theory and Insurance –Statistics) from the University of Karachi (UoK) and Ph.D. (Mathematics) from Strathclyde University, Glasgow-UK along with many courses in Population Studies of UN, Computing.
He has taught for more than 45 years at the UoK and extensively delivered lectures at other institutions and universities of Pakistan and abroad. He has been a Meritorious Professor at CS, UoK and HoD of Actuarial Sciences, Risk Management & Mathematics and Statistics at College of Computer Science & Information Systems, Institute of Business Management since 2014. His fields of interests are algorithmic analysis & design of multivariate time series, stochastic simulation and modelling, computer science, soft computing, risk theory, data sciences and fuzzy set theory.
Dr. Aqil Burney is Chairman (elect) of the National ICT Committee for Standard PSQCA – Ministry of Science & Technology Govt. of Pakistan and Member of the National Computing Education Accreditation Council (NCEAC), Member of IEEE (USA), Member of ACM (USA), and Fellow of the Royal Statistical Society (UK) for 30 years.
E-mail: aqil.burney@iobm.edu.pk