Big data clustering techniques based on Spark: a literature review

Mozamel M. Saeed [Corresp., 1] , **Zaher Al Aghbari** [2] , **Mohammed Alsharidah** [1]

[1] Department of Computer Science, Prince Sattam Bin Abdul Aziz, Riyadh, Saudi Arabia

[2] Department of Computer Science, University of Sharjah, Sharjah, United Arab Emirates

Corresponding Author: Mozamel M. Saeed
Email address: m.musa@psau.edu.sa

A popular unsupervised learning method, known as clustering, is extensively used in data mining, machine learning and pattern recognition. The procedure involves grouping of single and distinct points in a group in such a way that they are either similar to each other or dissimilar to points of other clusters. Traditional clustering methods are greatly challenged by the recent massive growth of data. Therefore, several research works proposed novel designs for clustering methods that leverage the benefits of Big Data platforms, such as Apache Spark, which is designed for fast and distributed massive data processing. However, Spark-based clustering research is still in its early days. In this systematic survey, we investigate the existing Spark-based clustering methods in terms of their support to the characteristics Big Data. Moreover, we propose a new taxonomy for the Spark-based clustering methods. . To the best of our knowledge, no survey has been conducted on Spark-based clustering of Big Data. Therefore, this survey aims to present a comprehensive summary of the previous studies in the field of Big Data clustering using Apache Spark during the span of 2010-2020. This survey also highlights the new research directions in the field of clustering massive data.

Big Data Clustering Techniques based on Spark: A literature review

Mozamel M. Saeed [1], Zaher Al Aghbari[2], Mohammed A Alsharidah [3]

[1] Department of Computer Science. Prince Sattam Bin Abdul Aziz -University, KSA
[2] Department of Computer Science, University of Sharjah, UAE
[3] Department of Computer Science. Prince Sattam Bin Abdul Aziz -University, KSA

Corresponding Author:
Mozamel M. Saeed[1]
ALbuhtry Street, Laylla, RIYADH, 11912, KSA
Email address: Mozamel8888@gmail.com

Abstract

A popular unsupervised learning method, known as clustering, is extensively used in data mining, machine learning and pattern recognition. The procedure involves grouping of single and distinct points in a group in such a way that they are either similar to each other or dissimilar to points of other clusters. Traditional clustering methods are greatly challenged by the recent massive growth of data. Therefore, several research works proposed novel designs for clustering methods that leverage the benefits of Big Data platforms, such as Apache Spark, which is designed for fast and distributed massive data processing. However, Spark-based clustering research is still in its early days. In this survey, we investigate the existing Spark-based clustering methods in terms of their support to the characteristics Big Data. Moreover, we propose a new taxonomy for the Spark-based clustering methods. . To the best of our knowledge, no survey has been conducted on Spark-based clustering of Big Data. Therefore, this survey aims to present a comprehensive summary of the previous studies in the field of Big Data clustering using Apache Spark during the

31 span of 2010-2020. This survey also highlights the new research directions in the field of clustering

32 massive data.

34 **Keywords** Spark-based clustering; Big Data clustering; Spark; Big Data

36
37
I. Introduction
39
40   With the emergence of 5G Technologies, a tremendous amount of data is being generated very

41 quickly, which turned into a massive amount that is termed as Big Data. The attributes of Big Data

42 such as huge volume, a diverse variety of data, high velocity and multivalued data makes data

43 analytics difficult. Moreover, extracting meaningful information from such volumes of data is not

44 an easy task [1]. As an indispensable tool of data mining, clustering algorithms play an essential

45 role in big data analysis. Clustering methods are mainly divided into density-based, partition-

46 based, hierarchical, and model-based clustering.

47   All these clustering methods are developed to tackle the same problems of grouping single and

48 distinct points in a group in such a way that they are either similar to each other or dissimilar to

49 points of other clusters. They work as follows: 1) randomly select initial clusters and 2) iteratively

50 optimize the clusters until an optimal solution is reached [2]. Clustering has an enormous

51 application. For instance, clustering is used in intrusion detection system for the detection of

52 anomaly behaviours [3],[4]. Clustering is also used extensively in text analysis to classify

53 documents into different categories [5],[6]. However, as the scale of the data generated by modern

54 technologies is rising exponentially, these methods become computationally expensive and do not

55 scale up to very large datasets. Thus, they are unable to meet the current demand of contemporary

56 data-intensive applications [7]. To handle big data, clustering algorithms must be able to extract

57 patterns from data that are unstructured, massive and heterogeneous.

58   Apache Spark is an open-source platform designed for fast-distributed big data processing.

59 Primarily, Spark refers to a parallel computing architecture that offers several advanced services

60 such machine learning algorithms and real time stream processing [8]. As such, Spark is gaining

61 new momentum, a trend that has seen the onset of wide adoption by enterprises because of its

62 relative advantages. Spark grabbed the attention of researchers for processing big data because of

63 its supremacy over other frameworks like Hadoop MapReduce [9]. Spark can also run in Hadoop

64 clusters and access any Hadoop data source.   Moreover, Spark Parallelization of clustering

65 algorithms is an active research problem, and researchers are finding ways for improving the

66 performance of clustering algorithms. The implementation of clustering algorithms using spark has

67 recently attracted a lot of research interests.

68   This survey presents the state-of-the-art research on clustering algorithms using Spark Platform.

69 Research on this topic is relatively new. Efforts started to increase in the last few years, after the

70 Big Data platform, such as Apache Spark, was developed. This resulted in a number of research

71 works that designed clustering algorithms to take advantage of the Big Data platforms, specially

72 Spark due to its speed advantage. Therefore, review articles are needed to show an overview of the

73 methodologies of clustering Big Data, highlight the findings of research and find the existing gaps

74 this area.

75   Consequently, few researchers have published review articles [10 − 20]. These review articles

76 are either before 2016 or do not present a comprehensive discussion on all types of clustering

77 methods. Therefore, a comprehensive review on clustering algorithms of big data using Apache

78 Spark is needed because it is conducted based on a scientific search strategy. To the best of our

79 knowledge, no survey has been conducted on Spark-based clustering of Big Data. For this purpose,

80 this survey aims to present a comprehensive summary of the previous studies in the field of Big

81 Data clustering using Apache Spark during the span of 2010-2020. The contributions of this review

82 are:

83 • This review includes quality literature from pre-defined resources and based on pre-

84   defined inclusion/exclusion criteria. Therefore, out of the 476 full-text articles studied, 91

85   articles were included.

86 • A taxonomy of Spark-based clustering methods that may point researchers to new

87   techniques or new research areas.

88 • A comprehensive discussion on the existing Spark-based clustering methods and the

89   research gaps in this area. Furthermore, we presented some suggestions for new research

90   directions.

91 We believe that researchers in the general area of cluster Big Data and specially those designing

92 and developing Spark-based clustering would benefit from the findings of this comprehensive

93 review.

94 The rest of this survey is organised as follows. Section II presents the related surveys to the topic

95 of clustering Big data. In Section III, we present a background on the Apache Spark. Section IV

96 explains the methodology used in this survey. Section V discusses the different Spark clustering

97 algorithms. In Section VI, we present our discussion the clustering big data using Spark and future

98 work. Finally, we conclude the paper in Section VII.

99

100

101

II. Background

Over the last decade, a huge amount of data has been generated. This increase in data volume is attributed to the growing adoption of mobile phones, cloud-based applications, artificial Intelligence and Internet of Things. Contemporary data come from different sources with high volume, variety and velocity, which make the process of mining extremely challenging and time consuming [10]. These factors have motivated the academic and the industrial communities to develop various distributed frameworks to handle the complexity of modern datasets in a reasonable amount of time.

In this regard, Apache spark, a cluster computing, is an emerging parallel platform that is cost-effective, fast, fault-tolerant and scalable. Thereby, such features make Spark an ideal platform for the dynamic nature of the contemporary applications. Spark is designed to support a wide range of workloads including batch applications, iterative algorithms, interactive queries, and streaming [11]. Spark extends the Hadoop model and support features such as in-memory computation and resilient distributed dataset, which make it significantly faster than the traditional Hadoop map-reduce for processing large data [12].

As shown in Fig. 1, at the fundamental level, spark consist of two main components; A driver which takes the user code and convert it into multiple tasks which can be distributed across the hosts, and executors to perform the required tasks in parallel. Spark is based on RDD, which is a database tables that is distributed across the nodes of the cluster. Spark supports two main operations; Transformations; and actions. Transformation preform operations on the RDD and generates new one; Action operations are performed on RDD to produce the output [13].

A. Spark Components

1) Spark Core

Spark core is the foundation of Apache Spark and contains important functionalities, including components for task scheduling, memory management, fault recovery, interacting with storage systems. Spark Core is also home to the API that defines resilient distributed datasets (RDDs), which are Spark's main programming abstraction. RDDs represent a collection of items distributed across many compute nodes that can be manipulated in parallel. Spark Core provides many APIs for building and manipulating these collections [14].

2) Spark Streaming

Spark streaming component provide scalable, high throughput API for processing of real-time stream data from various sources. Examples of data streams include logfiles generated by production web servers, or queues of messages containing status updates of a particular system [15].

3) Spark MLlib

Spark comes with a library MLlib which supports several common Machine Learning algorithms that include classification, regression, clustering, features extraction, transformation and dimensionality reductions [16].

4) Spark SQL

Spark SQL [17] is a module for processing structured data, which also enables users to perform SQL queries. This module is based on the RDD abstraction by providing Spark core engine with more information about the structure of the data.

149   **5) Spark Graphx**

150   GraphX[18] is a library for manipulating graphs (e.g., a social network's friend graph) and

151   performing graph-parallel computations. Like Spark Streaming and Spark SQL, GraphX extends

152   the Spark RDD API, allowing us to create a directed graph with arbitrary properties attached to

153   each vertex and edge. GraphX also provides various operators for manipulating graphs (e.g.,

154   subgraph and mapVertices) and a library of common graph algorithms.

155

156   **B. Clustering Big Data**

157   Clustering is a popular unsupervised method and an essential tool for Big Data Analysis.

158   Clustering can be used either as a pre-processing step to reduce data dimensionality before running

159   the learning algorithm, or as a statistical tool to discover useful patterns within a dataset. Clustering

160   methods are based on iterative optimization [19]. Although these methods are effective in

161   extracting useful pattern from datasets, they consume massive computing resources and come with

162   high computational costs due to the high dimensionality associated with contemporary data

163   applications [20].

164

165   **C. Challenges of Clustering Big Data**

166   The challenges of clustering big data are characterized into three main components:

167   **1) Volume:** as the scale of the data generated by modern technologies is rising exponentially,

168   clustering methods become computationally expensive and do not scale up to very large

169   datasets.

170   **2) Velocity:** this refers to the rate of speed in which data is incoming to the system. Dealing with

171   high velocity data requires the development of more dynamic clustering methods to derive

172   useful information in real time.

173   **3) Variety:** Current data are heterogeneous and mostly unstructured, which make the issue to

174   manage, merge and govern data extremely challenging.

175   Conventional clustering algorithms cannot handle the complexity of big data due the above

176   reasons. For example, k-means algorithm is an NP-hard, even when the number of clusters is

177   small. Consequently, scalability is a major challenge in big data. Traditional clustering methods

178   were developed to run over a single machine and various techniques are used to improve their

179   performance. For instance, sampling method is used to perform clustering on samples of the data

180   and then generalize it to the whole dataset. This reduces the amount of memory needed to process

181   the data but results in lower accuracy. Another technique is features reduction where the dimension

182   of the dataset is projected into lower dimensional space to speed up the process of mining [21].

183   Nevertheless, the constant growth in big data volume exceeds the capacity of a single machine,

184   which underline the need for clustering algorithms that can run in parallel across multiple

185   machines. For this purpose, Apache spark has been widely adapted to cope with big data clustering

186   issues. Spark provides in-memory, distributed and iterative computation, which is particularly

187   useful for performing clustering computation. It also provides advanced local data caching system,

188   fault-tolerant mechanism and faster-distributed file system.

189

III. Literature Review

191

192   The topic of clustering big data using Spark platform have not been adequately investigated by

193   academia. This suggests a comprehensive survey on research works in this regard. The literature in

194 this area has already come up with some surveys and taxonomies, but most of them are related to

195 Hadoop platform while others are outdated or do not cover every aspect of clustering big data

196 using Spark.

197 The work in [22] conducted a survey on Hadoop framework for big data processing. Different

198 features of Hadoop map-reduce are discussed to deal with the problems of scalability and

199 complexity for processing big data. [23] conducted a survey on k-means using map reduce model.

200 In this article the technical details of parallelizing k-means using Apache Hadoop is discussed.

201 According to this research, k-means method is regarded as a viable approach for certain

202 applications of big data clustering and has attracted many researchers than any other techniques.

203 On the other hand, [24] conducted a survey on the major challenges for big data processing using

204 Hadoop map- reduce. According to this survey, network latency is the main limitation of Hadoop.

205 The authors of [25] conducted a survey on big data and Hadoop architecture. The paper classifies

206 existing Hadoop based systems and discusses their advantages and disadvantages. The paper

207 explains the different technologies (Hbase, Hive, Pig, etc.) used with the Hadoop distributed file

208 system (HDFS). [26] conducted a survey on large scale data processing using Hadoop over the

209 cloud. The main components of Hadoop platform and their functionalities are discussed.

210 The work in [27] conducted a comprehensive survey on spark ecosystem for processing large-

211 scale data. In this article, spark architecture and programming model is introduced. The authors

212 discussed the pros and cos of spark platform as well as the various optimization techniques used

213 for improving spark performance for processing large scale data.

214 In [28], the authors discussed the advantages of spark over the Hadoop map-reduce model. [29]

215 conducted a survey on the parallelization of density-based clustering algorithm for spatial data

216 mining based on spark. The authors of [30] conducted a performance evaluation of k-means over

217 spark and map- reduce. On the other hand, a performance evaluation of three versions of k-means

218 clustering for biomedical data using spark was conducted in [31]. A performance evaluation of

219 parallel k-means with optimization algorithms for clustering big data using spark was conducted in

220 [32]. However, all the above surveys are either before 2016 or do not present a comprehensive

221 discussion on all types of clusters. Therefore, a comprehensive survey on clustering algorithms of

222 big data using Apache Spark is required to assess the current state-of-the-art and outline the future

223 directions of clustering big data.

224

225 ## IV. Survey Methodology

226

227 The subject matter reviewed in this article is based on a literature review in clustering methods

228 using Apache spark. We searched for the works regarding this topic and classify them into

229 different Clustering techniques. All these papers talk about optimizing clustering techniques to

230 solve the issues of big data clustering problems for various problems, viz., improve clustering

231 accuracy, minimize execution time, increase throughput and scalability. Particularly, we are

232 addressing the following questions:

233   • What are the types of Spark-based clustering methods?

234   • Which methods were used in the literature to cluster Big Data?

235   • What are the gaps in this research area?

236   • What optimization techniques were used in clustering?

237   • What are the pros and cons of the different Spark-based clustering methods?

238

239

240 ### A. Search strategy

241   To narrow the scope of the searching for relevant papers to be included in this study, we used the

242 "AND" and "OR" Boolean operators to combine the terms related to Spark-based clustering of Big

243 Data. The following terms are used to find the relevant papers.

244   • "Clustering big data using spark",

245   • "Apache Spark for Big data",

246   • "Clustering Big Data",

247   • "Clustering methods",

248   • "Data partitioning",

249   • "Big Data Partitioning",

250   • "Data segmentation".

251

252   The papers relevant to Spark-based clustering of Big Data were retrieved from the following

253 online sources.

254   • IEEE Explorer

255   • Springer,

256   • Elsevier

257   • ScienceDirect,

258   • Google Scholar,

259   • Researchgate,

260

261

262

263 **B. Paper filtering**

264   Initially 1230 and additional 43 reference books papers were identified through our search using

265   the previously explained research strategies. As shown in Fig. 2, 797 of these were eliminated via

266   our exclusion criteria. 476 papers were remaining. By reading and analysing the full-text articles,

267   385 of them were excluded. Irrelevant papers were removed by applying the exclusion criteria

268   (shown below). In addition, duplicate papers retrieved from multiple sources were removed.

269   Finally, 91 articles were included in this survey. The following inclusion/exclusion rules are

270   applied on these papers.

271   • Inclusion criteria:

272   ▪ papers published within the period from January 2010 to April 2020.

273   ▪ papers in the area of Spark-based Big data clustering.

274   ▪ papers written in English language.

275   • Exclusion criteria:

276   ▪ papers on clustering but not on Big data.

277   ▪ papers that are not using a Big data platform such as Spark. papers with no clear

278   publication information, such as publisher, year, etc.

279

280

V. Spark-based Clustering Algorithms

282

283   In this work, the taxonomy of Spark-based Big Data clustering is developed to cover all the

284   existing methods. Fig. 3 shows the developed taxonomy.

285   **Survey Findings:** The research questions (see Section IV) that we investigated in this survey

286   are addressed as shown below:

287 • Answer to Q1: The Spark-based clustering algorithms were divided into three main

288 categories: k-means based methods, hierarchal-based methods and density based methods.

289 Each of these main categories were divided further into subcategories as depicted in Fig. 3.

290 A detailed discussion of the Spark-based clustering methods in these subcategories is

291 presented in the subsection below (Subsection V.A, V.B and V.C, Fig. 3 and Table 1).

292 • Answer to Q2: We discuss the different methods that have been proposed in the literature

293 under each of the three main Spark-based clustering categories in Subsections V.A, V.B

294 and V.C. The methods in these subsections are grouped based on their similarities in the

295 approach. This grouping of the discussed methods is shown in Table 1.

296 • Answer to Q3: The gaps in the Spark-based clustering field are identified into two main

297 points. The first is the lack of utilizing AI tools in clustering data and lack of using Big

298 Data platforms. The second is that most current clustering methods do not support the

299 characteristics of variety and velocity of Big Data. More discussion on this issue is in

300 Section VI.

301 • Answer to Q4: Some existing works employed optimization techniques to improve

302 clustering results. These optimization techniques were mainly used with k-means methods

303 as discussed in subsection V.A.2 and V.D.

304 • Answer to Q5: The pros and cons of the different methods are discussed in the subsections

305 V.A, V.B and VC that discuss the different types of Spark-based clustering methods. We

306 also discuss our findings related to the Spark-based clustering methods in Section VI.

307

308

309

A. k-means based Clustering

This method divides the data into disjoint clusters of similar points. In each cluster, a central point is obtained via a distance function and is considered as the centroid of all other points within the cluster. The clusters are iteratively optimized until an optimal solution is reached.

k-mean is a framework of clustering or a family of distance functions, which provides the basis for different variants of k-mean algorithms. k means is extensively used in clustering big data due to its simplicity and fast convergence. One major backward of k-means is the priori setting of the number of clusters, which have significant effect on the accuracy of final classification [33]. In addition, k-means is not suited in situations where the clusters do not show convex distributed or vary in sizes [34]. Due to these limitations, several modifications of k -means have been proposed such as fuzzy k-means and k-means++ [35]. Several works have been conducted to execute k-means effectively under the Spark framework to improve its performance and scalability. Therefore, the Spark-based k-means methods can be divided into four subcategories: Machine Learning based methods, Fuzzy based methods, Statistics based methods and Scalable methods.

A.1. Machine Learning based Methods

The authors of [38] designed intelligent k-means based on spark. Intelligent k-means is a fully unsupervised learning that cluster data without any information regarding the number of clusters. A parallel implementation of biclustering using map-reduce over Spark platform was proposed by [41]. For improving the selection process of k-means, [47] combines Particle Swarm Optimization and Cuckoo-search to initiate better cluster centroid selections using spark framework. The work in [60] proposes a hybrid approach that integrate k-means and decision tree to cluster and detect anomaly in big data. At first, k-means is applied on the data to produce the clusters and then decision tree algorithm is applied on each cluster to classify normal and anomaly instances.

333   In [62] the author combines rule based and k-means algorithm for the detection of network

334   anomalies using apache spark. Rule based is used for the detection of known attacked, while k-

335   means is used as unsupervised learning for the detection of new unknown attacks. Kdd cup dataset

336   was used to evaluate the algorithm and 93% accuracy was achieved. A paralleled algorithm for the

337   evolving clustering method was proposed by [63]. EMC is an online method which process one

338   data sample on a single pass and there is no iteration required to process the same data again.

339   These features make the algorithm highly efficient for processing contemporary real time

340   applications where data arrive in a stream with high dimensionality. The authors evaluated the

341   proposed algorithm using massive credit card fraud dataset and the results show its superiority over

342   the traditional single EMC method.

343   **A.2. Fuzzy based Methods**

344   The authors of [45] proposed a parallel implementation of fuzzy consensus clustering for on the

345   Spark platform for processing large scale heterogenous data. The authors of [46] developed a crime

346   pattern-discovery system based on fuzzy clustering under Spark. The method uses L2 norm rather

347   than Euclidian distance to optimize the distance computations. In another paper, the fuzzy

348   clustering method is used under Spark to detect potential criminal patterns in large-scale

349   spatiotemporal datasets [53]. In [54] the authors developed a parallel Fuzzy based image

350   segmentation algorithm for handling big data in the agriculture field. At first, the images were

351   converted to RGB and distributed to the available nodes in cloud. Then, the membership of pixel

352   points to different cluster centroids were calculated. Finally, the centroids of the clusters are

353   updated iteratively until an optimal solution is obtained. The performance of the algorithm was

354   evaluated using the Spark platform and a significant reduction in execution time compared to

355   Hadoop-based approach. The authors of [55] proposed an algorithm of fuzzy c-Means. The

356   proposed algorithm is a modification of the Scalable Random Sampling with Iterative

357   Optimization (SRSIO-FCM). The highlighted characteristics of this research were the elimination

358   of the need for maintaining the membership matrix, which proved pivotal in reducing execution

359   time.

360

361

362   **A.3. Statistics based Methods**

363   The authors of [50] used Apache Spark to perform text clustering. Two algorithms were used: k-

364   means and LDA. LDA is Widely used technique for clustering high dimensional text data and it

365   produces considerably higher clustering accuracy than conventional k- means. In [57] the authors

366   used gaussian mixture model on spark MLlib to cluster the zika virus epidemic. The produced

367   clusters were useful to visualize the spread of the virus during the epidemic. The authors of [58]

368   implemented GMM clustering method under the framework of Spark. Gibbs sampling method is

369   used instead of Expectation Maximization algorithm to estimate the parameters of the model. The

370   efficiency of the algorithms was verified via multi-method comparison. The authors in [61]

371   presented a novel distributed gaussian based clustering algorithm for analysing the behaviour of

372   households in terms of energy consumption. Various factors such as weather conditions, type of

373   day and time of the day were considered. The proposed algorithm under Spark shows a higher

374   accuracy than other standard regression methods for energy consumption forecasting.

375   **A.4. Scalable Methods**

376   A parallel implementation of k means algorithm over spark is proposed in [36]. The proposed

377   algorithm involves three strategies for seeding: 1) a subset of data is selected randomly for

378   partitioning. 2)sequentially selecting k instance based on probability. 3) stochastically selecting

379   seeds in parallel. The efficiency of the proposed algorithm was demonstrated via experiments on

380   large scale text and UCI datasets. In another paper, the authors addressed the issue of pre-

381   determining the number of input clusters which is a present problem in most K-means methods by

382   automating the number of input clusters which resulted in better clustering quality when processing

383   large scale data [37].

384   In [39], the authors presented a scalable k-means algorithm based on spark streaming for

385   processing real time- data. The algorithm consists of two parts. One part runs an online algorithm

386   over the stream data and obtains only statistically relevant information and another part that uses an

387   offline algorithm on the results of the former to produce the actual clusters. Since the algorithm

388   only retains statistically relevant information, it's possible to save more physical spaces. In

389   addition, the proposed algorithm can explain the evolution of the data as all the needed information

390   is retrievable from the stored statistical information.

391   The authors of [40] used k-means under Spark to cluster students' behaviors into different

392   categories using information gathered from universities' information system management. It is a

393   powerful technique for performing simultaneous clustering of rows and Columns in a matrix data

394   format. This method is used extensively for the study of genes expression. The authors of [42]

395   clustered satellite images in an astronomy study using in k-means++ under the spark framework.

396   In this paper, the authors simultaneously apply k-means multiple times with different initial

397   centroids and value of k under each iteration. The optimal value of k is determined by clusters

398   validity index for all the executions. The work in [43] presented a Spark-based k-prototypes (SKP)

399   clustering method for mixed large-scale data analysis. The authors exploit the in-memory

400   operations of Spark to reduce the consumption time of MRKP method. The method was evaluated

401   using simulated and real datasets under Spark and Hadoop platform and the results show that

402   higher efficiency and scalability is achieved under Spark.

403

404   The authors of [44] implemented a Scalable Random Sampling for K-Prototypes Using Spark.

405   The algorithm randomly selects a small group of data points and approximate the cluster centers

406   from these data. As a result, the method perform computation for only small portion of the whole

407   data set, which result in a significant speedup of existing k-prototypes methods. A Parallel

408   Overlapping k-means algorithm (POKM) is proposed in [48]. This algorithm can perform parallel

409   clustering processes leading to non-disjoint partitioning of data. An implementation of parallel k-

410   means with triangle inequality based on spark is proposed in [49]. The method is an improved

411   version of k-means, which is supposed to speed up the process of analysis by skipping many point-

412   centre distance computations, which can be beneficial when clustering high dimensional data. The

413   authors of [51] implemented k-means with triangle inequality to reduce search time and avoid

414   redundant computation. The authors point out that the efficiency of k-means can be improved

415   significantly using triangle inequality optimisations. A distributed possibilistic c-means algorithm

416   is proposed in [52]. Possibilistic c means differ from other k-means techniques by assigning

417   probabilistic membership values in each cluster for every input point rather than assigning a point

418   to a single cluster.

419

420   The authors of [56] implemented an adaptive k-mean using Spark stream framework for real

421   time-anomaly detection in clouds virtual machines. The authors evaluated the method under Spark

422   and Storm in terms of the average delays of tuples during clustering and prediction and the results

423   indicate that Spark is significantly faster than Storm. [59] designed a framework for clustering and

424   classification of big data. The framework integrates k-means and decision tree learning (ID3)

425   algorithms. the authors evaluated the framework under Spark in cluster of 37 nodes. the results

426   show that the proposed algorithms outperform spark machine learning library but is slightly slower

427   than the approximate k-means.

428

429

B. Hierarchical Clustering

431   This clustering technique is composed of two approaches: agglomerative and divisive. The first

432   approach considers every data point as a starter in its singleton cluster and the two nearest clusters

433   are combined in each iteration until the two different points belong to a similar cluster. However,

434   the second approach performs recursive top-down splitting. The existing hierarchical clustering

435   methods can be divided into three subcategories: Data Mining based methods, Machine Learning

436   based methods and Scalable methods.

B.1. Data Mining based Methods

438   A weighted agglomerative hierarchical clustering algorithm is introduced in [65]. The algorithm

439   was developed to analyse residents' activities in China. The data was based on the mobile phone's

440   connection with the nearest stations, and within a week that data was collected and stored in Spark

441   for analysing. At first, hot areas where there are large population were identified, followed by an

442   analysis of pedestrian's flow for each hot area. Meaningful information was obtained at less cost

443   and higher accuracy than the traditional method of investigation. The work in [66] proposed a

444   distributed-hierarchical based clustering algorithm that combines the features of the divisive and

445   agglomerative methods. The method consists of two operations. The first operation performs a

446   division on the domain of the dataset using the definition of binary space partition, which yields a

447   set of coarse clusters that are then refined by identifying outliers and assigning remaining points to

448   nearest cluster. The second operation involves an agglomerative procedure over the previously

449   refined clusters. In [68], they proposed a Distributed-based Hierarchical Clustering System for

450   Large-Scale Semiconductor Wafers (DHCSSW) by applying the big data Spark framework to

451   existing hierarchical clustering algorithm.

452   **B.2. Machine Learning based Methods**

453   In [70], the authors designed clustering algorithms that can be used in MapReduce using Spark

454   platform. Particularly, they focus on the practical and popular serial Self-organizing Map clustering

455   algorithm (SOM). [71] proposed an algorithm called Spark-GHSOM, which scales to large real-

456   world datasets on a distributed cluster. Moreover, it also proposed a new distance hierarchy

457   approach for mixed attribute datasets and generated a multi-level hierarchy of SOM layers.

458

459

460   **B.3. Scalable Methods**

461   In [64], a parallel algorithm of Single-linkage Hierarchical Clustering was proposed by

462   formulating the problem as a Minimum Spanning Tree problem. The algorithm was evaluated

463   using two large datasets with different distributions. The authors observed that spark is totally

464   successful for the parallelization of linkage hierarchical clustering with acceptable scalability and

465   high performance. The work in [67] proposed a system to detect anomaly for multi-source

466   VMware-based cloud data center. The framework monitors VMware performance stream data

467   (e.g., CPU load, memory usage, etc.) continuously. The authors of [69] presented an incremental

468   hierarchical density-based stream clustering algorithm based on cluster stability.

469

470

471

C. Density-based Clustering

473   Density-based clustering approaches in comparison with other types of clustering algorithms

474   have some superiorities, such as clustering arbitrary shape groups of data regardless of the

475   geometry and distribution of data, robustness to outliers, independence from the initial start point

476   of the algorithm, and its deterministic and consistent results in the repeat of the similar algorithm.

477   Motivated by these features, several studies have been conducted on the parallelization of Density

478   clustering method over Spark. Density-based clustering methods can be divided into four

479   subcategories: Graph based methods, Data Mining based methods, Machine Learning based

480   methods and Scalable methods.

C.1. Graph based Methods

482   In [74], the authors proposed a parallel implementation of density peaks clustering algorithm

483   based on Spark's GraphX. The method was evaluated using spark and the results indicate that spark

484   can perform up to 10x time faster compared to Hadoop map-reduce implementation. [79] proposed

485   a distributed parallel algorithm of structure similarity clustering based on Spark (SparkSCAN) to

486   cluster directed graph. Similarly, the authors of [80] exploited the advantage of the in-memory

487   computation feature of spark to design a distributed network algorithm called CASS for clustering

488   large-scale network based on structure similarity. Optimization approaches such as Bloom filter

489   and shuffle selection are used to reduce memory usage and execution time. [82] designed a

490   distributed algorithm that produces an approximate solution to the exact DBSCAN clustering. The

491   method uses vertex-centric instead of Euclidean distance whereby a neighbourhood graph is

492   computed.

493 **C.2. Data Mining based Methods**

494 In another paper [76], the authors presented a fast parallel DBSCAN algorithm using Spark to

495 get around the shuffle operation. Each executor computes the partial clusters locally. The merging

496 process is deferred until all the partial clusters have been sent back to the driver. In [78] the authors

497 propose a scalable distributed density based hesitant fuzzy clustering for finding similar expression

498 between distinct genes. The proposed method benefits from the robustness of density-based

499 clustering against outliers and from the weighted correlation operators of hesitant fuzzy clustering

500 to measure similarity. The output clusters are based on the content of the neighbour graph. [83]

501 designed and implemented a scalable Shared Nearest Neighbours clustering called SparkSNN over

502 spark framework. Shared Nearest Neighbours is proven efficient for handling high-dimensional

503 spatiotemporal data. The algorithm was evaluated in terms of scalability and speed-up using

504 Marylanf crime data, the results demonstrated the effectiveness of the proposed algorithm.

505 **C.3. Machine Learning based Methods**

506

507 An algorithm based on adaptive density estimation is proposed for distributed big data approach

508 and tested on some prevalent datasets. This algorithm has no dependency, and every step of the

509 algorithm executes independently. Bayesian Locality Sensitive Hashing (LSH) is used to divide the

510 input data into partitions. The outliers are filtered out by locality preservation, which makes this

511 approach robust. The clusters are made very much homogenous via density definition on Ordered

512 Weighted Averaging distance [72]. A scalable distributed density-based clustering for performing

513 multi- regression tasks is proposed in [77]. In this work, locality sensitive hashing is used to enable

514 the algorithm to handle high dimensional data. A distributed clustering algorithm named

515 REMOLD is introduced in [84]. A two-step strategy has been applied in the REMOLD algorithm.

516   In the first step, it uses the LSH partitioning method for balancing the effect of runtime and local

517   clustering while in the second step the partitions are clustered locally and independently using

518   Kernel-density and Higher-density nearest neighbour. Gaussian distribution is used to model the

519   local clusters. These models are eventually assembled at a central server to form the global

520   clusters.

521   **C.4. Scalable Methods**

522   In [73], a parallel implementation of DBSCAN algorithm (S_ DBSCAN) based on spark is

523   proposed. The algorithm is divided into three stages; partitioning the input data based on random

524   sampling; perform local DBSCAN in parallel to generate partial clusters; merge the partial clusters

525   based on the centroid. The algorithm can quickly realize the mergers and divisions of clustering

526   results from the original data. The authors compared the performance of their parallel algorithm

527   with a serial version on the Spark platform for massive data processing and an improvement in

528   performance was demonstrated. [75] proposed a scalable parallel implementation of DBSCAN

529   algorithm in Apache spark by applying a partitioning strategy. The algorithm uses a kd-tree in

530   order to reduce the search time. To achieve better performance and scalability, a partitioning

531   technique is applied to produce balanced sub-domains, which can be computed within Spark

532   executors. An implementation of DBSCAN algorithm using spark is proposed in [81]. Initially, a

533   pre-processing step is applied on the dataset to produce a set of representative points while

534   retaining the original data distribution and density information. This enables the algorithm to scale

535   up to large scale data. The new set is then used as an input to the algorithm for clustering. A real-

536   time density-based clustering algorithm (RT-DBSCAN) is proposed in [85]. RT-DBSCAN is an

537   extension of dbscan for supporting streamed data analysis. The algorithm employs the concept of

538   spatiotemporal distance for clustering spatio-temporal data. The algorithm was implemented over

539   spark stream and evaluated using social media content.

540

541   **D. Clustering Optimization**

542   Some Spark-based clustering techniques, especially the k-means based methods, were supported

543   by optimization techniques to improve their clustering results. Due to the rise of AI based

544   computing in recent years, some research works have utilized AI tool in enhancing the clustering

545   methods while leveraging the benefits of Big Data platforms such as Spark. Other studies adapt

546   optimization techniques to improve the performance of clustering methods. [86] proposed a hybrid

547   method composed of PSO and k-means using apache spark. The diversity of the swarm ensures

548   that a global search is conducted, hence, the resulting cluster centroids are not dependent on the

549   initial choice. The approach was compared with stand-alone k-means and it showed better

550   performance in terms of convergence. [87] proposed an adaptive swarm-based clustering for

551   stream processing of twitter data. Initially, fuzzy c-means is applied as pre-processing step to

552   produce the initial cluster centres, then the clusters are further optimized using adaptive particle

553   swarm optimization.

554   The authors of [88] and [89] combined the robust artificial bee colony algorithm with the

555   powerful Spark framework for large scale data analysis. The characteristics of ABC makes the

556   algorithms avoid local minimum while Spark in memory computation accelerates the speed of

557   computation and convergence time. The KDD CUP 99 data was utilized to verify the effectiveness

558   of the method. The experimental results show that the algorithm produce high clustering quality

559   and nearly as fast as the serial algorithms. Other unsupervised learning such as self-organised map

560   has also been proposed [90]. To tackle high dimensional data, subspace clustering was proposed by

561   [91].

VI.   Discussion and Future Direction

563   From the discussion of the previous section, we note that most existing methods (see Table 1) have

564   addressed the *volume* characteristic of the Big Data used in their experiments. However, few

565   existing methods have shown that their methods support the *variety* and *velocity* characteristics of

566   the used Big Data. Additionally, most methods used *real* Big Data validate their proposed

567   methods as seen in Table 1. From Table 1, we conclude that there is a lot of room for research in

568   clustering methods to support the characteristics of variety and velocity of Big data since only few

569   works have addressed these issues.

570   A fundamental assumption of most clustering algorithms is that all data features are considered

571   equally important. However, such approach often fails in high dimensional space. A subspace

572   clustering overcome the issue of high dimensional data by establishing a set of features that it

573   supposes to be most significant for each cluster.

574   Since the Big data platforms were only developed in the last few years, the existing clustering

575   problems adapted to such platforms were extensions of the traditional clustering techniques.

576   Researchers are yet to develop clustering techniques that are native to the Big Data platforms such

577   as Spark. The research direction of adapting the optimization techniques such as PSA, Bee colony

578   and ABC to smoothly work with Spark is yet to be investigated by researchers who are interested

579   in clustering Big Data. Another area of research that is has not been fully investigated is adopting

580   Fuzzy-based clustering algorithms on Spark.

581   In general, due to the infancy of Spark-based clustering algorithms, only few researchers

582   attempted designing techniques that leverage the potential of parallelism of Spark in cluster Big

583   Data. In the coming years, we foresee a large influx of research works in this important area of

584   Spark-based clustering of Big Data. Particularly, there are ample opportunities in future research to

585   utilize AI tools in clustering data while leveraging the benefits of Big Data platforms such as

586   Spark.

587   In Table 2, we note that most of the papers used in this survey were extracted from the IEEE

588   Explorer. However, the other data sources shown in Table 2 were of great benefit to this survey.

589   An interesting finding was shown in Table 3, where most the existing Spark-based Clustering were

590   published in the years 2016-2019. This indicates that clustering methods that leverage Big Data

591   platforms is still in its early days and there is a lot of potential of research in this area.

592   In summary, we highlight three new research directions:

593   • Utilizing AI tools in clustering data while leveraging the benefits of Big Data platforms

594   such as Spark.

595   • Clustering methods to support the characteristics of variety and velocity of Big data.

596   Additionally, support new aspects of clustering such as concept drift, scalability,

597   integration, fault-tolerance, consistency, timeliness, load balancing, privacy, and

598   incompleteness, etc.

599   • Clustering methods to utilize Spark as it is an efficient Big Data platform.

600

601

602

603   **VII. Conclusions**

604

605   As a consequence of the spread of smart devices and appearance of new technologies such as

606   IoT, huge data have been produced on daily bases. As a result, the concept of Big Data has

607 appeared. Unlike the traditional clustering approaches, Big Data clustering requires advanced

608 parallel computing for better handling of data because of the enormous volume and complexity.

609 Therefore, this work contributes to the research in this area by providing a comprehensive

610 overview of existing Spark-based clustering techniques on Big data and outlines some future

611 directions in this area.

612   Due to the infancy of the Big data platforms such as Spark, the existing clustering techniques

613 that are based on Spark are only extensions of the traditional clustering techniques. There is still

614 big room for developing clustering techniques designed specifically for Spark making use of the

615 random distribution of data onto Spark partitions, called RDDs, and the parallel computation of

616 data in the individual RDDs. Through this survey we found that most existing Spark-based

617 clustering method support the volume characteristic of Big Data ignoring other characteristics.

618 Therefore, future research should focus on other characteristics as well such as variety and

619 velocity. Additionally, future Spark-based clustering method should investigate new features such

620 as concept drift, scalability, integration, fault-tolerance, consistency, timeliness, load balancing,

621 privacy, etc.

622

623

624 **Acknowledgements**

625 The authors would like to thank the Deanship of Scientific Research at Prince Sattam Bin
626 Abdulaziz University for their support of this research

627

628

629

630 **References**

631 [1] Bhadani, A. K., & Jothimani, D. (2016). Big Data: Challenges, Opportunities, and Realities. In Singh, M.
632 K., & G., D. K. (Ed.), Effective Big Data Management and Opportunities for Implementation (pp. 1-24).
633 [2] Dave, M. and Gianey, H. "Different clustering algorithms for Big Data analytics: A review," 2016
634 International Conference System Modeling & Advancement in Research Trends (SMART), Moradabad,
635 2016, pp. 328-333, doi: 10.1109/SYSMART.2016.7894544.

636
637
638
[3] Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T. Al-Hashida, A.Y.. Intrusion detection model using machine learning algorithm on Big Data environment. *J Big Data* **5,** 34 (2018). https://doi.org/10.1186/s40537-018-0145-4.

639
640
641
[4] Hu, S., Xiao, Z., Rao, Q., and Liao, R. "An anomaly detection model of user behavior based on similarity clustering," 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2018, pp. 835-838, doi: 10.1109/ITOEC.2018.8740748.

642
643
644
[5] Fasheng, L. and Xiong, L. "Survey on text clustering algorithm -Research present situation of text clustering algorithm," 2011 IEEE 2nd International Conference on Software Engineering and Service Science, Beijing, 2011, pp. 196-199, doi: 10.1109/ICSESS.2011.5982288.

645
646
647
[6] Baltas A., Kanavos A., Tsakalidis A.K. An Apache Spark Implementation for Sentiment Analysis on Twitter Data. In: Sellis T., Oikonomou K. (eds) Algorithmic Aspects of Cloud Computing. ALGOCLOUD 2016. Lecture Notes in Computer Science, vol 10230. Springer, Cham.

648
649
650
[7] Ajin. V. W. and Kumar, L. D., "Big data and clustering algorithms," 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS), Bangalore, 2016, pp. 1-5, doi: 10.1109/RAINS.2016.7764405.

651
652
[8] Shoro, S. A. G. and Soomro, T. R., "Big data analysis: Apache Spark perspective," Global Journal of Computer Science and Technology, vol. 15, no. 1, 2015.

653
654
655
[9] Verma, A., Mansuri, A. H., and Jain, N., "Big data management processing with Hadoop MapReduce and spark technology: A comparison," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-4, doi: 10.1109/CDAN.2016.7570891.

656
657
[10] Labrinidis, A. and Jagadish, H. V.,.. Challenges and opportunities with big data. Proc. VLDB Endow. 5, 12 (August 2012), 2032–2033. DOI:https://doi.org/10.14778/2367502.2367572.

658
659
660
[11] Gousios, G,. 2018. Big data software analytics with Apache Spark. In Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings(ICSE '18). Association for Computing Machinery, New York, NY, USA, 542–543. DOI:https://doi.org/10.1145/3183440.3183458.

661
662
663
[12] Aziz K., Zaidouni D., Bellafkih M. Big Data Optimisation Among RDDs Persistence in Apache Spark. In: Tabii Y., Lazaar M., Al Achhab M., Enneya N. (eds) Big Data, Cloud and Applications. BDCA 2018. Communications in Computer and Information Science, vol 872. Springer, Cham.

664
665
[13] Salloum, S., Dautov, R., Chen, X., Peng, P. X., and Huang, J. Z.,. Big data analytics on apache spark. International Journal of Data Science and Analytics, 1(3):145–164, 2016.

666
667
668
[14] Mishra, D. D., Pathan, S. and Murthy, C., "Apache Spark Based Analytics of Squid Proxy Logs," 2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Indore, India, 2018, pp. 1-6, doi: 10.1109/ANTS.2018.8710044.

669
670
[15] Kim, K.. Real-time Streaming Data Analysis using Spark. International Journal of Emerging Trends in Engineering Research. Volume 6, 2018.

671
672
673
[16] Assefi, M., Behravesh, E., Liu G. and Tafti, A. P., "Big data machine learning using apache spark MLlib," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 3492-3498, doi: 10.1109/BigData.2017.8258338.

674
675
676
677
[17] Armbrust, M. M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K., Meng, X., Kaftan, T., Franklin, M. J., Ghodsi, A. and Zaharia, M.,. Spark SQL: Relational Data Processing in Spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data(SIGMOD '15). Association for Computing Machinery, New York, NY, USA, 1383–1394.

678
679
680
[18] Xin, R. S., Gonzalez, J. E., Franklin, M. J., and Stoica, I., "Graphx: Aresilient distributed graph system on spark," in First InternationalWorkshop on Graph Data Management Experiences and Systems,2013, pp. 2:1–2:6.

681
682
[19] Xu, D. and Tian, Y., "A Comprehensive Survey of Clustering Algorithms,"Annals of Data Science, vol. 2, no. 2, pp. 165–193, 2015.

683   [20] Zerhari, B.; Lahcen, A.A.; Mouline, S. Big data clustering: Algorithms and challenges. In Proceedings of
684   the International Conference on Big Data, Cloud and Applications, Tetuan, Morocco, 25–26 May 2015

685   [21] Shirkhorshidi A.S., Aghabozorgi S., Wah T.Y., Herawan T. Big Data Clustering: A Review. In: Murgante
686   B. et al. (eds) Computational Science and Its Applications – ICCSA 2014. ICCSA 2014. Lecture Notes in
687   Computer Science, vol 8583. Springer, Cham

688   [22] Rotsnarani Sethy, Mrutyunjaya Panda, "Big Data Analysis using Hadoop: A Survey" in Proc. Conf.
689   International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5,
690   Issue 7, July 2015.

691   [23] Bhandari, Rujal & Dabhi, Dipak. (2016). Extensive Survey on k-Means Clustering using MapReduce in
692   Datamining. Conference: International Conference on Electronics and Communication Systems (ICECS)
693   At: Coimbatore, Tamilnadu, India.

694   [24] S Sood, Akshay & Singh, Ravinder. (2019). A Survey of Performance Improvement Techniques for
695   Hadoop. International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.55 (2015).

696   [25] M. Manwal and A. Gupta, "Big data and Hadoop — A technological survey," 2017 International
697   Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT), Dehradun,
698   2017, pp. 1-6.

699   [26] D. Jiang, B. Ooi, L. Shi, and S. Wu, "Big Data Processing Using Hadoop: Survey on Scheduling," Proc.
700   VLDB Endow. vol. 3, no. 10, pp. 272–277, 2010.

701   [27] Tang, Shanjiang & He, Bingsheng & Yu, Ce & Li, Yusen & Li, Kun. (2018). A Survey on Spark
702   Ecosystem for Big Data Processing.

703   [28] R. C. Maheshwar and D. Haritha, "Survey on high performance analytics of big data with apache spark,"
704   2016 International Conference on Advanced Communication Control and Computing Technologies
705   (ICACCCT), Ramanathapuram, 2016, pp. 721-725, doi: 10.1109/ICACCCT.2016.7831734.

706   [29] F. Huang, Q. Zhu, J. Zhou, J. Tao, X. Zhou, D. Jin, X. Tan, and L. Wang, "Research on the Parallelization
707   of the DBSCAN Clustering Algorithm for Spatial Data Mining Based on the Spark Platform," *Remote
708   Sensing*, vol. 9, no. 12, p. 1301, Dec. 2017.

709   [30] S. Ketu and S. Agarwal, "Performance enhancement of distributed k-Means clustering for big Data
710   analytics through in-memory computation," 2015 Eighth International Conference on Contemporary
711   Computing (IC3), Noida, 2015, pp. 318-324, doi: 10.1109/IC3.2015.7346700.

712   [31] A. Shobanadevi and G. Maragatham, "Studying the performance of clustering techniques for biomedical
713   data using spark," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam,
714   2017, pp. 58-65, doi: 10.1109/ISS1.2017.8389249.

715   [32] Santhi V., Jose R. (2018) Performance Analysis of Parallel k-Means with Optimization Algorithms for
716   Clustering on Spark. In: Negi A., Bhatnagar R., Parida L. (eds) Distributed Computing and Internet
717   Technology. ICDCIT 2018. Lecture Notes in Computer Science, vol 10722. Springer, Cham.

718   [33] Hartigan J.A, Wong M.A., Algorithm A.S., A k-means clustering algorithm. Journal of the Royal Statistical
719   Society Series C. 1979;28(1):100–108.

720   [34] Jain A.K. Data clustering: 50 years beyond k-means. Pattern Recognition Letters. 2010;31(8):651–666

721   [35] Huang Z., Extensions to the k-means algorithm for clustering large data sets with categorical values. Data
722   mining and knowledge discovery. 1998;2(3):283–304

723   [36] Wang, B., Yin, J., Hua, Q., Wu Z. and Cao, J., "Parallelizing k-Means-Based Clustering on Spark," 2016
724   International Conference on Advanced Cloud and Big Data (CBD), Chengdu, 2016, pp. 31-36, doi:
725   10.1109/CBD.2016.016.

726   [37] Sinha A., and Jana, P. K., "A novel k-means based clustering algorithm for big data," International
727   Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp.
728   1875-1879.

729   [38] Kusuma, I., Ma'sum, M. A., Habibie, N., Jatmiko W. and Suhartanto, H., "Design of intelligent k-means
730   based on spark for big data clustering," International Workshop on Big Data and Information Security
731   (IWBIS), Jakarta, 2016, pp. 89-96, doi: 10.1109/IWBIS.2016.7872895.

[39] Backhoff, O. and Ntoutsi, E., "Scalable Online-Offline Stream Clustering in Apache Spark," IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, 2016, pp. 37-44, doi: 10.1109/ICDMW.2016.0014.Parallel Implementation of Density Peaks Clustering Algorithm Based on Spark

[40] Ding, D., Li, J., Wang H. and Liang, Z., "Student Behavior Clustering Method Based on Campus Big Data," 13th International Conference on Computational Intelligence and Security (CIS), Hong Kong, 2017, pp. 500-503, doi: 10.1109/CIS.2017.00116.

[41] Sarazin, T., Lebbah M. and Azzag, H., "Biclustering using Spark-MapReduce," IEEE International Conference on Big Data (Big Data), Washington, DC, 2014, pp. 58-60, doi: 10.1109/BigData.2014.7004493.

[42] Sharma, T., Shokeen, V., Mathur, D.,, Multiple k-Means++ Clustering of Satellite Image Using Hadoop MapReduce and Spark. Published in International Journal of Advanced Studies in Computer Science and Engineering, IJASCSE volume 5 issue 4, 2016

[43] Ben HajKacem, M. A., Ben N'Cir, C. E. and Essoussi, N., "KP-S: A Spark-Based Design of the K-Prototypes Clustering for Big Data," IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet, 2017, pp. 557-563, doi: 10.1109/AICCSA.2017.94.

[44] Ben HajKacem M.A., Ben N'cir CE., Essoussi N. Scalable Random Sampling K-Prototypes Using Spark. In: Ordonez C., Bellatreche L. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2018. Lecture Notes in Computer Science, vol 11031. Springer, Cham.

[45] Wu, J., Wu, Z., Cao, J., Liu, H., Chen G. and Zhang, Y., "Fuzzy Consensus Clustering With Applications on Big Data," in IEEE Transactions on Fuzzy Systems, vol. 25, no. 6, pp. 1430-1445, Dec. 2017, doi: 10.1109/TFUZZ.2017.2742463.

[46] Win, K.N., Chen, J., Chen, Y. Fournier-Viger, P.. PCPD: A Parallel Crime Pattern Discovery System for Large-Scale Spatiotemporal Data Based on Fuzzy Clustering. *Int. J. Fuzzy Syst.* 21**,** 1961–1974 (2019). https://doi.org/10.1007/s40815-019-00673-3.

[47] Gao, Z. Q. and Zhang, L. J., "DPHKMS: An efficient hybrid clustering preserving differential privacy in spark," in International Conference on Emerging Internetworking, Data & Web Technologies, 2017, pp.

[48] Zayani, A., Ben N'Cir, C. and Essoussi, N., "Parallel clustering method for non-disjoint partitioning of large-scale data based on spark framework," IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 1064-1069.

[49] Chitrakar A. S., and Petrovic, S., "Analyzing Digital Evidence Using Parallel k-means with Triangle Inequality on Spark," IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 3049-3058, doi: 10.1109/BigData.2018.8622430.

[50] Shah, J.,"New Clustering Using Spark", International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS) Volume V, Issue IX, 2016.

[51] Fatta, G. and Al Ghamdi, S., Efficient Clustering Techniques on Hadoop and Spark. International Journal of Big Data Intelligence. Vol. 6, 3-4, 269-290, 2019.

[52] Zhang, Y., Liu, H., Chen, T. and Tang, Di., A Distributed PCM Clustering Algorithm Based on Spark. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC '19). Association for Computing Machinery, New York, NY, USA, 70–74. DOI:https://doi.org/10.1145/3318299.3318315

[53] Win, K. N., Chen, J., Xiao, G., Chen Y. and Viger, P. F., "A Parallel Crime Activity Clustering Algorithm Based on Apache Spark Cloud Computing Platform," IEEE 21st International Conference on High Performance Computing and Communications; Zhangjiajie, China, 2019, pp. 68-74, doi: 10.1109/HPCC/SmartCity/DSS.2019.00025.

[54] Liu, B., He, S., He, D., Zhang Y. and Guizani, M., "A Spark-Based Parallel Fuzzy c -Means Segmentation Algorithm for Agricultural Image Big Data," in IEEE Access, vol. 7, pp. 42169-42180, 2019, doi: 10.1109/ACCESS.2019.2907573.

[55] Bharill, N., Tiwari, A.,and Malviya, A., Fuzzy Based Scalable Clustering Algorithms for Handling Big Data Using Apache Spark. IEEE Transactions on Big Data. 2016, PP. 1-1. 10.1109/TBDATA.2016.2622288.

[56] Solaimani, M., Iftekhar, M., Khan, L., Thuraisingham B. and Ingram, J. B., "Spark-based anomaly detection over multi-source VMware performance data in real-time," IEEE Symposium on Computational Intelligence in Cyber Security (CICS), Orlando, FL, 2014, pp. 1-8, doi: 10.1109/CICYBS.2014.7013369.

[57] Lavanya, K., Sairabanu, J., & Jain, P. Clustering of Zika virus epidemic using Gaussian mixture model in spark environment. *Biomedical Research-tokyo, vol 30*, pp. 127-133, 2019.

[58] Pang H., Deng L., Wang L., Fei M. The Application of Spark-Based Gaussian Mixture Model for Farm Environmental Data Analysis. In: Zhang L., Song X., Wu Y. (eds) Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems. AsiaSim 2016, SCS AutumnSim 2016. Communications in Computer and Information Science, vol 645. Springer, Singapore

[59] Mallios X., Vassalos V., Venetis T., Vlachou A. A Framework for Clustering and Classification of Big Data Using Spark. In: Debruyne C. et al. (eds) On the Move to Meaningful Internet Systems: OTM 2016 Conferences. OTM 2016. Lecture Notes in Computer Science, vol 10033. Springer, Cham

[60] Thakur S., Dharavath R. KMDT: A Hybrid Cluster Approach for Anomaly Detection Using Big Data. In: Satapathy S., Tavares J., Bhateja V., Mohanty J. (eds) Information and Decision Sciences. Advances in Intelligent Systems and Computing, vol 701. Springer, Singapore, 2018.

[61] Chakravorty, A., Rong, C., Evensen P. and Wlodarczyk, T. W., "A distributed gaussian-means clustering algorithm for forecasting domestic energy usage," International Conference on Smart Computing, Hong Kong, 2014, pp. 229-236, doi: 10.1109/SMARTCOMP.2014.7043863.

[62] Lighari, S. N. and Hussain, D. M. A., "Hybrid model of rule based and clustering analysis for big data security," 2017 First International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT), Karachi, 2017, pp. 1-5, doi: 10.1109/INTELLECT.2017.8277627.

[63] Kamaruddin S., Ravi V., Mayank P. Parallel Evolving Clustering Method for Big Data Analytics Using Apache Spark: Applications to Banking and Physics. In: Reddy P., Sureka A., Chakravarthy S., Bhalla S. (eds) Big Data Analytics. BDA 2017. Lecture Notes in Computer Science, vol 10721. Springer, Cham.

[64] Jin, C., Liu, R., Chen, Z., Hendrix, W., Agrawal A. and Choudhary, A., "A Scalable Hierarchical Clustering Algorithm Using Spark," 2015 IEEE First International Conference on Big Data Computing Service and Applications, Redwood City, CA, 2015, pp. 418-426, doi: 10.1109/BigDataService.2015.67.

[65] Guo, Y., Zhang J. and Zhang, Y., "An Algorithm for Analyzing the City Residents' Activity Information through Mobile Big Data Mining," IEEE Trustcom/BigDataSE/ISPA, Tianjin, 2016, pp. 2133-2138, doi: 10.1109/TrustCom.2016.0328.

[66] Ianni, M., Masciari, E., Mazzeo, G.M., Mezzanzanica, M., Zaniolo, C., Fast and effective Big Data exploration by clustering. Future Generation Computer Systems Volume 102, Pages 84-94, 2020.

[67] Solaimani, M., Iftekhar, M., Khan, L., Thuraisingham, B., and Ingram, J. B., Spark-based anomaly detection over multi-source VMware performance data in real-time. In 2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS) (pp. 1-8).

[68] Lee, S., and Kim, D. Distributed-based hierarchical clustering system for large-scale semiconductor wafers. In 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1528-1532).

[69] Hassani, M., Spaus, P., Cuzzocrea, A., and Seidl, T., I-hastream: density-based hierarchical clustering of big data streams and its application to big graph analytics tools. In 2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid) (pp. 656-665).

[70] Sarazin, T., Azzag, H., and Lebbah, M., SOM clustering using spark-mapreduce. In 2014 IEEE International Parallel & Distributed Processing Symposium Workshops (pp. 1727-1734).

[71] Malondkar, A., Corizzo, R., Kiringa, I., Ceci, M., & Japkowicz, N., Spark-GHSOM: Growing Hierarchical Self-Organizing Map for large scale mixed attribute datasets. Information Sciences, 496, 572-591, 2019.

828   [72] Hosseini, B. and Kiani, K.., A Robust Distributed Big Data Clustering-based on Adaptive Density
829   Partitioning using Apache Spark. Symmetry., 2018. 10. 342. 10.3390/sym10080342.
830   [73] Luo, G., Luo, X., Gooch, T. F., Tian L. and Qin, K., "A Parallel DBSCAN Algorithm Based on Spark,"
831   IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and
832   Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-
833   SocialCom-SustainCom), Atlanta, GA, 2016, pp. 548-553, doi: 10.1109/BDCloud-SocialCom-
834   SustainCom.2016.85.
835   [74] Rui, L., Xiaoge, L., Liping, D., Shuting, Z. and Mian, W., Parallel Implementation of Density Peaks
836   Clustering Algorithm Based on Spark, procedia Computer Science,Volume 107, 2017, Pages 442-447.
837   [75] Han, D., Agrawal, A., Liao, W. and Choudhary, A., "Parallel DBSCAN Algorithm Using a Data
838   Partitioning Strategy with Spark Implementation," 2018 IEEE International Conference on Big Data (Big
839   Data), Seattle, WA, USA, 2018, pp. 305-312, doi: 10.1109/BigData.2018.8622258.
840   [76] Han D., Agrawal A., Liao W., Choudhary A. A Fast DBSCAN Algorithm with Spark Implementation.
841   In: Roy S., Samui P., Deo R., Ntalampiras S. (eds) Big Data in Engineering Applications. Studies in
842   Big Data, vol 44. Springer, Singapore, 2018.
843   [77] Corizzo, R., Pio, G., Ceci, M. Malerba, D.. DENCAST: distributed density-based clustering for multi-target
844   regression. J Big Data **6,** 43 (2019). https://doi.org/10.1186/s40537-019-0207-2
845   [78] Hosseini, B. and Kourosh K., "A big data driven distributed density based hesitant fuzzy clustering using
846   Apache spark with application to gene expression microarray." *Eng. Appl. Artif. Intell.* 79 (2019): 100-113.
847   [79] Zhou Q., Wang J. SparkSCAN: A Structure Similarity Clustering Algorithm on Spark. In: Chen W. et
848   al. (eds) Big Data Technology and Applications. BDTA 2015. Communications in Computer and
849   Information Science, vol 590. Springer, Singapore
850   [80] Kim, J., Shin, M., Kim, J., Park, C., Lee, S., Woo, J., ... & Park, S. "CASS: A distributed network
851   clustering algorithm based on structure similarity for large-scale network." *PloS one* vol. 13,10 e0203670.
852   10 Oct. 2018, doi:10.1371/journal.pone.0203670
853   [81] Baralis, E., Garza P. and Pastor, E., "A Density-based Preprocessing Technique to Scale Out Clustering,"
854   2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 2078-2087, doi:
855   10.1109/BigData.2018.8621870.
856   [82] Lulli, A., Dell'Amico, M. and Ricci, L., NG-DBSCAN: scalable density-based clustering for arbitrary data.
857   Proceedings of the VLDB Endowment. Vol 10, pp. 157-168, 2016.
858   [83] Aryal, A. M. and Wang, S., "SparkSNN: A density-based clustering algorithm on spark," IEEE 3rd
859   International Conference on Big Data Analysis (ICBDA), Shanghai, 2018, pp. 433-437, doi:
860   10.1109/ICBDA.2018.8367722.

861   [84] Liang, M., Li, Q., Geng, Y., Wang, J. and Wei, Z., "REMOLD: An Efficient Model-Based Clustering
862   Algorithm for Large Datasets with Spark," 2017 IEEE 23rd International Conference on Parallel and
863   Distributed Systems (ICPADS), Shenzhen, 2017, pp. 376-383.
864   [85] Gong, Y., Sinnott, R.O., Rimba, P. RT-DBSCAN: Real-Time Parallel Clustering of Spatio-Temporal Data
865   Using Spark-Streaming. In: Shi Y. et al. (eds) Computational Science – ICCS 2018. ICCS 2018. Lecture
866   Notes in Computer Science, vol 10860. Springer, Cham.
867   [86] Sherar, M., and Zulkernine, F., "Particle swarm optimization for large-scale clustering on apache spark,"
868   2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, pp. 1-8, doi:
869   10.1109/SSCI.2017.8285208.
870   [87] Hasan, R.A., Alhayali, R.A., Zaki, N.D., & Ali, A.H. An adaptive clustering and classification algorithm
871   for Twitter data streaming in Apache Spark. *TELKOMNIKA Telecommunication Computing Electronics
872   and Control, vol 17*, pp. 3086-3099, 2019.
873   [88] Wang, Y. and Q. Qian, "A Spark-Based Artificial Bee Colony Algorithm for Large-Scale Data Clustering,"
874   IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th

875   International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems
876   (HPCC/SmartCity/DSS), Exeter, United Kingdom, 2018, pp. 1213-1218, doi:
877   10.1109/HPCC/SmartCity/DSS.2018.00204.

878 [89] Bonab, M.B., Hashim, S.Z.M., Alsaedi, A.K.Z., and Hashim, U.R. Modified k-means Combined with
879   Artificial Bee Colony Algorithm and Differential Evolution for Color Image Segmentation. In: Phon-
880   Amnuaisuk S., Au T. (eds) Computational Intelligence in Information Systems. Advances in Intelligent
881   Systems and Computing, vol 331, 2015, Springer, Cham

882 [90] Sarazin, T., Azzag, H. and Lebbah, M., "SOM Clustering Using Spark-MapReduce," 2014 IEEE
883   International Parallel & Distributed Processing Symposium Workshops, Phoenix, AZ, 2014, pp. 1727-
884   1734, doi: 10.1109/IPDPSW.2014.192.

885 [91] Sembiring, R. W. and Jasni M. Z., & Embong, A., Clustering High Dimensional Data Using Subspace and
886   Projected Clustering Algorithms. International Journal of Computer Science & Information Technology,
887   2010. 2. 10.5121/ijcsit.2010.2414.

Figure 1

Apache Spark Architecture

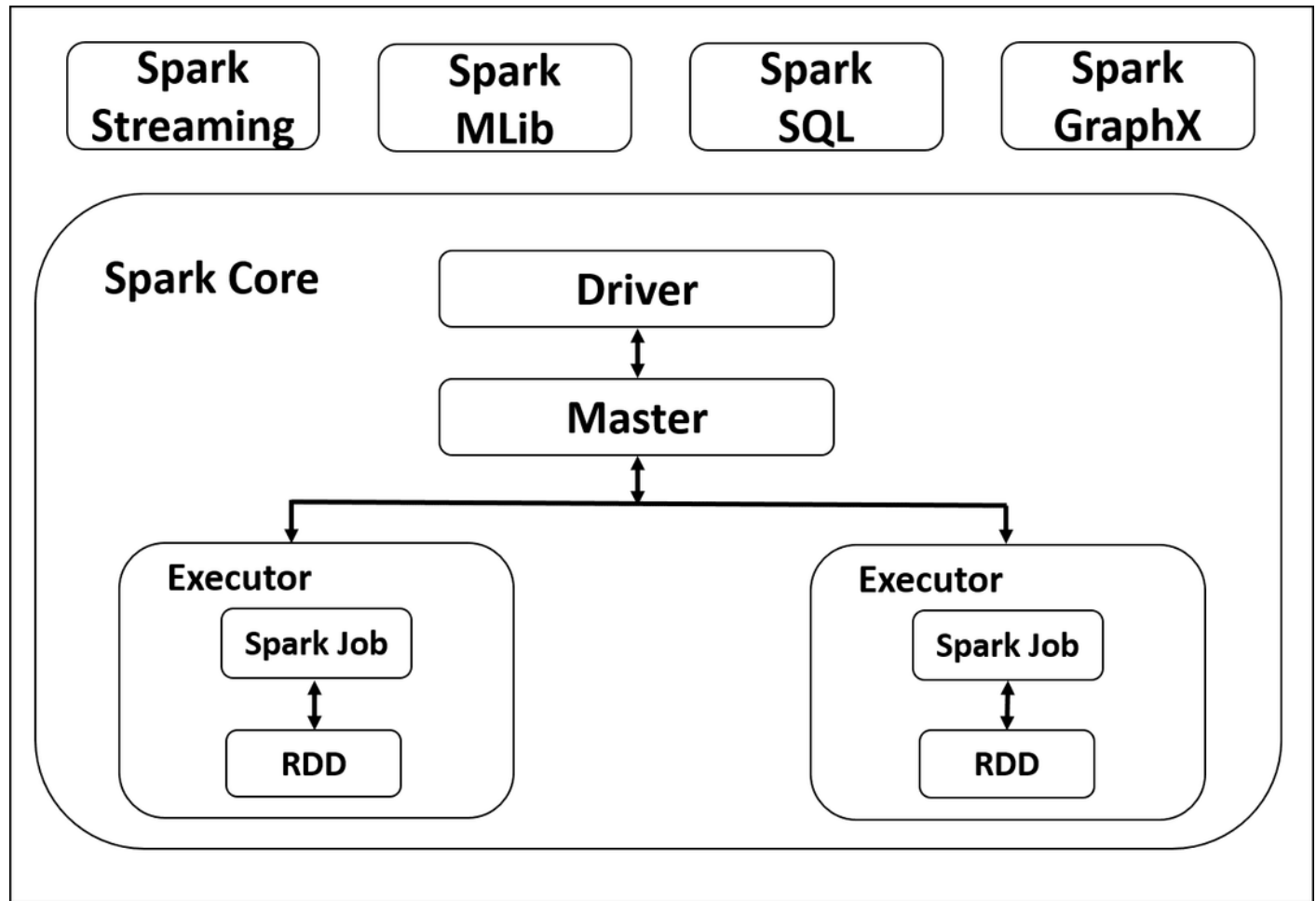



Figure 2

Flowchart for paper exclusion

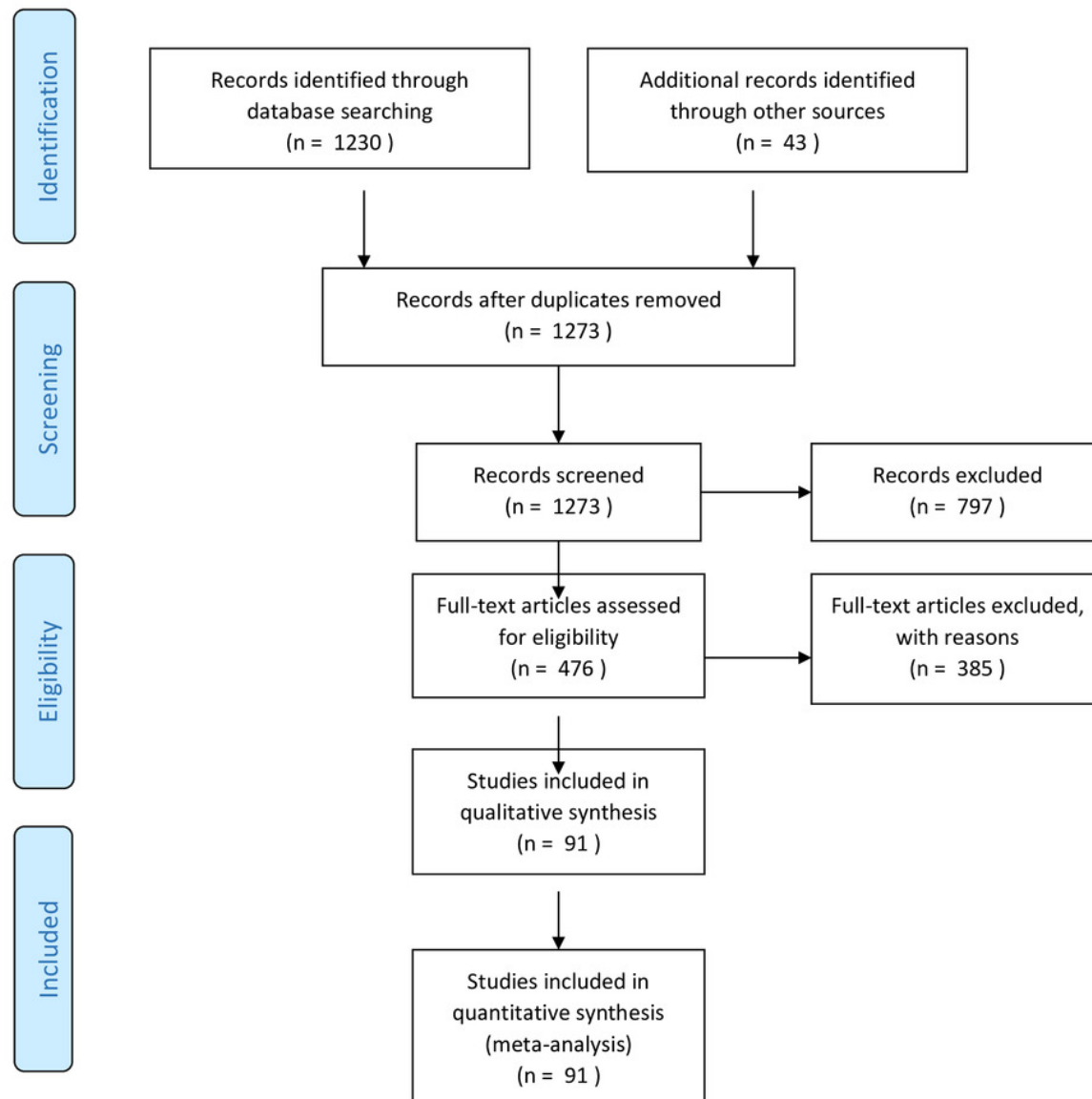
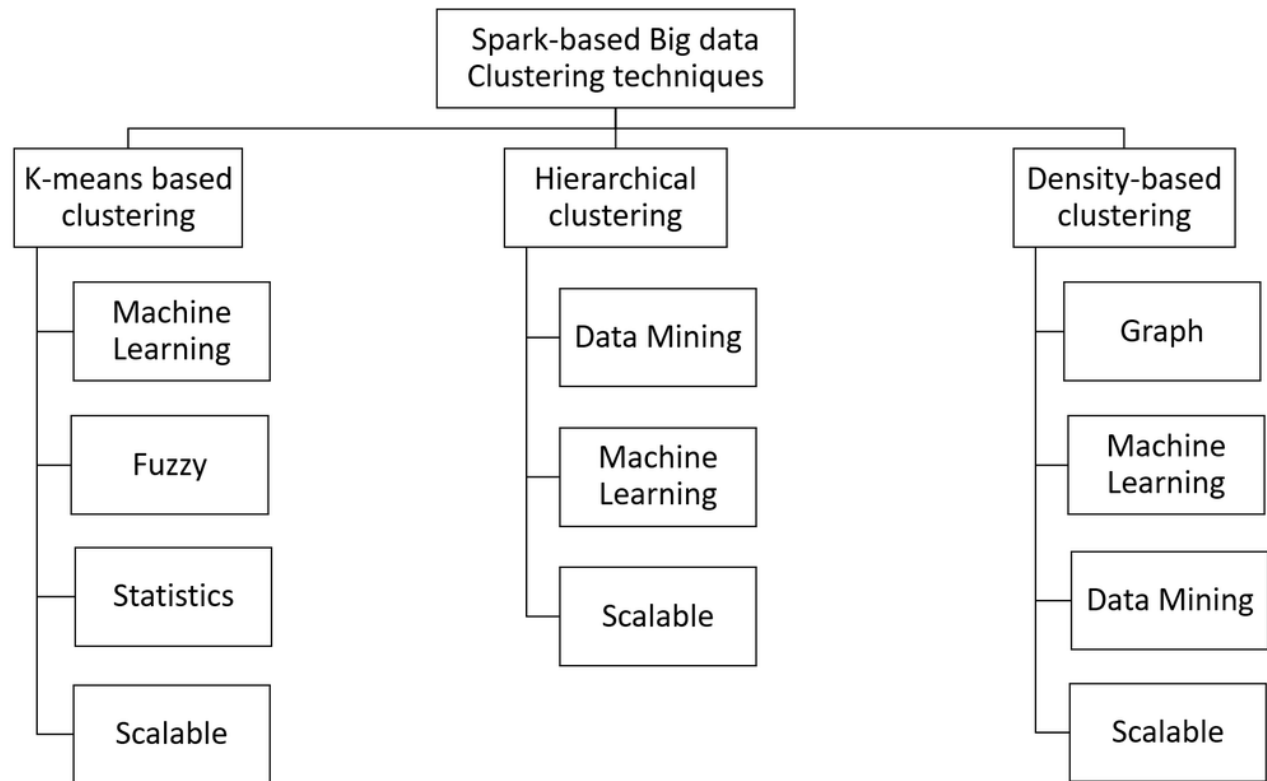



Figure 3

Taxonomy of Spark-based clustering methods

Comparison of Spark-based Clustering methods in terms of the supported Big Data characteristic (volume, variety and velocity) and in terms of the type of data (real and synthetic) the proposed method was validated.

3   Table 1: Comparison of Spark-based Clustering methods in terms of the supported Big Data

4   characteristic (volume, variety and velocity) and in terms of the type of data (real and synthetic) the

5   proposed method was validated.

Category	Sub-category	Paper	Supported Big Data Characteristic			Validated on	
			volume	variety	velocity	Real	Synthetic
K-means	Machine Learning	[38]	√		√		√
		[41]	√				√
		[47]	√			√	
		[60]	√				√
		[62]	√			√	
		[63]	√		√	√	√
	Fuzzy	[45]		√		√	
		[46]	√			√	
		[53]	√			√	
		[54]	√			√	
		[55]	√			√	
	Statistics	[50]	√			√	
		[57]	√			√	
		[58]	√			√	
		[61]	√	√	√	√	
	Scalable	[36]	√	√		√	
		[37]	√				√
		[39]		√	√	√	
		[40]	√			√	
		[42]	√			√	
		[43]	√	√		√	√
		[44]	√	√		√	√
		[48]	√			√	√
		[49]	√			√	
		[51]	√			√	
		[52]	√			√	
		[56]	√	√	√	√	
		[59]	√			√	
Hierarchical	Data Mining	[65]	√			√	
		[66]	√				√
		[68]	√				√
	Machine	[70]	√			√	√

	Learning	[71]	√	√	√	√	
	Scalable	[64]	√			√	
		[67]	√	√	√	√	
		[69]					√
Density	Graph	[74]	√			√	
		[79]	√			√	√
		[80]	√			√	
		[82]	√			√	√
	Data Mining	[76]	√				√
		[78]	√			√	√
		[83]	√			√	
	Machine Learning	[72]	√	√		√	
		[77]	√	√		√	√
		[84]	√			√	√
	Scalable	[73]	√			√	
		[75]	√			√	√
		[81]	√				√
		[85]			√	√	

Shows the data sources of the Spark-based clustering papers.

2   Table 2: Shows the data sources of the Spark-based clustering papers.

Data Source	Paper
IEEE Explorer	[2] [4] [5] [7] [9] [18] [19] [25] [27] [36] [37] [38] [39] [40] [41] [43] [45] [48] [49] [53] [54] [56] [61] [62] [64] [65] [67] [68] [69] [70] [73] [75] [83] [84] [86] [88] [90]
Elsevier	[1] [43] [74]
Springer	[3] [6] [20] [23] [24] [30] [32] [34] [35] [44] [46] [47] [58] [59] [60] [63] [76] [77] [79] [85] [89]
Google Scholar	[15] [42] [11] [12] [14] [26] [31] [57] [80] [82] [87]
ResearchGate	[8] [13] [16] [17] [72] [33] [50] [51] [90]
Science Direct	[10] [21] [22] [29] [52] [28] [66] [71] [74] [78]

Shows which papers in the survey were published in each of the last 6 years.

3   Table 3: shows which papers in the survey were published in each of the last 6 years.

Year of Publication	Papers	Number of Papers
2014	[36] [32][41] [56] [61] [67] [70] [90]	8
2015	[9] [8] [1] [22] [31] [64] [15] [89]	8
2016	[1] [2] [7] [6] [11] [16] [28] [36] [37] [38] [39] [42] [48] [50] [55] [58] [59] [65] [69] [79] [82]	21
2017	[6] [13] [17] [19] [21] [23] [40] [43] [45] [47] [62] [63] [74] [84] [86]	15
2018	[3] [4] [15] [18] [20] [24] [26] [27] [29] [30] [44] [49] [60] [68] [73] [75] [76] [80] [81] [83] [85] [88]	22
2019	[12] [46] [51] [52] [53] [54] [57] [72] [77] [78] [87] [91]	12
2020	[66]	1