Sayed *et al.*
Journal of Electrical Systems and Inf Technol (2023) 10:21
https://doi.org/10.1186/s43067-023-00090-5

Journal of Electrical Systems and Information Technology

REVIEW Open Access

A comparative study of big data use in Egyptian agriculture

Sayed A. Sayed[1*], Amira S. Mahmoud[1], Eslam Farg[1], Amany M. Mohamed[1], Ahmed M. Saleh[1], Mohamed A. E. AbdelRahman[1], Marwa Moustafa[1], Hisham M. AbdelSalam[2] and Sayed M. Arafat[1]

*Correspondence: sayed.ahmed@narss.sci.eg

[1] National Authority for Remote Sensing and Space Science (NARSS), Cairo, Egypt
[2] Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt

Abstract

The Egyptian economy relies heavily on the agricultural sector. As the population grows, arable land will diminish in the next decades. This makes food supply a priority. Big data could help the agriculture sector to address food security, especially in Egypt. In this paper, we examined the role of big data in agriculture in response to three questions: (1) What are the trend in peer-reviewed papers in the field of business development modeling and management? (2) What approaches were widely used especially in underdeveloped countries? (3) What is the current gap in terms of data sources, modeling, and analytic methods? As a result, 242 peer-reviewed articles have been studied. The contribution and findings of this study are summarized as. (1) A briefing on popular approaches which used frameworks was provided. (2) Publications based on the Internet of Things (IoT) in agriculture have increased dramatically by about 27%, 40%, and 44% in the years 2017, 2018, and 2019, respectively. (3) Around 37% of publications used Landsat and Sentinel-2 satellite images to build popular vegetation indices and land cover maps. (4) The challenges were identified as well as substantial opportunities that might serve as a roadmap for future growth. Therefore, by performing a comparative study in big data from this perspective, we explored the design principles using artificial intelligence and discussed a converged architecture to address the above-mentioned challenges.

Keywords: Agriculture, Big data, Systematic literature review (SLR), Geographical information systems (GIS), Internet of Things (IoT), Machine learning (ML)

Introduction

Several major issues including climate change, water storage, and crop variability threaten Egypt's agricultural sector [1]. Several factors, including soil erosion, water pollution, climate change, socio-cultural development, political laws, and market fluctuations, contribute to food insecurity [2]. Low soil fertility, pest illnesses, a lack of technological adaptation, and unpredictable weather are only a few of the obstacles that must be overcome to increase agricultural productivity [3]. Data are becoming not just valuable but also smart in the digital age [4]. During the middle of 2011, the term "Big data" (BD) was used to characterize the massive amounts of varied data that are difficult to manage and handle with traditional methods. Technically, the five main dimensions that characterize BD [5] are the massive amount of data, speed of data generation

and delivery, structured and unstructured multisources data, veracity, and value [6], as shown in Fig. 1. In [7], four use cases were discussed to relate BD five dimensions. Velocity and variety are key big data properties. Velocity is required in real-time decision-making. For example, a robot must select to pick a tomato. To make the proper decisions, robot sensors must process quickly. Combining data from different sources is defined as data diversity. Aquaculture monitoring which includes in-situ sensors, drone video, and feed management systems is an example of multimodal data. Volume is an important dimension in crop yield forecast use case. More space can be used for the forecast of organic soya production thanks to the processing of higher-resolution satellite images. Veracity is important for use cases with uncertain data, such as weather or multisource data.

The utilized tools, storage techniques, data processing mechanisms and data security techniques are collectively called as big data paradigm [8]. Big data paradigm enables researchers to analyze a vast amount of data into different modern practices. BD paradigm [9] involves four areas, namely methods, storage, processing, and representation. The approaches aim to uncover hidden trends and patterns within the multisource and large volumes of the acquired data. For both structured and unstructured data, the storage provides management systems and tools at a reasonable price.

The design and implementation of different cloud-based platforms exploit the processing capabilities to boost the overall performance. However, the main challenge lies in enhancing the data value and its accessibility for decision-makers. In Egypt, the recent governmental efforts have been directed to address the acquisition issues including cyber-infrastructure, automation and digitalization, data quality, integrity, reliability, and legal issues required in data collection, data access and usage and finally sharing and distribution of data. Several attempts for the private and public sectors had been focused on establishing BD cyber-infrastructures either individually or jointly. Recent research has concentrated on improving the government's ability to develop an action plan to promote the agriculture industry by integrating information and conducting predictive analysis. BD analytics and remote sensing (RS) technology open the door to improve agricultural sector productivity by extracting insights from the collected data to help farmers manage their farms and make on-farm management decisions.

Recently, a few papers had investigated the adaptation of BD technology in Egyptian framing. In [4], the authors introduced AgroSupportAnalytics, a cloud-based tool

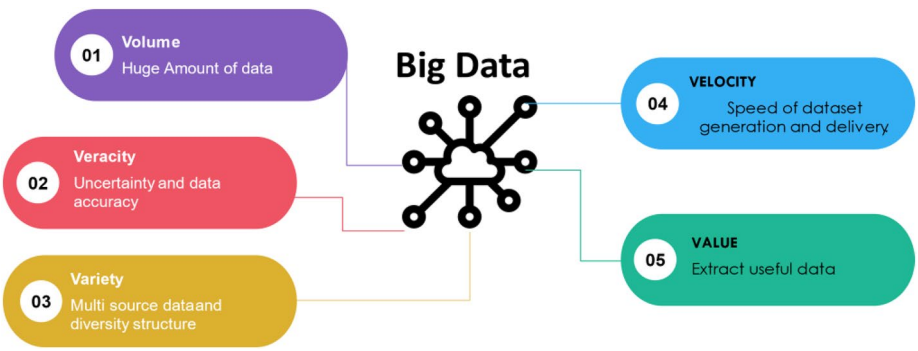

Fig. 1 The 5 main dimensions of big data after Fortune magazine

for managing complaints and making sound farming decisions in Egypt, to help and advice farmers and the discrepancies of the current manual method by agricultural specialists. Developed an automated complaint management and decision support method based on Egypt-specific requirement analysis. The solution uses knowledge discovery and analysis on agricultural data and farmers' concerns, deployed on a Cloud platform, to benefit Egyptian farmers. This article described the overall system architecture as well as the information and storage services based on the needs specification phases of the project and historical data sets of farmers' complaints and questions in Egypt. In [10], wireless sensor networks, their use in precision farming, and their relevance for Egypt's agriculture were discussed. By examining the use of a wireless sensor network in the cultivation of Egypt's potato crop, it was clear that the annual benefit from exporting the crop, after recovering the loss from its export prevention (estimated at 2 billion pounds, or the value of the potato export to Russia annually), after recovering the expected consequence of increasing the yield size and quality and after recovering the expected savings in the resource inputs, is greater than the cost of the system. The APTEEN protocol is the most ideal routing strategy for precision farming, and its network lifespan may reach 6.5 months, which is longer than the maximum potato crop lifetime of 120 days but shorter than the yearly cultivation duration of 6 months in Egypt. In [3], the authors presented a BD-analytics-based conceptual framework for Egyptian agriculture which could be applicable in monitoring, management, and forecasting.

In [11], a literature was introduced to assess smart farming to discover trends and opportunities using ProKnow-C methodology between 2015–2019 datasets and only 2401 articles were selected. Bibliometric analysis of the articles yielded a bibliographic portfolio of 39 works. The authors reach four main conclusions: (i) the necessity for universal information models to implement smart agriculture; (ii) the creation of standard IoT platforms for agriculture; (iii) the design of IoT devices with sophisticated encryption; and (v) predicting outcomes while considering relevant agricultural elements. In [7], the authors examined the circumstances for adopting big data technology in agriculture by analyzing twelve real-world use cases in precision agriculture and livestock. They employed a mixed method approach in Horizon 2020 project CYBELE, ranging from precision arable and animal farming to fisheries and food security, and a 56-person stakeholder survey. Large-scale deployments necessitate multidisciplinary methods and long-term project timeframes to solve big data concerns and avoid agricultural science compartmentalization. After studying use case challenges, solutions, and stakeholder viewpoints from all four angles, the authors conclude that big data solutions adoption is still small. These reviews indicate that even in developed countries the adaptation of BD in agriculture sector still in early stages. This research indicates the need to transfer technology and experience to the underdeveloped countries to integrate BD analytics in farming sector.

This paper introduces a systematic review of BD in agriculture applications to answer three questions. The first question highlights the rising tide of BD modeling and management publications. The second question manifests the trending topics while the last one identifies gaps present based on three key concepts, data source, modeling, and database. To better understand how to address the issues plaguing the Egyptian agricultural

sector, we also go over the structure of modern computing systems that store and process large amounts of data. Finally, challenges and further directions were highlighted.

The rest of this paper is organized as follow: Section II identifies Systematic Literature Review (SLR) methodology utilized in this study. Sections III briefly discusses non spatial and spatial big data frameworks. Section V discusses the major data sources incorporated in BD analytics. Section VI presents discussion. Finally, Section VII concludes the paper and provides future trends.

Systematic literature review (SLR) methodology

We selected the SLR as our research methodology. The key objective is to investigate and provide an extensive review of the existing BD analytics, applications, processing frameworks, and protocols related to the agriculture field. We followed the methodology in [12–14] to impartially select information and represent the results. The research methodology illustrated in Fig. 2 can be summarized into three phases: (1) review planning, (2) conducting the review, and (3) findings and reporting.

Research objectives

The main research objectives include:

O1: Defining cutting-edge research in BD field in agriculture.

O2: Characterization of the prevailing BD agriculture implementations, processing frameworks, as well as protocols.

O3: Identification of suggested taxonomy and supplementary highlights methods and approaches utilized in agriculture.

O4: Identification of gaps in research in the context of challenges and open issues.

Fig. 2 The adopted SLR methodology

Table 1 Research questions

No	Research question	Motivation
RQ1	What are the trend in peer-reviewed papers in the field of business development modeling and management?	Identify the recent development in BD frameworks utilizing ML and DL to boost smart agriculture. Define the challenges of applying modern technologies in Egypt
RQ2	What approaches were widely used especially in underdeveloped countries?	This question focuses on characterizing the current priorities in BD applications as well as the development throughout the past years
RQ3	What is the current gap in terms of data sources, modeling, and analytic methods?	Reporting the recent software, tools, technology, and data sources actively facilitate the shift to a smart agriculture environment to tackle the drawbacks of the Egyptian agriculture practice

Table 2 Search string

Sources	Search String	Context
IEEE Xplore, ScienceDirect, SpringerLink, and MDPI	("Big Data") AND ("Big data agricultural"), ("Egypt") or ("Egyptian") ("Big Data") AND ("Smart Farming")AND ("precision agriculture"), OR("Deep Learning"), ("Machine Learning") OR ("Agriculture supply-chain process")	Agriculture

Research questions

Defining the Research Questions (RQs) is the primary step in the SLR process. This research addressed three main questions with their respective motivation, as displayed in Table 1.

Search string

Next, we performed a pilot search on the basis of "BD in agriculture." Then, we performed a search on a wide range of keywords (Table 2) using multiple search engines and digital libraries. We chose SpringerLink, IEEE, MDPI, and ScienceDirect digital libraries due to their related scientific content to the paper objectives. Finally, we set up appropriate technical and scientific procedures to search in the aforementioned digital libraries.

Screening of relevant papers

We screened the papers to remove the irrelevant papers to the research questions. However, another detailed assessment to indicate the actual relevancy was carried out. First, we selected papers based only on the titles and excluded other papers irrelevant to the research questions. Next, a second screening round via reading each abstract was carried out to shortlist papers based on the following parameters.

Articles novelty.
Papers published other than conferences, journals.
Articles defining unclear data sources or data collection procedures.
Papers published between 2010 and 2020.

Quality assessment

We carried out a Quality Assessment (QA) questionnaire to evaluate each of the selected articles' quality quantitatively. In this SLR, a simple questionnaire based on [12] was designed to investigate the quantitative value of the selected papers' quality.

SLR QA questionnaire	Grading
(a) The paper introduces a novel idea to BD in agriculture	Yes (1) and No (0)
(b) The paper exemplifies a clear data sources and processing chain in the field of agriculture	"Yes (1)", "partially (0.5)", and "No (0)"
(c) Number of citations. Please state the number of citations	-No recorded citation (− 1), -citation count between 1 and 5 (0), -citation count is over 5 (+ 1)
(d) Is the paper published in	Quartile journal ranking Q1, Q2 (0.1), Q3, Q4(0.5)
(e) Is the paper answer the main research question	Yes (+ 1), No (0), please state the Question No.)

Selection process

Table 3 provides a quick summary of how many papers were chosen after a thorough search and selection procedure. Initially, 11,596 papers were chosen by the search protocol on the selected databases. The authors were split into two teams to conduct the above screening to exclude non-related articles based on titles, abstracts, keywords, and full articles. Next, duplicate elimination reduced the total relevant paper. In the end, the authors only select 242 out of 11,596 based on their abstracts and full paper scanning.

BD frameworks

Many studies compare the popular BD frameworks [15–18]. In this context we briefly discussed traditional DB, spatial BD frameworks and major BD sources.

Non-spatial BD frameworks

Several studies compare BD frameworks [3, 19] whose categorized into batch, stream, and hybrid [20] based on the data they handle. Unfortunately, the factors considered in comparing frameworks from the same category were hardly mentioned. An extensive comparison was conducted using the same approach [20] and considering the following aspects: computing cluster architecture, data flow, data processing model, scalability, fault tolerance, back-pressure mechanism, latency, and programming languages, beside ML-related libraries.

Table 3 Total numbers of papers through the selection process

Process	Selection criteria	IEEE Xplore	Springer	ScienceDirect	MDPI	Total
Search	Keywords	3145	779	5000	2672	11,596
Screening	Duplicate removal	237	198	381	567	1383
Screening	Title	235	181	261	208	885
Screening	Abstract	201	60	58	103	422
Inspection	Full paper	162	20	48	12	242

Batch BD frameworks

To process the data in a batch, it required to accumulate for several hours or even days. In order to process the information, the data had to be loaded into memory. Otherwise, it would have been kept in a different location, such as a database or a file system. Hadoop MapReduce and Spark are instances of batch BD frameworks for massive datasets. Informatica and Alteryx are two popular data-analysis tools for organizations of all sizes. Database management systems like Amazon Redshift and Google BigQuery are used for relational data.

Google unveiled the Hadoop framework [21], which includes the Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), and MapReduce. HDFS is the heart of Hadoop, and it's what makes Hadoop so useful for dependable data storage. NameNode and DataNode are HDFS's two designs, and YARN is the cluster management part of the Hadoop framework.

In conclusion, the MapReduce component has two primary operations: mapping and reducing. Users need just define map and reduce functions, with the framework handling administrative tasks including parallelization and failover. Hadoop MapReduce, in general, uses HDFS for data storage and YARN for managing resources and scheduling jobs. The overall structure of the Hadoop components [22] is displayed in Fig. 3.

Stream BD frameworks

Via stream frameworks, the data is processed in real time or in micro-batches [23]. A bunch of popular BD stream frameworks include Apache Storm and Apache Samza [20].

Apache storm

Twitter created Apache Storm to process huge, real-time structured and unstructured data [24–26]. A typical Apache storm topology [27] (Fig. 4) uses a directed acyclic network where edges represent data interchange and nodes represent computation

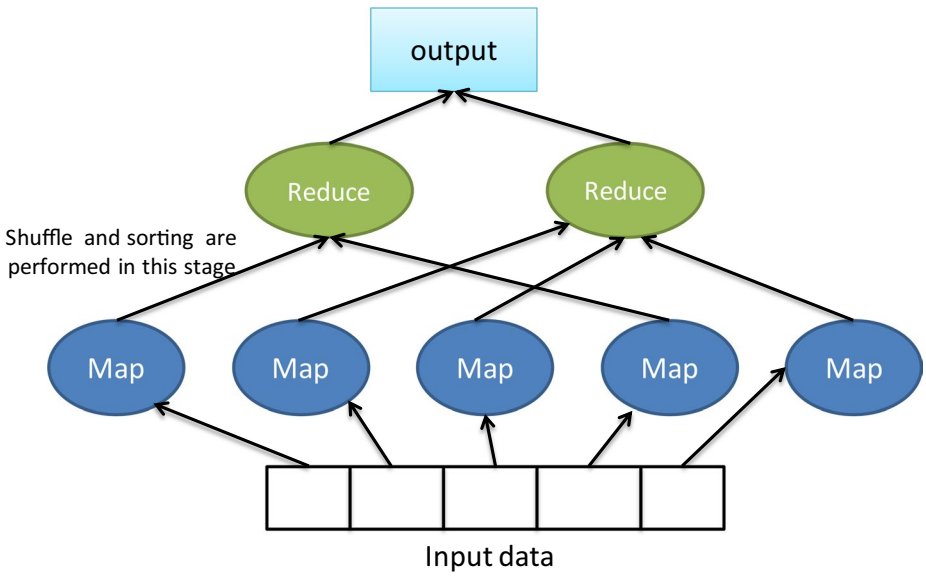

Fig. 3 Hadoop MapReduce overall structure [22]

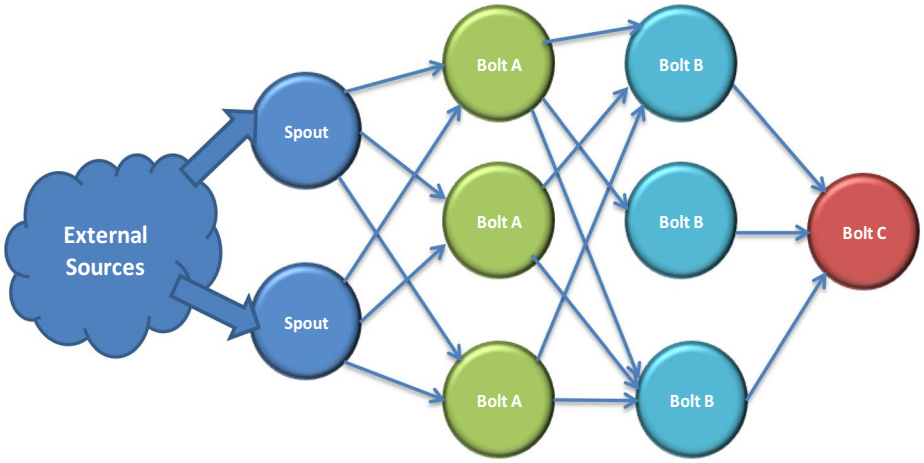

Fig. 4 Apache Storm Topology [27]

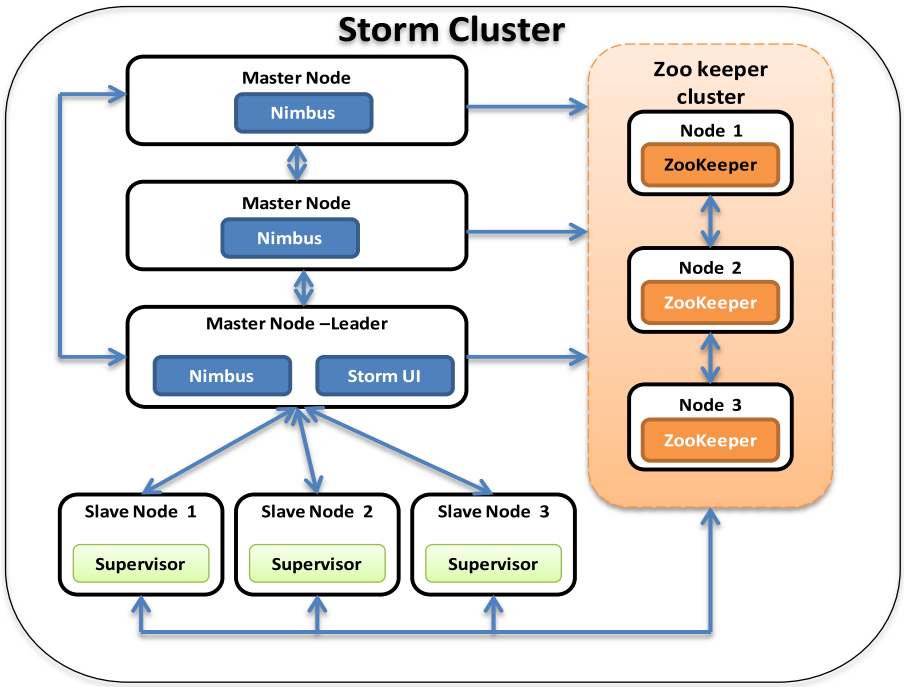

Fig. 5 Apache Storm Architecture [28]

resources. Nodes are master "Nimbus" or worker "Supervisor." Every node accepted streams (sequence of Tuples). First nodes only take Spouts, which can convert external messages to tuples and resend them without calculation. Bolts filter, calculate, join, and produce tuples. Stream grouping defines the bolt-to-spout protocol.

Figure 5 shows the Storm architecture [28] which consists of Nimbus, Supervisor, and ZooKeeper. Nimbus monitors worker and slave node progress and assigns tasks. Supervisor is a stateless daemon that monitors and restores topologies [28]. ZooKeeper manages configuration, synchronization, and group membership. Topology uses Trident

APIs, which offer high-level operators. Trident APIs Trident APIs divide work into micro-batches. Controlling throughput and delay with batch size. As directed acyclic graphs (DAGs), their topologies cannot run iterative algorithms [29].

Apache Samza

Apache Samza was created by LinkedIn to solve problems in stream processing, such as scalability, resource allocation, etc. [30]. Samza is built on Kafka and YARN [20, 31]. Figure 6 shows Apache Kafka's five primary components. Producer, Topics, Consumer, Partitions, Brokers. Producer writes Kafka topics. Topic describes every Kafka data stream. A consumer can read Kafka topics and must retain its failure offset. Brokers are Kafka's single nodes.

Hybrid BD frameworks

Batch and stream processing frameworks are needed for some applications. As a result, the utilization of hybrid processing frameworks is essential in these circumstances. Some of the most prominent examples include Apache Spark and Apache Flink.

Apache spark

Apache Spark is a Hadoop-based hybrid framework that boosts batch processing with fully in-memory computation [20]. The two cases for which the Apache Spark restricted storage layer is relevant are data loading into memory and result storage. Spark caches intermediate results, unlike Apache MapReduce. Apache Spark's central data structure, Resilient Distributed Datasets (RDDs), enables developers to reuse intermediate data. RDDs can optimize partitions and maintain stored data [17].

Apache Spark framework [33] comprises numerous core components and upper-level libraries, such as Spark's MLlib for machine learning [34], GraphX [35] for stream processing, and Spark SQL for stream processing and structured data processing [36].

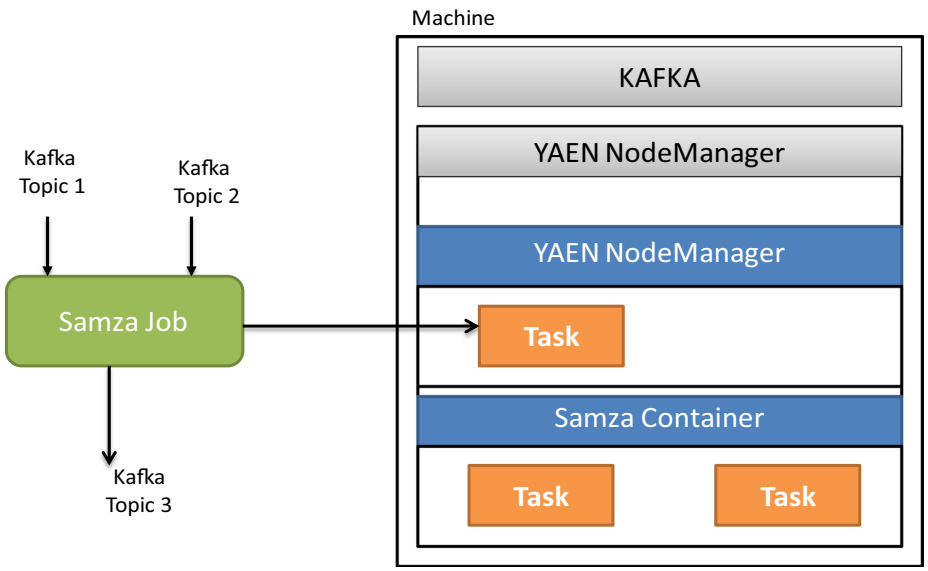

Fig. 6 Apache Samza Architecture [32]

Figure 7 depicts Apache Spark stack [37]. Scala-based Spark enables multiclusters. Spark supports Scala, Java, Python, and R and data visualization and analysis methods. Cluster managers request job-execution cluster resources. Spark's built-in cluster manager uses Hadoop YARN, Apache Mesos, and Amazon EC2. Spark supports HDFS, Cassandra, HBase, Hive, Alluxio, and other data sources.

Apache Flink

Real-time analytics, continuous data pipelines, batch processing, and iterative algorithms are all supported by Apache Flink [38], an open-source hybrid framework. The key benefit is great fault tolerance and low latency costs while processing massive amounts of data in a distributed setting. For limited data sets, the DataSet API is a common method of processing data in batches [38]. Figure 8 illustrates Apache Flink Ecosystem.

Spatial BD frameworks

GIS supports various activities for government especially in active sector like agriculture. Many BD processing frameworks, such as Hadoop MapReduce, were invented to process and analyze huge GIS data in order to extract geographic information for specific geographic operations such as distance-based queries, k-nearest neighbor (KNN) searches, filter-based queries, etc.

Hadoop-based

This section discusses the two most prevalent Hadoop MapReduce dependent on GIS data processing frameworks: Hadoop-GIS as well as Spatial-Hadoop.

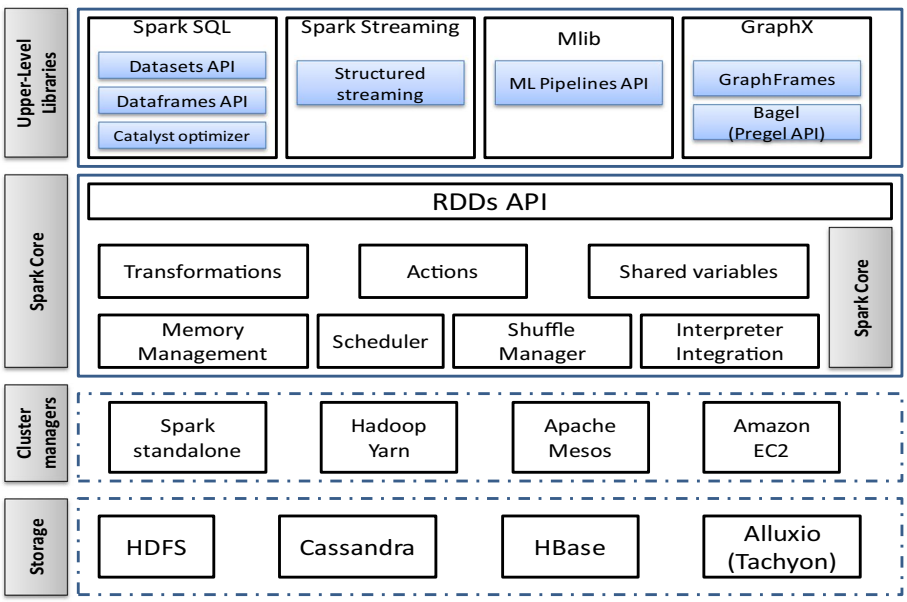

Fig. 7 The Apache Spark stack architecture [37]

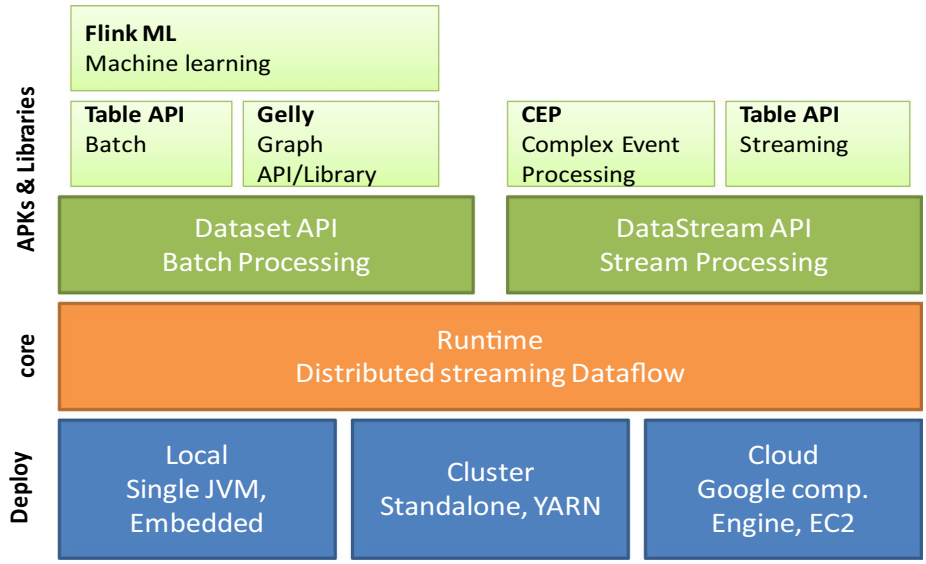

Fig. 8 The Apache Flink Ecosystem [38]

Hadoop-GIS

Hadoop-GIS is a MapReduce framework for handling massive amounts of vector data, partitioning, and geographic queries [39]. There is a wide variety of geographical (spatial) questions, including: distance-based queries; relationship-based queries; descriptive queries; and distance-based spatial mining and statistics queries, such as spatial clustering and spatial regression [35]. Hadoop-GIS uses SATO spatial partitioning and local spatial indexing to improve query speed. However, complicated geometries are not allowed. This includes things like convex/concave polygons, line strings, multipoint geometries, and multipolygon geometries. Hadoop-GIS, in fact, only works with two-dimensional data and provides support for two types of queries over geometric objects: box range and spatial joins.

Spatial-Hadoop

To address the shortcomings of Hadoop-GIS, the Spatial-Hadoop MapReduce architecture was developed. It includes SpatialRecordReader and SpatialFileSplitter, two new components for processing spatial data efficiently and scalable with spatial data, geographic indexes, and operations [40].

Points, multipoints, line strings, and polygons are just some of the geometry types that can be used with Spatial-Hadoop. Uniform grids, R-Trees, Quad-Trees, KD-Trees, and Hilbert curves are some of the spatial partitioning techniques used in spatial indexes [41]. It also allows for numerous predefined spatial operations like range queries, k-nearest neighbor queries, and spatial joins. In addition to skylines, convex hulls can also be generated from the many geometric objects it supports, such as segments and polygons. Spatial-Hadoop, a distributed platform for geospatial data analytics, is what makes the aforementioned features a reality.

Spark-based

This section introduced two of the most widespread Spark-based GIS data processing frameworks: Spatial-Spark and Geo-Spark.

Spatial-spark

In order to process geographic information system (GIS) data, the framework known as "Spatial-Spark" was developed. It was built atop Spark RDD to supply a wide variety of spatial operations like range query, spatial join, spatial filtering, R-Tree index, and R-Tree partitioning to speed up queries [42]. To handle both broadcast spatial join and partitioned spatial join, Spatial-Spark can be thought of as an in-memory BD framework [42].

Geo-spark

To analyze massive amounts of GIS data more quickly than Spatial-Hadoop, an in-memory cluster computing framework called Geo-Spark has been developed on top of Spark [43]. To better accommodate spatial data types, indexes, and geometric operations at scale, Geo-Spark broadens the idea of RDDs and SparkSQL. It's useful for k-nearest neighbor (KNN) queries and other geographic data partitioning systems like a uniform grid, R-Tree, Quad-Tree, or KDB-Tree. To find a happy medium in a cluster between execution time and memory/processor consumption, Geo-Spark is calibrated to pick an appropriate join algorithm [44]. By combining operationally fast programming languages (like Java and Scala) with declarative (like SQL) languages and spatial RDD APIs, Apache Spark's Geo-Spark enables developers to create effective spatial analysis applications. Finally, Tables 4 and 5 draw a comparison among the discussed BD processing frameworks either Non-Spatial or Spatial DB Framework, including different metrics.

Common data sources

To the best of our knowledge, satellite imagery, Wireless Sensor Web (WSW) and Internet of Things (IoT), crowdsourcing, Social Media records, GPS traces and mobile call detail record, simulation, Unmanned Aerial Vehicle (UAV) video, Airborne and Terrestrial Light Detection and Ranging (LiDAR), and Geographic Information System (GIS) are all common BD sources in agricultural applications (GIS).

Table 4 Comparison among popular Non-spatial DB Framework

Features	Hadoop	Spark
Processing type	Batch	Hybrid
Computing cluster architecture	YARN	YARN and Mesos
Data Flow	MapReduce data flow	A queue of RDDs called DStream processed one at-a-time using microbatching cluster
Data Processing Model	MapReduce	exactly-once
Fault Tolerance	Yes	Yes (using lineage)
Latency	low	High
Scalability	Yes	Yes (user demand)
Back-pressure Mechanism	No	Yes
Programming Languages	Java mostly	API for Scala, Java, Python, and R
Support for Machine Learning	Yes	Yes (Spark MLlib)

Table 5 Comparison among popular Spatial DB Framework

Features	Hadoop-GIS	Spatial-Hadoop	Spatial-Spark	Geo-Spark
DataFrame API	No	No	No	Yes
In-memory processing	No	No	Yes	Yes
Spatial Partitioning	SATO	Multiple	Multiple	Multiple
Spatial Indexing	R-Tree	R-/Quad-Tree	R-Tree	R-/Quad-Tree
KNN query	Yes	Yes	No	Yes
Query optimizer	No	No	No	Yes
Distance query	Yes	Yes	Yes	Yes
Distance join	Yes	Yes	Yes	Yes
Filter (Contains)	Yes	Yes	Yes	Yes
Filter (ContainedBy)	Yes	Yes	Yes	No
Filter (Intersects)	Yes	Yes	Yes	Yes
Filter (WithinDistance)	Yes	Yes	Yes	No

Satellite imagery

To investigate the planet's surface, satellites use either active or passive sensors to gather imagery of it [23]. Images captured by passive sensors are used to calculate how much sunlight is reflected from Earth's surface. In contrast, active sensors are typically used to acquire the images. Active sensors, such as the Synthetic Aperture Radar (SAR), are effectively used to address the shortcomings of passive sensors and expand the observational capacity for agricultural applications when there is thick cloud cover, rain, or when it is nighttime.

Wireless sensor web and IoT

In a Wireless Sensor Network (WSN), a wide variety of high-tech sensors, such as those that measure temperature, humidity, wind speed and direction, etc., are networked together. For better identification and visualization across various agricultural regions, WSN relies on Internet of Things (IoT) technology, which integrates and deploys a number of heterogeneous geographically distributed sensors [45]. When conventional lines of communication breakdown, the gathered data [46] may help farmers, specialists, and investors keep a tighter rein on day-to-day operations. Although WSNs have many applications in smart farming, they still lack the "Socio-techno-economic viewpoint" required for full coordination between the various data sources and protocol implementations [47].

Crowd-sourcing and social media

A number of tools have emerged in recent years to facilitate the gathering of information from the general public. Crowdsourcing is an example of an active platform, where contributors are aware of the data collecting [48], while social media is an example of a passive platform, where contributors are unaware [49–51]. For pest monitoring and information exchange, social media has largely replaced crowdsourcing systems [52].

Data collection [52], information extraction, analytical workflow, geo-location pattern/image/text analytics, and information sharing via social media services are all used in

agricultural growth [44]. Real-time analytics based on social media platforms [53] offer a lot of opportunities for automatic monitoring and detection of plant diseases, crop yields, and predictions [54]. Simplifying spatiotemporal analysis and generating a spatial-based choice for supporting environment, visual analytics can aid small farmers in meeting consumer demand using social media data. Text messages are important, but the movies and pictures that users upload are what really make social media what it is. Analyses that rely on images and videos, as well as visual analytics, mine social media posts for relevant data [55].

Mobile call detail record (CDRs) and GPS traces

When it comes to managing natural disasters like landslide monitoring, tsunami monitoring, earthquake management, forest fires, and floods, GPS traces and mobile CDRs data are essential resources. The data from GPS logs have proven useful in a variety of farming applications [56], such as determining patterns of agricultural machinery's mobility and tracking fuel use.

Simulation

One of the most important agricultural contributions to meteorological phenomena, land surface phenomena, and other types of pollution is numerical modeling, also known as forecasting [57, 58]. Water spray [59] and subsurface pipe parameter [60] estimates have also benefited from mechanistic modeling.

Management in agriculture can be improved with the use of various modeling and simulation techniques. As a result of the need to better comprehend the physical, chemical, and biological control parameters in crop and animal production systems, several mechanistic models were developed to enhance the scientific understanding of agriculture. The second set of simulation models was designed to aid in planning and decision-making.

UAVS, Drones, and LiDAR

Drones and other Unmanned Aerial Vehicles (UAVs) can provide high-resolution imagery useful for a wide range of agricultural applications [61] including livestock monitoring, crop production, yield prediction, fertilizer and pesticide spraying, and soil mapping [62]. A UAV or drone can be outfitted with a wide variety of sensors, including cameras, LiDAR, and even weather detectors. Many applications, such as pesticide spraying by drone, plant phenotyping, and yield production estimation, can benefit from incorporating the collected sensor data into real-time decision making [63].

LiDAR technology [64] allows for the generation of accurate topography maps and Digital Elevation Models (DEMs), both of which are used extensively in analyses of crop architectural factors, forests, and other agricultural settings. Yield prediction and monitoring, soil type identification, soil erosion estimation and prevention, land parceling, and crop analysis field management are all areas where LiDAR has proven useful [65]. The geospatial community places a high value on LiDAR technology due to the vast amounts of data it generates [66].

Vector-based GIS data

Geographical Information System (GIS) comprises computer hardware, software, in addition to various methods. Its utilization lies in collecting, managing, processing, analyzing, modeling, and displaying spatial data for solving multifaceted management and planning issues. Vector-based GIS data is a powerful addition to agriculture control systems [67] like farmland suitability analysis, figuring out how much fertilizer to use, and figuring out how much pesticide to use. Critical facility geospatial analysis (hospitals, schools, fire stations, etc.) [68], human impact assessment (based on age, gender, socioeconomic status, etc.), resource inventory (vehicles, supplies, equipment, etc.), and infrastructure assessment (location of buildings, roads, and utilities) (utility grids and transport networks) assist and strengthen the agricultural community [67].

To determine whether or not land is suitable for irrigation with reclaimed water, in [67] the authors created a methodology that combined multicriteria decision analysis with geographical information data (GIS-MCDA). With the use of Geographic Information Systems (GIS), in [69] the authors investigated the ideal soil-site characteristics for citrus to maximize yield. Satellite images, aerial photography and video from UAVs, Wireless Sensor Web and IoT, simulation, crowdsourcing, social media records, GPS traces and CDRs, LiDAR, and GIS data are all examples of BD sources mentioned above.

Discussion

To indicate the main trend among the 242 papers that have been analyzed, Fig. 9 demonstrates a consistently increasing trend in BD's popularity in the agriculture setting since 2014. Following the assessment of publications in the last few years, it indicated that those BD frameworks chosen in the early years were based on Hadoop and MapReduce. An increasing trend in the use and integration of Hadoop began around 2018. The papers reviewed in this work were published in 34 different journals over the four digital libraries indicating a broad paradigm of disciplines using BD engineering in agriculture and food security studies. Out of these, 23 journals published only one paper for BD with various agriculture practice applications. Journals such as Precision Agriculture, Computers and Electronics in Agriculture, and IEEE Access are where the most up-to-date and reliable information on biotechnology (BD usefulness) in the agricultural sector can be found.

According to the three RQs, the distribution of the 242 peer-reviewed studies is illustrated in Fig. 10 Overall, one can report that 46 articles were published

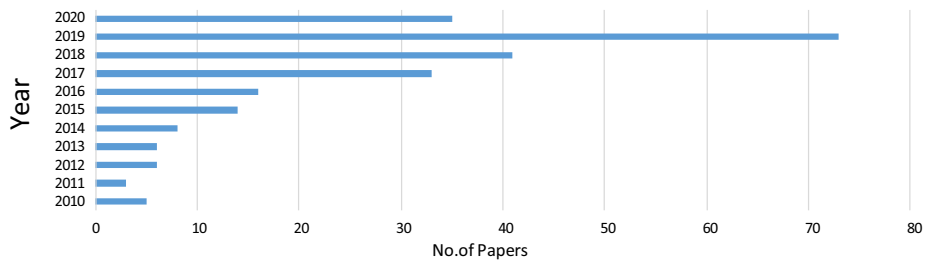

Fig. 9 The trend of selected published articles in agriculture between 2010 and 2020

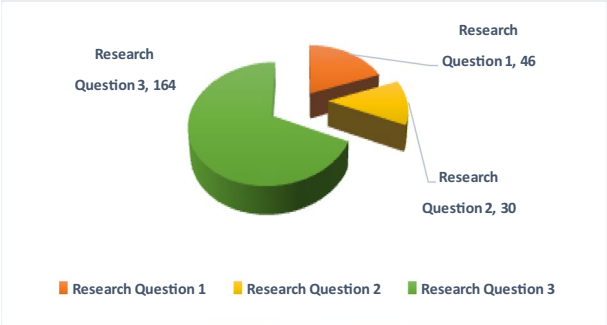

Fig. 10 The distribution of researches according to the three questions in last ten years

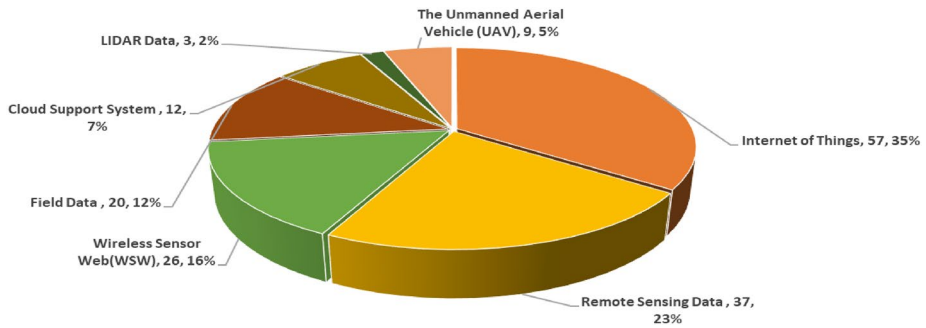

Fig. 11 The number of papers per technology adopted

motivated by RQ1. Only about 30 studies applied machine or deep learning in agriculture practice. Finally, 164 articles focused on new and trendy technologies in precision agriculture.

Figure 11 demonstrates the number of published papers per technology adopted, and it indicates that the largest number of papers utilized Internet of Things (IoT) (57). This was followed by a considerable number of studies adopted the RS data, including satellite images, optical or radar datasets, and photogrammetry (37). Studies focused on wireless sensors, collected field data, and Cloud Support System was 26, 20, and 12, respectively. Other data sources include The Unmanned Aerial Vehicle (UAV) (9) and Lidar data (3), respectively. Multidata sources were investigated in different studies in this context.

Figure 12 indicates the publication trends over the last ten years for each of the adopted technologies in 242 articles. It can be noted that there is a rapid shift in trend with the popularity of BD since 2014. In the light of research published over the past few years, it was clear that early researchers relied on more than just satellite data. After Sentinel-1 becomes widely accessible in 2017, a rising trend is observed toward combining radar and optical data. Over the past three years, we've seen a change in how widely technology like UAVs, Lidar, and wireless sensors can be used. Although cloud computing shows much promise, just a fraction of the data used in existing studies comes from the cloud.

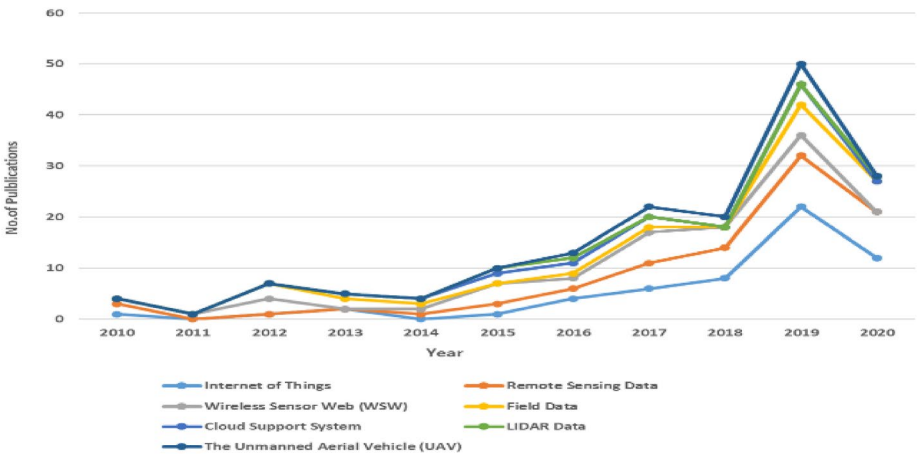

Fig. 12 The number of published studies utilizing technologies in precision agriculture per year

Conclusion

New findings in BD agriculture and farming were analyzed in this study. While BD analytics has the potential to aid the Egyptian agriculture sector in resolving a number of difficulties, doing so will necessitate a significant financial commitment. Egypt's farmers couldn't solve its food crisis without access to cutting-edge technology. This article summarizes the findings from a systematic assessment of 242 papers on BD in agriculture, demonstrating the relevance of this research to the field's challenges. As a result, there are numerous inferences that might be made:

Scientists are able to find solutions to farming issues with the aid of BD, free satellite imageries, massive computing capacity, and efficient machine learning approaches. In particular, the IoT (57), followed by RS Data (37) and WSW (26) were the top-three BD sources studies in the past few years.

Since 2017, there has been a dramatic growth in the number of farming operations that make use of the Internet of Things. Specifically, 2017 had seen the publication of 27% of the evaluated papers, 2018 saw 40%, and 2019 will see 44%. The rapid growth of Internet of Things used in farming is evidence of its usefulness and widespread adoption.

BD expanded on a wider variety of applications, including in the agricultural sector; the articles under examination appeared in 34 scholarly journals covering a wide range of disciplines.

37% of the articles cited the use of satellite imagery, most notably Landsat and Sentinel-2, to create several popular vegetation indices and land cover maps.

Numerous research used various machine learning techniques to handle RS data. Several research in the past few years have used deep learning, particularly in the fields of crop mapping and pest and disease identification.

Looking ahead, we aim to conduct additional in-depth research into the privacy, security, and consistency issues facing farm data suppliers. Recent government initiatives have highlighted the importance of investing in cyber-infrastructure and cloud-based

computing. Future research must pay close attention to the issue of privacy because of its relevance to the field of agriculture.

Acknowledgements
Not applicable.

Author contributions
MM. and SA.S contributed to conceptualization; SA.S, AM.M, and AS.M provided methodology; EF, AM.S, and MA.E. AR performed validation; MM and AS.M carried out formal analysis; MSM and AM.M performed writing—original draft preparation; SA, HM.AS, and SM.A performed writing—review and editing; MM and AS.M performed visualization. All authors read and approved the final manuscript.

Funding
This paper is based upon work supported by Science, Technology & Innovation Funding Authority (STDF) under grant (ESIP 2019) project ID (33547).

Availability of data and materials
The authors confirm that the data supporting the findings of this study are available within the article.

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 23 November 2022 Accepted: 23 March 2023
Published online: 04 April 2023

References

1. Gado TA, El-Agha DE (2021) Climate change impacts on water balance in Egypt and opportunities for adaptations. In: Agro-environmental sustainability in MENA regions. 2021, Springer, pp 13–47
2. Bank W (2020) World Development Indicators
3. Sayed A et al (2022) A conceptual framework for using big data in Egyptian agriculture. Int J Adv Comput Sci Appl 13(3):148
4. Munir K et al (2022) AgroSupportAnalytics: a cloud-based complaints management and decision support system for sustainable farming in Egypt. Egyp Inf J 23(1):73–82
5. Emani CK, Cullot N, Nicolle C (2015) Understandable big data: a survey. Comput Sci Rev 17:70–81
6. Kamilaris A, Kartakoullis A, Prenafeta-Boldú FX (2017) A review on the practice of big data analysis in agriculture. Comput Electron Agric 143:23–37
7. Osinga SA et al (2022) Big data in agriculture: between opportunity and solution. Agric Syst 195:103298
8. Chen CP, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf Sci 275:314–347
9. Villars RL, Olofson CW, Eastwood M (2011) Big data: what it is and why you should care. White Pap IDC 14:1–14
10. Abd El-kader SM, El-Basioni BMMJEIJ (2013) Precision farming solution in Egypt using the wireless sensor. Netw Technol 14(3):221–233
11. Iaksch J, Fernandes E, Borsato M (2021) Digitalization and big data in smart farming–a review. J Manag Anal 8(2):333–349
12. Fernandez A, Insfran E, Abrahão S (2011) Usability evaluation methods for the web: A systematic mapping study. Inf Softw Technol 53(8):789–817
13. Soualhia M, Khomh F, Tahar S (2017) Task scheduling in big data platforms: a systematic literature review. J Syst Softw 134:170–189
14. Sharma R, Kamble SS, Gunasekaran A (2018) Big GIS analytics framework for agriculture supply chains: a literature review identifying the current trends and future perspectives. Comput Electron Agric 155:103–120
15. Chandarana P, Vijayalakshmi M (2014) Big data analytics frameworks. In: 2014 international conference on circuits, systems, communication and information technology applications (CSCITA). 2014. IEEE
16. Inoubli W et al. (2016) Big data frameworks: A comparative study. CoRR, abs/1610.09962
17. García-Gil D et al (2017) A comparison on scalability for batch big data processing on Apache Spark and Apache Flink. Big Data Anal 2(1):1
18. Alkatheri S, Abbas SA, Siddiqui MA (2019) A comparative study of big data frameworks. Int J Comput Sci Inf Secur 17(1):498
19. Alkatheri S, Abbas S, Siddiqui MA (2019) A comparative study of big data frameworks. Int J Comput Sci Inf Secur 17(1):418
20. Gurusamy V, Kannan S, Nandhini K (2017) The real time big data processing framework: advantages and limitations. Int J Comput Sci Eng 5(12):305–312
21. Dittrich J, Quiané-Ruiz J-A (2012) Efficient big data processing in Hadoop MapReduce. Proc VLDB Endow 5(12):2014–2015
22. Kulkarni AP, Khandewal M (2014) Survey on Hadoop and introduction to YARN

23. Cumbane SP, Gidófalvi G (2019) Review of big data and processing frameworks for disaster response applications. ISPRS Int J Geo Inf 8(9):387
24. Kamburugamuve S et al (2013) Survey of distributed stream processing for large stream sources. Grids Ucs Indiana Edu 2:1–16
25. Toshniwal A et al. (2014) Storm@ twitter. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data
26. Iqbal MH, Soomro TR (2015) Big data analysis: apache storm perspective. Int J Comput Trends Technol 19(1):9–14
27. Allen ST, Jankowski M, Pathirana P (2015) Storm applied: strategies for real-time event processing. Manning Publications Co
28. Ficco M, Pietrantuono R, Russo S (2018) Aging-related performance anomalies in the apache storm stream processing system. Futur Gener Comput Syst 86:975–994
29. Wingerath W et al (2016) Real-time stream processing for big data. Inf Technol 58(4):186–194
30. Noghabi SA et al (2017) Samza: stateful scalable stream processing at LinkedIn. Proc VLDB Endow 10(12):1634–1645
31. Inoubli W et al. (2018) A comparative study on streaming frameworks for big data
32. Perwej Y et al (2017) An empirical exploration of the yarn in big data. Int J Appl Inf Syst 12:19
33. Zaharia M (2016) An architecture for fast and general data processing on large clusters. In: 2016 Association for Computing Machinery and Morgan & Claypool
34. Meng X et al (2016) Mllib: machine learning in apache spark. J Mach Learn Res 17(1):1235–1241
35. Chen X et al. (2014) High performance integrated spatial big data analytics. In Proceedings of the 3rd ACM SIGSPATIAL international workshop on analytics for big geospatial data
36. Armbrust M et al. (2015) Spark sql: relational data processing in spark. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data
37. Salloum S et al (2016) Big data analytics on apache spark. Int J Data Sci Anal 1(3–4):145–164
38. Carbone P et al (2015) Apache flink: stream and batch processing in a single engine. Bull IEEE Comput Soc Tech Comm Data Eng 36(4):489
39. Aji A et al. (2013) Hadoop-GIS: a high performance spatial data warehousing system over MapReduce. In: Proceedings of the VLDB endowment international conference on very large data bases. 2013. NIH Public Access
40. Eldawy A, Mokbel MF (2015) Spatialhadoop: a mapreduce framework for spatial data. In: 2015 IEEE 31st international conference on Data Engineering. 2015. IEEE.
41. Eldawy A, Mokbel MF, Jonathan C (2016) HadoopViz: a MapReduce framework for extensible visualization of big spatial data. In: 2016 IEEE 32nd international conference on data engineering (ICDE). IEEE
42. You S, Zhang J, Gruenwald L (2015) Large-scale spatial join query processing in Cloud. In: 2015 31st IEEE international conference on data engineering workshops
43. Lenka RK et al. (2016) Comparative analysis of SpatialHadoop and GeoSpark for geospatial big data analytics. In: 2016 2nd international conference on contemporary computing and informatics (IC3I). IEEE
44. Yu J, Zhang Z, Sarwat M (2019) Spatial data management in apache spark: the geospark perspective and beyond. GeoInformatica 23(1):37–78
45. Ben-Daya M, Hassini E, Bahroun Z (2019) Internet of things and supply chain management: a literature review. Int J Prod Res 57(15–16):4719–4742
46. Rotz S et al (2019) The politics of digital agricultural technologies: a preliminary review. Sociol Rural 59(2):203–229
47. Khanna A, Kaur S (2019) Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture. Comput Electron Agric 157:218–231
48. Stefanidis A, Crooks A, Radzikowski J (2013) Harvesting ambient geospatial information from social media feeds. GeoJournal 78(2):319–338
49. Tong Y, Cao CC, Chen L (2014) TCS: efficient topic discovery over crowd-oriented service data. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining
50. Loukis E, Charalabidis Y (2015) Active and passive crowdsourcing in government. In: Policy practice and digital science. 2015, Springer, pp 261–289
51. Qin H et al (2016) Geocrowdsourcing and accessibility for dynamic environments. GeoJournal 81(5):699–716
52. Büscher M, Liegl M, Thomas V (2014) Collective intelligence in crises. In: Social collective intelligence. Springer, pp 243–265
53. Balan T et al (2020) Smart multi-sensor platform for analytics and social decision support in agriculture. Sensors 20(15):4127
54. Akulwar P (2020) A recommended system for crop disease detection and yield prediction using machine learning approach. In: Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries, pp 141
55. Majumdar J, Naraseeyappa S, Ankalaki S (2017) Analysis of agriculture data using data mining techniques: application of big data. J Big data 4(1):20
56. Pandey PC, Tripathi AK, Sharma JK (2021) An evaluation of GPS opportunity in market for precision agriculture. In: GPS and GNSS Technology in Geosciences, Elsevier. pp 337–349
57. Jones JW et al (2017) Brief history of agricultural systems modeling. Agric Syst 155:240–254
58. Langhammer M et al (2019) Agricultural landscape generators for simulation models: a review of existing solutions and an outline of future directions. Ecol Model 393:135–151
59. Sedano CG, Aguirre CA, Brizuela AB (2019) Numerical simulation of spray ejection from a nozzle for herbicide application: comparison of drag coefficient expressions. Comput Electron Agric 157:136–145
60. Qian Y et al (2021) Experiment and numerical simulation for designing layout parameters of subsurface drainage pipes in arid agricultural areas. Agric Water Manag 243:106455
61. Raj R et al. (2020) Precision agriculture and unmanned aerial vehicles (UAVs). In: Unmanned aerial vehicle: applications in agriculture and environment. Springer, pp 7–23
62. Radoglou-Grammatikis P et al (2020) A compilation of UAV applications for precision agriculture. Comput Netw 172:107148

63. Panday US et al (2020) A review on drone-based data solutions for cereal crops. Drones 4(3):41
64. Haddeler G et al. (2020) Evaluation of 3D LiDAR sensor setup for heterogeneous robot team. J Intell Robot Syst
65. Zhou L et al (2020) Analysis of plant height changes of lodged maize using UAV-LiDAR data. Agriculture 10(5):146
66. Antonucci F, Costa C (2020) Precision aquaculture: a short review on engineering innovations. Aquacult Int 28(1):41–57
67. Paul M et al (2020) Assessment of agricultural land suitability for irrigation with reclaimed water using geospatial multi-criteria decision analysis. Agric Water Manag 231:105987
68. Praveen B, Sharma P (2020) A review: the role of geospatial technology in precision agriculture. J Public Aff 20(1):e1968
69. Debroy P et al. (2020) Characterization of the soil properties of citrus orchards in Central India using Remote Sensing and GIS. National Academy Science Letters, pp 1–4

Publisher's Note