A Conceptual Framework for using Big Data in Egyptian Agriculture

Sayed Ahmed[1], Amira S. Mahmoud[2], Eslam Farg[3], Amany M. Mohamed[4]
Marwa S. Moustafa[5], Mohamed A.E. AbdelRahman[6], Hisham M. AbdelSalam[7], Sayed M. Arafat[8]
National Authority for Remote Sensing and Space Science (NARSS), Cairo, Egypt[1, 2, 3, 4, 5, 6, 8]
Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt[7]

***Abstract*—Agriculture is a typical contributor to the Egyptian economy, which could benefit from the comprehensive capabilities of Big Data (BD). In this work, we review the BD role in the agriculture sector in responding to two main questions: 1) Which technique, frameworks and data types were adopted. 2) Identification of the existing gap associated with the data sources, modeling, and analysis techniques. Therefore, the contribution in this paper can be outlined in three main aspects. 1) Popular BD frameworks were briefed, and a thorough comparison was conducted between them. 2) The potential data sources were described and characterized. 3) A Conceptual framework for Egyptian agriculture practice based on BD analytics was introduced. 4) Challenges and extensive recommendations have been provided, which could guide future development.**

***Keywords*—Agriculture; big data (BD); big data paradigm; BD processing framework; conceptual BD framework; geographical information systems (GIS); Hadoop; spark**

I. Introduction

Climate change, water storage, and crop fluctuation are major issues in Egyptian agriculture. Variations in market prices and socio-cultural growth contribute to the volatility of food availability. Several challenges need to be tackled to improve agricultural productivity, such as low soil fertility, insect diseases, limited technical adaptation, and varied weather conditions. In the digital era, data become not only valuable but also intelligent. BD term has been introduced in mid-2011 to describe a broad set of heterogeneous large volumes of data that can hardly be managed and processed using conventional approaches [1, 2]. Massive amounts of data, rapid data generation and delivery, organized and unstructured data sources, validity, and value [3] are the five primary elements that define BD, as shown in Fig. 1.

The BD paradigm encompasses the tools, storage, processing, and security measures used [4]. An enormous quantity of data may be analyzed using BD paradigm. It has four parts: techniques, storage, processing, and representation (see Fig. 2). They seek to find hidden trends and patterns in vast amounts of data from several sources. The storage provides management methods and tools for storing organized and unstructured data.

A variety of cloud-based platforms are optimized for maximum processing power. Data value and accessibility for decision-makers are major BD challenges. Data quality, integrity, and legal concerns have recently been addressed by Egypt's government. Several private and public sector endeavours to develop BD cyber-infrastructures. Recent academic research has focused on combining data and predictive analytics to assist governments better develop agricultural action plans. BD analytics and Remote Sensing (RS) can assist farmers manage their fields by extracting insights from acquired data.

Several attempts have been made to employ BD in agriculture [5]. BD is used by the business sector to increase large-scale commercial agriculture efficiency [5, 6]. Meanwhile, agribusiness makes better use of new communication and data sources. BD tools and approaches are utilized to successfully address and organize farm development difficulties [7, 8]. Governments must plan for the transition to digital agriculture. Several recent studies have explored BD in agriculture. Herein, we also introduce the conceptual design of BD in the Egyptian agriculture sector.

In this paper, we introduce a brief review of the potential BD role in agriculture to answer two main questions. The first question indicates the trending non-spatial and spatial BD Framework. The second question manifests the growing number data sources integrated within BD in agriculture. Therefore, A conceptual framework to adopt BD in Egyptian agriculture sector was presented and the main challenges and further directions were highlighted.

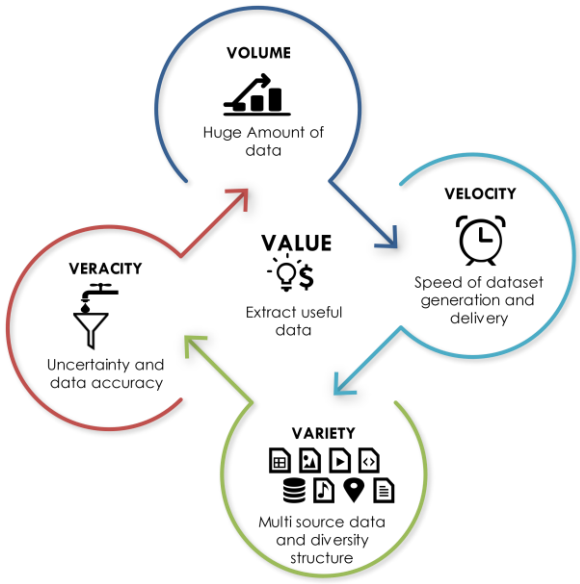

Fig. 1. Big data 5 V's Volume according to Fortune magazine.

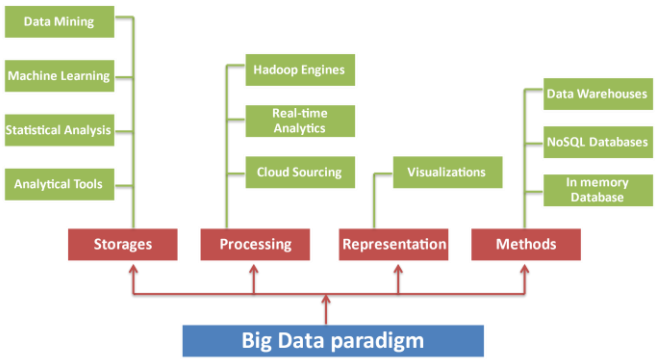

Fig. 2. The Big Data (BD) Paradigm.

The rest of this paper is organized as follows: Sections 2, and 3 briefly sum up the similarities of popular non-spatial and spatial BD framework. The potential BD data sources is discussed in Section 4. Section 5 presents the proposed conceptual framework for adopting BD in Egyptian agriculture sector. In Section 6, the BD challenges and future directions in the agriculture sector were discussed. Finally, Section 7 concludes the paper and provides future work.

II. Non-Spatial BD Frameworks

A. Batch BD Frameworks

The data had to pile up for hours or a few days to be processed in a batch setting. The data had to be loaded in memory processing time; otherwise, the data stored in database, or file system [9]. Examples of batch BD frameworks for large datasets include Hadoop Map Reduce and Spark. For smaller size, Informatica and Alteryx are widely used. For relational databases, Google BigQuery and Amazon Redshift are utilized.

Google introduced Hadoop framework [10], which comprised three elements, namely: Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), and MapReduce [11]. Typically, HDFS represents Hadoop's core component, which introduces reliable storage [11, 12]. HDFS has two architectures NameNode and DataNode [20]. YARN is considered the cluster management component in Hadoop framework [13]. Finally, MapReduce component performs two main functions, map and reduce. The users only define the map and reduce functions, and the framework is responsible for other administrative functions like parallelization and failover. Overall, Hadoop MapReduce employs HDFS for data storage, while YARN is employed for resources control and job scheduling [10, 13].

B. Stream BD Frameworks

Stream Frameworks process data as soon as it arrives at both micro-batches and real-time [9]. Examples of BD stream frameworks include Apache Storm and Apache Samza [11].

1) Apache storm: Twitter developed Apache Storm to process large-scale structured and non-structured data in real-time fashion [14-16]. A typical Apache storm topology [17] depends on a directed acyclic graph where the edge indicates the data exchange, and the node represents computation resources. A node is either a master node “Nimbus" or a worker node "Supervisor." All nodes could accept streams (sequence of Tuples). In contrast, first nodes only accept Spouts, which can read messages from external sources and convert them to tuples and resend them to other bolts nodes without any computation. Bolts receive, filter, compute, join, and create Tuples. The exchange protocol between bolts and spouts is defined by Stream grouping.

The Storm architecture [18] has three main components: Nimbus, Supervisor, and ZooKeeper. Nimbus oversees worker and slave nodes progress and assigns tasks in standard and failure cases. The supervisor is a stateless daemon responsible for initiating monitoring and restoring topologies execution [18]. ZooKeeper [19] maintains configuration information, distributed synchronization, and group membership.

Trident Application Programming Interfaces (APIs) were utilized in topology, which provides a wide range of high-level operators [14]. Trident APIs split the workload into micro-batches. The batch size is set as a parameter to control throughput and latency. However, their topologies are unfortunately inadequate to execute iterative algorithms due to their Directed Acyclic Graphs (DAGs) nature [20].

2) Apache samza: LinkedIn developed Apache Samza to tackle stream processing issues like scalability, resources allocation, etc. [21]. Apache Samza is built upon two other BD processing frameworks: Apache Kafka and Hadoop YARN [11, 21]. Apache Kafka is based on five main components: Producer, Topics, Consumer, Partitions, and Brokers. The Producer component is responsible for writing a topic for Kafka system. Every data stream entering Kafka system is called Topic. A consumer is an element with both reading ability to a Kafka topic and responsibility to maintain information with respect to its offset to be used in the case of failures. Brokers are the ingle nodes that form the Kafka cluster.

C. Hybrid BD Frameworks

Some applications require batch and stream processing frameworks. Therefore, it is mandatory to use hybrid processing frameworks in such cases. Apache Spark, as well as Apache Flink, are regarded as the most notable examples.

1) Apache spark: Apache Spark represents a hybrid framework constructed on top of Hadoop engine but optimizes processing through accelerating batch processing workloads using complete in-memory processing [11].

Apache Spark limited the creation of storage layer links to two cases: loading the data into memory to be processed and storing the final results. Unlike Apache MapReduce, Spark piles the intermediate results in memory. Resilient Distributed Datasets (RDDs) are the core data structure of Apache Spark, allowing developers to accumulate intermediate for reusability purposes. RDDs are fault-tolerant that could optimize partitions, maintaining the stored data [22].

Apache Spark framework [22] includes several main components combined with upper-level libraries such as Spark's MLlib for machine learning [23], GraphX [24] for

stream processing, and Spark SQL [25] for stream processing, and structured data processing.

Spark core is implemented in Scala and supports multi clusters. Spark supports upper-level APIs like Scala, Java, Python, and R and operates various data visualization and analysis algorithms. A cluster manager is utilized for requesting cluster resources for jobs' execution. Spark built-in cluster manager has many cluster managers used by Spark core, such as Hadoop YARN, Apache Mesos, and AmazonEC2. Besides, Spark enables data access in different data sources, such as HDFS, Cassandra, HBase, Hive, Alluxio, and many other data sources.

2) Apache flink: Apache Flink [26] is regarded as an open-source hybrid framework for applications such as real-time analytics, continuous data pipelines, batch processing, in addition to iterative algorithms. The main advantage is processing huge data volumes at an economic level of latency and high fault tolerance in a distributed environment. The DataSet API is used to process finite data sets and is often known as batch processing [26].

Finally, Table I compares the mentioned BD processing frameworks based on the following factors, including cluster architecture, data flow, data processing model, fault-tolerance, latency, scalability, back-pressure mechanism, programming languages, as well as different machine learning libraries.

TABLE I. A COMPARISON BETWEEN POPLAR NON-SPATIAL BD FRAMEWORKS

Framework	**Hadoop**	**Storm**	**Trident Storm**	**Samza**	**Spark**	**Flink**
Processing type	Batch	Stream	Stream	Stream	Hybrid	Hybrid
Computing cluster architecture	YARN	Nimbus	Nimbus	YARN and Kafka	YARN and Mesos	YARN and Kafka
Data Flow	MapReduce data flow	cyclic graph	DAGs	Kafka - Kafka job – Kafka	A queue of RDDs called DStream processed one-at-a-time using micro-batching cluster	stream -> system (operators) -> sinks
Data Processing Model	MapReduce	at-least-once	exactly-once	at-least-once	exactly-once	exactly-once
Fault-Tolerance	Yes	Yes	Yes	Yes	Yes (using lineage)	Yes (generating snapshots)
Latency	low	several milliseconds	several milliseconds for small batches	Several milliseconds	High	Low
Scalability	Yes	Yes	User-defined parallel processing	Yes	Yes (user demand)	Yes (only tasks that can be done in parallel)
Back-pressure Mechanism	No	Yes	Yes	No (buffering instead)	Yes	Yes
Programming Languages	Java mostly	Java API with adapters for Python, Ruby, and Perl	Java API with adapters for Python, Ruby, and Perl	Java mostly	API for Scala, Java, Python, and R	Java and Scala
Support for Machine Learning	Yes	compatible with SAMOA API	Trident-ML	compatible with SAMOA API	Yes (Spark MLlib)	Yes (FlinkML)

III. SPATIAL BD FRAMEWORKS

A. Hadoop-based

1) Hadoop-GIS: Hadoop-GIS is regarded as a MapReduce-based framework to process large-scale vector data, partitioning, as well as geographic queries [27]. Geographic (Spatial) queries can take many forms, such as descriptive, spatial relationship-based, distance-based queries, along with spatial mining and statistics techniques. In order to boost query performance, Hadoop-GIS utilize a spatial partitioning and local spatial indexing called SATO [41]. However, complex geometry forms, such as convex/concave polygons, line string, multi-point, as well as multi-polygon, are not supported. In fact, Hadoop-GIS supports only two-dimensional data and two query types over geometric objects, including box range as well as spatial joins.

2) Spatial-Hadoop: Spatial-Hadoop is a complete MapReduce framework that was introduced to overcome Hadoop-GIS limitations. It contains two new components for efficient and scalable spatial data processing: SpatialRecordReader and SpatialFileSplitter to support spatial data, spatial indexes, and operations [28].

Spatial-Hadoop supports different geometry types, such as points, multi-points, line strings, and polygons. In spatial indexes, spatial partitioning approaches were implemented, such as uniform grids, R-Tree, Quad-Tree, K-Dimensional Tree (KD-Tree), as well as Hilbert curves. Also, it supports many predefined spatial operations, such as box range queries, KNN queries, and spatial joins. Besides, it supports various geometric objects, including segments and polygons, and operations over them, producing convex hulls in addition to skylines. The mentioned capabilities are implemented in Spatial-Hadoop as distributed geometric data analytics framework.

B. Spark-based

1) Spatial-Spark: Spatial spark is a framework to process GIS data based on cluster computing. It was constructed on top of Spark RDD for providing a broad range of spatial operations, including range query, spatial join, spatial filtering, R-Tree index, and R-Tree partitioning to boost queries [29]. Spatial-Spark can be considered an in-memory BD framework intended for supporting two spatial join operators, including broadcast spatial join and partitioned spatial join [29].

2) Geo-Spark: Geo-Spark is regarded as an in-memory cluster computing framework constructed on Spark top to process large-scale GIS data faster than Spatial-Hadoop [30]. Geo-Spark expands the concept of RDDs as well as SparkSQL for supporting spatial data types, indexes, in addition to geometric operations at scale. It also helps spatial data partitioning systems, including a uniform grid, R-tree, Quad-Tree, KD B-Tree, as well as KNN queries. Geo-Spark is optimized to select a suitable join algorithm for achieving a balance in a cluster between run time as well as memory/CPU use [31]. Geo-Spark enables the Apache Spark developers for developing efficient spatial analysis applications utilizing operational quickly (for instance, Java and Scala) in addition to declarative (i.e., SQL) languages and spatial RDD APIs.

Toward more solid knowledge, principal differences and similarities among Hadoop-GIS, Spatial-Hadoop, Spatial-Spark, and Geo-Spark [30, 31] dependent on prevalent characteristics such as spatial partitioning, spatial indexing, DataFrame API, in-memory processing, etc., are summarized in Table II.

TABLE II. COMPARISON AMONG POPULAR BIG GIS DATA PROCESSING FRAMEWORKS

Feature	Hadoop-GIS	Spatial-Hadoop	Spatial-Spark	Geo-Spark
DataFrame API	×	×	×	√
In-memory processing	×	×	√	√
Spatial Partitioning	SATO	Multiple	Multiple	Multiple
Spatial Indexing	R-Tree	R-/Quad-Tree	R-Tree	R/Quad-Tree
KNN query	√	√	×	√
Query optimizer	×	×	×	√
Distance query	√	√	√	√
Distance join	√	√	√	√
Filter (Contains)	√	√	√	√
Filter (ContainedBy)	√	√	√	×
Filter (Intersects)	√	√	√	√
Filter (WithinDistance)	√	√	√	×

IV. BD MAJOR DATA SOURCES

A. Satellite Imagery

Satellite imagery is captured by active or passive sensors to study the Earth's surface [9]. The collected images using passive sensors estimate reflected sunlight emitted from the sun. In contrast, the images are usually acquired using active sensors. In heavy cloud cover, rain conditions, and at nighttime, active sensors, including the Synthetic Aperture Radar, are efficiently utilized to tackle limitations of passive sensors and increase the observational capability for agriculture applications.

B. Wireless Sensor Web and IoT

Wireless Sensor Network represents a collection of heterogeneous and sophisticated sensors responsible for collecting various data types, including temperature, humidity, wind, etc. WSW depends on Internet of Things technology (IoT), which integrates and deploys several heterogeneous spatially distributed sensors to enrich the identification and visualization capabilities of different agriculture areas [32]. The collected data [33] could facilitate the communication between the farmers, experts, and investors to maintain a closer day-to-day management when classical communication methods fail. Despite their wide usability in smart farming, WSW lacks the comprehensive coordination to different data sources as well as protocols from "Socio-techno-economic perspectives" [34].

C. Crowd-sourcing and Social-media

In the last few years, several platforms were developed to collect data from the public. These platforms either actively contributed where contributors are aware of the data collection [35], such as crowdsourcing, or passively contributed where contributors are not aware such as social media [36]. Unlike crowdsourcing platforms, social media are used to track pest and sharing weather information.

Social media is utilized in agriculture development for data gathering, information extraction, analytic workflow, geo-location pattern/image/text analytics, and information transferring over social media services [37]. Real-time analytics dependent on social media platforms [37] offer considerable chances for automatic detection and monitoring of plant disease, yield productivity, and forecasting [38]. For social media data, visual analytics can simplify Spatio-temporal analysis and generate a spatial-based decision for supporting environment, helping small farmers match end-user demand. Social media not only depends on text messages but also depends on posted videos and images by users. Analysis dependent on image/video, along with visual analytics, utilize social media posts to extract critical information.

D. Mobile CDRs and GPS Traces

GPS traces and mobile CDRs data are valuable resources, especially in natural disasters management such as landslide monitoring, Tsunami monitoring, earthquake management, forest fire, and flood management. GPS traces data had been a value-added in different agriculture applications [39], like identifying mobility patterns for agriculture machines and fuel consumption tracking.

E. Simulation

Numerical simulation, or referred to as forecasting, is regarded as one of the essential agricultural contributions to meteorological phenomena, land surface phenomena, and diverse pollutions kinds [40, 41]. Also, mechanistic modeling has helped estimate water spray [42] and parameter estimation of subsurface pipe [43].

The modeling and simulation tools for agriculture management focus on different aspects. Several mechanistic models were developed to enrich the scientific understanding of agriculture aspects to gain insights into physical, chemical, and biological control parameters in crop and animal production systems. Another group of simulation models was developed to plan and support decision-makers.

F. UAVS, Drones and LiDAR

UAVs, as well as drones, deliver images and videos with very high-resolution amenable to be utilized in various agriculture applications [44] such as live-stock monitoring, crop production, yield prediction, fertilizer, pesticide spraying, and soil mapping [45]. Many sensors can be embedded in a UAV or drone, such as weather sensors, cameras, and LiDAR sensors. The obtained sensors data can be integrated into real-time decision making in many fields such as spraying of pesticides through drone, plant phenotyping, and yield production estimation [46].

LiDAR technology [47] can create detailed topography maps and Digital Elevation Models (DEMs) necessary in crop architectural parameters, forests, and crop parameter analysis. LiDAR can also help in yield forecasting and monitoring, soil types, estimate and prevent soil erosion, land segmentation, and crop analysis field management [48]. LiDAR technology is highly valuable in geospatial community, with the massive data amounts amenable to utilization in a diversity of applications.

G. Vector-Based GIS Data

Vector-based GIS data provides powerful add-ons in agriculture management applications [49] like farmland suitability analysis, estimation fertilizer costs, and pesticide. Additional geospatial analysis for critical facilities (healthcare providers, schools, fire station, etc.) [52], estimation of the actual effect on human (age, gender, social and economic status, etc.), resources inventory (vehicles, supplies, equipment, etc.), and infrastructure (utility grids and transport networks) help and empower farmers community. Common GIS data sources enrich precision agriculture in developing countries [49].

In [49], the authors implemented a framework that integrated GIS with Multi-Criteria Decision Analysis (GIS-MCDA) to assess land suitability for irrigation with reclaimed water. In [53,54], the authors developed a GIS-based approach that studies the appropriate soil-site citrus features for enhancing productivity.

From the above discussion of various BD sources, it can be noted that satellite imagery, aerial imagery, crowdsourcing, social media records, simulation and GIS data could offer economic solution to enrich the Egyptian farming sector with valuable information. On the other hand, WSW, IoT, video and images from UAVs, GPS traces and CDRs, and LiDAR could be valuable and cost-effective data sources for private sector that could open new business opportunities.

V. A Conceptual Framework for using BD in Egyptian Agriculture

In Egypt, agriculture's contribution of real GDP growth fell from 16.5% in 2002 to 11.4% in 2018 [1]. It employs 14.5 percent of the active population [1, 50] and supplies 91.5 percent of the population's requirements. It also only exports 8.7% of total commodities. BD analytics, the "new oil," can help the agriculture industry [51, 53]. According to a recent study, the internet's ubiquity and suitable communication technologies might assist evaluate enormous volumes of gathered data to address critical concerns like desertification and global food costs. Two significant reasons [51] often influence BD analytics adoption in Egypt:

- Push factor determines the motivation opportunities such as the new governmental investments in technology and infrastructure.
- Pull factor analyses business factors such data quality, security, and availability.

Egyptian agriculture issues fall into three categories: monitoring, management, and forecasting. As illustrated in Fig. 3, we developed a conceptual framework to address Egyptian agriculture difficulties using BD analytics. The next parts describe data collecting, data analysis, and issue solutions.

A. Data Collection

Data collection, the first and primary step of BD applications, aims to collect a variety of structured and unstructured data, including soil, yield, climate data, satellite images, and other information sources. The basic modules implemented in this stage include filtering and harmonizing the captured data and nullifying unnecessary data. Also, metadata is generated for each dataset to identify how the data is rendered and analyzed.

B. Data Analysis

The collected data had to be prepared by extracting the vital information for further analysis. The valuable information is extracted by cleaning, interpreting, integrating, mining, analyzing, and warehousing data in this stage. Different analysis approaches could be performed to improve data understanding, such as visual analysis, prescriptive, diagnostic, and predictive analysis.

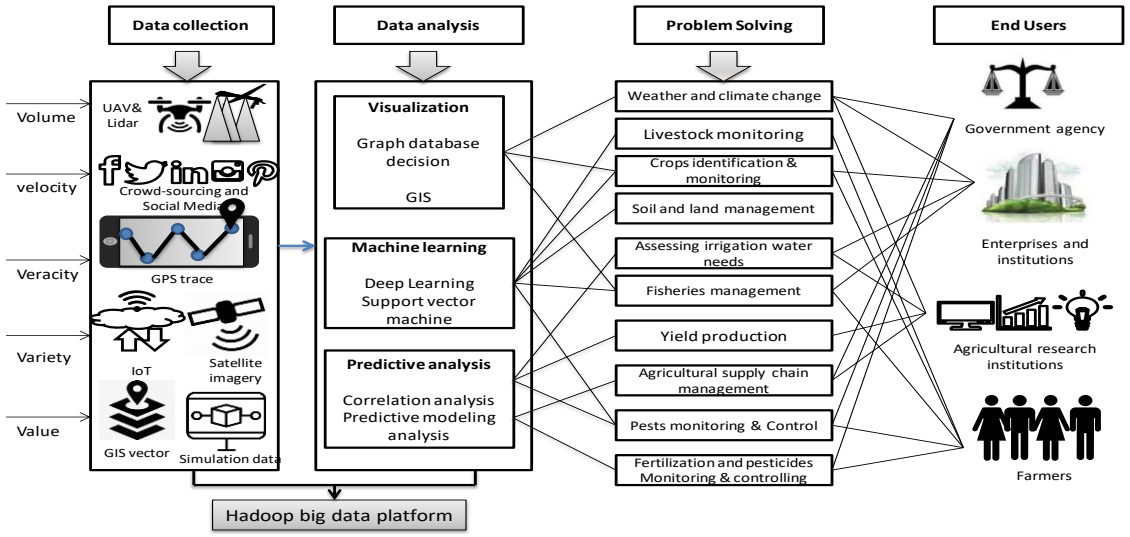

Fig. 3. The Proposed Conceptual Framework for BD in Egyptian Agriculture.

Generally, visual analysis can be utilized to gain insights about the uncovered relationships within massive datasets and empower investigators for obtaining more intuitive visual cognition as well as efficient decision-making assist. Now, government, as well as related policymakers, may utilize the aforementioned BD sources for conducting visual analysis of water resources monitoring, weather condition, soil condition, and close contact of scientific researchers for making decisions. Moreover, visualization is principally accomplished through GIS. By linking BD and GIS can help farmers, enterprises, and institutions better understand their spatial patterns as well as relationships. In this regard, GIS technology should first be provided with properly attained BD. The data collection is no longer restricted to conventional facilities and approaches, including stations, satellite RS, and field measurement. Besides, it still extends with IoT and UAV capabilities. For processing and analysis, this appears to happen primarily via batch processing technology like MapReduce and distributed system infrastructure like Hadoop.

The collected data is amenable to be transformed into an acceptable format for GIS systems. Decision-makers utilized tools such as GoogleEngine, which provide interactive digital maps with almost real-time visualization of much free satellite imagery and a continuous development platform for the intelligent integration of multi-sources information.

1) Descriptive analytics: Descriptive analytics enable visualization and interpreting of the collected data in order to answer the "what happened?" question [2, 7]. Several visualization tools and Ad-hoc were implemented for various data-related agriculture sources to tackle the complex nature of both structured and unstructured data. A basic summarization of these large volumes of data can be performed in diverse formats, including summaries, i.e., tables, charts, spreadsheets, etc.

2) Diagnostic analytics: Diagnostic analytics motivate analysts to perform a root cause analysis to discover key reasons for the events. A smart, well-designed dashboard combined with time-series data offers analysts mandatory tools to quickly summarize an overview that matches the business objectives. The question answered using diagnostic analytics is, "why did it happen?". Data mining, as well as correlation, can offer profound perception into defining the targeted problems and issues. DL is a full orientation shift in supervised machine learning, such as pattern recognition and natural language processing.

3) Predictive analytics: Predictive analytics aims to predict future by answering the "what is going to happen?" questions. Generally, ample statistical and machine learning approaches aim to correlate past and today data to assume the future.

4) Prescriptive analysis: Prescriptive analytics assess analysts for determining optimal actions and decisions on the basis of answers to a diversity of questions concerning "what might happen?". For instance, analysts might possess numerous choices for dispatching maintenance actions towards a specific asset. For maintenance actions, the time-varying expense required items to be repaired or substituted, while the risk linked to each of such decisions is capable of determining the optimal dispatch. Prescriptive analytics synthesize the BD, diverse sciences' principles, business rules, in addition to IoT disciplines for receiving the predictions merits, followed by taking the most optimal decisions. Prescriptive analytics goes beyond the prediction. Indeed, the "what will happen" and "when will happen" questions should be able to justify the "why it would happen" questions.

C. Problem Solving

Finally, the collected data is converted into actionable perceptions. Herein, the data captured from different sensors could be utilized to improve the monitoring, management, and prediction of agriculture sector activities in Egypt. Furthermore, BD is utilized to boost predictive insights in real-time for future outcomes in farming. Recently, the private Egyptian agricultural sector started implementing innovative technologies, particularly those requiring large scale of operations and costly initial investment. To implement new

technologies, this possesses a substantial influence on farmers' prospects. The modern visualization and analysis tools enable farmers, experts, and research institutions to easily connect and simplify data management in a cost-effective way. Such a shift to new technologies comes true by research as well as development in hardware and software services. Recently, agricultural innovation has caused auspicious new methods for boosting productivity. However, for Egyptian farmers, access to high quality and precise information at a reasonable expense is challenging.

Various efforts had been conducted to incorporate BD technologies in the agriculture sector. BD analytical help governments to establish policies and define mitigation plans toward climate change adaptation to secure food. BD integrated with IoT technology effectively supports a wide range of daily agriculture activities such as livestock monitoring, pest monitoring and fertilization control.

Several studies had investigated the power of BD in digital farming worldwide to effectively estimate the biophysical factors of different crops and yield prediction which adopted in crop monitoring, and the investigation of irrigation water needs.

VI. Challenges and Future Direction

This section discusses the open issues and challenges that face big data in agriculture sector. Some of the challenges were identified from the literature have been discussed previously.

A. Big Data Acquisition

Agriculture sector in Egypt requires different set of data from different sources in order to fill the gaps between current and required state by BD analytics. The integration of multiple data sources will improve data quality and data integrity, provided that individual data validation is conducted before data integration. In the context of agriculture management, data integration can benefit from the data semantics or properties related to the data itself. It is impossible to avoid noises and misinformation from big data as a lot of these are unintentional, especially from social media and crowdsourcing. Also, data privacy and accuracy issues associated with big data acquisition still represent significant challenges, despite the available protocols and analytical method which are crucially required during the acquisition process. One of the proposed solutions that can help to eliminate such noise and misinformation is to develop a framework that enables the integration of multiple sources of data, such as crowd-sourcing data sensor outputs. The framework could facilitate the detection to the anomaly or improper values caused by system failure or misleading data acquisition methods. Machine learning techniques can contribute to the integration filtration process to increase the data quality

B. Big Data Analytics

Due to the integration of multi-platform, multi-scale, and multi-discipline data, there is a must to enhance the predictive modeling capability for the farm management to become more efficient. Activities and research associated with using the integrated information and the results of predictive analysis are expected to better enhance our capability to efficiently handle livestock, farm clinic, and supply and chain process. It has been noted that crowd-sourced data provided by affected farmers have significant value during the management and decision making. However, analytical methods are still strongly required to integrate these crowd-sourced data reliably and precisely into the physical sensing data (e.g., satellite, UAV) and official data (e.g., terrain data, census data). Only in this case, the smart farming can be effectively depicted in terms of pests control, livestock managements, and yield production. Hence, the decision-making processes can benefit from the analytical results and build food security system that benefits populations and communities

C. Cyber-infrastructures

There is a critical necessity for the design and development of cyber-infrastructures so that big data can be effectively integrated into agriculture sector management agencies for real-time decision making. These cyber-infrastructure capabilities provide shared knowledge and communicating platforms to the decision-makers and responders from different organizations to conduct the process required in agriculture in an effective way. Research efforts and related activities are still needed to overcome the challenges emerging from big, sensed data, including efficient data management, fast data transfer, and intuitive data visualization.

VII. Conclusion

This paper conducted a systematic literature review to inspect the recent cutting-edge research of BD in the agricultural and farming field. BD analytics can help the Egyptian agriculture sector overcome several challenges; however, it required a hefty investment to be integrated. Egyptian Farmers had to adopt modern and new technology to balance the food gap and supply concerns. This paper reviews 242 peer-reviewed articles on BD in agriculture, indicating BD's prominent role in tackling the agriculture sector's challenges. Therefore, ample conclusions were drawn:

- The up-raising trend in adopting BD in different agriculture applications, the availability of free satellite imagery, and the massive computational capabilities and efficient machine learning algorithms help researchers gain insights and recommend solutions to agriculture challenges.
- BD tackled a more comprehensive range of applications, even in the agriculture sector.
- Satellite imageries were specifically employed to produce different popular vegetation indices and land cover maps, especially Landsat and Sentinel-2 imagery.
- Extensive studies adopted different machine learning methods for RS data processing. In the last five years, deep learning had been adopted in several studies, especially in crop mapping and pests and disease identification.

Acknowledgment

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

REFERENCES

[1] W. Bank. (2020). World Development Indicators.

[2] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: a survey," Computer science review, vol. 17, pp. 70-81, 2015.

[3] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, "A review on the practice of big data analysis in agriculture," Computers and Electronics in Agriculture, vol. 143, pp. 23-37, 2017.

[4] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information sciences, vol. 275, pp. 314-347, 2014.

[5] S. Himesh, E. Prakasa Rao, K. Gouda, K. Ramesh, V. Rakesh, and G. Mohapatra, "Digital revolution and Big Data: a new revolution in agriculture," CAB Rev, vol. 13, no. 21, pp. 1-7, 2018.

[6] C. Kempenaar et al., "Big Data analysis for smart farming: results of TO2 project in theme food security," Wageningen University & Research2016.

[7] A. Krishnan, K. Banga, and M. Mendez-Parra, "Disruptive technologies in agricultural value chains," 2020.

[8] S. Alkatheri, S. A. Abbas, and M. A. Siddiqui, "A Comparative Study of Big Data Frameworks," International Journal of Computer Science and Information Security (IJCSIS), vol. 17, no. 1, 2019.

[9] S. P. Cumbane and G. Gidófalvi, "Review of Big Data and Processing Frameworks for Disaster Response Applications," ISPRS International Journal of Geo-Information, vol. 8, no. 9, p. 387, 2019.

[10] J. Dittrich and J.-A. Quiané-Ruiz, "Efficient big data processing in Hadoop MapReduce," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2014-2015, 2012.

[11] V. Gurusamy, S. Kannan, and K. Nandhini, "The real time big data processing framework: Advantages and limitations," International Journal of Computer Sciences and Engineering, vol. 5, no. 12, pp. 305-312, 2017.

[12] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in 2010 IEEE 26th symposium on mass storage systems and technologies (MSST), 2010, pp. 1-10: Ieee.

[13] A. P. Kulkarni and M. Khandewal, "Survey on Hadoop and Introduction to YARN," 2014.

[14] S. Kamburugamuve, G. Fox, D. Leake, and J. Qiu, "Survey of distributed stream processing for large stream sources," Grids Ucs Indiana Edu, vol. 2, pp. 1-16, 2013.

[15] A. Toshniwal et al., "Storm@ twitter," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data, 2014, pp. 147-156.

[16] M. H. Iqbal and T. R. Soomro, "Big data analysis: Apache storm perspective," International journal of computer trends and technology, vol. 19, no. 1, pp. 9-14, 2015.

[17] S. T. Allen, M. Jankowski, and P. Pathirana, Storm Applied: Strategies for real-time event processing. Manning Publications Co., 2015.

[18] M. Ficco, R. Pietrantuono, and S. Russo, "Aging-related performance anomalies in the apache storm stream processing system," Future Generation Computer Systems, vol. 86, pp. 975-994, 2018.

[19] S. Haloi, Apache zookeeper essentials. Packt Publishing Ltd, 2015.

[20] W. Wingerath, F. Gessert, S. Friedrich, and N. Ritter, "Real-time stream processing for Big Data," it-Information Technology, vol. 58, no. 4, pp. 186-194, 2016.

[21] S. A. Noghabi et al., "Samza: stateful scalable stream processing at LinkedIn," Proceedings of the VLDB Endowment, vol. 10, no. 12, pp. 1634-1645, 2017.

[22] D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "A comparison on scalability for batch big data processing on Apache Spark and Apache Flink," Big Data Analytics, vol. 2, no. 1, p. 1, 2017.

[23] X. Meng et al., "Mllib: Machine learning in apache spark," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 1235-1241, 2016.

[24] R. S. Xin, D. Crankshaw, A. Dave, J. E. Gonzalez, M. J. Franklin, and I. Stoica, "Graphx: Unifying data-parallel and graph-parallel analytics," arXiv preprint arXiv:1402.2394, 2014.

[25] M. Armbrust et al., "Spark sql: Relational data processing in spark," in Proceedings of the 2015 ACM SIGMOD international conference on management of data, 2015, pp. 1383-1394.

[26] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache flink: Stream and batch processing in a single engine," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 36, no. 4, 2015.

[27] A. Aji et al., "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, vol. 6, no. 11: NIH Public Access.

[28] A. Eldawy and M. F. Mokbel, "Spatialhadoop: A mapreduce framework for spatial data," in 2015 IEEE 31st international conference on Data Engineering, 2015, pp. 1352-1363: IEEE.

[29] S. You, J. Zhang, and L. Gruenwald, "Large-scale spatial join query processing in Cloud," in 2015 31st IEEE International Conference on Data Engineering Workshops, 2015, pp. 34-41.

[30] R. K. Lenka, R. K. Barik, N. Gupta, S. M. Ali, A. Rath, and H. Dubey, "Comparative analysis of SpatialHadoop and GeoSpark for geospatial big data analytics," in 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), 2016, pp. 484-488: IEEE.

[31] J. Yu, Z. Zhang, and M. Sarwat, "Spatial data management in apache spark: The geospark perspective and beyond," Geoinformatica, vol. 23, no. 1, pp. 37-78, 2019.

[32] M. Ben-Daya, E. Hassini, and Z. Bahroun, "Internet of things and supply chain management: a literature review," International Journal of Production Research, vol. 57, no. 15-16, pp. 4719-4742, 2019.

[33] S. Rotz et al., "The politics of digital agricultural technologies: a preliminary review," Sociologia Ruralis, vol. 59, no. 2, pp. 203-229, 2019.

[34] A. Khanna and S. Kaur, "Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture," Computers and electronics in agriculture, vol. 157, pp. 218-231, 2019.

[35] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," GeoJournal, vol. 78, no. 2, pp. 319-338, 2013.

[36] H. Qin, R. M. Rice, S. Fuhrmann, M. T. Rice, K. M. Curtin, and E. Ong, "Geocrowdsourcing and accessibility for dynamic environments," GeoJournal, vol. 81, no. 5, pp. 699-716, 2016.

[37] T. Balan et al., "Smart Multi-Sensor Platform for Analytics and Social Decision Support in Agriculture," Sensors, vol. 20, no. 15, p. 4127, 2020.

[38] P. Akulwar, "A Recommended System for Crop Disease Detection and Yield Prediction Using Machine Learning Approach," Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries, p. 141, 2020.

[39] P. Debroy et al., "Characterization of the Soil Properties of Citrus Orchards in Central India using Remote Sensing and GIS," National Academy Science Letters, pp. 1-4, 2020.

[40] J. W. Jones et al., "Brief history of agricultural systems modeling," Agricultural Systems, vol. 155, pp. 240-254, 2017/07/01/ 2017.

[41] M. Langhammer, J. Thober, M. Lange, K. Frank, and V. Grimm, "Agricultural landscape generators for simulation models: A review of existing solutions and an outline of future directions," Ecological Modelling, vol. 393, pp. 135-151, 2019/02/01/ 2019.

[42] C. G. Sedano, C. A. Aguirre, and A. B. Brizuela, "Numerical simulation of spray ejection from a nozzle for herbicide application: Comparison of drag coefficient expressions," Computers and Electronics in Agriculture, vol. 157, pp. 136-145, 2019.

[43] Y. Qian, Y. Zhu, M. Ye, J. Huang, and J. Wu, "Experiment and numerical simulation for designing layout parameters of subsurface drainage pipes in arid agricultural areas," Agricultural Water Management, vol. 243, p. 106455, 2021/01/01/ 2021.

[44] R. Raj, S. Kar, R. Nandan, and A. Jagarlapudi, "Precision Agriculture and Unmanned Aerial Vehicles (UAVs)," in Unmanned Aerial Vehicle: Applications in Agriculture and Environment: Springer, 2020, pp. 7-23.

[45] P. Radoglou-Grammatikis, P. Sarigiannidis, T. Lagkas, and I. Moscholios, "A compilation of UAV applications for precision agriculture," Computer Networks, vol. 172, p. 107148, 2020.

[46] U. S. Panday, A. K. Pratihast, J. Aryal, and R. B. Kayastha, "A Review on Drone-Based Data Solutions for Cereal Crops," Drones, vol. 4, no. 3, p. 41, 2020.

[47] G. Haddeler, A. Aybakan, M. C. Akay, and H. Temeltas, "Evaluation of 3D LiDAR Sensor Setup for Heterogeneous Robot Team," Journal of Intelligent & Robotic Systems, 2020/08/12 2020.

[48] L. Zhou, X. Gu, S. Cheng, G. Yang, M. Shu, and Q. Sun, "Analysis of Plant Height Changes of Lodged Maize Using UAV-LiDAR Data," Agriculture, vol. 10, no. 5, p. 146, 2020.

[49] M. Paul, M. Negahban-Azar, A. Shirmohammadi, and H. Montas, "Assessment of agricultural land suitability for irrigation with reclaimed water using geospatial multi-criteria decision analysis," Agricultural Water Management, vol. 231, p. 105987, 2020.

[50] FAO, "The future of food and agriculture – Trends and challenges," 2017.

[51] E. D. Lioutas and C. Charatsari, "Big data in agriculture: Does the new oil lead to sustainability?," Geoforum, vol. 109, pp. 1-3, 2020.

[52] Russ, M., 2021. Knowledge management for sustainable development in the era of continuously accelerating technological revolutions: A framework and models. Sustainability, 13(6), p.3353.

[53] Fazelabdolabadi, B., Montazeri, M. and Pourafshary, P., 2021. A Data Mining Perspective on the Confluent IonsEffect for Target Functionality. HighTech and Innovation Journal, 2(3), pp.202-215.

[54] Habeeb, N.J. and Weli, S.T., 2021. Combination of GIS with Different Technologies for Water Quality: An Overview. HighTech and Innovation Journal, 2(3), pp.262-272.