MEASUREMENT AND ANALYSIS OF AGRICULTURAL FIELD STATE USING CLOUD-BASED DATA PROCESSING PIPELINE

Denys Shutka, MSc Student, Roman Prodan, MSc Student, Vasyl Tataryn, PhD, As.-Prof.,
Lviv Polytechnic National University, Ukraine;
e-mail: denys.shutka.mknuo.2022@lpnu.ua

Abstract. The increasing demand for precision agriculture has prompted the integration of advanced technologies to optimize agricultural practices. This article presents an approach to agricultural field data processing using a cloud-based data pipeline. The system leverages data from various sensors deployed in the fields to collect real-time information on key parameters such as soil moisture, temperature, humidity, etc. The collected data is transmitted to the cloud where it undergoes a series of data processing and analysis stages.

The article demonstrates the effectiveness of the cloud-based data pipeline in enhancing agricultural resilience. It facilitates prompt decision-making by farmers and stakeholders based on real-time data analysis. Additionally, the system offers a valuable tool for monitoring and optimizing irrigation strategies, resource allocation, and crop management practices. This research highlights the potential of cloud-based data pipelines in revolutionizing precision agriculture. The ability to measure and analyze agricultural field data accurately and efficiently opens new avenues for sustainable farming practices and mitigating risks related to wildfires and droughts.

Key words: Data Processing, ETL, Data Analysis, Monitoring and Measurement, Internet of Things, Agriculture, Cloud Technologies.

1. Introduction

As the economy grows, the demand for food products increases [1], making agricultural product control increasingly vital. The agricultural sector requires a combination of traditional agricultural practices, modern science and technology, and efficient management methods to enhance productivity and improve the quality and quantity of agricultural production.

Fields and farms covering vast areas are at risk of fire for various reasons, such as lightning strikes during thunderstorms or careless handling of fire. With the ongoing threat of infrastructure bombing in Ukraine, the need for efficient agricultural monitoring has become more urgent than ever.

Ukraine holds a prominent position as a leading producer of black soil, covering 27.8 million hectares or 8.7% of the world's black soil resources [2]. This fertile soil represents 67.7% of Ukraine's agricultural land and, when properly managed, provides the highest yield for various crops with adequate moisture.

By developing a prototype of an agricultural data processing pipeline for real-time monitoring of crop fields, data from various field sensors can be obtained from any sized field area at any time. Utilizing cloud technologies for the processing system presents numerous opportunities for expanding its functionality.

For instance, integrating an AI system [3] for continuous data analysis through anomaly detection is a modern approach facilitated by the data source provided by the cloud processing pipeline. Additionally, a notification system can send emails to stakeholders in case of system failures or critical indicators. Real-time access to information from various sensors opens the door to various solutions that can enhance agricultural productivity.

Leveraging big data [4], the Internet of Things [5], artificial intelligence, event processing systems, cloud notifications, non-relational databases, data visualization technologies, and more are essential for the success of modern businesses. The application of these technologies in agricultural practices can pave the way for increased productivity and global market presence. And the current lack of such systems leads to the necessity of manual monitoring which is not as quick and effective and requires more spending over time.

An agricultural data processing pipeline is an automated system designed to promptly alert users about field fires or indicate optimal fertilizer application times for plants. The demand for such a system is high as it eliminates the need for large teams inspecting expansive agricultural fields, thanks to the accessibility of data measurements through the internet.

Data processing is a novel and thriving IT domain applicable in different industries from banking to smart homes. As data operations become the future of every business, the agricultural sector can utilize these processed results to boost productivity and enhance its position in the Ukrainian market.

2. Drawbacks

The manual monitoring of agricultural fields poses significant drawbacks that hinder effective farming practices. One of the key limitations is the lack of real-time monitoring, as the data collected through manual inspections is often outdated by the time it reaches decision-makers. This delay in obtaining critical information can lead to missed opportunities and delayed responses to emerging challenges.

Moreover, relying on manual methods makes the monitoring process slow and labor-intensive. Specialists must physically visit each field, covering vast areas that could span tens or hundreds of hectares, resulting in considerable time and effort expenditure. This slow pace of data collection can impede timely decision-making and prevent rapid interventions during critical situations.

Another drawback of manual monitoring is the potential for inconsistency in data collection. Different personnel may interpret and record information differently leading to discrepancies in the gathered data. Such inconsistencies can create inaccuracies in the analysis and decision-making process, introducing uncertainty in agricultural planning.

The lack of real-time monitoring and inconsistent data can lead to losses in agricultural productivity. Without timely insights into crop health, soil conditions, and potential risks, farmers may miss crucial opportunities to address issues promptly leading to reduced yields and lower overall profitability.

Furthermore, manual monitoring is susceptible to human errors and oversights, further complicating the accuracy of data collection. These errors may result from fatigue, distractions, or varying levels of expertise among the monitoring personnel, adding another layer of uncertainty to the decision-making process.

3. Goal

The goal of the current article is to present and demonstrate the use of an automated cloud-based data pipeline for the measurement and analysis of agricultural field data, as well as to showcase the effectiveness and efficiency of this technology in enhancing precision agriculture by leveraging data from sensors deployed in the fields, processing it in the cloud, and enabling accurate and real-time insights into the state of agricultural fields.

4. Measurement and analysis of agricultural field data

In the pursuit of generating valuable insights from the growing volume and diversity of data, the significance of data engineering and its role in analytics, data science, and machine learning has escalated. Data engineering teams face mounting pressure to convert raw unstructured data into a clean and reliable format, a crucial step that precedes its effective utilization for addressing business challenges.

The data processing pipeline's integral component lies in constructing the Extract - Transform - Load (ETL) model [6]. ETL serves as the process that data engineers employ to extract data from various sources, transform it into a useful and reliable format, and subsequently load it into systems accessible to end users for problem-solving and data analysis.

Fig. 1 presents the architectural diagram of the agricultural data processing pipeline utilizing Google Cloud Platform [7]. Each solution component is depicted as a step-by-step process, illustrating how it processes data in streaming mode [8].

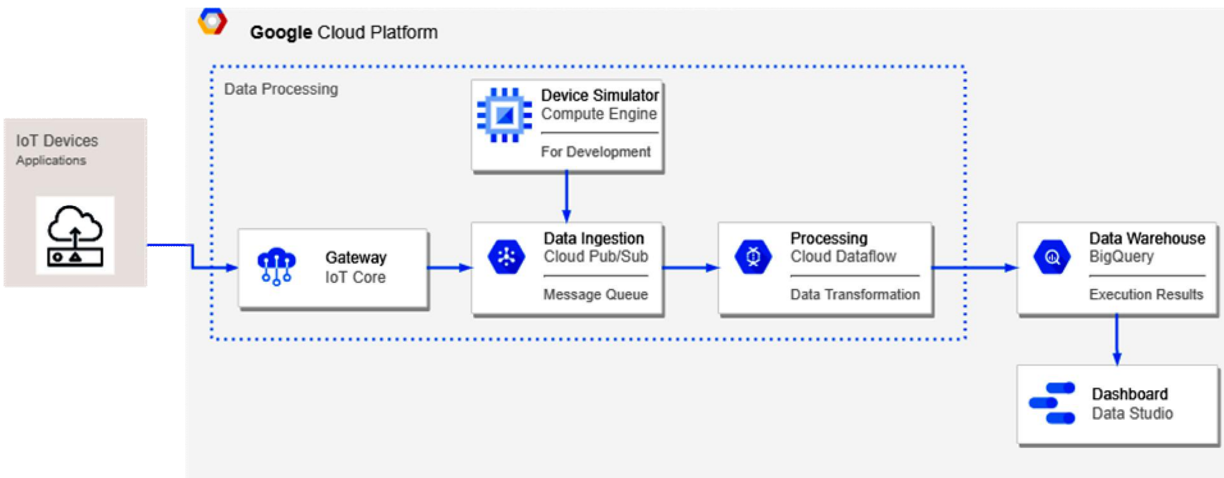

Fig. 1. Architectural diagram of the data pipeline

The pipeline consists of multiple components with varying functionality:

• IoT Devices: These encompass the Internet of Things sensors that collect essential data on various conditions, such as temperature, humidity, and solar radiation. Additionally, they transmit crucial metadata like sensor IDs, observation time, and geodata.

• Gateway: As part of the Google IoT Core service [9], the Gateway enables the management of remote sensors through registration tools, ensuring secure communication and data transmission across the network.

• Device Simulator: To test the cloud system's development, a Python script [10] called the Device Simulator generates simulated sensor data. The script runs in a Python 3 virtual environment on a Linux virtual machine configured using the Google Compute Engine service [11]. All-access is through an SSH channel [12] connected through the Cloud Console, as the virtual machine lacks a graphical interface. The simulator is used for accelerating development and making it easier to roll out new changes and enhancements.

• Data Ingestion: This component, part of the Pub/Sub service [13], receives data messages from sensors and stores them in a queue. It acts as an intermediate element in the pipeline, ensuring system stability during real-time operations. Even if there are delays or errors in subsequent processing components, data from sensors is preserved in this queue.

• Processing: The key ETL pipeline component, hosted in the Dataflow service of the GCP platform [14], directly processes the received data. It performs data transformation from the Pub/Sub topic and writes the processed data to storage tables.

• Data Warehouse: The repository for all pipeline data, implemented using Google BigQuery service [15]. Storage is organized into datasets, each containing tables where processed data from the previous component is stored.

• Dashboard: A data visualization panel created using the Data Studio service [16]. This serves as an option for displaying resulting data and generating essential reports on the system's state.

After storing data in tables, valuable insights can be extracted by constructing appropriate plots and applying suitable aggregations. These diverse diagrams offer insights into the system's functioning by providing parameters on individual sensor states, combined metrics, and data dependencies and distributions. An exemplary graph in Figure 2 illustrates the aggregation of temperature data from all sensors. The average value gives a general system characteristic, while the maximum and minimum values highlight critical parameters. Rapid temperature fluctuations forming spikes of data might indicate environmental changes such as field ignition where the sensor temperature data may spike or lead to a failure in parameters.

In Figure 3, a pie chart displays the illumination level on sensors, enabling statistics on natural conditions across all fields or specific parts of the system using different filters. Such data can act as a metric to understand the overall distribution of light during the day allowing for appropriate decision making such as the selection of the right types of crops and estimate of their growth. The definition of levels can be manually set for specific crops to represent accurate values for each type of plant.

These types of aggregation allow you to see statistics about natural conditions in all fields in general or on specific parts of the system, using various filters. The development of more complex types of graphs is done using other techniques of data analysis [17] and the development of filtering tools [18].

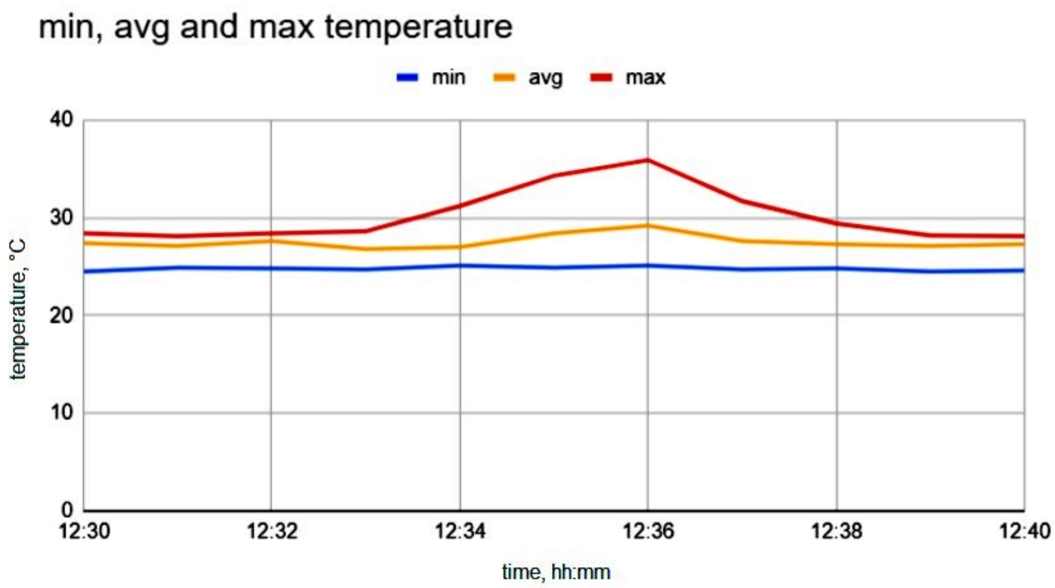

Fig. 2. Changes of the average, minimum, and maximum temperature values over time provided by sensors

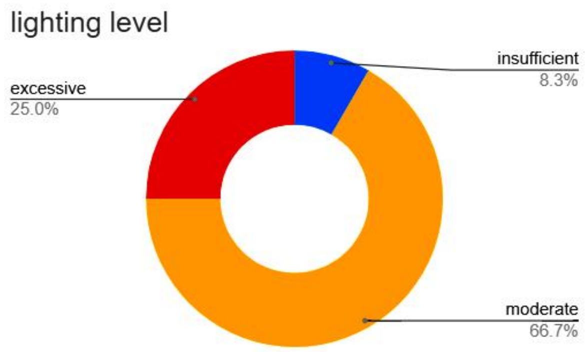

Fig. 3. Pie chart of sensor illumination level for the showcase. Illumination levels correspond to the following values: insufficient – below 2,000 lux; moderate – from 2,000 to 5,000 lux; excessive – over 5,000 lux.

The data can also be processed at wider time intervals to obtain long-term statistics and plan crops for the next seasons. Figure 4 shows a histogram of the humidity distribution that can be made by collecting data over a long period. Visualizations of this type help better understand the conditions of specific plots of land so that specialists have enough context about the weather conditions in specific fields during certain seasons to choose the right fertilizer types and make accurate amounts of crop selection.

In addition to data about general characteristics of the environment, many analytical operations can be performed on metadata – data that does not provide information about metrics but the state or placement of sensors. An example of the use of such metadata can be seen in Figure 5, where the availability of sensors is displayed using a pie chart. If certain sensors have failed or lost their connection to the network, this may indicate either their failure or some external factors that could have damaged them. Depending on the distribution of unavailable sensors at a given time, the prediction of the potential root cause can be made.

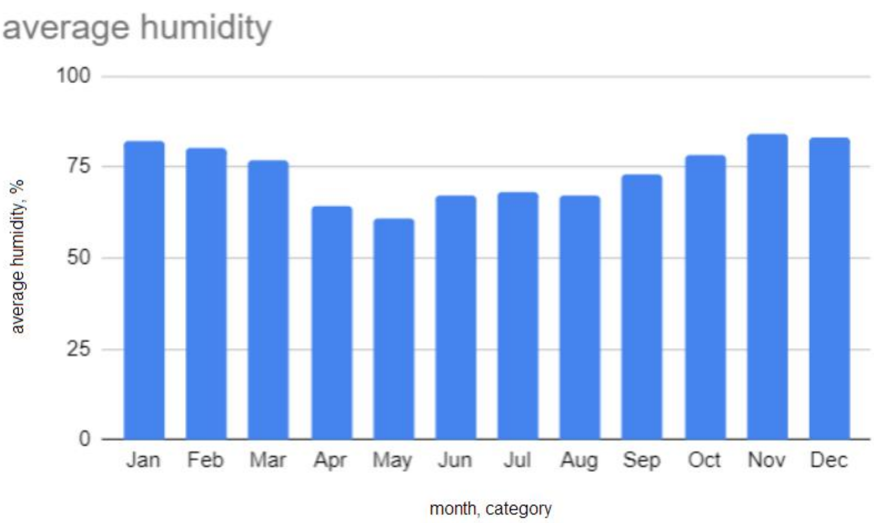

Fig. 4. Histogram of the distribution of humidity levels throughout the year

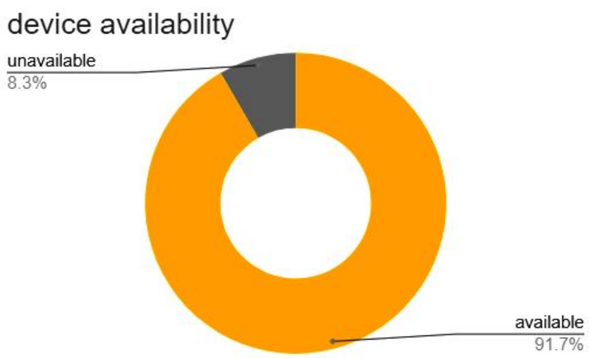

Fig. 5. Sensor availability pie chart

These data-driven approaches offer valuable answers to various business questions, providing only a glimpse of the multitude of examples that can be derived from this seemingly small set of information. The successful outcome of the cloud-based data processing pipeline lies in its ability to provide such data and deliver insights into numerous queries, making it an invaluable tool for leveraging cloud technologies to enhance agricultural practices.

The primary objective of this pipeline is to enable accurate measurement and analysis of agricultural field data. By harnessing the power of cloud computing, the data pipeline provides scalable and efficient data processing capabilities allowing for quick and timely analysis. The processed data is then utilized to gain insights into the state of the agricultural fields, including factors such as potential wildfires and droughts.

The potential of data-driven insights is vast, offering numerous solutions to tackle agricultural business questions. The examples presented in the article provide only a glimpse of the possibilities that can be derived from such datasets. The ability to access real-time data and make insights into various inquiries is a testament to the success of the pipeline data processing system utilizing cloud technologies.

The adoption of pipeline processing for agricultural data through cloud technologies marks a pivotal step toward the future of enhanced yields and stability for agricultural companies. These systems are gaining popularity among leading companies in numerous countries. Drawing inspiration from successful decisions taken by such companies in Europe and America, coverage of Ukraine's fertile land and the immense potential of utilizing cutting-edge technologies will significantly increase food product supplies both domestically and globally.

Several countries have already witnessed a remarkable 15% annual increase in agricultural land productivity, thanks to the active implementation of the Internet of Things and Big Data technologies, Leveraging data collected during planting seasons empowers specialists to better plan fertilizer use, identify fertile areas, and optimize agricultural practices.

The strength of data monitoring lies in its ability to assess various measurements across extensive fields of tens or hundreds of hectares without requiring manual inspections. In the event of wildfires or emergencies, data systems can promptly notify supervisors, mitigating business risks and enhancing operational stability during challenging times.

By developing a data pipeline, users gain the ability to address vital business expansion planning questions and conduct real-time monitoring of farmland to ensure optimal performance. Data processing technologies represent the future and are actively evolving. Many global companies are already deploying them and Ukrainian agricultural firms have the potential to be the next beneficiaries of these transformative advancements. Embracing these technologies will undoubtedly open new horizons for growth and expansion within the agricultural sector.

5. Conclusions

A prototype of a cloud-based data processing pipeline for the agricultural sector was implemented and successfully used to process data from sensors. The data warehouse is a storage of output that can be used to extract information from processed records and build visualizations as shown in the article.

The presented metrics are only the tip of the iceberg when it comes to the multitude of ways of interpreting the data. In provided visualizations, metrics such as temperature values, humidity and lighting values as well as device availability are used to showcase some of the key parameters that can be obtained by analyzing the data in real time. Such measurements are beneficial for experts in precision agriculture to do future planning and estimation of the output of a specific type of crop in the agricultural business.

6. Gratitude

The authors thank the Team of Editorial Board of the Scientific journal "Measuring Equipment and Metrology" for their support.

7. Mutual claims of authors

The authors have no claims against each other.

References

[1] E. Fukase, W. Martin, “Economic growth, convergence, and world food demand and supply”, World Development, Volume 132, 2020. DOI: https://doi.org/10.1016/j.worlddev.2020.104954

[2] S. P. Poznyak, “Black soils of Ukraine: geography, genesis and current state”, 2016. DOI: https://doi.org/10.15407/ugz2016.01.009

[3] [3] M. S. Alkatheiri, “Artificial intelligence assisted improved human-computer interactions for computer systems:”, Computers and Electrical Engineering, Volume 101, 2022. DOI: https://doi.org/10.1016/j.compeleceng.2022.107950

[4] H. E. Pence, “What is Big Data and Why is it Important?”, Journal of Educational Technology Systems, 43(2), 159–171, 2014. DOI: https://doi.org/10.2190/ET.43.2.d

[5] M. B. Hoy (2015), “The ‘Internet of Things’: What It Is and What It Means for Libraries”, Medical Reference Services Quarterly, 34:3, 353-358. DOI: 10.1080/02763869.2015.1052699

[6] [6] N. Biswas, "A new approach for conceptual extraction-transformation-loading process modeling,”, International Journal of Ambient Computing and Intelligence 10.1, 30-45, 2019. DOI: DOI: 10.4018/IJACI.2019010102

[7] A. Gupta, P. Goswami, N. Chaudhary, R. Bansal, "Deploying an Application using Google Cloud Platform," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 236-239. DOI: https://doi.org/10.1109/ICIMIA48430.2020.9074911

[8] G. van Dongen, D. van den Poel, "Evaluation of Stream Processing Frameworks," IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 8, pp. 1845-1858, 2020. DOI: https://doi.org/10.1109/TPDS.2020.2978480

[9] B. Grados, H. Bedon. “Software Components of an IoT Monitoring Platform in Google Cloud Platform: A Descriptive Research and an Architectural Proposal”, Communications in Computer and Information Science, vol 1193. 2019. DOI: https://doi.org/10.1007/978-3-030-42517-3_12

[10] Z. Dobesova, "Programming language Python for data processing," 2011 International Conference on Electrical and Control Engineering, 2011, pp. 4866-4869, DOI: 10.1109/ICECENG.2011.6057428

[11] J. Shah and D. Dubaria, "Building Modern Clouds: Using Docker, Kubernetes & Google Cloud Platform," 2019 IEEE

9th Annual Computing and Communication Workshop and Conference, 2019, pp. 0184-0189, DOI: https://doi.org/ 10.1109/ CCWC.2019.8666479

[12] F. Bergsma, B. Dowling, F. Kohlar, J. Schwenk, D. Stebila, “Multi-Ciphersuite Security of the Secure Shell (SSH) Protocol”, 2014 ACM SIGSAC Conference on Computer and Communications Security, 369–381, 2014. DOI https://doi.org/10.1145/2660267.2660286

[13] J. Gascon-Samson, F.-P. Garcia, B. Kemme, J. Kienzle, "Dynamoth: A Scalable Pub/Sub Middleware for Latency-Constrained Applications in the Cloud," 2015 IEEE 35th International Conference on Distributed Computing Systems, 2015, pp. 486-496, DOI: https://doi.org/10.1109/ ICDCS.2015.56

[14] J. Sreemathy, R. Brindha, M. Selva Nagalakshmi, N. Suvekha, N. Karthick Ragul and M. Praveennandha, "Overview of ETL Tools and Talend-Data Integration," 2021 7th International Conference on Advanced Computing and Communication Systems, 2021, pp. 1650-1654, DOI: https://doi.org/10.1109/ICACCS51430.2021.9441984

[15] D. Dzulhikam, M. E. Rana, "A Critical Review of Cloud Computing Environment for Big Data Analytics," 2022 International Conference on Decision Aid Sciences and Applications, 2022, pp. 76-81, DOI: https://doi.org/10.1109/ DASA54658.2022.9765168

[16] S. M. Ali, N. Gupta, G. K. Nayak, R. K. Lenka, "Big data visualization: Tools and challenges," 2016 2nd International Conference on Contemporary Computing and Informatics, 2016, pp. 656-660, DOI: https://doi.org/10.1109/ IC3I.2016.7918044

[17] B. Xia, P. Gong, “Review of business intelligence through data analysis”, Benchmarking: An International Journal. 21. 300-311, 2014. DOI: https://doi.org/10.1108/BIJ-08-2012-0050.

[18] M. S. Gounder, V. V. Iyer, A. Al Mazyad, "A survey on business intelligence tools for university dashboard development," 2016 3rd MEC International Conference on Big Data and Smart City, 2016, pp. 1-7, DOI: https://doi.org/10.1109/ ICBDSC.2016.7460347