Deep Science Publishing
https://doi.org/10.70593/978-81-988918-6-0_5

DeepScience
Open Access Books

Chapter 5: Designing data engineering pipelines for real-time agricultural insights

5.1. Introduction

Advanced data engineering techniques have now shown considerable potential in realizing the precision agriculture practice in crop production cycle, helping farmers take prompt and timely decisions using real-time farm data, relate it to their years of hands-on experience and monitoring, to forecast their yield and the quality of the produce. Data has also become one of the most valuable resources today in realizing realistic decision-making protocols for agriculture that draw insight from historical agricultural data, in-turn enhancing the domain knowledge with adequate modeling and analysis. All-in-all it is now possible to compute and store historical data from all seasons of a crop production life-cycle, in both on-cloud storage as well as local edge or IoT database, through advanced sensors technologies and decision-making pipelines. Further, leveraging the power of Artificial Intelligence for predictive analysis, big data tools can analyze both on-cloud and local data efficiently for providing insights into upcoming harvests (Kamilaris & Prenafeta-Boldú, 2018; Jha et al., 2019; Tsouros et al., 2019). Several data analysis frameworks also combine domain knowledge with Artificial Intelligence based models to ascertain towards improving not only the predictive yield based decisions, but also the day-to-day remedial measures to keep the yield in check. This has been further emphasized and evidenced by improvements in the farming eco-system that followed after the adaptation and understanding of the significance of data in the recent years, and the advantages that proper decisions based on right data can achieve in farming. Having adopted this approach, it is now imperative that data acquisition from modern and timely decision-making pipelines leading into accurate yields and quality of harvest, remains key to effective food production and agriculture-based research. In particular, the importance of adopting the significance of big data and data science in dealing with agricultural problems (Wolfert et al., 2017; Zhang et al., 2021).

5.1.1. Significance of Data in Agricultural Practices

Data in the context of agriculture refers to quantitative and qualitative information specific for the agricultural markets of nations, the functions of production, and the process of distribution and exchange of outputs. Since a large population of the world relies directly on agriculture and its products, the decisions taken on the agricultural markets are of wider interest for the society as a whole. Moreover, the increasing opening and globalization of different agricultural markets makes almost impossible the analysis of a single agricultural market without referring to international levels, as more and more actions taken by domestic subjects are influenced by the performance of other countries. Comprehensive, accurate, good quality and timely data are essential to inform sound agricultural investment and policy decisions.

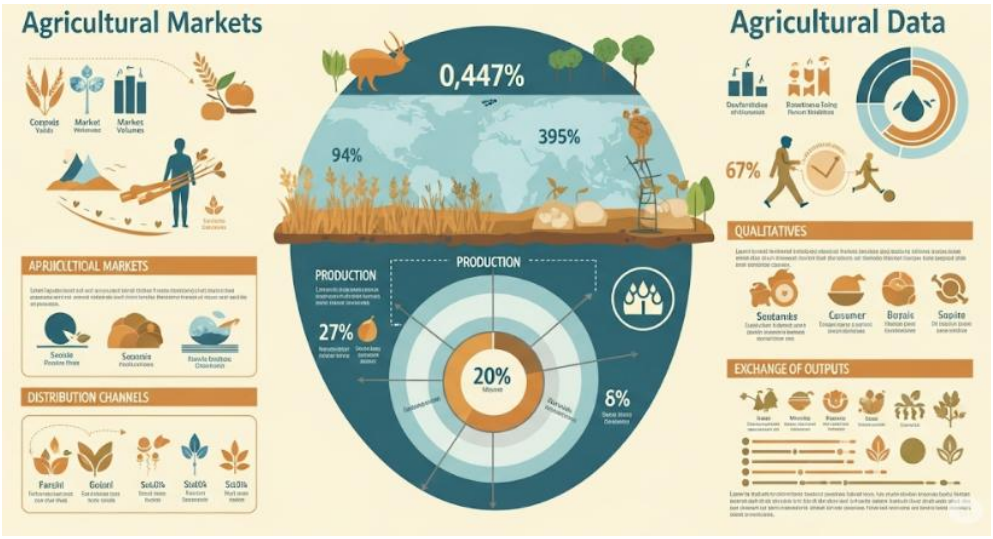

Fig 5 . 1 : The Agri-Data Ecosystem

The significance of the agricultural statistics arises not only from the social and economic implications proper to the agricultural sector but also from the specificity of its primary function. Farmers are responsible for producing the goods that are essential for human life and the agricultural output must satisfy some requirements in terms of availability, selection and prices. Both, the failure of the agricultural sector and the incorrect management of its investment decisions can produce serious consequences for the population. The preoccupation with the agricultural statistics motives a considerable body of research investigating the efficacy of these data and filing guidelines for their preparation. Moreover, the documents and recommendations prepared by international institutions in charge of formulating important decisions regarding economical and financial aspects advocate improvement in the domain of the agricultural statistics, in the terms of documentation adequacy and coordination of statistics.

5.2. The Importance of Data in Agriculture

Without farmers to provide us food, humanity would not exist. The importance of agriculture has been recognized and praised since the Time of Ancient Greece. In modern times, however, the world has seen an explosion in demand for food. Global population growth combined with a transition from subsistence to market-driven agriculture in the developing world and dietary changes in the developed world mean agriculture faces growing challenges for the coming century. Yet in part because of its early transition to modernity, labor is quickly leaving the agricultural sector in industrialized economies. A 1% annual increase in total factor productivity currently seems insufficient to meet food demand for the coming century. This has necessitated pushing the frontier of knowledge further outward to allow agriculture to meet the coming demand.

Farmers in emerging markets are increasingly turning to their own data to make the best decisions on how to develop their industry, stimulating the growth of the data-driven agriculture market, also known as precision agriculture. Using data, farmers can minimize their inefficiencies, reduce their use of fertilizers and pesticides, maximize their yield, and as a result, earn more money at less expense. In the developed world, meanwhile, the agricultural sector has matured; products and services are designed to facilitate the farmer's work. The use of data has become a habit; farmers rely on their data to drive informed decisions, and optimizing data operations has become a priority.

Data is only as good as what is done with it, though; in general, the more accurate and relevant the data, the better insight it can provide. For farmers, the implications of poor decisions based on faulty data can be disastrous; whole seasons can be lost by suboptimal planting, failure to treat pests, or harvesting too soon or too late. In addition, for farmers to equip themselves with data-driven agriculture, precision technologies and systems need to be applied across the diversity of non-traditional farming landscapes. This is particularly necessary for developing countries, which face rising world commodity prices yet have minimal support for their subsistence farmers.

5.2.1. The Critical Role of Data in Enhancing Agricultural Outcomes

Agriculture is undeniably a multi-observable system, consisting of both above-ground and below-ground resources. These light-exposed crops or trees consist of physical organisms that can be further classified as biological systems comprising microorganisms, living bushes or trees. These are viewed as valuable resources for sustainable agriculture. The below-ground systems are comprised of organic compounds consisting of carbon, hydrogen, and some oxygen. These compounds undergo continuous interaction in the nitrogen, carbon, phosphorus, silica, and sulphur cycles

throughout the environment. Additionally, above-ground agriculture must contend with weather conditions such as radiation, temperature, precipitation, air pressure, and humidity. Thus, massive fluctuations in annual yields are attributed to the evolution of climate change. The extraordinary increase in the population in the twentieth century, up to today, has resulted in extensive advances in agricultural and livestock production. Advanced technologies in the field almost doubled agricultural production in several countries by stimulating the use of synthetic fertilizers, agrochemicals, modified seeds, and irrigation systems. The global demand and food security challenge, on the one hand, and the issues with sustainable development and environmental impacts, on the other hand, are the two pillars of the current and future convergence of science and technology in fields related to food, nutrition, health, and agriculture.

In that regard, data and knowledge are envisaged as the essential milestones by which agricultural development policy and research become substituted by a conscious control mechanism. Education is at the center of the transition to a knowledge-based economy, and the ICT potential must be completely utilized to sustain the adaptation process of agriculture and rural areas in a knowledge-based economy. However, the economists are forbidden from classifying agriculture in the ordinary market economy. The results of economically evaluating these investments in knowledge are reflected in the tension between present and future possible rewards.

5.3. Overview of Data Engineering

Data engineering is a critical foundation of data-centered organizational structures across a variety of technologies and industries. Data drives the decisions made in all aspects of business operations - deciding which new products to introduce to which markets based on customer sentiment analysis, understanding how climate change affects agricultural output, using global weather data to manage disaster and famine relief budgets, deploying ad revenue based on audience views. Data engineering is a set of processes and systems used to acquire, store, and prepare data for downstream analytics, machine learning, and reporting tasks. The primary goal of data engineering is to create performant, maintained, reliable data systems and pipelines. Data engineering is itself a very large space, as the term data engineering refers to quite a few tasks. The typical tasks that data engineers use time to work on include building and maintaining data pipelines that move data from one or more sources into storage, orchestrating the movement of data from one storage system to another, and preparing and cleaning data in storage to create datasets for other people to use. More formally, data engineering broadly refers to the data movement, transformation, and storage systems required to build a data model of the world. Businesses ingest data through various methods from a variety of sources and combine these data sources and prepare them for modeling and

analysis based on their needs. Steps for working with data at a high-level begin with extracting data from various sources, confirming the correctness of the data, transforming the data so that it is in the proper format, configuration, and codes for downstream use cases, storing data into appropriate warehouses or storage systems, and then performing analyses on that data system using dashboards, reports, or other data-driven systems that use machine learning.

5.3.1. The Role of Data Engineering in Modern Agriculture

Agriculture generates enormous amounts of data each year. The agricultural data value chain is currently in a phase of maturity in which more and more integrated technologies make it possible to extract more value from this data. Formerly, companies simply sold sensors and variables from which relevant measures could be deduced: humidity, temperature, etc. to estimate the efficiency of a given crop. Lately, and in many sectors, the response has developed to a different level of data granularity: from a varietal point of view, genomics data is provided to provide a genotype-level resolution. Data on molecular markers are also developed for precision agriculture to guide farmers in their decision-making. These data are complemented with additional data from complementary biochemical tests within intelligent and interconnected platforms. These tools are integrating in addition to irrigation systems with a reservoir of moisture and water needs at root level synchronized with climate data, drone monitoring do provide with RGB, multispectral, and hyperspectral images highlighting areas of stress and progressive monitoring of pollen drift during flowering to detect blows of pollen by the identification of pollen farm markers.

Additionally, recent research proposes various statistical approaches to minimize the errors in the estimation of crop parameters and models and the importance of developing private-public partnerships is discussed to optimize the generation of climatic and satellite remote sensing services that are useful for the prediction of pest outbreaks in one region. Emerging technologies are expected to cover the critical aspects of pest monitoring, such as real-time pest detection and classification or weather monitoring at low costs. Unmanned aerial vehicle systems carrying multispectral cameras are now insuring costs and flexibility in pest monitoring.

5.4. Real-Time Data Processing

Real-time data processing refers to the continuous input and processing of a never-ending stream of data, which shows the state of the world at any given point in time and has to be processed with minimal latency. This is different from traditional data processing, which acts on a batch of data to generate results, whose latency can be high.

For example, if a final yield estimation is done based on the harvest data, that might be performed after months of the crop growing season, with the latency to produce that yield estimate based on data actions spanning that amount of time. Real-time processing generally uses a stream processing framework to capture the continuous stream of events and perform computations on the events as they flow through the different stages of the pipeline. The traditional processes in the field of data engineering have focused on enabling processing after collection of the data, preparing a storage or data lake layer, and batch processing using high-performance computing hardware to produce useful insights. However, for the use cases in the area of precision agriculture, developing a real-time data pipeline is important to produce timely and actionable insights. The variation across different locations and sensors adds to the volume and variety of the data being generated while the nature of the events being sent and the high stakes involved require a low latency and high availability of the pipeline, which are classical challenges in the development of real-time data engineering pipelines. There is existing work in building scalable stream data pipelines for different event types, including example use cases in healthcare and social network analysis.

5.4.1. Definition and Importance

Real-time data processing refers to the need to process input data and produce the intended output in a very short time window. The exact time window will vary depending on the application and use case. For example, in a stock trading application, several milliseconds may be far too long. In other applications, such as critical alerts on an intelligent transport system, a minute may be within an acceptable range. In the context of the current topic, which is more realtime focused, we adopt a definition of real time as a matter of seconds. The reason for needing real-time data processing in some applications is due to new incoming data providing new insights that can lead to better optimization, demand, ordering, and allocation. In addition, providing urgent alerts will prevent resource wastage, population discontent, and prevent accidents leading to loss of human life or assets.

Emerging intelligent systems around the globe across various domains are producing trillions of data points regularly in a variety of formats. In the transportation domain, for example, sensors provide hidden relationships between vehicle speed, weather conditions, and pollution levels across a city. This clustering of information exposes the demand for real-time insights necessary in order to enable quicker turnarounds. Several factors, including vehicle connectivity, social media support, low-cost sensors, increasing user expertise, and sensor incentive schemes, have led to an explosion in the volume of generated data. This data is also growing in variety and has also seen changes toward the veracity of existing data coming across other identified systems, given that

large amounts of it originate from low-cost, volatile sensors. Some of these support being physically located or available on mobile devices used to connect to wireless networks.

5.4.2. Challenges in Real-Time Processing

The need for timely results is the main obstacle in real-time data processing. Solutions differ on how they address the need, mainly by accepting small latencies, or complex live query management with the danger of losing freshness, correctness, or data integrity at query time. Real-time processing enables the generation of decisions based on actual and relevant ground truth. In this context, the main problems addressed are latency, correctness, data throughput, and resource management efficiency.

The notion of low-latency computing is variously quantified from systems offering one switching delay to several hours up to one day of end-to-end delay. The integration of batch and stream processing provides convenient frameworks for solving data-intensive problems with large training sets and strict prediction time limits. Adding low-latency and real-time batch-and-stream pipeline is the only way to enable data-driven predictive analytics.

However, not all data is equal – something that stream practitioners have learned over a decade of production system usage. The first thing to remember is that facts and events of interest in streams are, at best, semi-structured. Traditionally, fixed-format binary encoding has been avoided because bytes are difficult to debug, rely on formats that may not change often, and may prevent high-performance processing. Further, the variety of sources and speed of change suggest a schema-less solution. Although strings are easier to debug because they are text, generic parsers for flat files, CSV, or XML, JSON in streams don't scale well to high data and service rates. Finally, the absence of metadata-based event descriptions doesn't allow for efficient polling of only a few of the file data elements. The one success of semi-structured "data" has been web crawling, where the metadata is the URL.

5.5. Data Sources in Agriculture

While there are several data sources to be explored, some are more common than the others because of their operating and other advantages. In this section, we will introduce the data sources that we believe have applications in precision agriculture. Since we are focusing primarily on the APM model, to bias our annotations, we will focus more on the data and its availability as compared to the machine learning models.

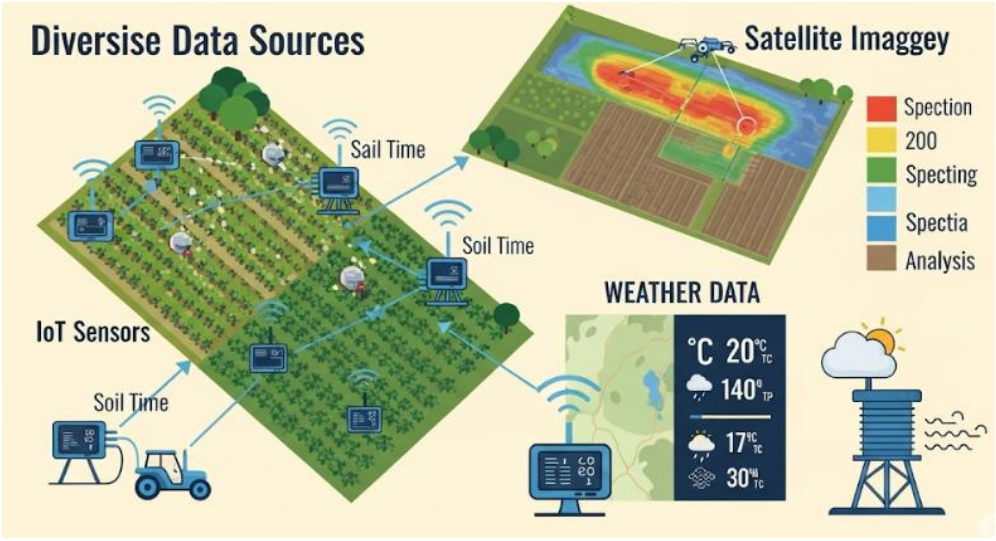

Fig 5 . 2 : Precision Ag Data Streams

1. IoT Sensors

Internet of Things (IoT) technology refers to instruments and sensors that are used for measuring and predicting various agronomic variables. IoT has become an integral part of agriculture partly due to the decreased costs of implementing IoT and Wireless Sensor Networks. IoT sensors are capable of collecting several use case-specific features at a high spatio-temporal resolution. For instance, some remote sensors can be installed on farming equipment/vehicles for tracking movement patterns. Using IoT for precision agriculture has gained momentum due to the advantages it offers, as false data can be detected, and actual events can be processed, which allow farmers to avoid crop loss at the harvest stage. However, a limitation to the use of IoT sensors is also that they often have a limited scope (only a single weather condition, or soil nitrogen indication).

2. Satellite Imagery

To assist in tasking and transitioning of crops, large satellite constellations can be tasked to assist in determining what crops are grown and at what maturity level. By focusing on the features extracted from data obtained from satellite constellations, researchers have showcased the capability of effectively determining and monitoring crop growth stages, as such methodologies are capable of covering large areas and are accurate. Multispectral satellite imagery can also help improve crop health for agricultural land. However, unlike IoT, the satellite images face high costs, longer data acquiring processes, and segmentation issues.

3. Weather Data

Weather is one of the most important features for a large number of agricultural tasks. They impact the modeling performance. Weather data can either be live feeds or can be obtained from data repositories. These repositories also allow researchers to create different models to test or restructure them according to their requirements. The limitations to weather data are that they often cannot cover at a high granularity using satellite data.

5.5.1. IoT Sensors

Data-driven Agriculture 4.0 is being considered a game-changer in developing sustainable agriculture practices, given the extensive data collection capability of emerging technologies such as Internet of Things, robotics, and artificial intelligence. However, these data sources need to be comprehensively understood in order to implement data-driven practices in agricultural supply chains. Sensor networks are among the direct sensors that collect a plethora of data, capturing different aspects of the agriculture supply chains. While sensors have been used extensively at the field and farm-monitoring levels, the techno-economic viability for deploying such systems in post-harvest supply chains is still not entirely clear. This chapter explores the different aspects related to the use of sensors for the food-supply chain.

IoT devices connected through Wireless Sensor Networks consist of a high density of sensor nodes monitoring a certain environment. Low cost, low power consumption, low weight, small size, precise detection ability, and easy deployment make Wireless Sensor Networks a valuable solution in many applications. Wireless Sensor Networks have been used in sensing applications, including battlefield surveillance, natural disaster management, and environment monitoring. Wireless Sensor Networks have seen an applications boom in agricultural data collection, including crop monitoring, soil monitoring, water management, livestock monitoring, and precision horticulture. Sensors deploy Wireless Sensor Networks designed specifically for agriculture are capable of capturing soil moisture, soil macro- and micro-nutrients, solar radiation, humidity, ambient, foliage and leaf temperatures, crop height, and other relevant information for automation and decision-making in agriculture.

Many researchers have integrated Wireless Sensor Networks with decision-making algorithms to develop smart sector-specific applications in agriculture. Large-scale sensor networks have gained importance in Smart Agriculture 4.0, which is considered the fourth evolution stage of smart agriculture. Smart Agriculture 4.0 aims to achieve optimization while meeting the economic, environmental, and social sustainability goals through paradigm shifts with respect to the widgets employed, production techniques used, and dimensioning of food systems. Such techniques can improve water, land, and capital efficiency; reduce greenhouse gas emissions; enhance product quality and safety;

increase net economic returns; improve the welfare of farmers, workers, and consumers; protect biodiversity, landscapes, and cultural heritage; and ensure food security and safety.

5.5.2. Satellite Imagery

Unmanned Aerial Vehicles (UAV) and IoT sensors are increasingly becoming a common source of data to supplement, corroborate, and enhance satellite imagery data. The data collected in combination with the Earth Observation data is private, often has more capability for addressing the hard problems of tissue testing detection prior to harvest but is much more costly if the task has to be repeated over time and space. The specialized sensors used only take measurements during the flowering and pod fill periods which are the dates of interest and need to match with satellite images. For oil palm detection to be used, timeliness is one of the key concerns for both the remote sensing and the onsite methods used. UAV data is a good alternative and is affordable for a particular site area. Satellite imagery is one of the best sources of Earth observation data for providing continuous and uniformed spatial-temporal data and has been available for long historical time periods at a global scale. The optical image quality is sufficiently high to be able to distinguish between oil palms from the traditional forms of agriculture. Satellite data provides information on a variety of different agricultural features over a period of time. Satellite imagery has particular advantages and disadvantages when comparing to UAVs and in situ sensors. Satellite imagery is restricted from cloud cover but does have some edge processing and cloud detection and masking capabilities available, especially in the commercial satellite area.

5.5.3. Weather Data

Weather data plays a significant role in agricultural operations, such as planting, irrigation, and harvesting. Farmers make millions of decisions each season that could be based on the weather data. Phenomena such as frost can lead to huge losses for farmers, while heaving rain, hail, and winds can create crop damage and losses during crucial crop stages of production, which can lead to those crops being rejected at the market point. Precise weather data can help farmers improve the timing of cotton and wheat planting, as well as peanut irrigation. It has also been documented that potato yield can be positively influenced by capturing the historical temperature information during the main growth season using weather data. Farmers are increasingly resorting to the use of weather and climate information in making decisions, which have resulted in higher crop yields.

Several studies have shown a positive relationship between agriculture production and weather patterns and have explored the impact of extreme weather on agricultural production. The links between agriculture production and weather data have typically been identified using micro-level data merged with large numbers of weather observations. Using this more detailed merged data, micro-level econometric models can be applied to estimate the crop growing season. For each observation, weather data at each grid cell for the time period of interest can be downloaded, defined, and summarized. At a roughly hemispherical scale, similar but less detailed weather data for agricultural production can be generated using meteorological models. Weather information is most often needed at a fine time scale, specifically during the vegetative and other specialized growth stages. Hence, using historical weather data can improve agricultural production by quickly predicting and delivering real-time warnings for disasters and natural hazards at critical times. However, farmers have trouble accessing real-time weather data for their specific locations.

5.6. Data Pipeline Architecture

While the previous chapter elaborated on the steps needed to ingest data for real-time agricultural insights, a core question remains: how should data pipeline architecture be designed? This chapter presents the data pipeline architecture and discusses the trade-offs considered to optimize it to power the use cases in the previous chapter. Based on the type of pipeline, the data pipeline architecture makes different assumptions on latency as well as consistency and fault-tolerance guarantees.

Data ingestion pipelines are usually designed for either batch or streaming processing. Batch processing is designed around reducing operating costs by combining work into larger batches. With batch systems, data at rest is periodically processed in large batches while at times there may be little or no access to processed data. In contrast, streaming processing is designed around making data available for processing as soon as it arrives. In practice, streaming systems take actions on new data as it arrives, in smaller batches, but much closer in time to when data enters the system. With the low cost of operating servers, batch jobs are also increasingly being scheduled to run every few minutes instead of hours or days.

The integrated approach followed is to weaken the consistency guarantee of stream processing. Many systems rely on idempotent operations and making some writes to fault-tolerant caches in order to support popular techniques such as delayed job triggering while achieving high levels of parallelism. Alternatively, streaming ingestions can publish coarse, aggregate events that batch consumers receive and process at a deterministically predictable trigger interval. In addition to being less resource-intensive,

these workflows often produce cleaner results because both batch and stream operations in the system are designed to run with much more similar workloads.

5.6.1. Batch vs. Stream Processing

When designing a data processing pipeline, the first architectural design decision is whether to use batch or stream processing. Batch processing relies on the ingested data being collected into batches before being processed. Therefore, batch processing cannot provide fast near real-time data insights. Stream processing processes the incoming data serially and typically provides low latency, microsecond-level intermediate, and final insights. Stream processing is a better design decision when the data insights need to be cataloged in final form as quickly as possible, typically in the seconds to minutes time range after arrival. Many data processing systems need to take advantage of both batch and stream processing, since very often people want final answers but also want those answers as quickly as possible.

One way to think of batch and stream processing is that batch processing might answer the question of whether any crops were affected by an observed early spring frost. The final batch insights would come out days to weeks later after the damage from the frost is assessed. Stream processing could provide an answer to the question of whether any crops were affected by an observed early spring frost, that would be delivered minutes to hours after the frost occurred using nearest neighbor search on satellite images to assess the newly visible leafy tops of each crop. Closely timing the query with the satellite overpass and only thinking about crops nearby the thermometer monitoring the frost event would allow if not an accurate answer, at least a low-latency one.

5.6.2. Microservices Architecture

A microservices-based architecture provides us with a modular data engineering pipeline where we can orchestrate services for specific tasks in a sequence or in parallel. Choosing the right architecture is essential; a modular architecture reduces friction in model development and enables us to scale independently based on demand. The components in our data pipeline are fairly independent of each other but have clear inputs and outputs. For the current problem, we develop several microservices to ingest data from IoT sensors, store the data in a staging area, process the data for raw insights, and finally, develop automated models that can classify the tasks that are to be run on the field.

We have built these data services using a serverless computing service that allows code to be deployed as functions, triggered by an event, without worry among developers

about the environments, scaling, or failover. A wide variety of services can trigger a function, such as an event in a storage service, a change in a database, an event in a queue, a cron job, or an event-driven by an external API call. With no maintenance, simple resource-based policies for security, and a charge-per-execution billing model, the service is particularly suitable for agile environments and start-ups. The functions are stateless, and scaling is instantaneous; multiple functions are executed in parallel on different servers. Multiple triggers through different events can be defined for the same function, and the functions can be composed in a simple linear work-chain (in case of file processing) or complex workflows with possible parallel execution. Our architecture is modular so that we can decouple billing based on class of event and independently charge for these.

5.7. Data Ingestion Techniques

Data ingestion is the operation that allows us to collect and import data into a pipeline, system, or database. Before being analyzed, data should be ingested into a specific platform. These operations can be carried out either in a batch mode or in a streaming mode. Batch operations correspond to extract-transform-load methods in data engineering paradigms. The data is collected and stored in the data origin, periodically loaded into the pipeline, and transformed before consumption. ETL processes are the most common and historical method. Streaming operations correspond to continuous processing of data feeds, where sources generate and publish events that are automatically consumed and analyzed by the system. Stream-based architectures are core components used for real-time processing in modern data engineering paradigms.

As mentioned above, ETL processes are responsible for data extraction and preparation processes in a data pipeline. Extraction is the process of retrieving, collecting, and writing data from a source to a destination. In a data processing framework, it represents loading raw data into distributed file storage or databases. Data is extracted from traditional sources and structured data in tables. Preparation is responsible for cleaning and transforming the data in a data processing pipeline. The output is either raw or processed data used in later consumption tasks, such as machine learning inference and visualization. ETL operations are predominantly used in classical data engineering frameworks, but are still important tools for new data-centric pipelines. Modern data engineering pipelines also manage data betting, a new class of processes used to continuously refine data that consumes greater computational resources than the actual analysis tasks.

5.7.1. ETL Processes

Data Engineering (DE) pipelines become more complex the more disparate the sources of data become. They become even more complex if processing must be done in real-time. Thus, we turn to another widely used DE technique, the ETL processes, which we will explore in this section and which allows us to mitigate the pressure of burdening real-time processing for some of the data we want to integrate and which flows closest to an analytic procedures for agricultural monitoring and intervention. In our case, an ETL process is used to ingest external weather data to train agricultural models.

Although the philosophy of ETL processes has been around for many years now, probably equal to the origin of the concept of Business Intelligence, which ETL processes aim to support by collating and preprocessing data from disparate sources, ETL tools and frameworks have become vastly popular recently. They have gained implementational maturity in the last decade, and many platforms providing this functionality have appeared. Popular tools include various data integration solutions. Salient characteristics of these frameworks are usually:

1. Prepackaged functionality: ETL tools are basically function libraries that help you with reading from disparate data sources and writing to disparate targets, transforming the data along the way.

2. Low-code or no-code approaches: Most popular ETL products of all types offer user-friendly graphical drag-and-drop interfaces that make it easy to create and deploy complex data workflows. This saves time compared to hand-coding these applications using a general-purpose programming language.

5.7.2. Real-Time Data Streams

Designing reusable and extensible ETL processes requires embedding data processing logic within specific function calls that have well-defined configurable parameters, and pushing data from disparate sources into a queue-like system. While traditional ETL processes are heavily time-bounded, allowing periodic execution, increasing amounts of event data being collected have necessitated a shift to real-time data streams as one of the primary ingestion mechanisms. Batch ingestion results in data latency since data is not immediately available for insight generation. For example, if there are 100,000 telemetry messages coming every minute from the sensors deployed on a farm, a batch ingestion may have an input time window of 5–10 minutes and result in an early dataset with 500,000 records. While this sample dataset may be used to develop a model to estimate the stigma flower number and understand the correlation of environmental factors towards pollination at the sample time, it cannot be used to predict the stigma

flower number at the 1000-hour minute mark, which might contribute to a significant proportion of the yield logistic equation.

However, an increasing number of cloud IoT services enable more dynamic data capture. These services allow components to subscribe to a given stream of incoming data based on a set of criteria that collaborates directly with the sensor deployment. This allows for data-driven direct triggers on microservices deployed on the cloud and functions that call insight generation models for a specific region directly. Since there is a logic in the cloud that can be invoked, the data that is triggered is pushed efficiently at real time at the point of time when the event occurs. The pulled data can be used to monitor the operations and check for any numeric anomalies. Data latencies might still be an issue, depending on the actual services deployed.

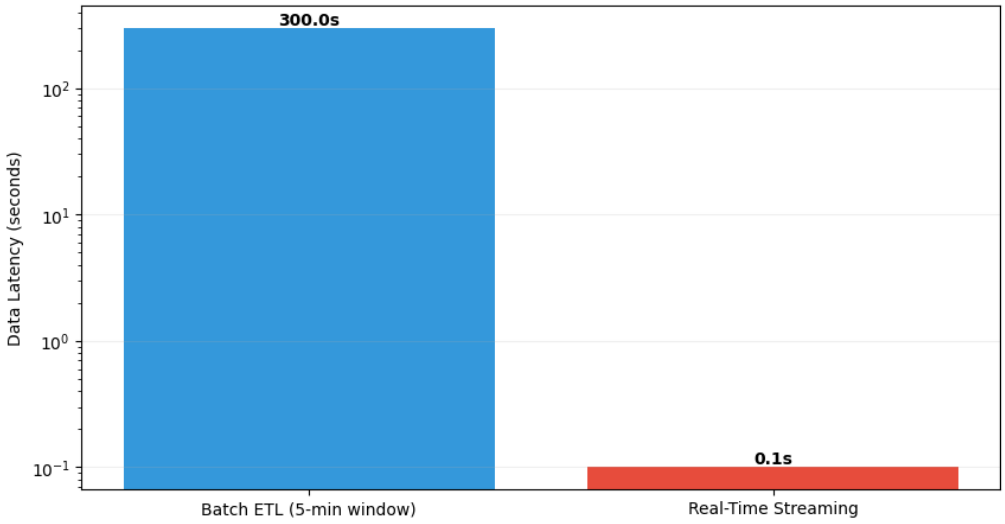

Fig 5 . 3 : ETL Approach Latency Comparison

5.8. Conclusion

This paper presented an approach to automate the collection of, and access to, agricultural and environmental data for real-time monitoring and decision support. In developing our work, we draw insights from the socio-technical systems perspective. Designing and deploying a cloud-based data engineering pipeline for agriculture necessitates consideration of on-the-ground realities of local and regional agricultural systems. The study explores the opportunities and details around sourcing agricultural and environmental data from diverse and creative sources, design and development

considerations for data pipelines in the cloud, and the trade-offs around data ownership and sharing.

The paper contributes to data-centric development by proposing a pipeline for data interoperability. How agricultural and other natural systems are observed and understood is increasingly reliant on data. New opportunities to create high-dimensional datasets grow in both size and social significance as computer-, satellite-, and sensor-based observations of the world gather speed and geolocation precision, and new methods for modeling systems and classifying the unstructured elements within them are disseminated widely. Yet translating this flood of data into durable and readily sharable codebooks and ontologies is a task frequently hindered by the same challenges that plague data science itself. It is this challenge of sharing data and understanding the meaning of the data, especially the hidden context around utilizing particular types of data in understanding systems accurately, which forms the motivation around developing a simple and minimally technical solution for addressing the challenge of data interoperability in discipline-specific domains. In conclusion, we believe that the application of large amounts of data will be the future of agriculture.

5.8.1. Final Thoughts on the Impact of Data in Agriculture

Data is revolutionizing the field of agriculture. It has been a challenge to predict crop yield and how climatic as well as non-climatic variables affect agriculture growth across various regions. Achieving data-driven decisions for agricultural management could not only help in increasing crop yield but could also help in deciding the cycle of crop cultivation and help with precision irrigation resources in areas of scarcity or drought-like weather. Using data it is also possible to predict soil moisture and monitor water resources cycle that could greatly help in agricultural management practices. Sensors for soil and environment, satellite and aerial coverage have been strategically deployed to capture relevant information such as weather, soil type, moisture levels, temperature. With the present research that shows the impact of climate on crop yield, across multiple crops, because of these studies being temporal in nature as more and more data is collected what is today science can become a practice for society tomorrow.

All said and done, the socio-economic and cultural impact of agriculture is huge in our societies just like in every other society across the globe. With the population forecast to cross about 10 billion by 2050, in countries like India, China, Brazil and Japan for example such practices can help enhance sustainable development practices apart from increasing crop yields. Ensuring food security and eliminating poverty is of utmost importance. The central question we leave the readers with is that do we want to wait for science to validate the decisions or allow the developments in technology to take centerpiece and possibly become the core of data-driven decision-making practices

across agriculture. This question is likely to have different answers across the globe but as technologies advance and pave the way to become more integrated and technologically agnostic this could likely become a reality soon.

References

Wolfert S., Ge L., Verdouw C., Bogaardt M.J. (2017). Big Data in Smart Farming – A Review. Agricultural Systems, 153, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

Jha K., Doshi A., Patel P., Shah M. (2019). A Comprehensive Review on Automation in Agriculture Using Artificial Intelligence. Artificial Intelligence in Agriculture, 2, 1–12. https://doi.org/10.1016/j.aiia.2019.05.004

Kamilaris A., Prenafeta-Boldú F.X. (2018). Deep Learning in Agriculture: A Survey. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Tsouros D.C., Bibi S., Sarigiannidis P.G. (2019). A Review on UAV-Based Applications for Precision Agriculture. Information, 10(11), 349. https://doi.org/10.3390/info10110349

Zhang Y., Wang G., Wang J. (2021). Applications of AI and Big Data in Smart Agriculture: A Review. IEEE Access, 9, 75999–76020. https://doi.org/10.1109/ACCESS.2021.3083991