RUDN University

Faculty of Physics, Mathematics and Natural Sciences

Foreign Languages Department

Professional translation

Amine Roukh, Fabrice Nolack Fote, Sidi Ahmed Mahmoudi, Saïd Mahmoudi, Faculty of Engineering - ILIA / Infortech University of Mons, Belgium. Procedia Computer Science. Volume 177, 2020, Pages 78-85 https://doi.org/10.1016/j.procs.2020.10.014

Выполнил

Студент группы НФИмд-01-22

Логинов Сергей Андреевич

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phrases | dictionary | 10 sentences | 3 questions | translation | Σ |
| 2 | 1 | 1 | 1 | 1 | 6 |
|  |  |  |  |  |  |

**Dictionary**

|  |  |  |
| --- | --- | --- |
| word | transcription | translation |
| Internet of Things (IoT) | [ˈɪntɜːnət əv θɪŋz] | Интернет вещей |
| Cloud computing | [klaʊd kəmˈpjuːtɪŋ] | Облачные вычисления |
| Big data | [bɪɡ ˈdeɪtə] | Большие данные |
| Layer | [ˈleɪə] | Уровень/слой |
| Perfomance | [pəˈfɔːməns] | Производительность |

**Sentences for translation**

|  |  |  |
| --- | --- | --- |
| № | Предложение | Перевод |
| 1 | In the era of Big data, data-driven farming is changing the agricultural businesses thanks to the use of modern technologies such as the Internet of Things (IoT) sensors, drones, and farm monitoring. | В эпоху Больших данных, сельское хозяйство, основанное на данных, меняет сельскохозяйственные предприятия благодаря использованию современных технологий, таких как сенсоры интернета вещей (IoT), дроны и мониторинг фермы. |
| 2 | IoT devices produce a massive amount of precious agri-data, which are collected and analyzed in real-time using innovative application tools. | Устройства IoT производят огромное количество ценных сельскохозяйственных данных, которые собираются и анализируются в режиме реального времени с использованием инновационных приложений. |
| 3 | Smart farming, enabled by the combination of technology, allows stakeholders in the agri-ecosystem to monitor crops in real-time and maximize productivity and profitability with minimum efforts. | Умное сельское хозяйство, обеспечиваемое комбинацией технологий, позволяет заинтересованным сторонам в агроэкосистеме отслеживать урожаи в реальном времени и максимизировать производительность и прибыльность с минимальными усилиями. |
| 4 | Many Smart farming solutions have been introduced, but universal applicability for other farms is not feasible due to non-standard Big data processing architecture. | Было представлено множество решений для умного сельского хозяйства, однако их универсальность для других ферм невозможна из-за нестандартной архитектуры обработки больших данных. |
| 5 | The paper proposes WALLeSMART, a cloud-based Smart farming management system, applied to the Wallonia region of Belgium, which addresses the challenges of acquiring, processing, storing, and visualizing large amounts of data. | В статье предлагается система управления умным сельским хозяйством на основе облачных технологий под названием WALLeSMART, применяемая в регионе Валлония в Бельгии, которая решает проблемы сбора, обработки, хранения и визуализации больших объемов данных. |
| 6 | An initial prototype of WALLeSMART has been developed and tested with various farms, showing prominent results. | Был разработан и протестирован первоначальный прототип системы WALLeSMART с использованием различных ферм, что показало заметные результаты. |
| 7 | Precision agriculture and smart farming are necessary to cater to the growing population, provide livelihoods for farmers, and protect the environment. | Точное сельское хозяйство и умное сельское хозяйство необходимы для удовлетворения растущей численности населения, обеспечения средств к существованию для фермеров и защиты окружающей среды. |
| 8 | The emergence of IoT and cloud computing technologies is pushing smart farming to the next level, with agricultural sensors expected to reach 12 million installations globally by 2023. | Появление технологий интернета вещей (IoT) и облачных вычислений поднимает умное сельское хозяйство на новый уровень, и ожидается, что сельскохозяйственные сенсоры будут установлены более чем на 12 миллионов объектов по всему миру к 2023 году. |
| 9 | Big data in agriculture provides in-depth insights about the entire agri-ecosystem, enabling predictive algorithms and advanced analysis techniques for proactive problem-solving. | Большие данные в сельском хозяйстве обеспечивают глубокое понимание всей агроэкосистемы, позволяют использовать прогнозные алгоритмы и продвинутые методы анализа для превентивного решения проблем. |
| 10 | Managing big data in agriculture is complex, requiring new architectural conceptions and application tools for data acquisition, processing, storage, analysis, and visualization. | Управление большими данными в сельском хозяйстве является сложной задачей, требующей новых архитектурных концепций и приложений для сбора, обработки, хранения, анализа и визуализации данных. |

**Questions for answering**

**What are the advantages of using the Lambda architecture in big data processing for smart farming?**

The Lambda architecture in big data processing for smart farming offers several advantages. It provides a generic, scalable, and fault-tolerant real-time data processing architecture. It allows for the processing of both batch and real-time data, ensuring that the most recent data is available for analysis. The architecture supports incremental views in real-time processing and batch views for historical data. It enables comprehensive analysis of data by merging views from different layers. Additionally, the Lambda architecture guarantees linear scalability, fault-tolerance against hardware failures, and improved performance.

**What are the main components of the WALLeSMART system architecture?**

The WALLeSMART system architecture consists of three main components:

* Master Dataset: Data from dairy farms and weather stations is collected and transmitted to the cloud platform. Data ingestion tools capture and store the real-time messages, which are then processed and stored using storage tools such as Apache Cassandra and PostgreSQL.
* Batch Layer: Batch processing is performed using Apache Hadoop, which allows for distributed processing of large data sets using the MapReduce programming model. This layer filters, aggregates, and prepares the data for analysis.
* Speed Layer: Real-time processing is carried out using Apache Storm, a low-latency framework for stream processing. It consumes, processes, and outputs results from real-time streaming data sources, enabling fast processing of large volumes of data.

**What tools and technologies are used for data ingestion, storage, and visualization in the WALLeSMART system architecture?**

The WALLeSMART system architecture utilizes various tools and technologies for data ingestion, storage, and visualization:

* Data Ingestion: Apache Kafka is used as the data ingestion tool to capture and store real-time messages from different sources.
* Storage: The architecture uses Apache Cassandra for NoSQL data storage and PostgreSQL for relational data storage. These tools are used as output destinations to capture real-time and batch data for archiving and further processing.
* Visualization: The system employs a web-based application developed using Node.js for visualization purposes. Interactive maps are created using OpenStreetMap, and interactive charts are generated using the Chart.js library. The platform also provides a Web API, implemented using GraphQL, to allow researchers to access farms and weather data for analysis and modeling.

Translation

|  |  |
| --- | --- |
| **English** | **Russian** |
| **Big Data Processing Architecture for Smart Farming** | **Архитектура обработки больших данных для современного земледелия** |
| **Abstract**  In the era of Big data, data-driven farming is changing the agricultural businesses thanks to the use of modern technologies such as the Internet of Things (IoT) sensors, drones, and farm monitoring. IoT devices produce a massive amount of precious agri-data, which are collected and analyzed in real-time using innovative application tools. This combination of technology, known as “Smart farming”, helps various stakeholders in the agri-ecosystem to monitor crops in real-time, as well as maximize productivity and profitability in farm and business operations with the minimum efforts. Although many Smart farming solutions have been introduced, both from industry and academia, universal applicability of these approaches for other farms, unfortunately, is not feasible. Most of these solutions are based on a home-made non-standard Big data processing architecture. In this paper, we propose WALLeSMART, a cloud-based Smart farming management system, applied to the Wallonia region of Belgium. The framework introduces a general architecture to address the challenges of acquisition, processing, storing, and visualization of very large  amounts of data, both in batch and real-time basis. An initial prototype has been developed and tested with various farms showing  prominent results. | **Аннотация**  В эпоху больших данных фермерство, управляемое данными, меняет сельскохозяйственный бизнес благодаря использованию современных технологий, таких как датчики Интернета вещей (IoT), беспилотники и мониторинг ферм. IoT-устройства производят огромное количество ценных агрикультурных данных, которые собираются и анализируются в режиме реального времени с помощью инновационных прикладных инструментов. Такое сочетание технологий, известное как "Умное земледелие", помогает различным заинтересованным сторонам агроэкосистемы контролировать урожай в режиме реального времени, а также максимизировать производительность и прибыльность фермерских и бизнес-операций с минимальными усилиями. Несмотря на то, что было представлено множество решений в области интеллектуального земледелия как в промышленности, так и в научных кругах, универсальное применение этих подходов для других фермерских хозяйств, к сожалению, не представляется возможным. Большинство этих решений основано на самодельной нестандартной архитектуре обработки больших данных. В данной работе мы предлагаем WALLeSMART, облачную систему управления умным сельским хозяйством, применимую к региону Валлония в Бельгии. Система представляет собой общую архитектуру для решения задач сбора, обработки, хранения и визуализации очень больших объемов данных.  объемов данных, как в пакетном режиме, так и в режиме реального времени. Первоначальный прототип был разработан и протестирован на различных фермах, показав  выдающиеся результаты. |
|  |  |
| **1. Introduction**  The United Nations’ Population Division reports that the world population is currently (2020) growing by approximately 81 million people each year. The median estimate for future growth sees the world population reaching 9.7 billion in 2050 [18]. This means that agriculture is faced with multiple challenges: catering to the growing number of people, providing a livelihood for farmers, and protecting the environment. Precision agriculture and the acronym smart farming will be necessary to face and solve those challenges. Precision agriculture is the concept of observing, measuring and responding to variability in crops. Whereas, smart farming extends the precision agriculture concept by adding the notions of data context, situation and location awareness to the tasks of management and decision making in real-time [21]. The emergence of the Internet of Things and Cloud Computing technologies are set to push the future of smart farming to the next level. Today, IoT-based agriculture is becoming more familiar between farmers, and smart farming is rapidly becoming the standard as a result of agricultural drones and sensors usage. Business Insider Intelligence projects there to be nearly 12 million agricultural sensors installed globally by 2023 [14]. Those machines generate huge amounts of data every second. IBM company estimates that the average farm can generate half a million data points per day [14]. Big data has proven to be fruitful in the agriculture sector, it provides all the stakeholders with in-depth insights about the entire agri-ecosystem. This can be achieved using predictive algorithms that can alert even before a problem occurs. However, the task of managing this Big data, ranging from acquisition, processing, storing, analyzing, and visualizing is very complex and therefore requires new architectural conception and application tools [11]. The complexity resides in the facts that (i) large, complex, heterogeneous data is coming from a variety of sources encountered in agriculture, (ii) data must be processed in both real-time and batch basis using high efficient tools, (iii) advanced analysis techniques such as predictive or prescriptive analytics must be applied, (iv) data analyzed needs to be visualized in user-friendly interactive Web and mobile application. Resolving those complexities might unlock the potential for full automation of the agri-food chain. In this paper, we propose a generic smart farming cloud-based framework, called WALLeSMART1, which helps farmers, researchers, and administrators managing their farms and making the right decisions. To achieve this goal, we fix four main objectives: (i) the study of the state-of-the-art big data processing architecture, (ii) the proposition of a generic architecture to the agriculture sector, (iii) the development and deployment of such solution applied our specific use case, using a real data sets coming from dairy farms and weather stations, in batch and real-time basis, and (iv) the visualization of those data via an interactive Web application. These four objectives will be discussed in the next sections.  The remainder of this paper is organized as follows. Section 2 reviews the most important works on smart farming platforms and tools. Section 3 gives a background view of big data processing system architectures. In Section 4, our methodology in developing a novel real-time architecture for agriculture is given and all its required components are  described. Section 5 highlights our implementation and deployment on real farms conducted to evaluate the quality of our architecture. Section 6 concludes the paper by summarizing the important results and suggests future work. | **1. Введение**  Отдел народонаселения Организации Объединенных Наций сообщает, что в настоящее время (2020 год) население мира увеличивается примерно на 81 миллион человек в год. Согласно медианной оценке будущего роста, в 2050 году численность населения мира достигнет 9,7 миллиарда человек [18]. Это означает, что перед сельским хозяйством стоит множество задач: удовлетворение потребностей растущего числа людей, обеспечение средств к существованию для фермеров и защита окружающей среды. Точное сельское хозяйство и аббревиатура "умное земледелие" будут необходимы для решения этих задач. Точное сельское хозяйство - это концепция наблюдения, измерения и реагирования на изменчивость урожая. В то время как интеллектуальное сельское хозяйство расширяет концепцию точного сельского хозяйства, добавляя понятия контекста данных, осведомленности о ситуации и местоположении к задачам управления и принятия решений в режиме реального времени [21]. Появление Интернета вещей и технологий облачных вычислений подтолкнет будущее умного земледелия на новый уровень. Сегодня сельское хозяйство на основе IoT становится все более привычным для фермеров, а "умное" земледелие быстро становится стандартом благодаря использованию сельскохозяйственных дронов и датчиков. По прогнозам Business Insider Intelligence, к 2023 году в мире будет установлено около 12 миллионов сельскохозяйственных датчиков [14]. Эти машины ежесекундно генерируют огромное количество данных. По оценкам компании IBM, средняя ферма может генерировать полмиллиона точек данных в день [14]. Большие данные доказали свою плодотворность в сельскохозяйственном секторе, они дают всем заинтересованным сторонам глубокое понимание всей агроэкосистемы. Этого можно достичь с помощью алгоритмов прогнозирования, которые могут предупреждать еще до возникновения проблемы. Однако задача управления этими большими данными, начиная от сбора, обработки, хранения, анализа и визуализации, очень сложна и поэтому требует новой архитектурной концепции и прикладных инструментов [11]. Сложность заключается в том, что (i) большие, сложные, гетерогенные данные поступают из различных источников, встречающихся в сельском хозяйстве, (ii) данные должны обрабатываться как в режиме реального времени, так и пакетно с использованием высокоэффективных инструментов, (iii) должны применяться передовые методы анализа, такие как предиктивная или предписывающая аналитика, (iv) проанализированные данные должны быть визуализированы в удобных интерактивных веб- и мобильных приложениях.  Решение этих сложных проблем может открыть потенциал для полной автоматизации агропродовольственной цепочки. В данной работе мы предлагаем общую облачную систему интеллектуального земледелия под названием WALLeSMART1, которая помогает фермерам, исследователям и администраторам управлять своими хозяйствами и принимать правильные решения. Для достижения этой цели мы ставим четыре основные задачи: (i) изучение современной архитектуры обработки больших данных, (ii) предложение общей архитектуры для сельскохозяйственного сектора, (iii) разработка и развертывание такого решения на нашем конкретном примере, используя реальные наборы данных, поступающие с молочных ферм и метеостанций, в пакетном режиме и в режиме реального времени, и (iv) визуализация этих данных с помощью интерактивного веб-приложения. Эти четыре задачи будут рассмотрены в следующих разделах.  Остальная часть данной работы организована следующим образом. В разделе 2 приводится обзор наиболее важных работ по платформам и инструментам интеллектуального земледелия. В разделе 3 дается общее представление об архитектурах систем обработки больших данных. В разделе 4 представлена наша методология разработки новой архитектуры реального времени для сельского хозяйства и описаны все ее необходимые компоненты.  описаны все необходимые компоненты. В разделе 5 описывается наша реализация и развертывание на реальных фермах, проведенное для оценки качества нашей архитектуры. Раздел 6 завершает статью, обобщая важные результаты и предлагая дальнейшую работу. |
| **2. Related Work**  The applications of big data in the agriculture field are numerous, we can cite for instance: precision farming, yield prediction, risk mitigation, loss reduction, supply chain management, farm-to-fork traceability, and sustainable farming [10, 8, 21, 13]. Besides, agricultural big data systems can be broadly divided into three categories: (i) advanced sensor technology systems, (ii) risk management systems, and (iii) agricultural management systems [16]. In this section, we focus mainly on the third category, since our work belongs to this type of system. Recently, initiatives are ongoing to create agricultural platforms that collect the data needed by smart farming decision support tools. They are often created by private companies or public-private partnerships [5]. Among the former, Monsanto’s Integrated Farming Systems platform collects data on information such as soil health and pest pressures and provide them for farmers. The Climate Corporation proposes the Climate FieldView platform to aggregate data of different sources in one place and provides diagnostic and applicative tools to farmers [5]. However, since 2018 only four companies dominate the market: DowDuPont, Syngenta-ChemChina, BASF, and Bayer-Monsanto due to a series of mergers and acquisitions [20]. Regards public-privates partnerships, many countries are building systems to advance the use of modern agricultural technologies. For example, Akkerweb is a Dutch web-based platform that provides access to external data sources such as weather, parcel boundaries, satellite, and data from commercials farm management system. It stores geo-referenced data, including soil maps and drone imagery. It allows combining data sources and processing of data, through a set of application modules to provide farmers decision support and recommendations [19]. Another example is the platform Barto2 in Switzerland. As a stock company, Barto brings together public and private actors to build up a smart-farming platform that also aims to digitize operational and production farms data while avoiding duplication, which speeds up on-farm processes, reduces administrative tasks.  Barto itself is based on 365FarmNet, a German-based farm management software provider. Nearly 45.000 farmers are already active on this platform in Germany, Poland, Bulgaria, Austria, and France 3. The platform operates as SaaS (Software as a service) and provides solutions for managing and recording all activities on a farm. However, those platforms are proprietary, and therefore the internal system architecture is not available. In the academic sector, many kinds of researches have  been conducted in the smart farming field. The authors of [2] from Spain, demonstrate the advantage of a tool, called PLATEM, that applies real-time decisions from data such as variable rate irrigation, and selected parameters from field and weather conditions. In [17], the authors proposed a Web-based service-oriented herd manage application by leveraging occasionally connected mobile database synchronization framework. The goal is to provide improved  accessibility, more up-to-date information and lower cost of ownership for the herd management software system. An initial prototype has been deployed and tested for over six months on a three thousand cow dairy farm in central Washington State, USA. The work of [1] presents a platform, called SmartDairyTracer, designed to monitor the state of dairy cattle and feed grain in real time, as well as ensure the traceability and sustainability of the different processes involved in the production. SmartDairyTracer uses global edge computing architecture which consists of three principal layers: IoT, edge, and business solution layers [15]. | **2. Предыдущие работы**  Области применения больших данных в сельском хозяйстве многочисленны, например: точное земледелие, прогнозирование урожайности, снижение рисков, сокращение потерь, управление цепочками поставок, прослеживаемость от фермы до вилки и устойчивое земледелие [10, 8, 21, 13]. Кроме того, сельскохозяйственные системы больших данных можно разделить на три категории: (i) передовые сенсорные системы, (ii) системы управления рисками и (iii) системы управления сельским хозяйством [16]. В этом разделе мы сосредоточимся в основном на третьей категории, поскольку наша работа относится к этому типу систем. В последнее время предпринимаются инициативы по созданию сельскохозяйственных платформ, которые собирают данные, необходимые для инструментов поддержки принятия решений в умном сельском хозяйстве. Они часто создаются частными компаниями или государственно-частными партнерствами [5]. Среди первых, платформа Integrated Farming Systems компании Monsanto собирает данные о такой информации, как здоровье почвы и давление вредителей, и предоставляет их фермерам. Корпорация Climate предлагает платформу Climate FieldView, которая объединяет данные из разных источников в одном месте и предоставляет фермерам диагностические и прикладные инструменты [5]. Однако с 2018 года на рынке доминируют только четыре компании: DowDuPont, Syngenta-ChemChina, BASF и Bayer-Monsanto благодаря серии слияний и поглощений [20]. В рамках государственно-частного партнерства многие страны создают системы для продвижения использования современных сельскохозяйственных технологий. Например, Akkerweb - это голландская веб-платформа, которая обеспечивает доступ к внешним источникам данных, таким как погода, границы участков, спутник и данные из коммерческой системы управления фермой. Она хранит данные с географической привязкой, включая почвенные карты и снимки с беспилотников. Она позволяет объединять источники данных и обрабатывать их с помощью набора прикладных модулей для предоставления фермерам поддержки принятия решений и рекомендаций [19]. Другим примером является платформа Barto2 в Швейцарии. Являясь акционерным обществом, Barto объединяет государственных и частных субъектов для создания платформы "умного" сельского хозяйства, которая также направлена на оцифровку оперативных и производственных данных ферм, избегая дублирования, что ускоряет процессы на ферме, сокращает административные задачи.  Barto базируется на 365FarmNet, немецком поставщике программного обеспечения для управления фермерскими хозяйствами. Около 45 000 фермеров уже работают на этой платформе в Германии, Польше, Болгарии, Австрии и Франции 3. Платформа работает как SaaS (программное обеспечение как услуга) и предлагает решения для управления и учета всех видов деятельности на ферме. Однако эти платформы являются собственными, и поэтому внутренняя архитектура системы недоступна. В академическом секторе было проведено множество исследований в области интеллектуального земледелия.  проводились в области интеллектуального сельского хозяйства. Авторы [2] из Испании демонстрируют преимущества инструмента под названием PLATEM, который в режиме реального времени принимает решения на основе таких данных, как переменная скорость орошения и выбранные параметры поля и погодных условий. В [17] авторы предложили веб-приложение для управления стадом, ориентированное на сервис, используя периодически подключаемую мобильную систему синхронизации баз данных. Цель состоит в том, чтобы обеспечить улучшенную  доступность, более актуальную информацию и снизить стоимость владения программной системой управления стадом. Первоначальный прототип был развернут и тестировался более шести месяцев на молочной ферме на три тысячи коров в центральной части штата Вашингтон, США. В работе [1] представлена платформа под названием SmartDairyTracer, предназначенная для мониторинга состояния молочного скота и кормового зерна в режиме реального времени, а также обеспечения отслеживаемости и устойчивости различных процессов, участвующих в производстве. SmartDairyTracer использует глобальную архитектуру вычислений на границе, которая состоит из трех основных уровней: IoT, пограничный уровень и уровень бизнес-решений [15]. |
| **3. Big Data Processing for Smart Farming**  Developments and usage of network technologies, IoT, and cloud computing in smart farming allow generating big data. Big data is described as both structured and unstructured data that is too large to be processed by traditional data processing tools and techniques. The “big” part of this term is characterized by the 5 Vs, which are volume, velocity, variety, veracity and, value [3]. Additionally, some companies contributed other dimensions to big data. This large quantity of data must be managed and analyzed to capture its full value. To this end, first, raw data from different sources need to be ingested, integrated, and stored into a Data Lake, a centralized storage repository. Second, big  data solutions must be used to process this data, based on long-running batch jobs as well as real-time jobs to filter, aggregate, and otherwise prepare the data for analysis. The last step consists in providing insights from data through analysis and reporting tasks. In general, there exist two basic requirements for big data systems. First, receiving a  massive real-time data stream from different sources and second, performing an analysis of this data to get results right away [4]. Based on this, three state-of-the-art software architectures are often used for big data systems, called Lambda, Kappa, and a Hybrid of the two.  3.1. Lambda Architecture  Lambda architecture describes a generic, scalable and fault-tolerant real-time data processing architecture [12]. It has two branches for batch and real-time processing. As shown in Figure 1a, data is provided to both branches. In the batch layer, data is appended to a storage area, called the master dataset. Usually, this data is processed later using a  batch processing tool such as Apache Hadoop, resulting in batch views. In the speed layer, data is processed in real-time using a stream processing tool such as Apache Storm, which results in an incremental view. Since processing massive data sets in the batch layer takes a lot of time, the results are usually not up-to-date. The speed layer is used to  close this gap by querying the most recent data. Finally, the serving layer merges views of the above layers, in which end-user can send his queries.  3.2. Kappa Architecture  Kappa architecture aims at simplifying the development process of the system [9]. It uses only one branch; the speed layer as shown in Figure 1b. That is, incoming data is processed immediately in real-time using only one stream  Изображение выглядит как текст, диаграмма, снимок экрана, линия  Автоматически созданное описание  processing tool. Results are saved in incremental views. Afterward, the data goes to the serving layer that queries any required results. Due to this change, the Kappa architecture supports processing the entire data in the master dataset according to needs, as well as the incoming real-time data in parallel. The advantage of the Kappa architecture is its single data processing engine, which can simplify the overall system. On the other hand, processing large data sets and real-time data streams in parallel may increase computational power and storage capacity.  3.3. Hybrid Architecture  While Lambda and the Kappa architectures support the processing of historical data and real-time data. However, the processing mode is separated. That is, the user has to do a lot of work manually to couple the results of both processing modes, in order to enable a comprehensive analysis of the data [6]. The work of [6] introduces a hybrid processing architecture, called BRAID. BRAID intertwines the processing of historical data and real-time data by adding communication channels between the batch engine and the stream engine. This enables to carry out comprehensive analyses automatically at a reasonable overhead. The authors discussed different implementation techniques for their proposed architecture. In the same direction, the study of [7] introduced a conceptual architecture for big data streaming integrated with complex event processing, called BiDCEP. The system expands the Lambda and Kappa architectures to fit the complex event processing and event management domains of enterprise IT. The authors trigger a technical discussion on the merits of combining both architectures, as well as present a motivational example | **3. Обработка больших данных для современного земледелия**  Развитие и использование сетевых технологий, IoT и облачных вычислений в интеллектуальном сельском хозяйстве позволяет генерировать большие данные. Большие данные описываются как структурированные и неструктурированные данные, которые слишком велики для обработки традиционными инструментами и методами обработки данных. "Большая" часть этого термина характеризуется 5 Vs, которые представляют собой объем, скорость, разнообразие, достоверность и ценность [3]. Кроме того, некоторые компании придают большим данным другие измерения. Этим большим количеством данных необходимо управлять и анализировать, чтобы извлечь всю их ценность. Для этого, во-первых, необработанные данные из различных источников должны быть получены, интегрированы и сохранены в Data Lake - централизованном хранилище данных. Во-вторых, большие  Во-вторых, для обработки этих данных должны использоваться решения для работы с большими данными, основанные на длительных пакетных заданиях, а также заданиях реального времени для фильтрации, агрегации и другой подготовки данных к анализу. Последний этап заключается в получении информации из данных с помощью задач анализа и отчетности. В целом, существует два основных требования к системам больших данных. Во-первых, получение  массивный поток данных в реальном времени из различных источников и, во-вторых, выполнение анализа этих данных для получения результатов сразу же [4]. Исходя из этого, для систем больших данных часто используются три современные программные архитектуры, называемые Lambda, Kappa и гибрид этих двух архитектур.  3.1. Архитектура Lambda  Архитектура Lambda описывает общую, масштабируемую и отказоустойчивую архитектуру обработки данных в реальном времени [12]. Она имеет две ветви для пакетной обработки и обработки в реальном времени. Как показано на рисунке 1a, данные предоставляются в обе ветви. На уровне пакетной обработки данные добавляются в область хранения, называемую основным набором данных. Обычно эти данные обрабатываются позже с помощью  Инструмент пакетной обработки, такой как Apache Hadoop, в результате чего получаются пакетные представления. На скоростном уровне данные обрабатываются в режиме реального времени с помощью инструмента обработки потоков, такого как Apache Storm, что приводит к инкрементным представлениям. Поскольку обработка массивных наборов данных на пакетном уровне занимает много времени, результаты обычно не являются актуальными. Скоростной уровень используется для того, чтобы  восполнить этот пробел путем запроса самых последних данных. Наконец, обслуживающий слой объединяет представления вышеуказанных слоев, в которые конечный пользователь может отправлять свои запросы.  3.2. Архитектура Kappa  Архитектура Kappa направлена на упрощение процесса разработки системы [9]. Она использует только одну ветвь; скоростной слой, как показано на рисунке 1b. То есть, входящие данные обрабатываются немедленно в режиме реального времени с использованием только одного инструмента обработки потоков. Результаты сохраняются в инкрементных представлениях. После этого данные поступают на обслуживающий слой, который запрашивает все необходимые результаты. Благодаря этому изменению архитектура Kappa поддерживает параллельную обработку всех данных основного набора данных в соответствии с потребностями, а также входящих данных в режиме реального времени. Преимуществом архитектуры Kappa является наличие единого механизма обработки данных, что позволяет упростить систему в целом. С другой стороны, параллельная обработка больших наборов данных и потоков данных реального времени может увеличить вычислительную мощность и емкость хранилища.  3.3. Гибридная архитектура  Архитектуры Lambda и Kappa поддерживают обработку исторических данных и данных реального времени. Однако режим обработки разделен. То есть пользователю приходится проделывать большую работу вручную, чтобы сопрячь результаты обоих режимов обработки, чтобы обеспечить всесторонний анализ данных [6]. В работе [6] представлена гибридная архитектура обработки, названная BRAID. BRAID объединяет обработку исторических данных и данных реального времени путем добавления каналов связи между пакетным и потоковым механизмом. Это позволяет проводить комплексный анализ автоматически при разумных накладных расходах. Авторы обсудили различные методы реализации предложенной ими архитектуры. В том же направлении в работе [7] была представлена концептуальная архитектура для потоковой обработки больших данных, интегрированная с комплексной обработкой событий, названная BiDCEP. Эта система расширяет архитектуры Lambda и Kappa, чтобы соответствовать доменам комплексной обработки событий и управления событиями в корпоративных ИТ. Авторы проводят техническую дискуссию о достоинствах объединения обеих архитектур, а также представляют мотивирующий пример |
| **4. WALLeSMART System Architecture**  In this section, we describe the architecture of the WALLeSMART framework. Based on the previous discussions, we adapt a Lambda architecture to implement our system. Our choice is justified by the fact that batch and real-time analysis are different in our case. Thus, we use separate tools for each job. Moreover, the Lambda architecture guarantees linear scalability, and fault-tolerance against hardware failures while improving the performance considerably. Our system implementation is composed of the three Lambda layers, the batch, the speed, and the serving layer. Tools used in our implementation are carefully chosen based on several criteria: are open source, scalable, distributed, extensible, and proven in production uses. In the following sections, we will present our underlying system layers.  4.1. Master Dataset  Data flowing in the platform goes through several steps. First, data is coming from two main sources (i) dairy farms and (ii) weather station observations. Incoming data can be in various formats, such as XML, JSON, or XLS files received in different ways, such as Web service, Email attachment, or MQTT server, which all depend on the sensor manufacturer being used. Next, data is transmitted to our cloud platform via different network protocols, such as Wi-Fi, 3G/4G, or LoRa. We use EMQ X as an MQTT message broker. Second, when the data arrives at our platform, it will be processed by a data ingestion tool that is responsible for capturing and storing real-time messages to be consumed by a stream/batch processing consumer. In our study, we use Apache Kafka as a data ingestion tool. It can scale to handle millions of messages per second. Third, and after capturing real-time messages, the platform processes them to gain knowledge about the data. Finally, storage tools are used as an output destination to capture real-time and batch data for archiving, or for further processing. In our architecture, we use Apache Cassandra and PostgreSQL for NoSQL and relational data, respectively.  4.2. Batch Layer  In the case of batch processing, our solution process data files using long-running batch jobs to filter, aggregate, and prepare the data for analysis. To this matter, we use Apache Hadoop, which allows for a distributed processing of large data sets across clusters of computers using the MapReduce programming model. MapReduce allows the system to compute the batch views based on the three steps (Map, Shuffle, Reduce) in parallel in order to significantly improve the performance.  4.3. Speed Layer  In the case of real-time processing, our solution uses Apache Storm, which is a low latency framework for stream processing that uses a topology of spouts and bolts to consume, process, and output the results from real-time streaming data sources. Storm runs in-memory and is, therefore, able to process large volumes of data at in-memory speed.  4.4. Serving Layer  The speed layer helps to manipulate WALLeSMART and to provide insights into the data through analysis, reporting, and visualizing in real-time. Our platform provides a Web-based application developed using Node.js, which is a runtime built on Google’s V8 JavaScript engine for easily building fast and scalable network applications. Farms data are plotted on interactive maps using OpenStreetMap, and rendered into interactive charts using Chart.js library. To help researchers accessing farms and weather data for further analysis and modeling, we provide a Web API (Application Programming Interface). We use GraphQL query language for fulfilling API queries. GraphQL provides a complete description of the data in the API and gives users the power to ask for exactly what they need. The API can also be used to develop a mobile version of the platform. Moreover, our application has many other features such as dashboard view, user’s management, roles and permissions management, scheduling batch and real-time jobs, access  tokens management, emails, notifications, etc | **4. Архитектура системы WALLeSMART**  В этом разделе мы описываем архитектуру фреймворка WALLeSMART. Основываясь на предыдущих обсуждениях, мы адаптируем архитектуру Lambda для реализации нашей системы. Наш выбор обоснован тем, что в нашем случае пакетный анализ и анализ в реальном времени отличаются. Таким образом, мы используем отдельные инструменты для каждого задания. Более того, архитектура Lambda гарантирует линейную масштабируемость и отказоустойчивость к аппаратным сбоям при значительном повышении производительности. Наша реализация системы состоит из трех уровней Lambda - пакетного, скоростного и обслуживающего. Инструменты, используемые в нашей реализации, были тщательно отобраны на основе нескольких критериев: с открытым исходным кодом, масштабируемые, распределенные, расширяемые и проверенные на практике. В следующих разделах мы представим наши базовые слои системы.  4.1. Основной набор данных  Данные, поступающие в платформу, проходят через несколько этапов. Во-первых, данные поступают из двух основных источников (i) молочных ферм и (ii) наблюдений метеостанций. Поступающие данные могут быть в различных форматах, таких как XML, JSON или XLS файлы, полученные различными способами, такими как веб-сервис, вложение электронной почты или сервер MQTT, которые зависят от производителя используемых датчиков. Далее данные передаются на нашу облачную платформу через различные сетевые протоколы, такие как Wi-Fi, 3G/4G или LoRa. Мы используем EMQ X в качестве брокера сообщений MQTT. Во-вторых, когда данные поступают на нашу платформу, они обрабатываются инструментом ввода данных, который отвечает за захват и хранение сообщений в реальном времени для последующего потребления потребителем потоковой/пакетной обработки. В нашем исследовании мы используем Apache Kafka в качестве инструмента ввода данных. Он способен обрабатывать миллионы сообщений в секунду. В-третьих, после захвата сообщений в реальном времени платформа обрабатывает их для получения знаний о данных. Наконец, инструменты хранения данных используются в качестве места вывода для сбора данных в реальном времени и пакетных данных для архивирования или дальнейшей обработки. В нашей архитектуре мы используем Apache Cassandra и PostgreSQL для NoSQL и реляционных данных, соответственно.  4.2. Уровень пакетной обработки  В случае пакетной обработки наше решение обрабатывает файлы данных с помощью длительно выполняемых пакетных заданий для фильтрации, агрегации и подготовки данных к анализу. Для этого мы используем Apache Hadoop, который позволяет распределенно обрабатывать большие массивы данных на кластерах компьютеров с помощью модели программирования MapReduce. MapReduce позволяет системе вычислять пакетные представления на основе трех этапов (Map, Shuffle, Reduce) параллельно, чтобы значительно повысить производительность.  4.3. Слой скорости  В случае обработки данных в реальном времени наше решение использует Apache Storm, который представляет собой фреймворк с низкой задержкой для обработки потоков, использующий топологию патрубков и болтов для потребления, обработки и вывода результатов из источников потоковых данных в реальном времени. Storm работает в памяти и поэтому способен обрабатывать большие объемы данных со скоростью памяти.  4.4. Слой обслуживания  Слой скорости помогает манипулировать WALLeSMART и обеспечивать понимание данных посредством анализа, отчетности и визуализации в режиме реального времени. Наша платформа представляет собой веб-приложение, разработанное с использованием Node.js, который является средой выполнения, построенной на движке V8 JavaScript от Google для легкого создания быстрых и масштабируемых сетевых приложений. Данные о фермах наносятся на интерактивные карты с использованием OpenStreetMap и отображаются в виде интерактивных графиков с помощью библиотеки Chart.js. Чтобы помочь исследователям получить доступ к данным о фермах и погоде для дальнейшего анализа и моделирования, мы предоставляем веб-интерфейс API (интерфейс прикладного программирования). Для выполнения запросов к API мы используем язык запросов GraphQL. GraphQL предоставляет полное описание данных в API и дает пользователям возможность запрашивать именно то, что им нужно. API также может быть использован для разработки мобильной версии платформы. Кроме того, наше приложение имеет множество других функций, таких как просмотр приборной панели, управление пользователями, управление ролями и разрешениями, планирование пакетных заданий и заданий в реальном времени, управление маркерами доступа, электронная почта, управление доступом и т.д. |
| **5. Implementation and Results**  As stated above, our platform is developed for the dairy farms sector of the Belgium Wallonia region. The objective is to help farmers (i) better control their dairy cows such as health care, reproduction, milk production and movement, and to (ii) follow weather conditions forecast as well as historical observations in the easiest manner. The data used in our system come from two main sources: Elev ́eo4 for cows behavior and Pameseb5 for weather conditions. Elev ́eo data are gathered via SenseHub6 and SmartVel7 sensors and it contains measurements from 4-1-2007 till 21-3-2020. This data set contains: 30 dairy farms, 7139 unique cows, and 55940 measurement-sets, consisting of 23 parameters. The Pameseb data set contains data for a network of 45 agrometeorological stations covering all the Wallonia region. The period of historical measurements ranges from 1-1-2008 till 21-3-2020, and forecasting data is for 7 future days. It is characterized by: 5 parameters: temperature, precipitations, humidity, sunshine, and wind speed,  Изображение выглядит как текст, снимок экрана, Шрифт, диаграмма  Автоматически созданное описание  more than 3 million historical measurements on an hourly basis, and 6879 forecasting measurements on an hourly basis.  5.1. The Data Value Chain in WALLeSMART  The implementation of WALLeSMART components is described in Figure 2. The platform is deployed on a multicore Linux server using Docker container technology to isolate software from its environment and ensure that it works uniformly anywhere. In practice, the data flow steps through the platform are: (i) data acquisition from Elev ́eo and Pamesed, (ii) data ingestion using Kafka, (iii) data processing using Storm and Hadoop, and (v) data visualization in the Web interface.  5.2. Performance Evaluation  In this section, we present some experiments and results conducted to test the effectiveness of our platform. In Figure 3a, the results of a real-time job executed by the speed layer, to calculate the average, maximum and minimum daily temperature for the next seven days are shown. Since we are receiving estimated data on an hourly basis, it is necessary to determine the forecasting on a daily basis. Datastream coming to Kafka topics is directly forwarded to Storm spouts, computations are done by Storm bolts, results are resent to Kafka, then stored in Cassandra by the end. On the other hand, Figure 3b depicts the results of a batch job execution triggered by the batch layer. The task consists in calculating the average, maximum and minimum daily temperature for a range of one year (the year of 2019) for every weather station. This batch process is scheduled to be executed at a regular interval on the entire master dataset. As we aforementioned, the batch job is executing using the MapReduce model via Hadoop, in order to speed-up calculation. Results are saved to the database when finishing.  Those results are shown in the speed layer through a user-friendly Web interface. This can help farmers having a closer look at weather data and timely decision making, for example, to take an estimation of irrigation needs and adjust irrigation schedules accordingly. In another set of experiments, we evaluated the performance of our architecture in executing batch and real-time jobs. In Figure 3c, we measured the execution time of the same previous real-time job for a period of one month. The average time is typically small (an average of 1 minute and 18 seconds) overall days, this is because we are getting the same amount of data each time. Note that the reported value includes the time of reading data from the source, processing it, and then saving it the storage area. In the same way, we evaluated the execution time of a batch job, that read data of dairy farms, conduct some data analysis, and save results in the database. As we can see from Figure 3d,  Изображение выглядит как текст, График, снимок экрана, линия  Автоматически созданное описание  execution time is considerable (the average is 2 hours and 31 minutes). Moreover, the curve tends to increases every day. This a normal phenomenon, as we are getting more dairy data each time. However, we can improve our system by adding more server nodes to do the computing tasks in a parallel mode. | **5. Реализация и результаты**  Как было сказано выше, наша платформа разработана для сектора молочных ферм бельгийского региона Валлония. Цель состоит в том, чтобы помочь фермерам (i) лучше контролировать своих молочных коров, например, в вопросах здравоохранения, воспроизводства, производства молока и перемещения, и (ii) наиболее простым способом отслеживать прогноз погодных условий, а также исторические наблюдения. Данные, используемые  в нашей системе поступают из двух основных источников: Elev ́eo4 для поведения коров и Pameseb5 для погодных условий. Данные Elev ́eo собираются с помощью датчиков SenseHub6 и SmartVel7 и содержат измерения с 4-1-2007 по 21-3-2020. Этот набор данных содержит: 30 молочных ферм, 7139 уникальных коров и 55940 наборов измерений, состоящих из 23 параметров. Набор данных Pameseb содержит данные для сети из 45 агрометеорологических станций, охватывающих весь регион Валлонии. Период исторических измерений - с 1-1-2008 по 21-3-2020, а прогнозные данные - на 7 будущих дней. Она характеризуется: 5 параметров: температура, осадки, влажность, солнечная активность и скорость ветра, более 3 миллионов исторических измерений на почасовой основе и 6879 прогнозных измерений на почасовой основе.  5.1. Цепочка создания ценности данных в WALLeSMART  Реализация компонентов WALLeSMART описана на рисунке 2. Платформа развернута на многоядерном сервере Linux с использованием контейнерной технологии Docker для изоляции программного обеспечения от окружающей среды и обеспечения его одинаковой работы в любом месте. На практике этапы потока данных через платформу следующие: (i) получение данных с Elev ́eo и Pamesed, (ii) ввод данных с помощью Kafka, (iii) обработка данных с помощью Storm и Hadoop, и (v) визуализация данных в веб-интерфейсе.  5.2. Оценка производительности  В этом разделе мы представляем некоторые эксперименты и результаты, проведенные для проверки эффективности нашей платформы. На рисунке 3a показаны результаты задания в реальном времени, выполняемого скоростным слоем, для расчета средней, максимальной и минимальной дневной температуры на следующие семь дней. Поскольку мы получаем расчетные данные на почасовой основе, необходимо определить прогноз на ежедневной основе. Датастрим, поступающий на темы Kafka, напрямую направляется на носики Storm, вычисления выполняются болтами Storm, результаты повторно отправляются в Kafka, а затем хранятся в Cassandra. С другой стороны, на рисунке 3b показаны результаты выполнения пакетного задания, инициированного пакетным слоем. Задание заключается в вычислении средней, максимальной и минимальной дневной температуры для диапазона в один год (год 2019) для каждой метеостанции. Этот пакетный процесс планируется выполнять с регулярным интервалом для всего основного набора данных. Как мы уже упоминали выше, пакетное задание выполняется с использованием модели MapReduce через Hadoop, чтобы ускорить вычисления. По завершении результаты сохраняются в базе данных.  Эти результаты отображаются в скоростном слое через удобный веб-интерфейс. Это может помочь фермерам более детально изучить погодные данные и своевременно принять решение, например, оценить потребности в орошении и соответствующим образом скорректировать графики полива. В другой серии экспериментов мы оценили производительность нашей архитектуры при выполнении пакетных заданий и заданий в реальном времени. На рисунке 3c мы измерили время выполнения одного и того же предыдущего задания в реальном времени за период в один месяц. Среднее время обычно небольшое (в среднем 1 минута и 18 секунд) в целом, это объясняется тем, что мы каждый раз получаем одинаковое количество данных. Обратите внимание, что в указанное значение включено время чтения данных из источника, их обработки, а затем сохранения в области хранения. Таким же образом мы оценили время выполнения пакетного задания, которое считывает данные о молочных фермах, проводит анализ данных и сохраняет результаты в базе данных. Как видно из рисунка 3d, время выполнения значительно (в среднем 2 часа 31 минута). Более того, кривая имеет тенденцию к увеличению с каждым днем. Это нормальное явление, так как с каждым разом мы получаем все больше молочных данных. Однако мы можем улучшить нашу систему, добавив больше серверных узлов для выполнения вычислительных задач в параллельном режиме. |
| **6. Conclusion**  If there is one sector that affects every individual on the planet, it has to be agriculture. The challenges of agricultural production are increasing, making the need to understand the complex agricultural ecosystems more imperative than ever. The richness of big data in the agri-ecosystem, generated by the emerging agri-tech, is leveraged to provide  unique solutions for diverse today’s worldwide challenges, and ensure a secure and healthy future for all people and the entire planet. In this paper, we introduced a new cloud-based smart farming management platform called WALLeSMART. Before presenting our solution, we give an investigation of the most important state-of-the-art platforms and big data architecture initiatives. Based on this, a general big data architecture for smart farming is introduced. The framework borrows the approach of Lambda architecture to solve the problems of acquisition, processing, storing and visualization of real-time big data. Furthermore, we gave a brief description of the most advanced technologies for  real-time stream processing. An initial prototype has been developed and tested with 30 dairy farms and 45 weather stations in the Wallonia region of Belgium. The backend part of the platform showed robust data management, while the frontend part gained users satisfaction. Moreover, examples of batch and real-time job execution have been dis-  cussed in the paper. Future work will include expanding the platform among more farms, and improving the user interface with further features such as time-series data exploration, real-time cows positioning, and spatial-temporal information, as well as ensuring the security aspects. In the long term, we are planning to use ontologies and linked data technologies for combining different data sources to enable big data analysis and reasoning in order to generate new insights. Moreover, we are exploring the application of big data machine learning techniques to help users make meaningful, forward-looking decisions. | **6. Заключение**  Если и существует отрасль, которая затрагивает каждого человека на планете, то это сельское хозяйство. Проблемы сельскохозяйственного производства возрастают, что делает необходимость понимания сложных сельскохозяйственных экосистем как никогда актуальной. Богатство больших данных в агроэкосистеме, генерируемых развивающимися агротехнологиями, используется для предоставления  уникальные решения для различных современных мировых проблем и обеспечения безопасного и здорового будущего для всех людей и всей планеты. В этой статье мы представили новую облачную платформу управления умным сельским хозяйством под названием WALLeSMART. Прежде чем представить наше решение, мы даем обзор наиболее важных современных платформ и инициатив в области архитектуры больших данных. На основе этого представлена общая архитектура больших данных для интеллектуального сельского хозяйства. Структура заимствует подход архитектуры Lambda для решения проблем сбора, обработки, хранения и визуализации больших данных в реальном времени. Кроме того, мы дали краткое описание самых передовых технологий для  обработки потоков данных в реальном времени. Первоначальный прототип был разработан и протестирован на 30 молочных фермах и 45 метеостанциях в регионе Валлония в Бельгии. Бэкэнд платформы продемонстрировал надежное управление данными, а фронтэнд - удовлетворенность пользователей. Кроме того, в статье были приведены примеры выполнения заданий в пакетном и реальном времени.  рассмотрены в статье. Будущая работа включает в себя распространение платформы на большее количество ферм, улучшение пользовательского интерфейса с помощью дополнительных функций, таких как исследование временных рядов данных, позиционирование коров в реальном времени, пространственно-временная информация, а также обеспечение безопасности. В долгосрочной перспективе мы планируем использовать онтологии и технологии связанных данных для объединения различных источников данных, чтобы обеспечить анализ больших данных и рассуждения для получения новых знаний. Кроме того, мы изучаем возможности применения методов машинного обучения больших данных, чтобы помочь пользователям принимать осмысленные, перспективные решения. |
|  |  |

English

Chitosan-Starch Films with Natural Extracts: Physical, Chemical, Morphological and Thermal Properties

Abstract

In the era of Big data, data-driven farming is changing the agricultural businesses thanks to the use of modern technologies such as the Internet of Things (IoT) sensors, drones, and farm monitoring. IoT devices produce a massive amount of precious agri-

data, which are collected and analyzed in real-time using innovative application tools. This combination of technology, known as “Smart farming”, helps various stakeholders in the agri-ecosystem to monitor crops in real-time, as well as maximize productivity and profitability in farm and business operations with the minimum efforts. Although many Smart farming solutions have been introduced, both from industry and academia, universal applicability of these approaches for other farms, unfortunately, is not feasible. Most of these solutions are based on a home-made non-standard Big data processing architecture. In this paper, we propose WALLeSMART, a cloud-based Smart farming management system, applied to the Wallonia region of Belgium. The framework introduces a general architecture to address the challenges of acquisition, processing, storing, and visualization of very large

amounts of data, both in batch and real-time basis. An initial prototype has been developed and tested with various farms showing

prominent results.

Introduction

The United Nations’ Population Division reports that the world population is currently (2020) growing by approximately 81 million people each year. The median estimate for future growth sees the world population reaching 9.7 billion in 2050 [18]. This means that agriculture is faced with multiple challenges: catering to the growing number of people, providing a livelihood for farmers, and protecting the environment. Precision agriculture and the acronym smart farming will be necessary to face and solve those challenges. Precision agriculture is the concept of observing, measuring and responding to variability in crops. Whereas, smart farming extends the precision agriculture concept by adding the notions of data context, situation and location awareness to the tasks of management and decision making in real-time [21]. The emergence of the Internet of Things and Cloud Computing technologies are set to push the future of smart farming to the next level. Today, IoT-based agriculture is becoming more familiar between farmers, and smart farming is rapidly becoming the standard as a result of agricultural drones and sensors usage. Business Insider Intelligence projects there to be nearly 12 million agricultural sensors installed globally by 2023 [14]. Those machines generate huge amounts of data every second. IBM company estimates that the average farm can generate half a million data points per day [14]. Big data has proven to be fruitful in the agriculture sector, it provides all the stakeholders with in-depth insights about the entire agri-ecosystem. This can be achieved using predictive algorithms that can alert even before a problem occurs. However, the task of managing this Big data, ranging from acquisition, processing, storing, analyzing, and visualizing is very complex and therefore requires new architectural conception and application tools [11]. The complexity resides in the facts that (i) large, complex, heterogeneous data is coming from a variety of sources encountered in agriculture, (ii) data must be processed in both real-time and batch basis using high efficient tools, (iii) advanced analysis techniques such as predictive or prescriptive analytics must be applied, (iv) data analyzed needs to be visualized in user-friendly interactive Web and mobile application. Resolving those complexities might unlock the potential for full automation of the agri-food chain. In this paper, we propose a generic smart farming cloud-based framework, called WALLeSMART1, which helps farmers, researchers, and administrators managing their farms and making the right decisions. To achieve this goal, we fix four main objectives: (i) the study of the state-of-the-art big data processing architecture, (ii) the proposition of a generic architecture to the agriculture sector, (iii) the development and deployment of such solution applied our specific use case, using a real data sets coming from dairy farms and weather stations, in batch and real-time basis, and (iv) the visualization of those data via an interactive Web application. These four objectives will be discussed in the next sections.

The remainder of this paper is organized as follows. Section 2 reviews the most important works on smart farming platforms and tools. Section 3 gives a background view of big data processing system architectures. In Section 4, our methodology in developing a novel real-time architecture for agriculture is given and all its required components are

described. Section 5 highlights our implementation and deployment on real farms conducted to evaluate the quality of our architecture. Section 6 concludes the paper by summarizing the important results and suggests future work.

Related Work

The applications of big data in the agriculture field are numerous, we can cite for instance: precision farming, yield prediction, risk mitigation, loss reduction, supply chain management, farm-to-fork traceability, and sustainable farming [10, 8, 21, 13]. Besides, agricultural big data systems can be broadly divided into three categories: (i) advanced sensor technology systems, (ii) risk management systems, and (iii) agricultural management systems [16]. In this section, we focus mainly on the third category, since our work belongs to this type of system. Recently, initiatives are ongoing to create agricultural platforms that collect the data needed by smart farming decision support tools. They are often created by private companies or public-private partnerships [5]. Among the former, Monsanto’s Integrated Farming Systems platform collects data on information such as soil health and pest pressures and provide them for farmers. The Climate Corporation proposes the Climate FieldView platform to aggregate data of different sources in one place and provides diagnostic and applicative tools to farmers [5]. However, since 2018 only four companies dominate the market: DowDuPont, Syngenta-ChemChina, BASF, and Bayer-Monsanto due to a series of mergers and acquisitions [20]. Regards public-privates partnerships, many countries are building systems to advance the use of modern agricultural technologies. For example, Akkerweb is a Dutch web-based platform that provides access to external data sources such as weather, parcel boundaries, satellite, and data from commercials farm management system. It stores geo-referenced data, including soil maps and drone imagery. It allows combining data sources and processing of data, through a set of application modules to provide farmers decision support and recommendations [19]. Another example is the platform Barto2 in Switzerland. As a stock company, Barto brings together public and private actors to build up a smart-farming platform that also aims to digitize operational and production farms data while avoiding duplication, which speeds up on-farm processes, reduces administrative tasks.

Barto itself is based on 365FarmNet, a German-based farm management software provider. Nearly 45.000 farmers are already active on this platform in Germany, Poland, Bulgaria, Austria, and France 3. The platform operates as SaaS (Software as a service) and provides solutions for managing and recording all activities on a farm. However, those platforms are proprietary, and therefore the internal system architecture is not available. In the academic sector, many kinds of researches have

been conducted in the smart farming field. The authors of [2] from Spain, demonstrate the advantage of a tool, called PLATEM, that applies real-time decisions from data such as variable rate irrigation, and selected parameters from field and weather conditions. In [17], the authors proposed a Web-based service-oriented herd manage application by leveraging occasionally connected mobile database synchronization framework. The goal is to provide improved

accessibility, more up-to-date information and lower cost of ownership for the herd management software system. An initial prototype has been deployed and tested for over six months on a three thousand cow dairy farm in central Washington State, USA. The work of [1] presents a platform, called SmartDairyTracer, designed to monitor the state of dairy cattle and feed grain in real time, as well as ensure the traceability and sustainability of the different processes involved in the production. SmartDairyTracer uses global edge computing architecture which consists of three principal layers: IoT, edge, and business solution layers [15].

Big Data Processing for Smart Farming

Developments and usage of network technologies, IoT, and cloud computing in smart farming allow generating big data. Big data is described as both structured and unstructured data that is too large to be processed by traditional data processing tools and techniques. The “big” part of this term is characterized by the 5 Vs, which are volume, velocity, variety, veracity and, value [3]. Additionally, some companies contributed other dimensions to big data. This large quantity of data must be managed and analyzed to capture its full value. To this end, first, raw data from different sources need to be ingested, integrated, and stored into a Data Lake, a centralized storage repository. Second, big

data solutions must be used to process this data, based on long-running batch jobs as well as real-time jobs to filter, aggregate, and otherwise prepare the data for analysis. The last step consists in providing insights from data through analysis and reporting tasks. In general, there exist two basic requirements for big data systems. First, receiving a

massive real-time data stream from different sources and second, performing an analysis of this data to get results right away [4]. Based on this, three state-of-the-art software architectures are often used for big data systems, called Lambda, Kappa, and a Hybrid of the two.

3.1. Lambda Architecture

Lambda architecture describes a generic, scalable and fault-tolerant real-time data processing architecture [12]. It has two branches for batch and real-time processing. As shown in Figure 1a, data is provided to both branches. In the batch layer, data is appended to a storage area, called the master dataset. Usually, this data is processed later using a

batch processing tool such as Apache Hadoop, resulting in batch views. In the speed layer, data is processed in real-time using a stream processing tool such as Apache Storm, which results in an incremental view. Since processing massive data sets in the batch layer takes a lot of time, the results are usually not up-to-date. The speed layer is used to

close this gap by querying the most recent data. Finally, the serving layer merges views of the above layers, in which end-user can send his queries.

3.2. Kappa Architecture

Kappa architecture aims at simplifying the development process of the system [9]. It uses only one branch; the speed layer as shown in Figure 1b. That is, incoming data is processed immediately in real-time using only one stream processing tool. Results are saved in incremental views. Afterward, the data goes to the serving layer that queries any required results. Due to this change, the Kappa architecture supports processing the entire data in the master dataset according to needs, as well as the incoming real-time data in parallel. The advantage of the Kappa architecture is its single data processing engine, which can simplify the overall system. On the other hand, processing large data sets and real-time data streams in parallel may increase computational power and storage capacity.

3.3. Hybrid Architecture

While Lambda and the Kappa architectures support the processing of historical data and real-time data. However, the processing mode is separated. That is, the user has to do a lot of work manually to couple the results of both processing modes, in order to enable a comprehensive analysis of the data [6]. The work of [6] introduces a hybrid processing architecture, called BRAID. BRAID intertwines the processing of historical data and real-time data by adding communication channels between the batch engine and the stream engine. This enables to carry out comprehensive analyses automatically at a reasonable overhead. The authors discussed different implementation techniques for their proposed architecture. In the same direction, the study of [7] introduced a conceptual architecture for big data streaming integrated with complex event processing, called BiDCEP. The system expands the Lambda and Kappa architectures to fit the complex event processing and event management domains of enterprise IT. The authors trigger a technical discussion on the merits of combining both architectures, as well as present a motivational example

4. WALLeSMART System Architecture

In this section, we describe the architecture of the WALLeSMART framework. Based on the previous discussions, we adapt a Lambda architecture to implement our system. Our choice is justified by the fact that batch and real-time analysis are different in our case. Thus, we use separate tools for each job. Moreover, the Lambda architecture guar-

antees linear scalability, and fault-tolerance against hardware failures while improving the performance considerably. Our system implementation is composed of the three Lambda layers, the batch, the speed, and the serving layer. Tools used in our implementation are carefully chosen based on several criteria: are open source, scalable, distributed, extensible, and proven in production uses. In the following sections, we will present our underlying system layers.

4.1. Master Dataset

Data flowing in the platform goes through several steps. First, data is coming from two main sources (i) dairy farms and (ii) weather station observations. Incoming data can be in various formats, such as XML, JSON, or XLS files received in different ways, such as Web service, Email attachment, or MQTT server, which all depend on the sensor manufacturer being used. Next, data is transmitted to our cloud platform via different network protocols, such as Wi-Fi, 3G/4G, or LoRa. We use EMQ X as an MQTT message broker. Second, when the data arrives at our platform, it will be processed by a data ingestion tool that is responsible for capturing and storing real-time messages to be consumed by a stream/batch processing consumer. In our study, we use Apache Kafka as a data ingestion tool. It can scale to handle millions of messages per second. Third, and after capturing real-time messages, the platform processes them to gain knowledge about the data. Finally, storage tools are used as an output destination to capture real-time and batch data for archiving, or for further processing. In our architecture, we use Apache Cassandra and PostgreSQL for NoSQL and relational data, respectively.

4.2. Batch Layer

In the case of batch processing, our solution process data files using long-running batch jobs to filter, aggregate, and prepare the data for analysis. To this matter, we use Apache Hadoop, which allows for a distributed processing of large data sets across clusters of computers using the MapReduce programming model. MapReduce allows the system to compute the batch views based on the three steps (Map, Shuffle, Reduce) in parallel in order to significantly improve the performance.

4.3. Speed Layer

In the case of real-time processing, our solution uses Apache Storm, which is a low latency framework for stream processing that uses a topology of spouts and bolts to consume, process, and output the results from real-time streaming data sources. Storm runs in-memory and is, therefore, able to process large volumes of data at in-memory speed.

4.4. Serving Layer

The speed layer helps to manipulate WALLeSMART and to provide insights into the data through analysis, reporting, and visualizing in real-time. Our platform provides a Web-based application developed using Node.js, which is a runtime built on Google’s V8 JavaScript engine for easily building fast and scalable network applications. Farms data are plotted on interactive maps using OpenStreetMap, and rendered into interactive charts using Chart.js library. To help researchers accessing farms and weather data for further analysis and modeling, we provide a Web API (Application Programming Interface). We use GraphQL query language for fulfilling API queries. GraphQL provides a complete description of the data in the API and gives users the power to ask for exactly what they need. The API can also be used to develop a mobile version of the platform. Moreover, our application has many other features such as dashboard view, user’s management, roles and permissions management, scheduling batch and real-time jobs, access

tokens management, emails, notifications, etc

5. Implementation and Results

As stated above, our platform is developed for the dairy farms sector of the Belgium Wallonia region. The objective is to help farmers (i) better control their dairy cows such as health care, reproduction, milk production and movement, and to (ii) follow weather conditions forecast as well as historical observations in the easiest manner. The data used

in our system come from two main sources: Elev ́eo4 for cows behavior and Pameseb5 for weather conditions. Elev ́eo data are gathered via SenseHub6 and SmartVel7 sensors and it contains measurements from 4-1-2007 till 21-3-2020. This data set contains: 30 dairy farms, 7139 unique cows, and 55940 measurement-sets, consisting of 23 parameters. The Pameseb data set contains data for a network of 45 agrometeorological stations covering all the Wallonia region. The period of historical measurements ranges from 1-1-2008 till 21-3-2020, and forecasting data is for 7 future days. It is characterized by: 5 parameters: temperature, precipitations, humidity, sunshine, and wind speed, more than 3 million historical measurements on an hourly basis, and 6879 forecasting measurements on an hourly basis.

5.1. The Data Value Chain in WALLeSMART

The implementation of WALLeSMART components is described in Figure 2. The platform is deployed on a multicore Linux server using Docker container technology to isolate software from its environment and ensure that it works uniformly anywhere. In practice, the data flow steps through the platform are: (i) data acquisition from Elev ́eo and Pamesed, (ii) data ingestion using Kafka, (iii) data processing using Storm and Hadoop, and (v) data visualization in the Web interface.

5.2. Performance Evaluation

In this section, we present some experiments and results conducted to test the effectiveness of our platform. In Figure 3a, the results of a real-time job executed by the speed layer, to calculate the average, maximum and minimum daily temperature for the next seven days are shown. Since we are receiving estimated data on an hourly basis, it is necessary to determine the forecasting on a daily basis. Datastream coming to Kafka topics is directly forwarded to Storm spouts, computations are done by Storm bolts, results are resent to Kafka, then stored in Cassandra by the end. On the other hand, Figure 3b depicts the results of a batch job execution triggered by the batch layer. The task consists in calculating the average, maximum and minimum daily temperature for a range of one year (the year of 2019) for every weather station. This batch process is scheduled to be executed at a regular interval on the entire master dataset. As we aforementioned, the batch job is executing using the MapReduce model via Hadoop, in order to speed-up calculation. Results are saved to the database when finishing.

Those results are shown in the speed layer through a user-friendly Web interface. This can help farmers having a closer look at weather data and timely decision making, for example, to take an estimation of irrigation needs and adjust irrigation schedules accordingly. In another set of experiments, we evaluated the performance of our architecture in executing batch and real-time jobs. In Figure 3c, we measured the execution time of the same previous real-time job for a period of one month. The average time is typically small (an average of 1 minute and 18 seconds) overall days, this is because we are getting the same amount of data each time. Note that the reported value includes the time of reading data from the source, processing it, and then saving it the storage area. In the same way, we evaluated the execution time of a batch job, that read data of dairy farms, conduct some data analysis, and save results in the database. As we can see from Figure 3d, execution time is considerable (the average is 2 hours and 31 minutes). Moreover, the curve tends to increases every day. This a normal phenomenon, as we are getting more dairy data each time. However, we can improve our system by adding more server nodes to do the computing tasks in a parallel mode.

6. Conclusion

If there is one sector that affects every individual on the planet, it has to be agriculture. The challenges of agricultural production are increasing, making the need to understand the complex agricultural ecosystems more imperative than ever. The richness of big data in the agri-ecosystem, generated by the emerging agri-tech, is leveraged to provide

unique solutions for diverse today’s worldwide challenges, and ensure a secure and healthy future for all people and the entire planet. In this paper, we introduced a new cloud-based smart farming management platform called WALLeSMART. Before presenting our solution, we give an investigation of the most important state-of-the-art platforms and big data architecture initiatives. Based on this, a general big data architecture for smart farming is introduced. The framework borrows the approach of Lambda architecture to solve the problems of acquisition, processing, storing and visualization of real-time big data. Furthermore, we gave a brief description of the most advanced technologies for

real-time stream processing. An initial prototype has been developed and tested with 30 dairy farms and 45 weather stations in the Wallonia region of Belgium. The backend part of the platform showed robust data management, while the frontend part gained users satisfaction. Moreover, examples of batch and real-time job execution have been dis-

cussed in the paper. Future work will include expanding the platform among more farms, and improving the user interface with further features such as time-series data exploration, real-time cows positioning, and spatial-temporal information, as well as ensuring the security aspects. In the long term, we are planning to use ontologies and linked data technologies for combining different data sources to enable big data analysis and reasoning in order to generate new insights. Moreover, we are exploring the application of big data machine learning techniques to help users make meaningful, forward-looking decisions.