농업 디지털 데이터 수집 및 관리 시스템 설계 연구

이현조, 채철주*

한국농수산대학교 교양학부

o2near@gmail.com, chae.cheoljoo@gmail.com

A Study on the Design of Data Collection and Management System for Agricultural Digital Data

Lee Hyun Jo, Cheol-Joo Chae

Dept. of General Education, Korea National College of Agriculture and Fisheries

요 약

최근 스마트팜을 통해 다양한 농업 디지털 데이터가 수집되고 있다. 수집되는 데이터의 활용도를 향상시키기 위해서는 데이터 개방·공유 및 메타데이터 도출, 분석서비스 제공이 가능한 농업 디지털 데이터 플랫폼이 요구된다. 본 연구에서는 스마트팜을 통해 수집되는 데이터를 저장 및 관리하기 위해 농업 디지털 데이터 수집 및 관리 시스템을 제안한다.

Ⅰ. 서 론

최근 작물의 생장 환경을 최적화하여 생산성을 향상시키기 위해 스마트팜 보급이 활성화되었으며, 이에 따라 생육 정보, 환경 정보 및 시설제어 정보 등 다양한 농업 디지털 데이터가 수집되고 있다. 수집된 농업 디지털 데이터를 활용하여 다양한 영농 의사결정 지원 서비스를 제공할 수 있다. 이러한 농업 디지털 데이터의 가용성·접근성 향상 및 체계적인 관리를 위해 통합 농업 디지털 데이터 플랫폼이 요구된다[1].

본 연구에서는 하둡 에코시스템을 기반으로 농업 디지털 데이터 수집 및 관리 시스템을 설계한다. 제안하는 시스템은 데이터 수집, 데이터 분산·연계, 데이터 저장 및 자원관리, 분산처리, 스케쥴관리 등을 지원하며, 수집·분석한 데이터를 외부에 연계하여 다양한 서비스를 제공할 수 있도록 전체 프로세스를 구성하였다.

Ⅱ. 본론

본 연구에서는 스마트팜에서 수집되는 데이터를 저장·관리하기 위해 농업 디지털 수집 및 저장 시스템을 설계한다. 제안하는 플랫폼은 하둡 에코시스템[2]을 기반으로 설계되었으며, 데이터 수집 컴포넌트, 데이터 분산·연계 컴포넌트, 데이터 저장 및 자원관리 컴포넌트, 분산처리 컴포넌트, 스케쥴관리 컴포넌트 등으로 구성된다.

1. 데이터 수집 컴포넌트 : 실시간 생성 데이터(센서 정보, 시설 제어 정보), 주기적 생성 데이터(생육정보) 등의 내부 데이터 및 기상정보, 물류유통 정보 등의 외부 데이터를 수집하고, 수집한 데이터에 대한 표준화 및 전처리를 수행한다. 내부 수집 데이터의 경우, 설비 및 센서 이상 등으로 인해 결측치, 이상치가 발생할 수 있다. 이상치가 발생한 경우, 데이터 평활화를 수행하여 극단값이나 이상치를 완화한다. 데이터 결측치의 경우, 데이터 항목의 특성을 고려하여 회귀함수, kNN, GAN 인공신경망 등을 활용한 결측치 예측·생성을 수행하여 결측값을 대체한다.

2. 데이터 분산·연계 컴포넌트 : 전처리된 수집 데이터를 효율적으로 통합 관리하기 위해, 하둡 Sqoop 및 Flume을 활용한다. Sqoop은 대용량 데이터들을 효율적으로 변환해주는 CLI 애플리케이션이다. Sqoop을 통해 하둡 뿐만 아니라, Oracle, MySQL 등의 관계형 데이터베이스에 내보내기·가져오기를 수행할 수 있기 때문에 외부에서 제공하는 의사결정 도구를 사용할 수 있다. Flume은 로그 데이터를 효율적으로 수집, 취합, 이동하기 위한 분산형 소프트웨어로써, 데이터의 수집 위치와 전송방식을 동적으로 변경하여 전체 데이터 흐름을 관리한다.

3. 데이터 저장 및 자원관리 컴포넌트

• 데이터 저장 : 농업 디지털 데이터 저장을 위해 하둡 분산 파일 시스템 HDFS를 사용한다. HDFS는 데이터 회복성 및 안정성, 병렬 처리 성능을 제공하며, 데이터 노드를 클러스터에 추가·관리하기 용이하여 확장성이 높아 농업 디지털 데이터 저장 관리에 효율적이다. 한편, 사용자 질의처리 지원을 위해 Hbase를 사용한다. Hbase는 대용량 데이터에 대해 실시간 랜덤 조회 및 업데이트, 전체 데이터에 대한 일관성 관리를 지원한다.

• 자원관리 : 컴포넌트에서 요구하는 자원을 관리하고 워크로드 분산할당을 지원하기 위해, 하둡 Yarn을 사용한다. Yarn은 CPU, 메모리, 디스크, 네트워크와 같은 다양한 시스템 자원 및 자원의 라이프사이클을 모니터링하며, 어플리케이션의 리소스 요청에 따라 가용한 자원을 분배하고 관리한다. Yarn은 먼저 들어오는 어플리케이션에게 자원을 먼저 할당해주는 FIFO, 추가적인 어플리케이션이 요청되었을때 이전에 사용하던 자원을 절반을 반납하고 새로운 어플리케이션에 할당하는 Fair, 요청 자원이 적은 어플리케이션을 위해 자원과 큐를 미리 예약하는 Capacity 등 다양한 자원관리 방식을 제공하여 작업 효율성을 향상시킨다.

4. 분산처리 컴포넌트

• MapReduce 기반 분산처리 : 필터링, 그룹화, 정렬을 통해 Key-Value 쌍을 생성하는 Map, Key-Value쌍을 바탕으로 분류된 데이터를 병합하는 Reduce를 통해 대용량 농업 디지털 데이터에 대한 분산처리를 수행한다. 병렬 데이터 처리 엔진 Pig, 데이터 요약·질의·분석 시스템 Hive, 분류·클러스터링·패턴 마이닝·회귀 분석 등 다양한 마이닝 도구를 제공하는 Mahout 등과 연계하여 농가에 데이터 분석 서비스를 제공한다.

• Spark 기반 분산처리 : 인-메모리 기반 데이터 처리를 수행하는 Spark를 활용하여 온라인 분석처리를 지원한다. 스트림 데이터 처리 Spark Streaming, SQL 지원 Spark SQL, 머신러닝 라이브러리 MLlib 및 그래

프처리 GraphX와 연계하여 데이터 분석 서비스를 제공한다.

5. 스케쥴 관리 컴포넌트 : 분산 환경을 구성하는 서버들의 환경설정을 통합적으로 관리하기 위해 하둡 ZooKeeper를 사용한다. ZooKeeper는 특정 서버에 서비스가 집중되어 서비스가 정체되는 병목현상을 방지하기 위해 서비스 분산 및 어플리케이션 동시 처리를 지원하며, 하나의 서버에서 처리한 결과를 다른 서버들과도 동기화하여 데이터 안정성을 보장한다. 운영(active) 서버에서 문제가 발생해 서비스를 제공할 수 없는 경우, 다른 대기중인 서버를 운영 서버로 변경하여 서비스를 중지없이 제공한다.

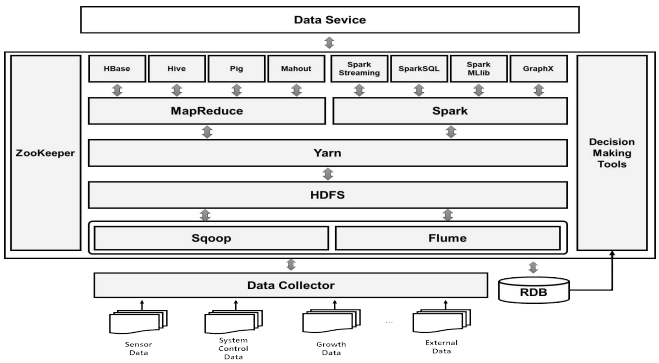

그림 1 농업 디지털 데이터 수집 및 관리 시스템

그림2는 제안하는 농업 디지털 수집 및 저장 플랫폼을 활용한 데이터 수집 및 연계 프로세스를 나타낸다.

1. 내·외부 데이터 수집

• 내부 수집 데이터 : 시설원예 스마트팜에서는 농가경영체정보, 시설정보, 영농정보, 스마트팜 내외부 환경 정보, 시설 제어 정보, 작물의 생육정보 및 작물의 판매·유통·물류 정보 등 7개 그룹의 데이터를 수집한다. 표1은 내부 수집 데이터에 대한 분류, 개요, 수집 방법·주기를 나타낸다[3].

• 외부 수집 데이터 : AgriX, KADX, 스마트팜 코리아 등에서는 농업경영체정보, 물류·유통정보, 환경·시설제어·생육 정보와 같은 다양한 농업 디지털 데이터를 파일, OpenAPI, 웹서비스 등을 통해 제공 중이다. 농업 디지털 데이터 분석 성능 및 제공하는 서비스의 품질을 향상시키기 위해 외부 연계 API를 설계하여 외부 데이터를 수집한다.

표 1. 시설원예 스마트팜 내부 수집 데이터

분류	개요	수집방법·주기
경영체	▪ 농장을 경영하는 업체 관련 정보	초기/변동시 수동
시설	▪ 농가 시설 유형 및 설비 정보	초기/변동시 수동
영농	▪ 재배작물에 대한 개요	발생시 수동
환경	▪ 센서로 측정하는 환경 정보	센서 자동 수집
제어	▪ 환경조절을 위한 장치 동작 정보	동작시 자동 저장
생육	▪ 작물 생장도 측정을 위한 변화도	주기별 수동 수집

2. 데이터 수집 및 전처리

• 데이터 수집기 : 농가별로 설비가 다르기 때문에, 데이터 구조 및 포맷이 상이하거나 품질이 고르지 않다. 양질의 의사결정 서비스를 제공하기 위해서는, 수집되는 데이터에 대한 일관성있는 관리가 필수적이다. 이를 위해 데이터 수집기에서는 데이터 표준화 및 품질관리를 수행한다.

• 데이터 전처리 : 수집된 내·외부 데이터에는 사람의 실수, 측정 장비의 정밀도, 데이터 손실, 중복 입력 등 다양한 원인에 의해 데이터에 손상이 발생할 수 있다. 따라서 데이터 분석 및 처리를 수행하기 전, 적합한 형태로 교정해야 한다. 데이터 전처리기에서는 중복된 데이터 및 오류를 제거하는 데이터 필터링, 다양한 형식으로 수집된 원데이터를 일관성 있는 형식으로 포맷팅하는 데이터 변환, 데이터의 불일치성을 교정하기 위한 데이터 정제 등을 수행하여 데이터의 품질을 유지한다.

3. 데이터 분석

• 데이터 분석 : 데이터 분석은 환경 및 생산 정보에 관한 통계분석, Hive, Pig, SparkSQL 등을 활용한 사용자 SQL 질의처리 뿐만, SparkMLlib, Tensorflow, Mahout 등을 활용한 인공지능 기반 데이터 마이닝 및 분석 서비스로 구성된다. 또한 영농 의사결정 지원 서비스와 같은 스마트 농업 서비스를 개발·제공하기 위해서 Sqoop, Flume을 이용한 외부 DB연계를 통해 개발도구, 라이브러리 등을 활용할 수 있다.

4. 데이터 연계

• 외부 데이터 연계 : 스마트팜 내부에서 수집된 환경·시설제어·생육 정보는 스마트팜 코리아에 제공하여 농가의 작물재배를 위한 최적 환경설정 서비스에 활용할 수 있다. 경영체정보, 물류·유통정보 등의 수집된 데이터를 다양한 기관 및 농가에 제공하여 스마트 농업 서비스를 개발하는데 활용할 수 있다. 이를 위해 외부연계 API 및 농업디지털데이터 분석 서비스와 같은 웹서비스 API를 구현하여 데이터 연계를 수행한다.

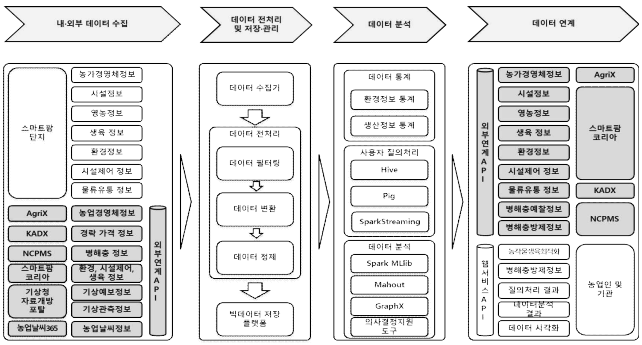

그림 2 농업 디지털 데이터 수집 및 연계 프로세스

Ⅲ. 결론

본 논문에서는 스마트팜에서 수집되는 데이터를 관리하기 위해, 농업 디지털 데이터 수집 및 관리 시스템을 제안하였다. 제안하는 플랫폼은 데이터 수집 및 전처리, 데이터 분산 처리, 연계, 자원 및 스케쥴 관리 등을 수행하며, 내·외부 데이터 수집 및 웹기반 데이터 분석 결과 서비스, 외부 데이터 연계 서비스를 제공하도록 설계하였다. 본 논문의 결과물을 좀 더 최적화하여 적용한다면 스마트팜 분야에서 여러 방면으로 활용될 수 있을 것으로 판단된다.

ACKNOWLEDGMENT

본 연구는 (2022년도) 한국과학기술정보연구원(KISTI) '울주형 스마트팜 산업 활성화를 위한 데이터 수집·분석 및 활용 지원사업' 과제로 수행한 것입니다.

참 고 문 헌

[1] 스마트농업 육성사업 추진현황과 개선과제. 2022.06.15. 2023.01.12. 접속, https://nabo.go.kr/q/4OxCA7Z5

[2] 정재화, "시작하세요! 하둡 프로그래밍", 2014년

[3] SPS-X KOAT-0009-7470, 시설원예 분야 스마트팜 수집 데이터 규격. 2022.04.11. 2023.01.12. 접속, https://www.koat.or.kr/action.do?action=stpdts%24view%24form&mgt_no=101&seq=14&page=2