발전소 고장 예측 AI 모델 학습 및 추론을 위한 센서 빅데이터 질의 처리 시스템 구현

엄정호*, 유찬희*,**, 김유선*,**, 박경석*,**
*한국과학기술정보연구원
**UST 빅데이터과학과
jhum@kisti.re.kr, rbyche@kisti.re.kr, yskblue@kisti.re.kr, gspark@kisti.re.kr,

Implementation of Sensor Big Data Query Processing System for AI model training and inference of Power Turbine Equipment Failure Estimation

Jung-Ho Um*, Chan Hee Yu*,**, Yuseon Kim*,**, Kyongseok Park*,**
*Korea Institute of Science and Technology Information
**Department of Big Data Science, UST

요 약

발전시설 장비는 이상이 생기면 큰 경제적 피해를 발생시키기 때문에, 장비의 계통마다 수십만 개의 센서들이 부착되어 장비의 정상 작동 여부를 모니터링 한다. 장비의 이상 감지를 위해서, 최근 활발히 연구되고 있는 딥러닝 등의 기술을 활용한 AI 모델을 생성하여 장비의 고장을 예측한다. AI 모델을 학습하고 추론하기 위해서는 수많은 센서 중에서 AI 모델을 생성할 센서들을 선택하고, 지속적으로 모니터링 되는 값들을 비교하여 이상 감지 여부를 스트리밍 환경에서 추론할 수 있는 센서 빅데이터 질의 처리 및 스트리밍 추론 시스템이 필요하다. 본 논문에서는 AI 모델을 학습하고 스트리밍 추론할 수 있는 빅데이터 질의 처리 시스템을 설계 및 구현한다.

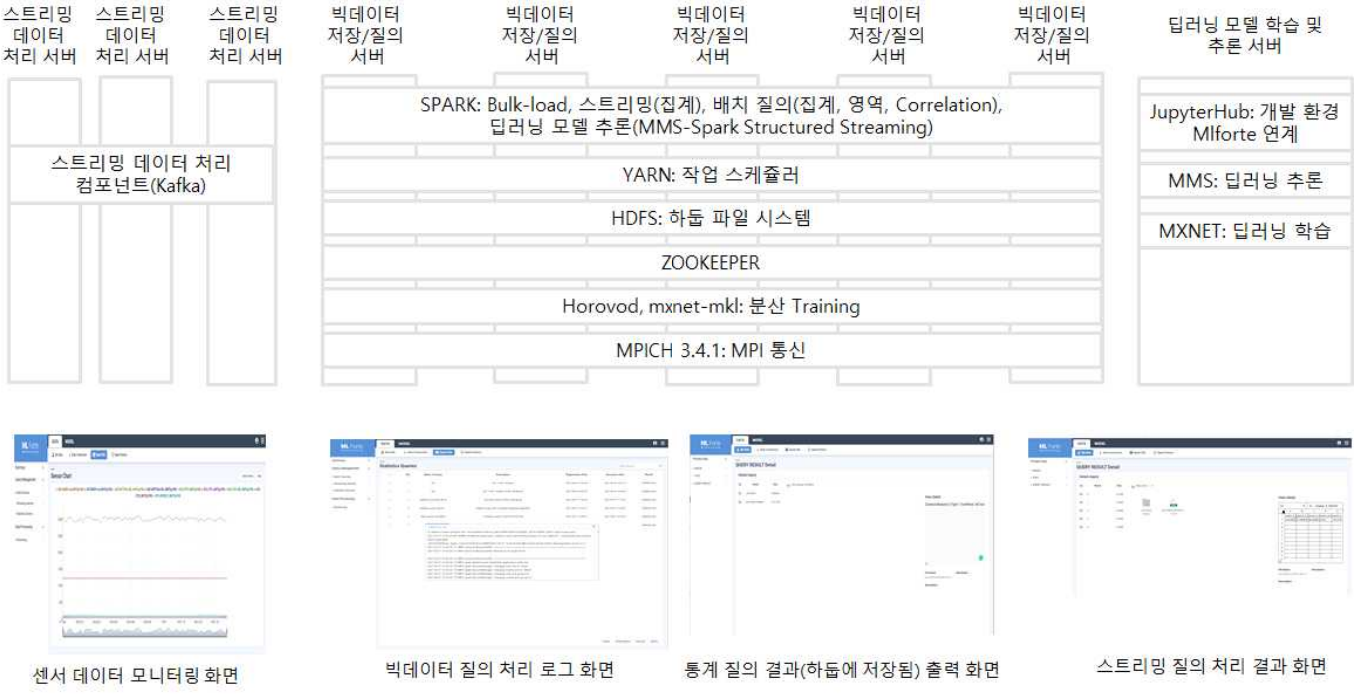
1. 서론

발전시설에는 장비의 장애에 대비하기 위해 수십만 개의 센서가 부착되고, 실시간으로 데이터를 적재 및 모니터링하는 시스템이 구축되어 있다. 최근 딥러닝 기술 등의 머신러닝 기법을 활용한 이상 감지 기술이 발전함에 따라 발전소 모니터링 시스템에서도 이러한 기술을 적용하여 이상을 감지하고자 노력하고 있다[1]. 이러한 인공지능 모델을 학습하기 위해서는 기존의 모니터링 데이터에 대한 질의를 통해서 학습 데이터셋 구축이 필요하고, 또한 인공지능 모델이 만들어지면 이를 지속적으로 스트리밍 환경에서 추론할 수 있는 질의 처리 시스템이 필요하다. 이에 따라, 본 논문에서는 인공지능 모델을 학습하기 위한 빅데이터 질의 처리와 고장 감지를 위한 인공지능 모델 추론을 스트리밍 환경에서 수행할 수 있는 빅데이터 질의 처리 시스템을 설계 및 구현한다.

2. 시스템 설계 및 구현

발전소 고장 예측 AI 모델 학습 및 추론을 위한 센서 빅데이터 질의 처리 시스템의 전체구조는 그림 1과 같다. 시스템에서는 수십만 개의 센서를 효율적으로 저장할 수 있는 빅데이터 저장 시스템을 도입하고, 또한 각 센서마다 동작하는 방식이 다르므로 다양한 인공지능 모델을 생성하고 추론할 수 있는 환경을 구축하는 것을 목표로 한다.

시스템은 지속적으로 입수되는 센서 데이터를 처리하기 위한 스트리밍 데이터 처리 컴포넌트(Kafka 2.15[2], Spark Structured Streaming 3.0[3] 활용), 입수되는 데이터를 저장하는 분산 파일 시스템(Hadoop 3.3 [4] 활용), 적재된 데이터를 질의 처리하기 위한 배치 질의 처리 컴포넌트(Spark SQL 3.0[5] 활용), 적재된 데이터 등을 활용하여 고장 예측을 지원하는 딥러닝 학습 컴포넌트(MXNET 1.6[6] 활용), 다양한 인공지능 모델에 대해서 REST API 방식으로 제공하는 다중 모델 추론 서비스(Multi Model Service 1.14[7]), 추론 서비스와 연결



(그림 1). 발전소 고장 예측 AI 모델 학습 및 추론을 위한 센서 빅데이터 질의 처리 시스템 구조

하여 스트리밍 방식으로 고장 예측을 진단하는 스트리밍 AI 추론 모델 실행 컴포넌트(Spark Structured Streaming 활용), 마지막으로 이러한 빅데이터 질의 처리 시스템을 다중 사용자가 웹에서 활용할 수 있는 웹 인터페이스로 구성되어 있다. 데이터 구조는 크게 모니터링 데이터가 적재되는 디렉토리와 타 시스템에서 실시간 데이터로 적재된 데이터베이스에서 추출한 데이터를 적재하는 디렉토리, 배치 질의를 위해 벌크 로딩되어 빅데이터 저장형 파일 포맷으로 변환된 디렉토리, 그리고 사용자마다 질의를 처리하기 위해 질의를 정의하는 파일과 질의 처리 결과 파일을 관리하는 디렉토리로 구성된다. 이와 같은 구성을 지니는 시스템에서 각 서비스별 흐름은 다음과 같다.

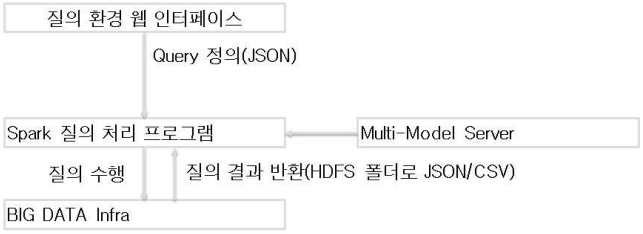
① 시스템의 입력 데이터는 발전설비에 부착된 센서들을 TimeStamp 순으로 동기화하고, 모든 센서에 대하여 같은 TimeStamp에 기록된 값을 저장하는 레코드를 고려한다.

② 해당 레코드는 실시간 처리 컴포넌트를 통해 입수되며, 파일 시스템에 연/월/일 별로 적재된다. 적재된 파일은 사용자 인터페이스를 통해 모니터링 될 수 있다.

③ 지속해서 입수되는 데이터는 실시간 처리 컴포넌트에서 MIN/MAX/AVG 등의 집계 또는 SELECT 검색 질의를 수행할 수 있다.

④ 빅데이터 질의 처리 컴포넌트를 통해 파일 시스템에 적재된 데이터를 배치 질의 처리할 수 있으며, 또는 다른 데이터베이스로부터 데이터를 업로드하면, 마찬가지로 질의 처리를 수행할 수 있다. 현재 구현한 질의는 집계 질의, 시간 및 센서별로 탐색할 수 있는 SELECT 질의, 그리고 Correlation 등의 통계 질의 처리를 Spark SQL 및 Spark ML을 활용하여 구현 및 제공한다.

⑤ 시간별, 센서별 선택 질의를 통해 탐색한 데이터는 딥러닝의 학습 데이터로 활용될 수 있으며, 학습을 통해 생성된 인공지능 모델은 다중 모델 서비스를 통해 추론 API를 제공하고, 스트리밍 데이터 질의 처리 컴포넌트에서 이를 호출하여 실시간으로 장애를 추론할 수 있는 기능을 제공한다.



(그림 2). 질의 처리 프로세스

빅데이터 질의 처리는 그림 2와 같이 YARN 스케줄

러를 통해서 이루어지며, 각 질의는 JSON의 형태로 질의 타입, 선택한 시작 및 종료 탐색 시간, 선택할 센서 리스트, 질의를 호출한 사용자를 기술하고 이를 질의 처리 모듈에 전달한다. 질의 처리 모듈에서는 사용자와 질의타입을 조합하여 실행할 질의 처리 프로그램의 이름을 지정하고, 해당 프로그램을 YARN에 제출하여 질의를 스케쥴링한다.

3. 결론

본 논문에서는 발전소 장비 고장 예측 등을 수행하는 인공지능 모델의 학습 및 추론을 지원하기 위해 센서 빅데이터를 수집, 저장, 배치 질의, 스트리밍 질의 등을 지원하는 시스템을 설계하였으며, 이를 기존 빅데이터 처리 소프트웨어로 잘 알려진 Kafka, Spark, Hadoop을 활용하여 구현하였다. 향후에는 [8]의 연구에서 제안한 Spark Streaming 기반 센서 데이터 통신을 위한 구조 최적화 부분을 Spark Structured Streaming에도 적용하여 시스템의 성능을 고도화하기 위해 연구를 수행할 예정이다.

사사

이 논문은 산업통상자원부의 재원으로 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구입니다. (No. 20181110100420)

참고문헌

[1] C. H. Yu, K. Park, K. A. Sarda, J. Lim, J. Um, “Automated fault prediction system for power plants using deep learning model”, ICIC express letters : an international Journal of research and surveys, vol 14, No.7, 711-719, 2020.07.

[2] Kreps, Jay, Neha Narkhede, and Jun Rao. "Kafka: A distributed messaging system for log processing." Proceedings of the NetDB. Vol. 11. 2011.

[3] Armbrust, Michael, et al. "Structured streaming: A declarative api for real-time applications in apache spark." Proceedings of the 2018 International Conference on Management of Data. 2018.

[4] Hadoop 3.3 https://hadoop.apache.org/docs/r3.3.0/, (2021.10.11. 접근).

[5] Armbrust, Michael, et al. "Spark sql: Relational data processing in spark." Proceedings of the 2015 ACM SIGMOD international conference on management of data. 2015.

[6] Chen, Tianqi, et al. "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems." arXiv preprint arXiv:1512.01274, 2015.

[7] Multi model server, https://github.com/awslabs/multi-model-server, (2021.10.11.접근).

[8] 엄정호, 유찬희, Komal Sarda, 박경석, “실시간 발전소 시설 장비 센서 데이터에 대한 빅데이터 스트리밍 질의 처리 시스템 설계 및 구현” 정보처리학회 추계학술대회, 2020.11.