스파크 기반의 대용량 데이터 압축을 이용한 실시간 데이터 분석 기법

박수용[O], 신용태[*]
[O]숭실대학교 컴퓨터학과,
[*]숭실대학교 컴퓨터학부
e-mail: waterdragon@soongsil.ac.kr[O], shin@ssu.ac.kr[*]

Real-time data analysis technique using large data compression based spark

Soo-Yong Park[O], Yong-Tae Shin[*]
[O]Dept. of Computer Science, Soongsil University,
[*]Dept. of Computer Science, Soongsil University

● 요 약 ●

스파크는 데이터 분석을 위한 오픈소스 툴이다. 스파크에서는 실시간 데이터 분석을 위하여 스파크 스트리밍이라는 기술을 제공한다. 스파크 스트리밍은 데이터 소스가 분석서버로 데이터 스트림을 전송한다. 이때 전송하는 데이터의 크기가 커질 경우 전송과정에서 지연이 발생할 수 있다. 제안하는 기법은 전송하고자 하는 데이터의 크기가 클 때 허프만 인코딩을 이용하여 데이터를 압축하여 전송시키므로 지연시간을 줄일 수 있다.

키워드: 스파크(spark), 데이터 분석(data analysis), 실시간(real time), 데이터 압축(data compression)

I. Introduction

데이터 분석은 여러 애플리케이션 또는 인터넷에서 생성된 데이터에 대해 분석을 수행하여 의미있는 정보를 추출해 내는 것을 의미한다. 데이터 분석에는 배치 방식과 스트리밍 방식이 있다. 배치 방식은 데이터가 일정량 쌓이게 되면 분석 서버로 전송하여 분석 서버에서 데이터 분석을 수행하는 방식을 사용한다. 배치 방식은 구현이 간단하다는 장점이 있지만 데이터가 수집된 후에 분석을 시작하기 때문에 실시간성이 떨어진다. 스트리밍 방식은 데이터가 발생하는 곳과 분석 서버가 소켓으로 연결되어 있다. 스트리밍 방식은 데이터가 발생하는 즉시 소켓을 통하여 분석 서버로 전송되어 분석을 시작한다. 스트리밍 방식은 구현이 복잡하나 실시간성을 가지고 있다. 인터넷의 발전으로 분석하고자 하는 데이터의 양이 증가하면서 소켓을 통하여 전송될 때 지연이 발생할 수 있다. 제안하는 기법은 실시간 분석이 필요한 대용량의 데이터를 압축하여 전송하고 분석서버에서 압축을 풀어 분석을 진행한다.

본 논문의 구성은 다음과 같다. 2장 관련 연구에서는 스파크에 대해 연구한다. 3장에서는 압축기법이 적용된 스트리밍 데이터 분석 기법을 제안한다. 4장에서는 결론 및 향후 연구 과제를 제시한다.

II. Related works

1. Spark Streaming

스파크는 빅데이터 분석을 위한 오픈 소스 병렬 처리 프레임워크다. 스파크는 데이터소스인 HDFS로부터 데이터를 가져와서 고속의 병렬 분산 처리한다. 스파크에서는 데이터를 배치 처리하기 위한 Map Reduce 방식도 지원하지만 실시간 분석을 위한 Streaming 방식도 지원한다. 스파크에는 데이터 처리를 위해 필터, 결합, 집계처리에 사용할 수 있는 API를 제공하고 있다. Fig. 1은 스파크에서 데이터를 처리하는 과정을 나타낸다.

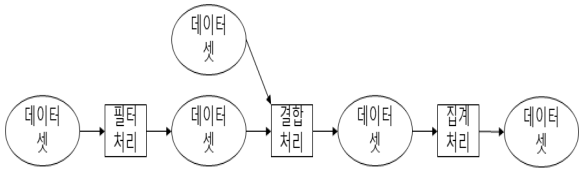

Fig. 1. 스파크에서 데이터를 처리하는 과정

스파크에서는 API를 사용하여 데이터를 처리할 때마다 새로운 데이터셋이 생성된다. 또한 대화식으로 처리할 수 있는 CLI를 제공한다.

III. The Proposed Scheme

스파크에서의 실시간 데이터 분석은 데이터소스와 분석서버를 소켓으로 연결한다. 데이터 소스는 소켓을 통하여 데이터 스트림을 스파크 분석서버로 전송한다. 분석 서버에서는 데이터 스트림을 배치 형태로 분할하고 각각 분석을 한 다음에 최종 결과물을 출력한다. Fig 2는 스파크에서의 실시간 분석 과정을 나타낸다.

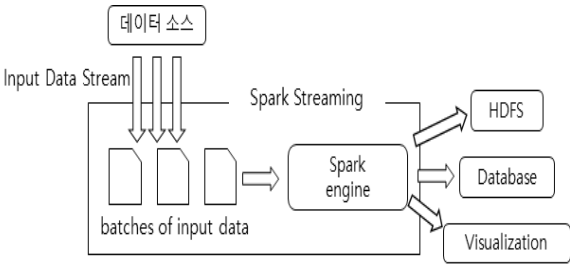

Fig. 2. 스파크에서의 실시간 데이터 분석 과정

Fig. 2에서 데이터 소스로부터 스트림이 전송될 때 보내야하는 데이터의 양이 크다면 지연이 발생할 수 있다. 전송해야 하는 데이터가 클 경우 큐에 저장하고 순서대로 전송하는 방식을 사용할 수 있지만 데이터가 큐에서 대기해야하기 때문에 실시간성이 떨어진다. 제안하는 기법은 데이터 소스가 전송해야하는 데이터의 양이 클 경우 압축을 진행하여 데이터의 양을 축소시켜 전송한다. Fig. 3.은 제안하는 실시간 분석 기법을 나타낸다.

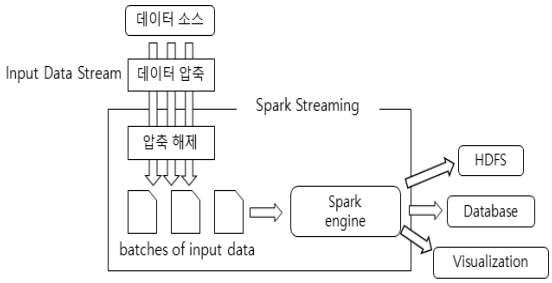

Fig. 3. 제안하는 분석 기법

분석에 사용되는 데이터는 대부분 텍스트 형식이다. 제안하는 기법에서 압축 알고리즘은 허프만 인코딩 방식을 사용한다. 허프만 인코딩 방식은 텍스트 형식의 압축 알고리즘 중 압축률이 뛰어난 호프만 인코딩을 사용한다. 허프만 인코딩은 문자의 빈도수에 따라 트리를 구성하고 트리의 가지에 0 또는 1의 값을 부여하여 분자를 압축한다. 따라서 압축을 해제 할 때 또한 트리가 필요하다. 제안하는 기법에서는 데이터 압축시에 사용한 트리를 같이 전송하여 분석서버에서 압축을 해제할 수 있도록 한다.

IV. Conclusions

본 논문에서는 스파크를 활용한 실시간 데이터 분석에서 데이터가 클 경우 데이터를 압축하여 전송하는 방식을 제안하였다. 제안하는 기법은 전송해야하는 데이터의 크기가 클 경우 압축을 통하여 크기를 줄여서 전송 지연시간을 줄일 수 있다. 그러나 데이터의 전송 시간보다 압축 시간이 길 경우에는 비효율적이기 때문에 추후 구현과 실험을 통하여 데이터 압축을 시작하는 임계점을 찾아야한다.

ACKNOWLEDGEMENT

본 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.IITP-2019-0-00135 ,ICT 기반 환경 모니터링 센서 신뢰성 검증 및 평가 플랫폼)

REFERENCES

[1] Jun-Seok Song, Sang-Young Kim, Byung-Hoo Song, Kyung-Tae Kim, Hee-Yong Youn, “A Performance Comparison of Machine Learning Library based on Apache Spark for Real-time Data Processing”, Proceedings of the Korean Society of Computer Information Conference, Vol. 25, No. 1, pp.15-16, January 2017

[2] Eunseong Lee, Jungho Kim, Seongsoo Hong, “Selective Resource Managing Technique for Real-Time Processing in Spark Streaming System”, The Korean Institute of Information Scientists and Engineers, pp.64-66, July 2017