한국컴퓨터정보학회 동계학술대회 논문집 제24권 제1호 (2016. 1)

K-means를 이용한 아파치 스파크 및 맵 리듀스 성능 분석

정영교*, 정동영O, 송준석*. 윤희용*
*O성균관대학교 전자전기컴퓨터공학과
e-mail:{joung0491, jungdy, alskpo, youn7147}@skku.edu*O

Apache Spark and Map Reduce with Performance Analysis using K-Means

Young-Gyo Jung*, Dong-Young JungO, Jun-Seok Song*, Hee-Yong Youn*
*ODept. of Electrical and Computer Engineering, Sungkyunkwan University

● 요 약 ●

빅 데이터의 데이터 수집 및 분석 기술에 대한 연구는 컴퓨터 과학 분야에서 각광 받고 있다. 또한 소셜 미디어로 인한 대량의 비정형 데이터 분석을 요구하는 다양한 분야에 접목되어 효용성을 인정받고 있다. 그러나 빅 데이터 개념을 기반으로 하는 하둡과 스파크는 유즈케이스에 따라 성능이 크게 달라진다는 문제점이 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 하둡의 맵리듀스를 줄이고 아파치 스파크를 이용한 빅 데이터 분석을 위하여 머신러닝 알고리즘인 K-Means 알고리즘을 이용하여 프로세싱 모델의 성능을 비교한다.

키워드: 빅 데이터(Big data), 맵 리듀스(Map Reduce), 스파크(Spark), K-Means

I. Introduction

빅 데이터는 데이터 수집 및 분석 기술에 대한 연구는 컴퓨터 과학 분야에서 각광 받고 있으며 아파치 하둡(Apache Hadoop)의 HDFS(Hadoop Distributed File System)[1]와 맵 리듀스 (Map Reduce)[2] 프레임워크로 구성된 빅 데이터를 처리 분석과 아파치 스파크(Apache Spark)[3]의 분산 메모리를 활용하여 대량의 데이터 볼륨의 빠른 처리를 가능하게 RDDs(Resilient distributed Datasets)의 개념을 사용하는 연구가 활발히 진행되고 있다. 하지만 빅 데이터를 기반으로 하는 하둡과 스파크는 유즈케이스에 따라 성능이 달라진다는 문제점이 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 하둡의 맵 라듀스를 줄이고 스파크를 이용한 빅 데이터 분석을 위하여 머신러닝 알고리즘인 K-Means 알고리즘을 이용하여 프로세싱 모델의 성능을 비교한다.

II. Preliminaries

1. Related works

1.1 K-Means

K-Means 알고리즘은 비 계층적 클러스터링 기법으로, 문서와 클러스터의 중심값을 나타내는 centroid와의 유사도를 측정하여 문서를 적합한 클러스터에 재배치하는 방법이다. centroid는 클러스터에 속하는 문서들의 평균 벡터값을 이용한다.

K-Means 알고리즘은 특성상 생성된 클러스터 중심에 따라 클러스터링 결과가 달라진다. 특히 초기 클러스터중심을 어떻게 선택하는가에 따라 빠른 시간에 최적의 클러스터링 결과가 나오는 경우와 그렇지 않은 경우가 존재한다. 클러스터링에 영향을 미치는 또 다른 요소는 클러스터링 과정에 서 발생하는 새로운 클러스터링 중심(cluster centroid)을 결정하는 것이다.

III. The Proposed Scheme

K-Means 알고리즘을 사용하여 성능 평가를 했다. 성능평가 결과 스파크가 메모리 처리를 하는데 뛰어난 것으로 나왔다.

아파치 스파크와 맵 리듀스를 비교하기 위해 K-Means 알고리즘을 이용하여 클러스터링을 수행 할 수 있도록 data set의 프레임 워크를 이용하여 비교 분석을 수행 하였다. data set은 각각의 레코드의 위도 및 경도 값을 포함한 1240MB의 센서 데이터를 이용했다.

데이터 레코드의 샘플은 다음과 같이 구성 되어 있다.

1. Date
2. Device Name
3. Device ID
4. Status
5. Latitude
6. Longitude

data set을 K-Means 알고리즘을 이용하여 실행했다.

다양한 분석을 얻기 위해 싱글노드에 64MB, 1240MB와 두 개의 노드에 1240MB로 설정한 뒤 K-Means 알고리즘을 이용하여 필요조건에 따라 클러스터링 소요 시간을 성능평가 하였다.

성능평가 환경은 다음과 같이 구성하였다.

- 4GB RAM
- Linux Ubuntu
- 500GB Hard Drive

그림 1과 그림 2는 스파크를 와 Map Reduce를 이용한 K-Means 성능 평가 결과이다. 성능평가 결과 스파크의 성능이 맵리듀스 보다 최대 3배의 처리 시간 감소로 시간의 측면에서 상당히 높은 것으로 나타났다.

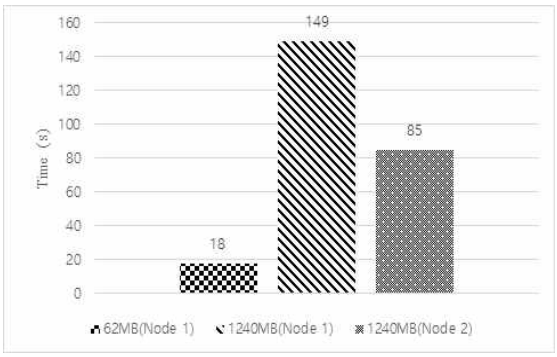

Fig. 1. Results for K-Means using Spark(MLlib)

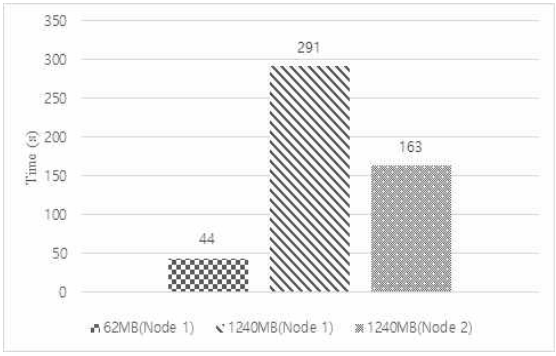

Fig. 2. Results for K-Means using Map Reduce(Mahout)

IV. Conclusions

본 논문에서는 하둡의 맵 리듀스를 줄이고 스파크를 이용한 빅데이터 분석을 위하여 하둡과 스파크 두 프레임워크를 K-Means 알고리즘을 이용하여 성능 비교를 하였다. 성능 평가 결과 스파크 성능이 뛰어나다는 것을 알 수 있었다.

향후 연구로는 맵 리듀스를 기반으로 하고 있는 Mahout의 대부분의 알고리즘을 기존의 맵 리듀스 보다 뛰어난 성능을 내는 스파크로 대체하는 연구가 필요하다.

Acknowledgment

본 연구는 BK21Plus 사업, 한국연구재단 기초연구사업 (2013R1A1A2060398), 삼성전자, 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업 (1391105003), 미래부/정보통신방송연구개발사업의 일환으로 수행하였음.

References

[1] Shvachko K., Hairong Kuang, Radia S, Chansler, R The Hadoop Distributed File System Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium.

[2] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In OSDI'04: Sixth Symposium on Operating System Design and Implementation, 2004.

[3] Apache Spark Research, https://spark.apache.org/research.html.