빅데이터 처리를 위한 PC와 라즈베리파이 클러스터에서의 Apache Spark 성능 비교 평가

서지혜*, 박미림*, 양혜경*, 용환승*
*이화여자대학교 컴퓨터공학과
e-mail: yc99603@naver.com

Performance Evaluation Between PC and RaspberryPI Cluster in Apache Spark for Processing Big Data

Ji-Hye Seo*, Mi-Rim Park*, Hye-Kyung Yang*, Hwan-Seung Yong*
*Dept of Computer Science and Engineering, Ewha Womans University

요 약

최근 IoT 기술의 등장으로 저전력 소형 컴퓨터인 라즈베리파이 클러스터가 IoT 데이터 처리를 위해 사용되고 있다. IoT 기술이 발전하면서 다양한 데이터가 생성되고 있으며 IoT 환경에서도 빅데이터 처리가 요구되고 있다. 빅데이터 처리 프레임워크에는 일반적으로 하둡이 사용되고 있으며 이를 대체하는 솔루션으로 Apache Spark가 등장했다. 본 논문에서는 PC와 라즈베리파이 클러스터에서의 성능을 Apache Spark를 통해 비교하였다. 본 실험을 위해 Yelp 데이터를 사용하며 데이터 로드 시간과 Spark SQL을 이용한 데이터 처리 시간을 통해 성능을 비교하였다.

1. 서론

최근 IoT(Internet of Things) 디바이스들의 발달로 인하여 다양한 형태의 IoT 디바이스들이 출시되고 있다. 특히 가장 주목받고 있는 IoT 디바이스로 라즈베리파이가 있다. 라즈베리파이(Raspberry Pi)는 저전력 싱글보드 컴퓨터로 저렴하고 데비안 계열의 리눅스 체제를 이용하기 때문에 리눅스에서 사용하는 라이브러리 설치가 용이하다는 장점이 있다[1][2]. 올해 라즈베리파이2 모델 B버전이 출시되면서 소형 컴퓨터지만 쿼드코어와 1GB의 메모리 탑재로 인해 좋은 성능을 보이기 때문에 라즈베리파이를 이용한 연구가 더욱 주목받고 있다. 본 논문에서는 저전력 소형 컴퓨터인 라즈베리파이를 이용한 빅데이터 환경을 구축하여 일반 컴퓨터와의 성능 비교 평가하였다. 아파치 spark는 인 메모리 컴퓨팅 기반의 데이터 분산 처리 시스템[3][4]으로 반복적인 데이터 연산에 있어 디스크 I/O 비용을 효율화하기 위해 등장하게 되었다. 본 논문에서는 빅데이터 처리 엔진으로 아파치(Apache)재단에서 진행하는 Spark를 이용하였다. 반복적인 연산에 약했던 기존 하둡(Hadoop)시스템 보다 아파치 spark는 연산 속도가 빠르며, HADOOP[5]과 다르게 Spark를 이용하면 HBASE[6]와 HIVE[7]와 같은 SQL 질의를 제공하는 빅데이터 시스템을 별도로 설치하지 않고도 SQL문이 사용가능하다. 그렇기 때문에 본 연구에서는 이런 장점을 갖고 있는 Spark를 사용하여 일반 컴퓨터와의 성능 평가를 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 빅데이터 처리 시스템의 설계 및 구현에 대해 설명 하고, 3장에서는 라즈베리파이와 일반 컴퓨터와의 성능 비교를 평가하였다. 마지막으로 4장에서 결론과 향후 연구를 제시하였다.

2. 시스템 환경 구축

<표 1>과 <표 2>는 본 논문에서 사용한 라즈베리파이 2 모델 B의 사양과 PC사양을 정리한 표이다.

<표 1> Raspberry Pi 2 Version B Spec

category	Spec of Raspberry Pi 2
CPU	900 MHz ARM Cortex-A7 QuadCore
Memory	1GB LPDDR2
Network	10/100 Mbit
OS	Raspbian

<표 2> PC spec

category	Spec of PC
CPU	Xeon E3-1220v3 3.1GHz QuadCore
RAM	8GB
Internal Cache	8MB L3 Cache
Hard Disk	500GB 7.2K 6Gbps NL SATA 3.5″ SS
OS	Linux mint 17 cinnamon 64bit

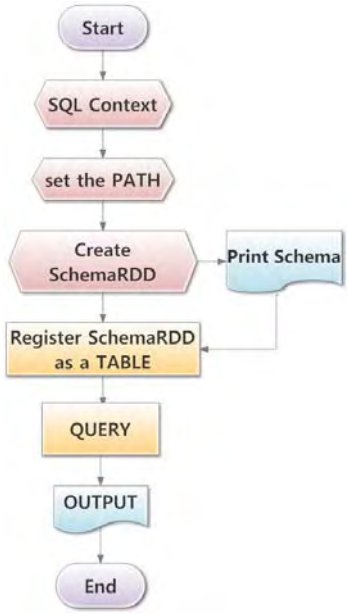
본 논문에서는 라즈베리파이에 Spark 버전 1.2.1를 사용하여 완전 분산모드 환경을 구축하였다. 완전분산모드 환경을 위해 1대의 마스터노드와 2대의 워커 노드로 총 3대의 라즈베리파이를 사용하였다. 데이터 저장장치는 라즈베리파이 보드 별로 16GB의 SD카드 메모리를 사용하였다.

본 논문에서 라즈베리파이와 일반 컴퓨터의 빅데이터 처리 성능을 비교를 위해 일반 컴퓨터 환경은 1대일 경우와 완전분산모드일 경우로 구성하였다. 완전분산모드인 경우 라즈베리파이와 동일하게 마스터 노드 1대 워커 노드 2대로 구성하였다. 일반 컴퓨터 환경은 Linux mint 17 cinnamon 64bit 운영체제에 Spark 1.2.1으로 구축하였으며, 컴퓨터의 RAM 사양은 8GB이다.

3. 성능 비교 평가

라즈베리파이와 일반 컴퓨터와 성능을 비교하기위해 미국 비즈니스 평가 사이트인 YELP.COM[8] 에서 제공하는 JSON형식의 파일 데이터를 이용하였다. 본 연구에서는 사용가능한 JSON데이터 중 비즈니스(business)와 후기(review), 사용자(user) 데이터만을 이용하였다. 비즈니스 관련 데이터 크기는 52.9MB, 사용자 데이터 크기는 158.5MB, 그리고 후기 데이터는 1.3GB이다.

본 연구에서는 Spark에서 제공되는 SQL문을 사용하기 위해 JSON 데이터를 스키마로 변환하는 작업을 수행한다. (그림 1)은 Spark SQL에서 JSON 데이터를 스키마로 만들고 질의(query)를 실행하는 과정을 보여주는 순서도이다.



(그림 1) Process JSON File in Spark SQL(Flow Chart)

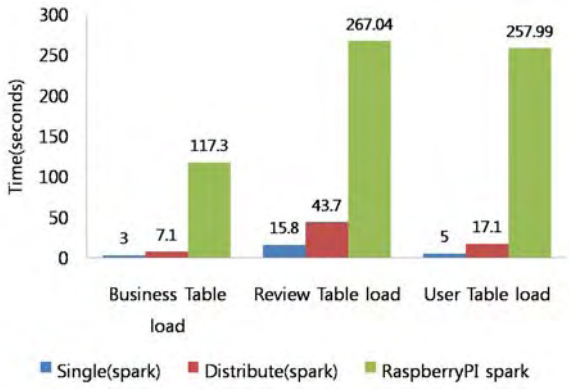
비즈니스 데이터에는 해당 사업에 관한 정보들을 포함하고 있다. 예를 들어, 사업에 대한 비평가들(reviewers)의 평가, 사업장의 주차 가능 유무, 위치, 여업시간, 애완동물 허용 등이 있다. 후기 데이터에는 (그림 2)의 스키마에 나온 정보처럼 별의 개수, 투표 현황 등이 있다. 사용자 데이터는 사용자가 이용하는 ID와 친구 목록 등이 있다.

```
root
 |-- business_id: string (nullable = true)
 |-- date: string (nullable = true)
 |-- review_id: string (nullable = true)
 |-- stars: integer (nullable = true)
 |-- text: string (nullable = true)
 |-- type: string (nullable = true)
 |-- user_id: string (nullable = true)
 |-- votes: struct (nullable = true)
 |    |-- cool: integer (nullable = true)
 |    |-- funny: integer (nullable = true)
 |    |-- useful: integer (nullable = true)
```

(그림 2) Schema of Review Schema

3.1 데이터 로딩 시간(성능) 비교

(그림 3)은 빅데이터 처리 엔진인 Spark에 JSON 데이터를 스키마로 변환하여 각각 비즈니스, 후기, 사용자 스키마에 데이터를 로딩 하는데 소요된 시간을 보여주는 그래프이다.



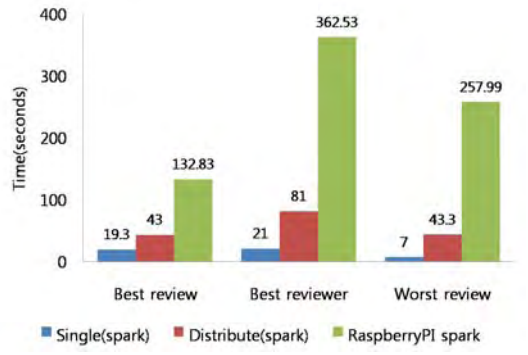
(그림 3) Data Loading Time Among Single, Distributed Environment and Raspberry Pi Cluster

(그림 3)을 보면 세 가지 환경에서 데이터를 로딩해본 결과는 한 대의 일반 컴퓨터가 소요된 시간이 가장 짧았고, 분산된 환경, 라즈베리파이 클러스터 순이었다. 여기서 한 대의 일반 컴퓨터가 분산 환경인 일반 컴퓨터보다 빠른 이유는 분산 처리를 할 경우 데이터를 모으기 위해 각 데이터 노드에 접근을 하여 데이터를 모으는 작업을 하게 된다. 따라서 네트워크 영향을 받아 처리 속도가 느리게 된다. 하지만 하나의 노드로 처리하기 힘든 데이터의 크기라면 분산 처리 속도가 빠르게 나타나게 된다. 또한 라즈베리파이 클러스터가 일반 컴퓨터와 비교하여 속도가 느린 이유는 일반 컴퓨터의 RAM사양이 8GB이기 때문에 라즈베리파이와의 성능 차이가 있기 때문에 이러한 결과를 얻게 되었다.

3.2 데이터 처리 시간(성능)비교

본 논문에서는 로드한 Yelp 데이터에서 좋은 후기를 갖는 상위 후기와 상위 비평가들 그리고 하위의 후기를 찾는 질의문을 작성하여 데이터 처리시간을 비교하였다.

(그림 4)는 Spark에서 세 가지 질의를 수행하고 소요된 시간을 보여주는 그래프이다.



(그림 4) Processing Query Time Among Single, Distributed Environment and RaspberryPI Cluster

데이터 로딩을 한 결과와 같이, 질의를 수행할 때도 한 대의 일반 컴퓨터에서 가장 빠르게 수행했으며, 라즈베리파이 클러스터에서는 일반 컴퓨터보다는 수행시간이 오래 걸렸다. 이 결과 또한 앞서 설명한 이유와 같이 네트워크의 영향과 일반 컴퓨터 사양이 라즈베리파이 클러스터보다 좋기 때문에 이러한 결과가 나타났다.

4. 결론

본 연구에서는 라즈베리파이2 클러스터를 이용한 빅데이터 처리환경을 구축하여 일반 컴퓨터 1대와 일반 컴퓨터 3대를 연결한 완전분산 환경에서의 성능을 평가하였다. 데이터 로딩과 질의 수행을 모두 실험해본 결과 일반 컴퓨터 1대, 분산된 환경, 라즈베리파이 클러스터 순으로 나타났다. 이는 현재 라즈베리파이가 일반 컴퓨터보다 사양적인 측면에서 떨어지기 때문에 이러한 결과를 얻게 되었다. 하지만 3대보다 더 많은 라즈베리파이를 이용하여 클러스터를 구성하게 된다면 일반 컴퓨터보다 좋은 성능을 보일 것이다. 따라서 향후 연구로 더 많은 라즈베리파이를 이용하여 일반 컴퓨터와 성능 비교를 진행해야한다. 또한 작은 개수의 라즈베리파이를 이용한 클러스터에서도 데이터 처리 속도를 향상시키는 방안에 대한 연구가 필요하다.

Acknowledgement

이 논문은 2014년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초사업 지원을 받아 수행된 것임 (2012R1A1A2003764)

참고문헌

[1] K. M. Ji and U. S. Kim, “Raspberry Pi using the Private Cloud Service,” The KSII Transactions: vol.14, no. 2, pp.155-156, 2013.

[2] Raspberry Pi Wikipedia, https://en.wikipedia.org/wiki/Raspberry_Pi, (2015.08.27)

[3] Apache Spark, http://spark.apache.org, (2015.09.02)

[4] Matei Z, Mosharaf C, Tathagata D, Ankur D, Justin M, Murphy M, Michael J,F, Scott S and Ion S, "Resilient Distributed Dataset: Fault-Tolerant Abstraction for In-Memory Cluster Computing," in Proceedings of the 9th USENIX conference on Networked System Design and Implementation(NSDI'12), 2012.

[5] Apache Hadoop, http://hadoop.apache.org, (2015.05.24)

[6] Apache HBase, http://hbase.apache.org, (2015.05.24)

[7] Apache Hive, http://hive.apache.org, (2015.05.24)

[8] Yelp, http://www.yelp.com(2015.06.02)