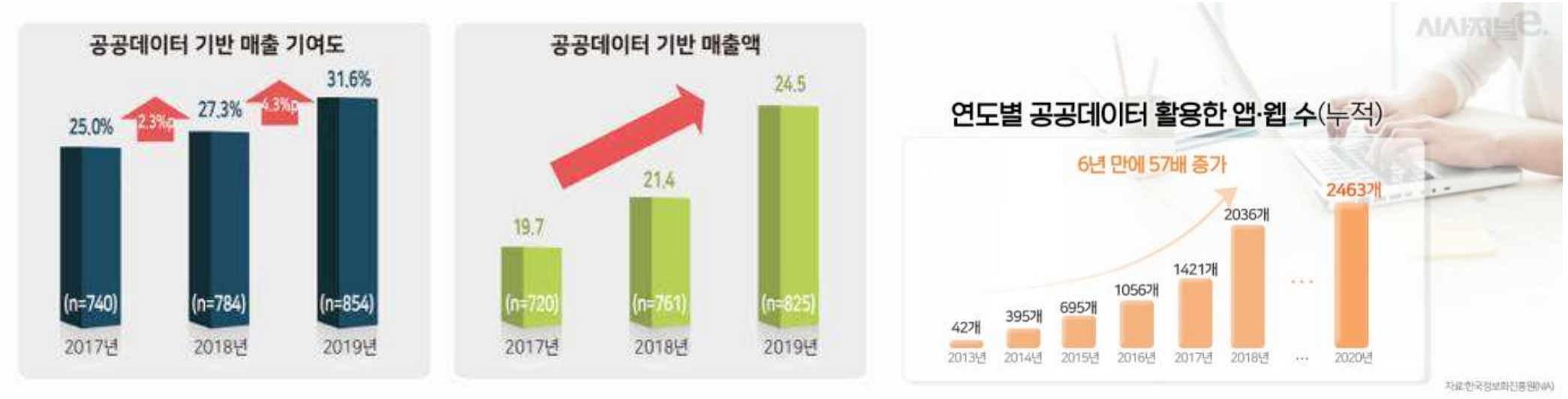
빅데이터 처리를 위한 소형 클러스터 구축과 분석

조이
1716851 이은서
1713941 조혜민

목차

연구 배경과 목적

개인에게 더 낮은 비용으로 높은 성능을 가진 데이터 처리 시스템을 제공하기 위해 빅데이터 처리를 위한 소규모 클러스터를 구축하고, 성능을 분석한다.



Raspberry Pi

- 교육 목적으로 개발된 초소형 컴퓨터
- 다양한 활용법
- 가격과 확장성에서 유리

Apache Hadoop

· 빅데이터 처리를 위해 개발된 분산 시스템

· **분산 파일 시스템**과, **맵 리듀스**로 구성

· 대용량 데이터를 값싸고 빠르게 분석할 수 있게 해줌

Apache Spark

· 오픈 소스 클러스터 컴퓨팅 프레임워크

· **고속 병렬 분산처리**

· 동일한 데이터에 대한 변환 처리가 연속으로 이루어지는 작업에 매우 효율적

실험 환경

· 하드웨어

- **Raspberry pi 3B**(Quad Core 1.2GHz Broadcom BCM 2837 64bit CPU, 1GB RAM)
- 16G Micro SD카드
- 8포트 스위치 허브

· OS / Software

- Ubuntu 18.04
- Apache Hadoop Version 3.1.3
- Apache Spark Version 2.3.0

· 네트워크

- 1000mbps LAN Port

· 데스크탑 사양

- Intel(R) Core(TM) i5-9600K CPU @ 3.70GHz
- 8GB RAM

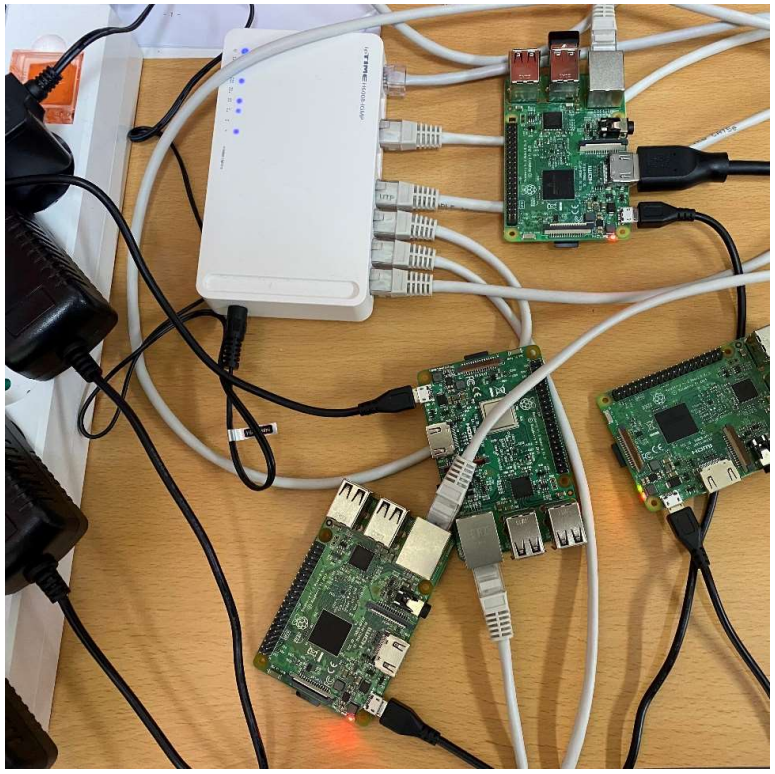
연구 방법 및 내용

① **5대의 라즈베리파이**를 1000Mbps 스위치 허브와 랜선으로 병렬 연결하여 소형 클러스터 구축

② 각각의 라즈베리파이에 **Ubuntu 18.04 운영체제**와 **Hadoop 3.1.3**, **Spark 2.3.0**를 설치

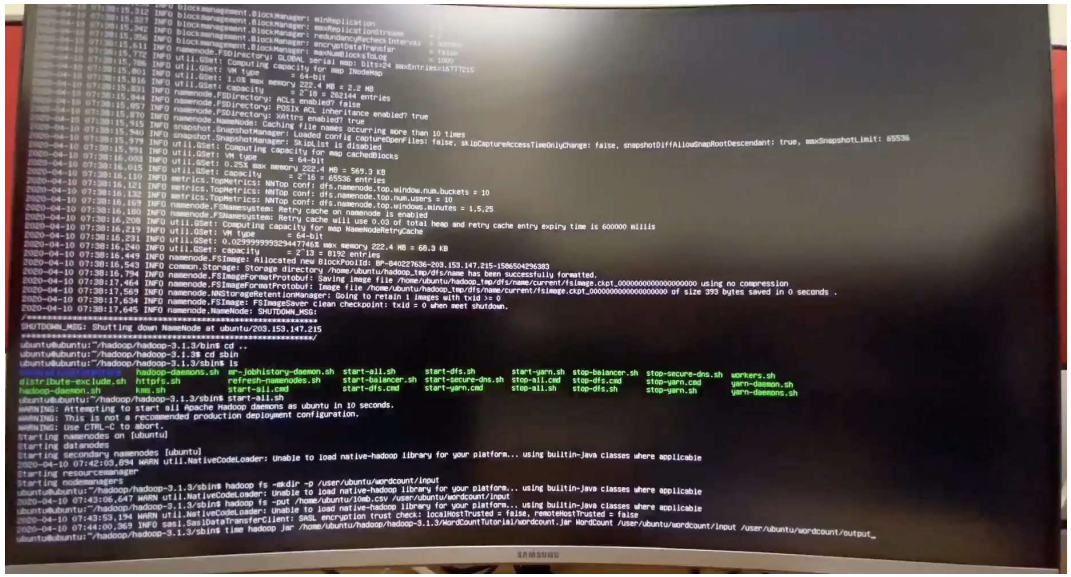
③ **노드 개수, 데이터 크기**를 변화시키며 각각의 실험에 알맞게 네트워크 구성과 클러스터 환경을 설정하고 하둡과 스파크에서 **Wordcount**를 수행

④ 데스크탑 환경에서의 수행 시간과 비교하여 **가격 당 성능 분석**

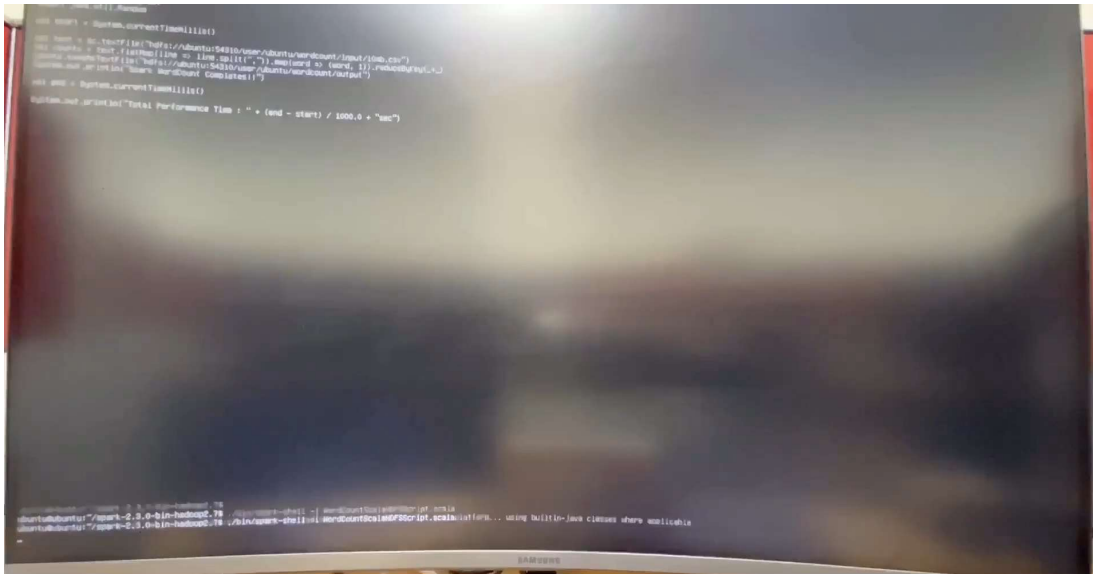


▲ 실험에 사용한 라즈베리파이 클러스터

시연 영상

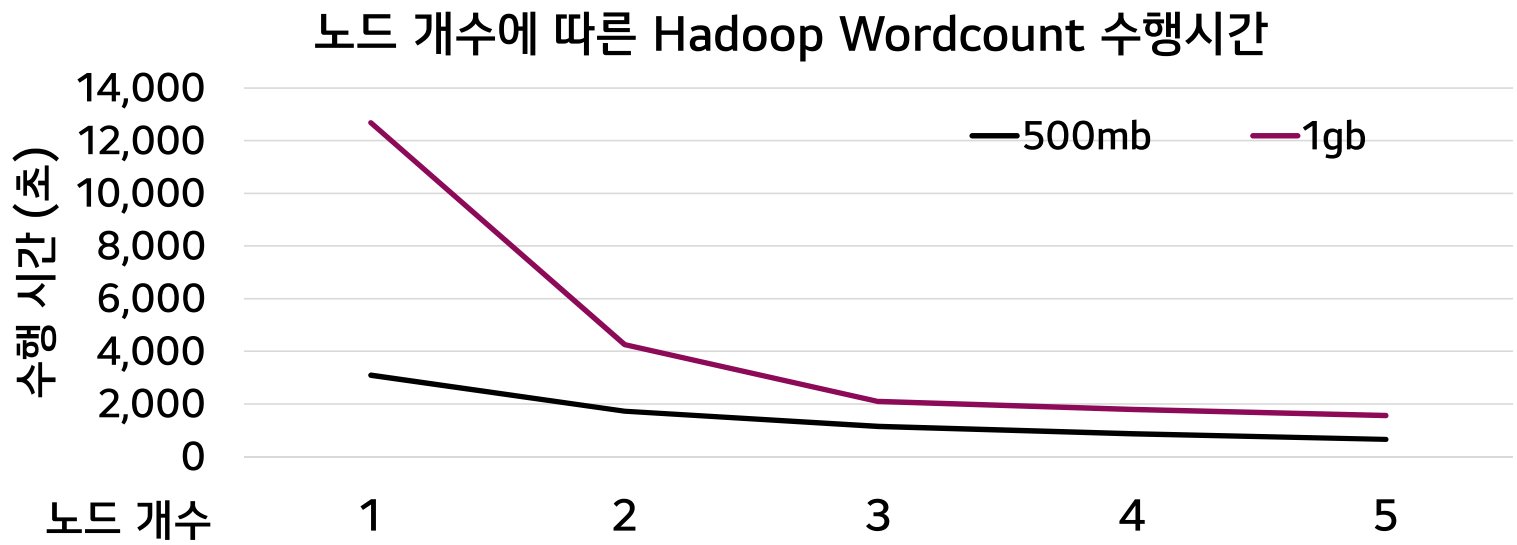


Hadoop 시연 영상



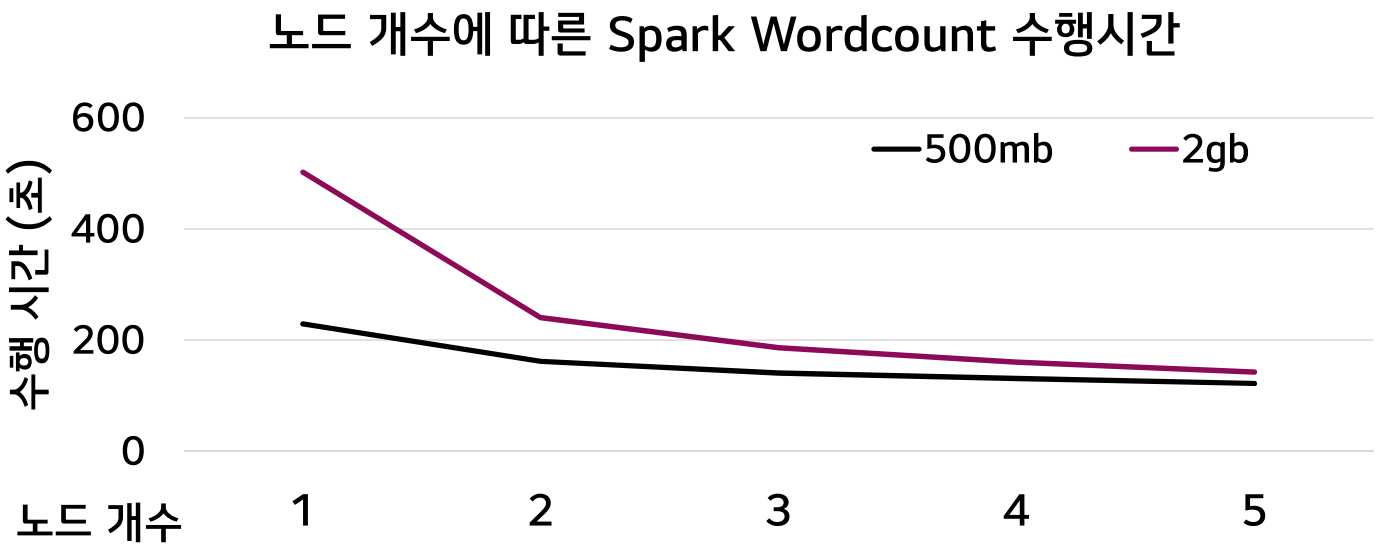
Spark 시연 영상

실험 결과 (Hadoop)



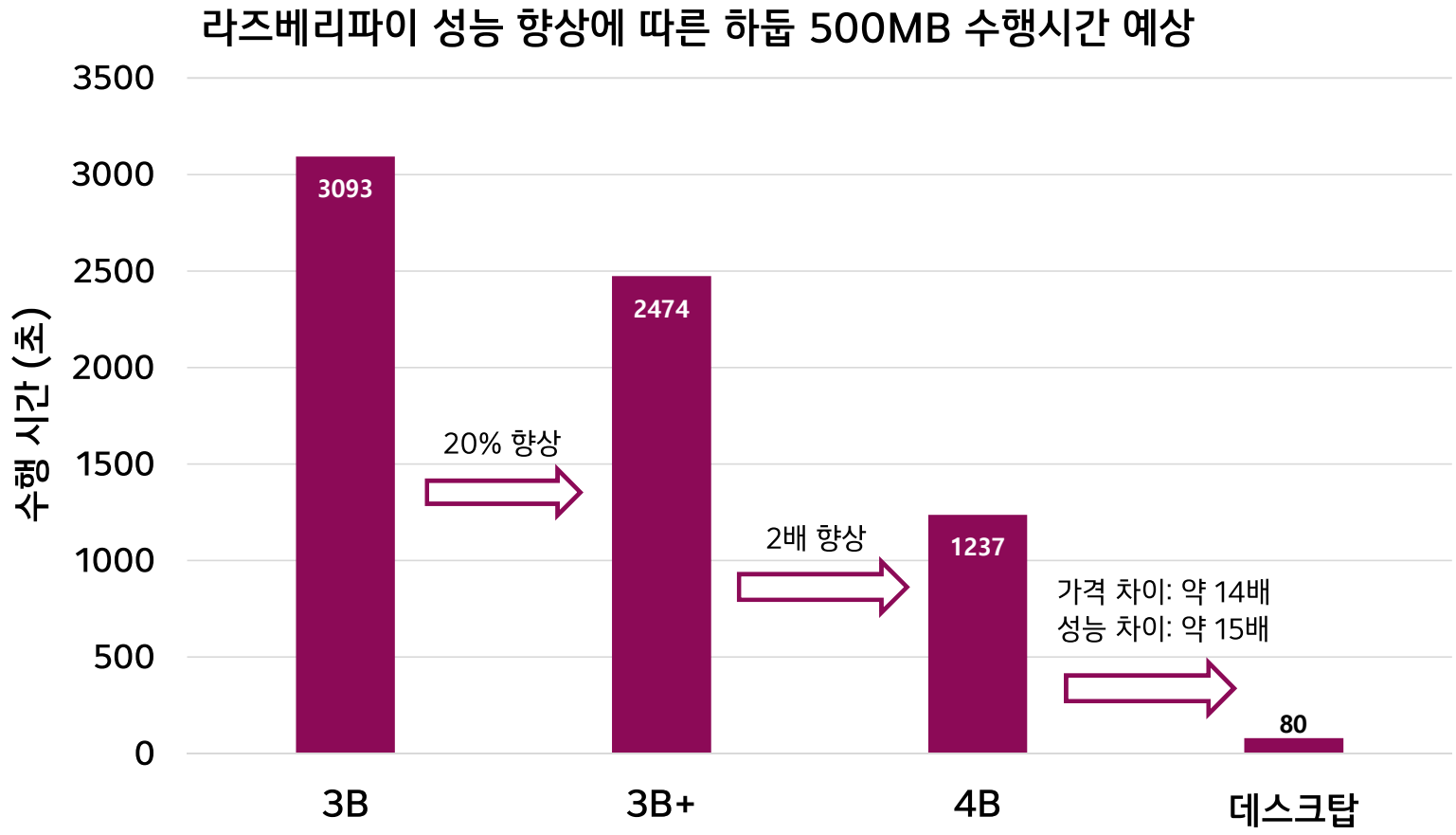
단위(초)	1	2	3	4	5	데스크탑
500MB	3,093	1,732	1,147	867	654	80
1GB	12,680	4,263	2,094	1,794	1,560	154

실험 결과 (Spark)



단위(초)	1	2	3	4	5	데스크탑
500MB	229.298	162.112	141.112	131.277	122.203	4.492
2GB	502.264	240.609	186.399	160.749	142.821	10.203

가격 당 성능과 개선 가능성



어려웠던 점

- **적절한 OS 선택** : 실행 환경에 적합한 OS를 찾기 힘들었던 문제
  → 여러 OS를 직접 설치해보고 비교하여 선택

- **힙 메모리 설정 문제** : 큰 데이터로 수행할 때 힙 메모리 설정 발생
  → 메모리와 관련된 환경 변수들을 다양하게 설정해보고 가장 적합한 크기로 선택

- **스왑 파일** : 램 용량이 작기 때문에 스왑 파일 할당 필요
  → 적절한 스왑 파일 용량을 여러 번 테스트해보고 설정

- **효율적인 병렬처리 문제** : 클러스터의 병렬 처리가 효율적으로 되지 않는 문제
  → Mapper와 Reducer의 개수를 적정하게 조정하여 해결

- **수행 시간 문제** : 같은 환경에서도 수행 시간의 편차가 크게 생기는 문제
  → Reduce 시작 시점에 따라 수행 시간이 달라지는 것을 발견하고 조정하여 해결

더 연구해볼 점

- 같은 환경의 클러스터에서 **Wordcount가 아닌 어플리케이션**을 실행하였을 경우 어떤 차이가 있는지 비교해보기

- 네트워크 성능이 중요하게 작용하는 하둡에서 **네트워크를 개선하였을 경우 얼만큼의 성능 향상이 가능한지 분석**해보기

- **Watt당 성능 비** 비교해보기

참고 자료

- 라즈베리 파이 모델 성능 비교표

	Raspberry Pi 3B	Raspberry Pi 3B+	Raspberry Pi 4
CPU	Quad Core 1.2GHz Broadcom BCM2837 64bit CPU	Broadcom BCM2837B0, Cortex-A53 (ARMv8) 64-bit SoC @ 1.4GHz	Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
Memory	1GB RAM	1GB LPDDR2 SDRAM	1GB, 2GB or 4GB LPDDR4-3200 SDRAM (depending on model)
Ethernet	100 Base Ethernet	Gigabit Ethernet over USB 2.0 (maximum throughput 300Mbps)	Gigabit Ethernet