J. Chosun Natural Sci.
Vol. 13, No. 3 (2020) pp. 114 – 119
https://doi.org/10.13160/ricns.2020.13.3.114

IoT 기반 빅데이터 효율성 향상을 위한 하둡기반 플랫폼 설계

장경성†, 배상현

Design for Haddop-based Platform to Improve Io T-based Big Data Processing Efficiency

Jang Kyungsung† and Sang Hyun Bae

Abstract

IoT 및 사물인터넷 기반 빅데이터 시스템을 구축하는 경우 발생하는 빈번한 전송에 따른 데이터 오류율과 자원의 비효율적 이용율을 극복하기 위하고 오픈소스기반 하둡시스템의 문제점을 극복하기 위한 본 연구에서는 순수 하둡을 기반으로 적용된 결과를 분석하고 하둡 2.x대 버전을 기준으로 빅데이터 시스템의 용량을 산정한 가이드를 제시하고 용량 산정의 기준을 에코 소프트웨어 적용 플랫폼을 제안한다.

Keywords : 빅데이터, 하둡, IoT, 인공지능

1. 서 론

4차산업시대 IoT의 궁극적인 잠재력은 이러한 시스템의 각 단계(step)에서 획득된 데이터로부터 가치를 도출할 능력을 갖고 실현에 대한 요구도 급증하고 있다. 또한 빅데이터 산업의 방대한 양의 데이터가 축적되고, 이를 분석함으로써 이전에는 얻을 수 없었던 새로운 가치(insight)를 창출하고자 하는 시도로, IoT가 빅데이터에 중요한 기여를 할 것이며, IoT 기반 빅데이터 기술은 센서네트워크(sensor networks)를 포함, Wi-Fi, 휴대전화망(cellular networks), 무선 망(wired networks) 상에 수많은 물리적 사물들이 연결되어 서로 통신하며, 지능서비스를 위한 데이터를 생성하여, 방대한 빅 데이터를 통해 새로운 가치를 창출할 것이다.[2]

빅데이터는 광범위한 에코 소프트웨어로 이루어져 있다. 예를 들어 비정형 데이터를 효과적으로 저장하고 분석하기 위한 기반이 되는 하둡의 경우 아파치 소프트웨어재단(Apache Software Foundation)을 중심으로 한 오프소스 하둡과 엔터프라이즈 기반의 클라우데라(Cloudera), 호튼웍스(Hortonworks), 맵알(MapR) 등의 기업에서 만든 하둡이 있다. 이러한 하둡은 각기 데이터 및 서비스 기반의 특성이 있기 때문에 에코 소프트웨어와 함께 세밀하게 검토 및 적용이 이루어져야 한다.[1]

본 연구에서는4차산업시대에서 실시간 처리를 위한 빅데이터 프로세싱 모델을 디자인한다. 원천 데이터는 각종 센서 및 모바일 장비에서 나오는 비정형 데이터가 중심이다. 이러한 데이터의 특성은 실시간적으로 생성이 되며, 데이터가 유실되었을 경우 분석의 결과치에 대한 어려움 때문에 기계의 오작동을 유발하기도 한다.

이러한 부분을 방지하기 위해서 큐(Queue)를 활용하게 되며, 오픈소스인 카프카(Kafka)를 사용하면 효과적으로 실시간 데이터 처리가 가능해진다. 이외에도 연계할 부분은 스파크(Spark), 스톰(Storm) 같은 실시간 처리 소프트웨어와 함께 솔(s ol r) 또는 일래스틱서치(Elastic Search) 같은 이벤트 분석 시스템도 필요하게 된다. [그림 2]의 아키텍처를 활용하게 되면 수집에서부터 실행까지 다양한 제조 공정에서 활용 및 모니터링이 가능하게 된다.

효과적인 스마트 공장을 구축하기 위해서 빅데이터는 가장 기본적인 요소이면서도 중요한 핵심부분이라고 할 수 있다. 이러한 스마트 공장은 급변하는 비즈니

SW중심대학, 조선대학교
자연과학대학 전산통계학과, 조선대학교

†Corresponding author : ksjang2007@chosun.ac.kr
(Received : August 19, 2020, Revised : August 29, 2020, Accepted : September 4, 2020)

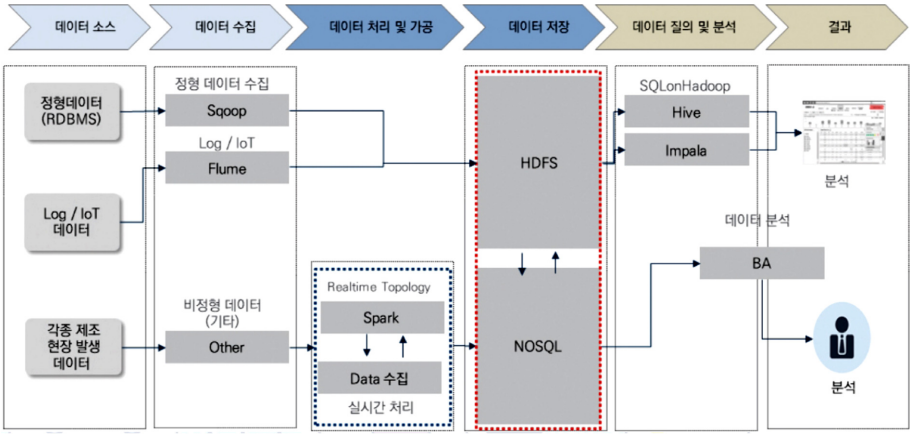

그림 1. 빅데이터 처리 절차

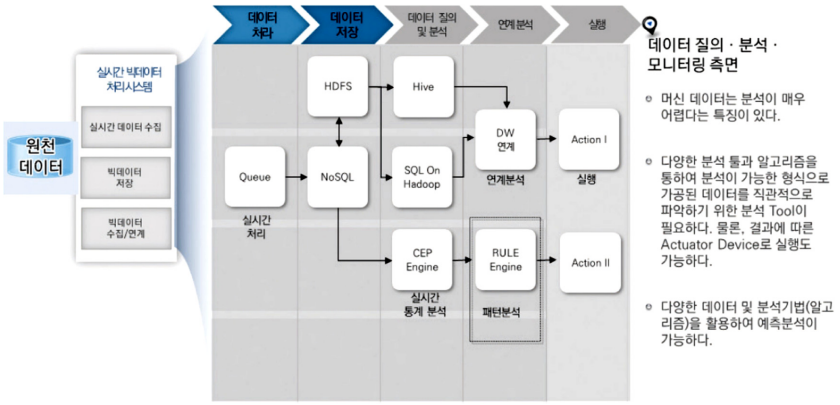

그림 2. 실시간 빅데이터 시스템 모델

스 환경에서 고객의 니즈를 파악하고, 제조현장에 빠른 적용을 통해서 기업의 경쟁력을 강화하는 요소라고 생각한다.

2. 실시간 비정형 빅데이터 플렛폼 설계

빅데이터 플랫폼은 데이터의 형태와 서비스의 형태에 따라서 다양한 인프라 스트럭쳐와 에코소프트웨어로 설계를 하여야 한다. 이러한 이유로 빅데이터 시스템을 설계할 때 표준 프레임이 없고 해당 요건에 따라서 구축이 되기 때문에 다양한 경험적 노하우에 의해서 구축이 되고 있는 것이 현실이다. 따라서 빅데이터 플랫폼은 다음과 같은 전체 조건에 의해서 에코 소프트웨어야 설계되어야 한다.

첫째, 데이터를 기반으로 하는 경우이다.

데이터를 기반으로 하는 빅데이터 시스템의 경우 데이터의 형태가 정형, 비정형으로 구분이 되고 정형은 일반적인 OLTP(Online Transaction Processing)에 적합한 형태의 데이터이며, 비정형 데이터의 경우는 기존의 OLAP(On Line Analytical Processing)에 적합한 형태의 비정형 구조를 가지고 있다. 이러한 비정형 데이터의 형태는 사진, 동영상, 음성, 텍스트등 다양한 형태의 속성을 가진 데이터의 경우를 의미한다.

이러한 데이터의 속성에 기반을 둔 처리 중심의 빅데이터 시스템인 경우는 그의 구현 형태가 데이터 중심적이어야 효율적인 형태의 시스템의 구현이 가능하다.

특히, 사물인터넷 즉 기계데이터를 기반을 둔 스마

트 팩토리등의 경우에는 데이터의 분할 처리가 매우 중요한 요소일 것이기 때문이다.

둘째, 서비스를 기반으로 하는 경우이다.

서비스기반의 빅데이터 플랫폼은 B2B, B2C등 다양한 비즈니스 방식을 기반으로 운영이 되는 형태를 의미한다. 이러한 운영은 중계업을 기반으로 하는 비즈니스 모델을 따르는 경우가 대부분이다. 이러한 비즈니스 서비스의 경우는 업종을 중심으로 구성이 되어 있고 업종 중에서 영위하고자 하는 형태의 사용 속성도 포함하는 경우가 대부분이다.

예를 들어 플랫폼 기업으로 대변되고 있는 우버, 에어비엔비등이 그의 예라고 할 수 있다.

이러한 플랫폼 기업의 프로덕트 경제에서 서브스크립션의 경제로 경제의 패러다임이 급격하게 변하는 경우에 더욱 중요한 요소라고 있다.

셋째, IT자원의 비용을 절감하기 위한 경우에 활용할 것인지를 대한 정의가 전제 되어야 할 것이다.

기존 전산 시스템은 소프트웨어의 경우에는 엔터프라이즈 기업들이 출시한 제품을 중심으로 라이선스의 개념이 적용되어 기술을 이끌고 있는 반면, 빅데이터와 그에 속한 에코 소프트웨어의 경우는 대부분 오픈소스를 기반으로 하고 있다.

그렇기 때문에 빅데이터 중심의 생태계는 오픈 소스 중심으로 가지고 있는 경험에 의존하고 있는 현실이다. 따라서 기존의 데이터를 사용하는 빈도가 높을 경우의 핫데이터(Hot Data)는 고가의 시스템 속에 그리고 사용빈도가 떨어지는 데이터(Cold) 데이터의 경우에는 오픈소스를 중심으로 한 하둡 영역에 저장하여 운영하게 되면 자원 및 비용을 획기적으로 절감되는 것을 의미한다. 이러한 것을 일반적인 데이터 및 정보생명주기 (ILM; Information Life Cycle Management)같은 형태로 운영하기도 하는데 비용과 자원의 효율화라는 두 가지 잇점이 있다. 또한, 한국에서는 하이브리드 빅데이터 시스템이라 하여 기존의 데이터웨어하우스와 빅데이터 시스템을 혼합한 형태로 운영하기도 한다.

이러한 세가지 관점에서 효율적인 빅데이터 시스템을 구축하게 된다면 효율적인 시스템의 구현이 가능하다.

따라서 본 논문의 목적은 일반적인 빅데이터 시스템이 아닌 사물인터넷 즉, 기계데이터 중심의 데이터 처리에 특화된 빅데이터 시스템의 설계에 기반한 것으로 이는 향후 기계 중심의 데이터 즉, 스마트 팩토리, 헬스케어, 장치산업, 자율주행 자동차등 다양한 산업군에 적용할 수 있는 효과적이며 특화된 빅데이터 시스템의 설계가 가능한 프레임 워크를 기술 하는데 있다.

3. 사물인터넷 기반의 빅데이터 프레임워크

사물인터넷 기반의 데이터를 효율적으로 처리하기 위해서는 기존의 오픈소스 기반의 하둡 및 에코 소프

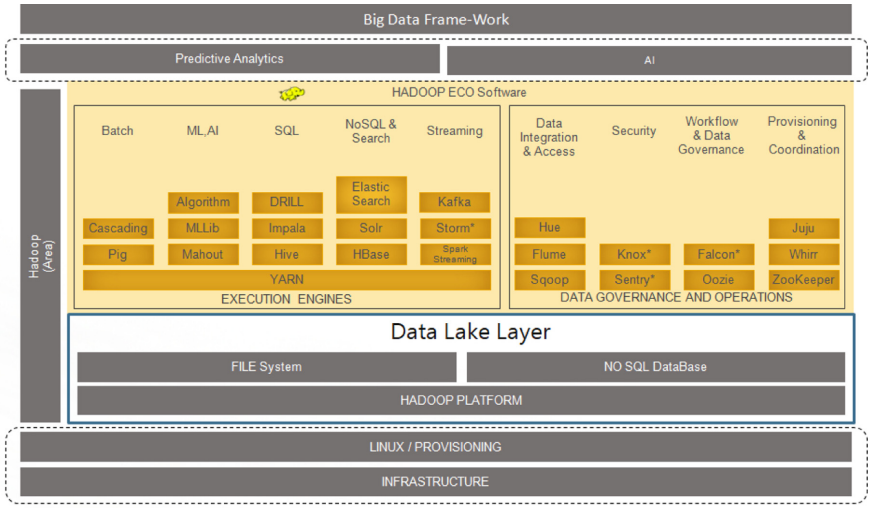

그림 3. 하둡 프레임워크

트웨어로서는 비효율적으로 구성된다. 특히, 오픈 소스 기반의 하둡은 다음과 같은 3가지 특성을 가지고 있다.

첫째, 네임 노드와 데이터 노드로 구성이 되어 있다.

둘째, 데이터를 64M- 128M로 데이터를 나누어 처리한다.

셋째, 3 Replication으로 데이터를 처리하여 장애 및 데이터를 보호한다.

하지만 사물인터넷의 데이터의 특징은 데이터의 속성상 시계열 중심의 Numeric 중심의 비정형 데이터가 주류를 이루고 있다. 이러한 데이터는 사이즈는 매우 작은 형태를 취하고 있지만 데이터의 양은 방대하기도 하다. 이러한 사물인터넷 데이터의 속성상 다음과 같은 형태를 요구한다.

데이터를 64~128Mb의 데이터가 아닌 8Kbyte 와 같은 작은 형태로 데이터를 분할하여 처리하게 된다면 자원을 획기적으로 절약할 수 있게 된다. 뿐만 아니라 네임 노드와 데이터 노드가 구분이 없는 통합된 형태의 구조를 취하는 특징 또한 함께 가지고 있어야 한다. 이는 배포판의 경우에는 상관이 없지만 엔터프리즈 비용의 경우에는 라이선스를 절감하게 되는 특징 또한 가지고 있다. 다음은 사물인터넷에 최적화된 프레임워크를 나타낸 것으로서 다음과 같은 5개의 Layer로 구분을 한다.

첫째, 최고 아래층에는 X.86으로 구성된 물리적 하드웨어 부분이 있다. 이러한 하드웨어 부분은 오픈소스를 사상을 하는 하둡의 경우에는 많은 CPU, Memory, Cache, Disk등 많은 자원에 효율적인 구성에 따라서 성능 및 장애를 최적화 할 수 있게 된다.

둘째, 두번층에는 OS의 영역으로 Linux가 있다. 리눅스의 경에는 상용화되거나 프리 소프트웨어 등 즉, Redhat, Suse, CentOS등 다양한 형태의 사용이 가능하다.

셋째, 데이터 레이크의 레이어로서 HDFS, NOSQL 영역이 포함된다. HDFS는 데이터의 저장을 8Kbyte 이하로 저장이 되는 오픈소스 기반의 하둡 배포판으로 구성을 하여 DISK 사이즈를 절감하게 되며 데이터의 형태에 따라 적합한 NOSQL(Not Only SQL)을 선택하여 사용하게 된다. NOSQL의 경우에는 Colume Family, Key-Value, Document, Graph중 적합한 것을 선택하여 사용한다.

넷째, Eco 소프트웨어 영역으로서 다음과 같은 수집, 저장, 처리, 분석의 공정별로 다양한 소프트웨어가 존재한다. 이러한 소프트웨어는 하둡의 1.x, 2.x, 3.x에 따라서 동일한 소프트웨어라고 하더라도 버전의 디펜던시가 존재한다. 현재, 빅데이터 랜드스케이프에 체계적으로 정리되어 있는 소프트웨어의 이미 수 백 여가지가 존재하고 버전별로 구분한다면 수천가지가 존재하기 때문에 효율적인 에코 소프트웨어를 선정하기는 매우 어려운 일이다.

4. 빅데이터 시스템 용량 산정

본 연구는 순수 하둡을 기반으로 적용된 결과이다.

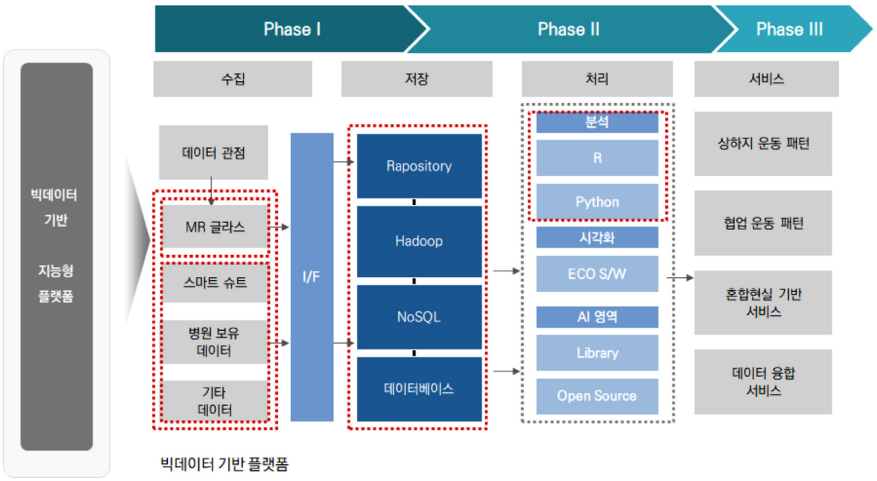

그림 4. 각종 데이터 처리 공정 단계

산정 요소	상황	사이즈 및 수량	산정 방식	산정 결과
기존 비정형 데이터	100TB	100		100
복제 개수	3개	3	A=(기존 비정형 데이터*복제 개수)	300
연 데이터 증가율	5년에 5%	0.25	B=(A*(1+연 데이터 증가율(1+년*%))	375
Swap space(Mapreduce data)	30%	0.3	C=(B+(B*Swap space))	487.5
예비율(Non HDFS)	30%	0.3	D=(C*(1+예비율))	633.75
압축율	압축 없음	1	E=(D/압축율)	633.75
네임노드 서버당 디스크 개수	서버	8	F=((E/네임노드갯수)/스토리지 베이 개수)	9.90234375
	스토리지 베이	8		

그림 5. 하둡 2.x대의 용량 산정 근거

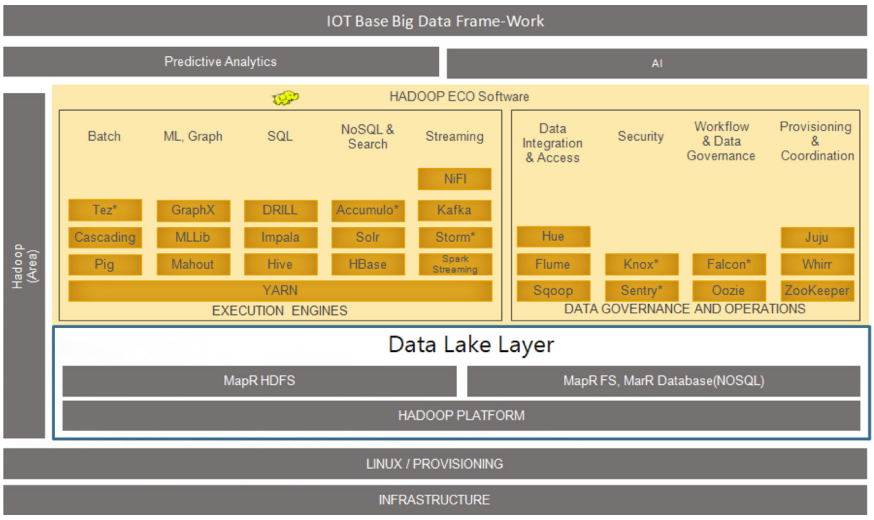

그림 6. 하둡 시스템 아키텍쳐 구성도

하둡 2.x대 버전을 기준으로 빅데이터 시스템의 용량을 산정한 가이드는 다음과 같다. 용량 산정의 기준은 일반적인 오픈소스 하둡을 기준으로 작성한 것이며 에코 소프트웨어는 배제하고 순수 하둡 기준으로 작성하였다.

본 산정의 근거는 하둡 2.x기준으로 다음과 같다. 원천 비정형 데이터를 100Tb로 구성하였다.

(1) 복제 갯수 : 3 Replication
(2) 연평균 데이터 증가율 : 1%
(3) Swap Space : 30%
(4) 예비율 : 30%
(5) 압축율 없음으로 시스템 적용

총 소요 디스크 용량 결과 : 633.75TB

사물인터넷에 특화된 하둡 시스템의 아키텍쳐는 다음과 같이 구성된다.

본 구성에 적용된 하둡 및 에코시스템의 구성을 다음과 같은 특징이 있다.

(1) Core : C
(2) No Name Node
(3) Data 분할 처리 : 8Kb

적용 조건은 동일 Fig3의 조건과 동일

(1) 복제 갯수 : 3 Replication
(2) 연평균 데이터 증가율 : 1%

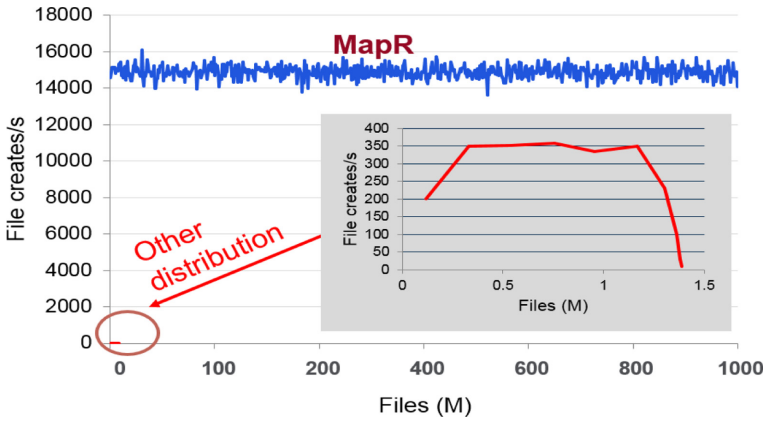

그림 7. 아파치하둡과 제안시스템 성능비교

(3) Swap Space : 30%
(4) 예비율 : 30%
(5) 압축율 없음으로 시스템 적용

조건 결과의 경우

최소 분할 값인 64Mbyte는 65.536Kb와 동일하기 때문에 Numeric으로 출력되는 사물인터넷의 센싱 장비의 특성상 원천 시계열 데이터는 Kb 단위의 작은 데이터 무수히 출력이 된다. 따라서 하둡 내의 데이터를 64Mb가 아닌 8Kb로 데이터를 나누어 저장하여 처리한다면 빅데이터 리소스의 자원과 함께 성능의 향상을 이룰 수 있다.

산술적 표현하면

Apache 하둡 구성 : 64Mb * 3 = 192Mb

MapR 하둡 구성 : 8Kb * 3 = 24Kb 로 구성

그 외에 하둡 고유 기능으로 제공하는 압축률을 적용하면 더 많은 자원을 절약할 수 있게 된다. 또한, 데이터의 특성을 고려하여 NoSQL을 적용하면 더 많은 리소스를 절약하는 것이 가능하다.

다음은 그림은 파일 생성율에 대한 결과를 나타낸 것으로서 일반 오픈소스 기반의 아파치 하둡 과 MapR 하둡 및 에코 소프트웨어로 구성한 시스템의 성능을 비교한 결과치 이다.

5. 결 론

본 연구는 사물인터넷과 IoT기반 빅데이터 수집, 처리 및 분석의 성능 향상을 추구하기 위한 플렛폼을 디자인하고 이를 실현하기 위한 세부 조건들을 설계 및 제안하였다.

제안 설계 플랫폼을 실현하기 위해서는 기존의 오픈소스기반 시스템이 가지고 있는 물리적 시스템의 의존성 및 세부적인 특성을 고려하여야 하는 경험적인 접근성이 요구됨으로 다양한 환경에서 요구분석을 통한 최선의 플랫폼 설계를 위한 분석 및 자료의 연구가 지속적으로 진행되어야 할 것이다.

향후 특정 시스템에서 제안된 시스템을 설계 및 구현하고 성능평가를 통하여 효율성과 최선의 방법론을 제시할 예정이다.

References

[1] 문영상, “빅데이터 기반의 스마트 팩토리 구현”, CAD&Graphics, 2018년 11월.
[2] 손진승, 최규현, “IoT기반 Big Data 기술동향”, 韓國電磁波學會誌, 2013년 7월.
[3] Telecommunications Technology Association. (2017). Information and communication terminology dictionary. Telecommunications Technology Association. http://terms.tta.or.kr