S·Core

빅데이터에 대한 의미 및 전통적인 데이터 처리 구조에 대한 이해

손병창 이사
에스코어㈜ 소프트웨어사업부 플랫폼사업팀

들어가며

현재 우리는 정보의 홍수처럼 폭발적으로 데이터가 증가하고 있는 빅데이터 시대에 살고 있다. 2017년 4차산업혁명시대의 도래와 함께 이전에는 기술적 한계로 활용이 어려웠던 대용량 데이터가 현재의 발전된 정보 처리 및 분석 기술을 만나면서 새로운 비즈니스 및 가치를 창출할 수 있는 시대가 본격화 된 것이다.

본 리포트에서는 빅데이터에 대해 깊이 있는 이해를 시작으로 빅데이터 처리를 위한 전통적인 아키텍처와 특징에 대해 소개하고자 한다.

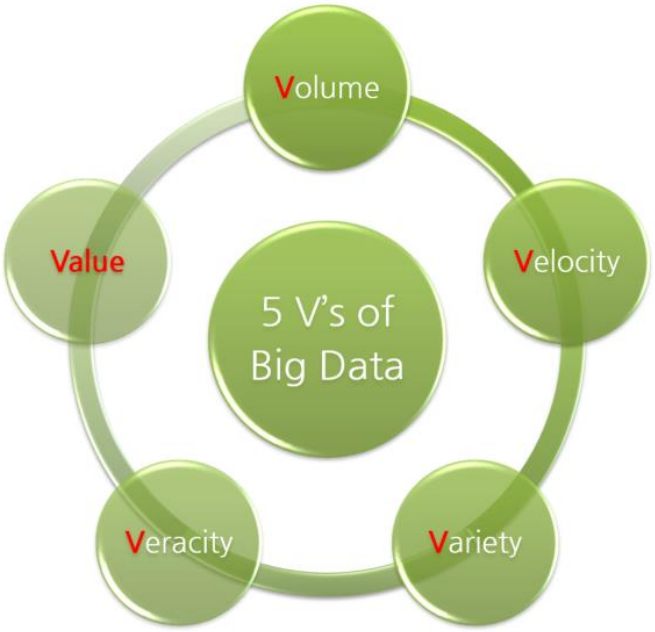
빅데이터란?

빅데이터는 전형적인 RDBMS 소프트웨어가 저장, 조회, 분석 가능한 범위를 초과하는 다양한 성질의 방대한 자료이다. 특히, 기하급수로 폭증하는 대규모 데이터로부터 비즈니스 예측과 같이 고급 분석용으로 활용되기도 한다. 이러한 빅데이터는 수집, 가공, 분석 등의 과정을 거쳐 사람의 현재 행동패턴을 즉각 판단하고 미래 행동을 유도하기도 한다. 인터넷 쇼핑몰에서의 고객 맞춤형 상품 추천 서비스가 그 예다.

빅데이터 특징

빅데이터를 '특성'과 '유형'으로 설명하면 다음과 같다.

아래 [그림1]은 Doug Laney가 정의한 5가지의 빅데이터 핵심 특성을 표현한 것이다.



[그림 1] 5V by Douglas B. Laney

각각의 'V'특성은 다음의 의미를 갖는다.

- Volume: 생성/저장된 데이터 양. 빅데이터는 일반적으로 TeraByte, PetaByte 보다 큼. 크기는 가치와 잠재적 통찰력을 갖춘 데이터로 간주될지 여부를 결정
- Velocity: 데이터 생성/처리/기록/게시의 빈도/속도(지속적 생산, 실시간 제공 척도)
- Variety: 데이터의 다양한 유형 (정형/반정형/비정형)의 이질성이 강한 데이터 멀티미디어 파일, 소셜플랫폼의 게시물, 트윗 등 내/외부 데이터 소스
- Veracity: 데이터의 지저분함 정도, 신뢰성/진실성에 대한 품질보증을 의미
- Value: 데이터 세트의 처리 및 분석을 통해 얻은 유익한 정보 가치. 비즈니스 통찰력 및 의사 결정을 위한 정보의 수익성 지표. 5가지 요소 중 가장 중요

데이터의 성격을 두 가지 관점으로 분류하면 아래 [그림 2]와 같이 구분된다.

[그림 2] 빅데이터 성격

빅데이터가 수집되는 규모(Volume)의 관점으로 본다면 보통 불규칙하게 수집되는 정성적인 데이터가 상대적으로 더 많은 것이 일반적이다.

구분	정량적	정성적
유형	정형/반정형 데이터	비정형 데이터
특징	숫자/규칙이 명확한 문자	불규칙 문자
분석	분석 용이	분석 난해

데이터를 담는 포맷 규격은 다양한데 3가지 유형으로 구분하면 다음과 같다.

• **정형(Structured)**: 열/행 관계식 테이블 구조로 잘 정의된 표준화 형식(e.g: 관계형 DB 테이블,

Excel, CSV)

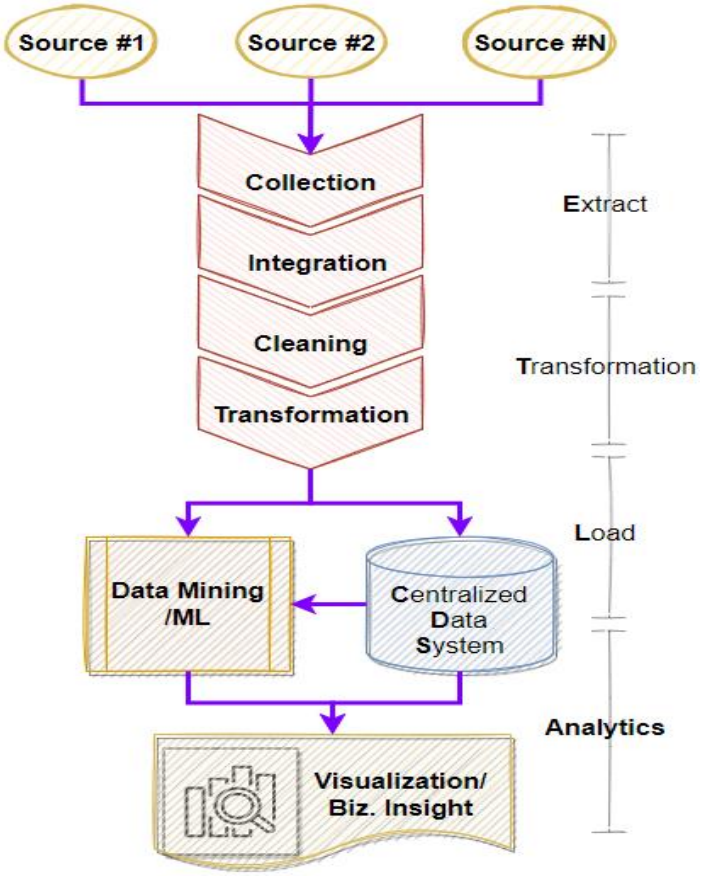
- **반정형(Semi-Structured):** 정렬 불가, 정확한 기능적 구조는 없지만 고유한 구조적 특성이 있음. 그룹화 가능, 보통 프로그래밍 언어에 적합 (e.g: XML, HTML, JSON, YAML)
- **비정형(UnStructured):** 사전에 정의된 구조나 형태가 없는 정성적 데이터로서 비즈니스 통찰력 얻기 위해서는 많은 계산이 필요 (e.g: Video, Audio, Image, Log/Text)

현실 세계에 존재하는 모든 데이터 포맷(파일)은 위 3가지 유형 중 하나에 포함된다. 다음은 원하는 정보를 얻기 위해 필요한 일반적인 빅데이터 처리 단계를 설명하겠다.

빅데이터 처리 단계

빅데이터 프로세싱이란 무엇일까? 여러 매체를 통해 언급되지만, 간단히 한 줄로 요약한다면 **"정보화 시대에 다양한 성질로 폭증하는 방대한 데이터를 관리 및 분석하여 비즈니스의 가치 창출을 위한 처리 기술"** 이다.

기업에서 운영중인 서비스에 대한 종합적인 통찰력을 확보하기 위해서는 먼저, 다양한 소스에서 데이터를 수집하고 품질 확보를 위해 정제 단계를 거치게 된다. 또한 미래에 대한 결과 예측은 통계 분석 및 기계학습 단계까지 필요로 한다. 데이터의 수집부터 활용까지의 일반적인 과정을 도식화하면 아래 [그림 3]과 같다.



[그림 3] 빅데이터 처리 4단계

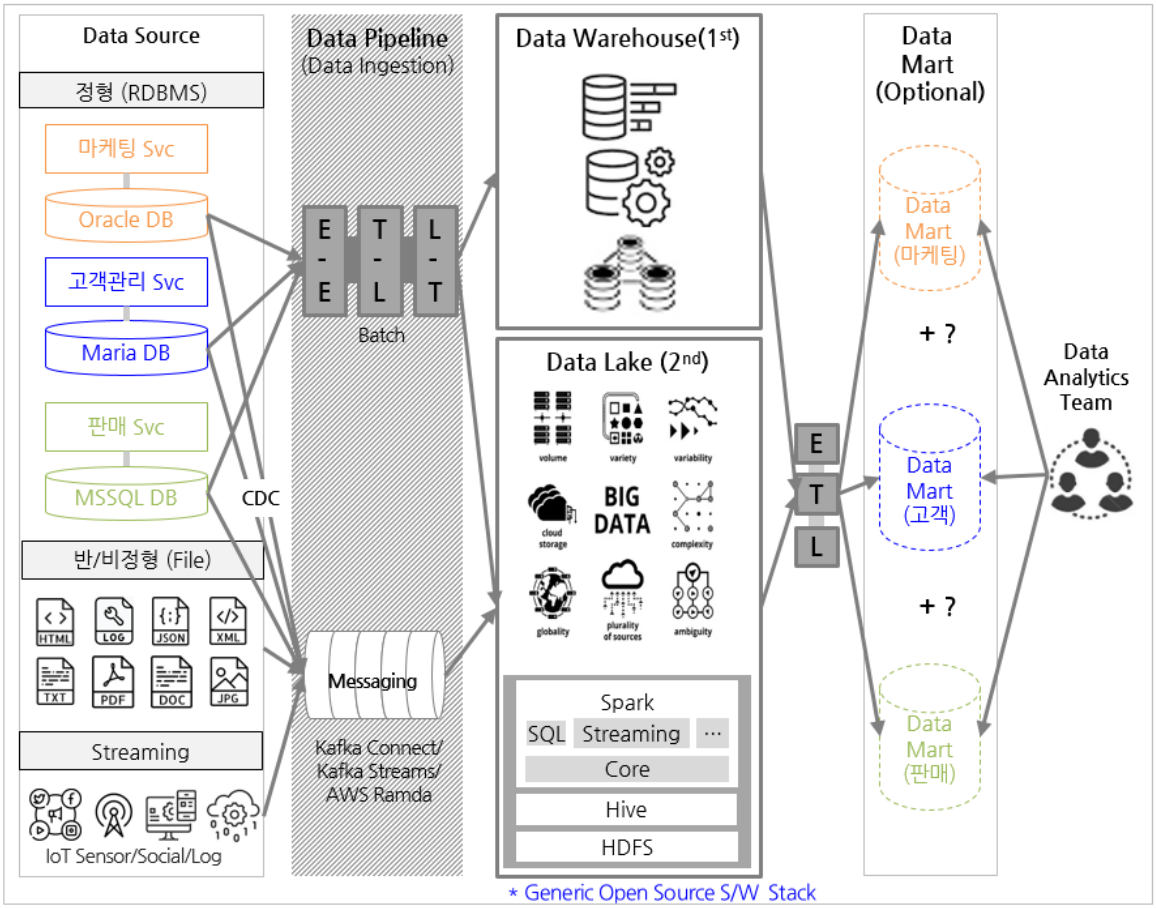
각 단계별 의미는 다음과 같다.

단계	설명
추출	• 이기종 원시 데이터로부터 수집하는 과정 (소셜 사이트, 운영 DB 등) ※ Streaming/Log, API, ETL/CDC 툴로 RDBMS의 데이터를 수집 • 이 단계에서 데이터의 융합/통합 수행, 종종 수집된 데이터에 레이블이 지정됨
변환	• 비정형은 정형화 형태로, 정형 데이터는 더욱 사용자 친화적인 데이터로 변환 ※ 필터링, 정렬, 데이터 클러스터링(그룹화), 중복된 데이터 제거, 정규화 등
로딩	• 정제/변환된 데이터를 중앙 대규모 시스템으로 로딩하거나 기계학습(ML: Machine Learning) 혹은, 데이터 마이닝 가속을 위한 분석 시스템으로 전송
분석	• 정제된 데이터 세트의 분석으로 비즈니스 통찰력 및 빠른 의사 결정 확보 ※ ML(지도/준지도/비지도/강화학습)은 패턴식별, 이상감지, 데이터 예측에 활용

이처럼 변화 확대된 빅데이터는 처리 결과(분석)를 얻는 과정까지 대형 데이터 시스템 및 저장소를 필요로 하고 고효율 분석을 위해서는 고성능 컴퓨팅 파워의 시스템이 요구 된다. 일반적인 데이터 프로세싱 아키텍처를 통해 특징과 한계점에 대해 살펴보겠다.

전통적 데이터 처리 구조의 특징

아래 [그림 4]는 데이터 복제 도구(ETL/ELT/CDC)를 활용하여 운영 Plane에 존재하는 다양한 데이터 소스로부터 중앙 데이터 저장소로 데이터를 수집 및 분석하는 흐름을 표현한 것이다. 데이터는 파이프라인을 통해 마이크로 서비스들이 생산한 이기종 관계형 데이터부터 구조화 되지 않은 파일과 스트리밍 소스들까지 다양하게 수집될 수 있다.



[그림 4] 데이터 처리 아키텍처

1세대, 정형 데이터 특화용 데이터 웨어하우스(DW: Data Warehouse)

Data Warehouse는 정보(Data)와 창고(Warehouse)의 의미가 합성되어 만들어진 어휘다(이하 DW). DW는 기업의 의사 결정에 도움을 주기 위해 조직 내 분산된 운영 데이터를 공통의 형식으로 변환해서 통합 관리하는 데이터 시스템이다. 주요 특징과 한계를 요약하면 다음과 같다.

구분	설명
처리	• 관계형 테이블 구조 기반의 방대한 정형 데이터 처리에 유용 • 통상 OLTP(On-Line Transaction Processing)용 데이터를 ETL/CDC로 수집 • 보통 데이터 적재 이후 배치 이관 외에 별도 삽입/삭제/변경은 하지 않음
특징	• 운영 시스템과는 분리되고 기본 자료 구조 상이, 분석용 정보로 재구조화 • 시간적 히스토리 특성으로 기간에 연관되어 저장 (특정 기간의 데이터를 대변) • 정형화된 통계형 배치 보고, BI 시각화 가능
한계	• 반/비정형 성격의 기계학습, 정보 예측 등의 빅데이터 프로파일링은 사실상 부적합 • 빅데이터 저장소로는 인프라 비용이 상승하여 저장소의 횡적 스케일이 한계

2세대, 비정형 데이터 처리가 가능한 데이터 레이크(DL: Data Lake)

Data Lake는 저장을 목적으로 구조화 여부와 상관없이 다양한 데이터 소스로부터 정제되지 않은(No Filtering & No Packaging) 기본 형식의 데이터를 저장하는 대규모 Repository이다. 줄여서 “DL”라고 부르기도 한다(이하 DL). 이는 미국에 Pentaho라 불리는 BI 전문 기업의 CTO(최고 기술 책임자), James Dixon에 의해 최초 소개되었다. Dixon은 DL을 데이터 호수라는 이름에 맞게 여러 경로의 수로를 통해 수집된 하나의 대규모 데이터 저장소로 정의하였다.

구분	설 명
처리	• 다양한 대규모 비즈니스 데이터의 Read/Write Workload에 최적화 • OLAP(On-Line Analytical Processing) 데이터 처리 가능→고효율 BI 제공
특징	• 데이터 사용 준비 전까지 원시 상태로 보관, 분석 필요시 변환(ELT) 가능 • DW 대비 기계학습 기반으로 End-User에 대한 미래예측 가능
한계	• 비즈니스 기반 데이터 통찰력/민첩한 의사 결정에는 여전히 한계 - 중앙의 '단일' 분석팀의 비즈니스 도메인 부족→업무 완화/효율화 필요 - 여전한 데이터 Silo화 발생→ Cross Domain 기반 데이터 통합 분석 필요

DL은 로그 스트림이나 모바일앱 및 서버 이벤트로그와 같은 실시간성 데이터에 대한 처리가 가능한 것이 주요 특징이다. DL은 OLAP를 통해 실시간 분석처리 기능에 대해 제공 가능하다. 다음은 OLAP가 제공하는 주요 기능이다.

• Pivoting: 데이터를 분석하는 Dimension을 사용자의 요구에 맞게 다양한 '기준'으로 '변환'하는 기능

• Filtering: 전체 데이터 중 사용자의 필요 정보만 **'걸러서'** 보여주는 기능

• Reporting: **대화식** 조작을 통해 원하는 형태의 **'보고서'**로 표현

• Slicing & Dicing: 다차원 **데이터 모델**에서 한 차원씩 잘라가며 데이터 **범위**를 좁혀 사용자가 원하는 방향에 따라 분석에 대한 차원 및 **'관점'**을 바꾸는 기능

• Drilling: 데이터의 **'깊이'**와 분석 차원을 마음대로 바꿔가며 **심도** 있는 **분석**을 제공하는 기능

OLAP는 위와 같이 사용자가 다차원 정보에 직접 접근하여 대화식(마치 컴퓨터와 직접 대화하는 것처럼)으로 정보를 분석하고 의사결정을 할 수 있도록 도움을 준다.

정제되지 않은 데이터를 다룰 경우에 주목해야 할 사항이 있다. 앞서 언급된 빅데이터의 5V 특성 중에 'Veracity'는 데이터의 품질 및 신뢰도 측면에서 매우 중요하다. 정제가 안된 데이터가 과하게 쌓인다면 자칫 데이터 '늪'으로 변질될 수 있기 때문이다. 데이터 늪이란 지속적으로 방대한 데이터를 무작정 수집할 때 관리 미흡으로 인해 원하는 데이터에 대해 엑세스가 불가능한 수준의 저장소를 의미한다. 결국, 이러한 데이터는 거의 아무 쓸모 없는 쓰레기와 같게 된다. 특히, 하둡(Hadoop)을 저장소로 사용할 경우 비용측면에서 무료로 생각할 수 있기 때문에 이러한 현상이 발생할 수가 있다. 변별력 있는 데이터 저장소를 구축하려면 수집된 데이터에 Metadata기반의 레이블을 지정을 하는 것이 방법이 될 수 있다. BI 전문 기업인 피라미드 애날리틱스(Pyramid Analytics)의 CTO 아비 페레즈는 호수가 늪으로 더러워지는 문제점을 방지하기 위해 3가지 원칙을 언급하였다.

1) DL 구축 초기에는 마이닝에 필요한 적은 데이터부터 수집하라

2) 분석 자동화를 위한 머신러닝 전략을 도입하라

3) 사전에 비즈니스 대상 문제를 명확히 파악하고 적합한 전략을 수립하라

시사점

지금까지 빅데이터에 대한 의미와 데이터의 전통적 처리 구조에 대해 살펴봤다.

데이터 저장소를 선정함에 있어서 기업의 데이터 규모와 유형에 따라 DL과 DW 중 특정 저장소를 채택하거나 두 저장소를 별도로 분리하여 DL에서 DW로 데이터를 전송하여 분석하는 구조로 가져갈 것인지도 고민하면 좋다.

위에서 언급한 데이터 늪에 대한 주의사항 외에 우리는 DL의 한계점에 대한 두 가지 문제점에 주목할 필요가 있다.

첫째, 기업의 분석 업무를 중앙의 단일 분석팀이 전담할 경우 업무별 도메인에 대한

세밀한 이해, 신속하고 정확한 분석 결과를 도출하기에는 어려움이 있다.

둘째, 단순히 데이터 관리를 탈 중앙화로 분리할 경우 업무간 연계성을 고려한 통합 분석은 한계가 있기 마련이다.

이러한 문제점을 개선하기 위해 데이터 처리(관리/활용) 기술 부문의 Gartner Report에서는 Data MSA 중심(Organization Centric)의 Data Mesh와 데이터 통합 기술 중심(Technology Centric)의 Data Fabric에 대해 최신 기술의 트렌드로 제시하고 있다. 이와 같은 신 기술 패러다임은 전통적 데이터 처리 구조의 한계점 극복에 도움을 줄 것으로 예상한다.

손병창 이사
에스코어㈜ 소프트웨어사업부 Knox기술그룹
sonbc121.son@samsung.com

CDC 기반의 프리미엄 데이터 동기화 및 대용량 분산 메시징 처리 엔진을 담당하고 있습니다.

S-Core
에스코어 주식회사
서울특별시 송파구 올림픽로35길 123 향군타워 13층 Tel: 02-6411-4115 Email: s-core@samsung.com
www.s-core.co.kr