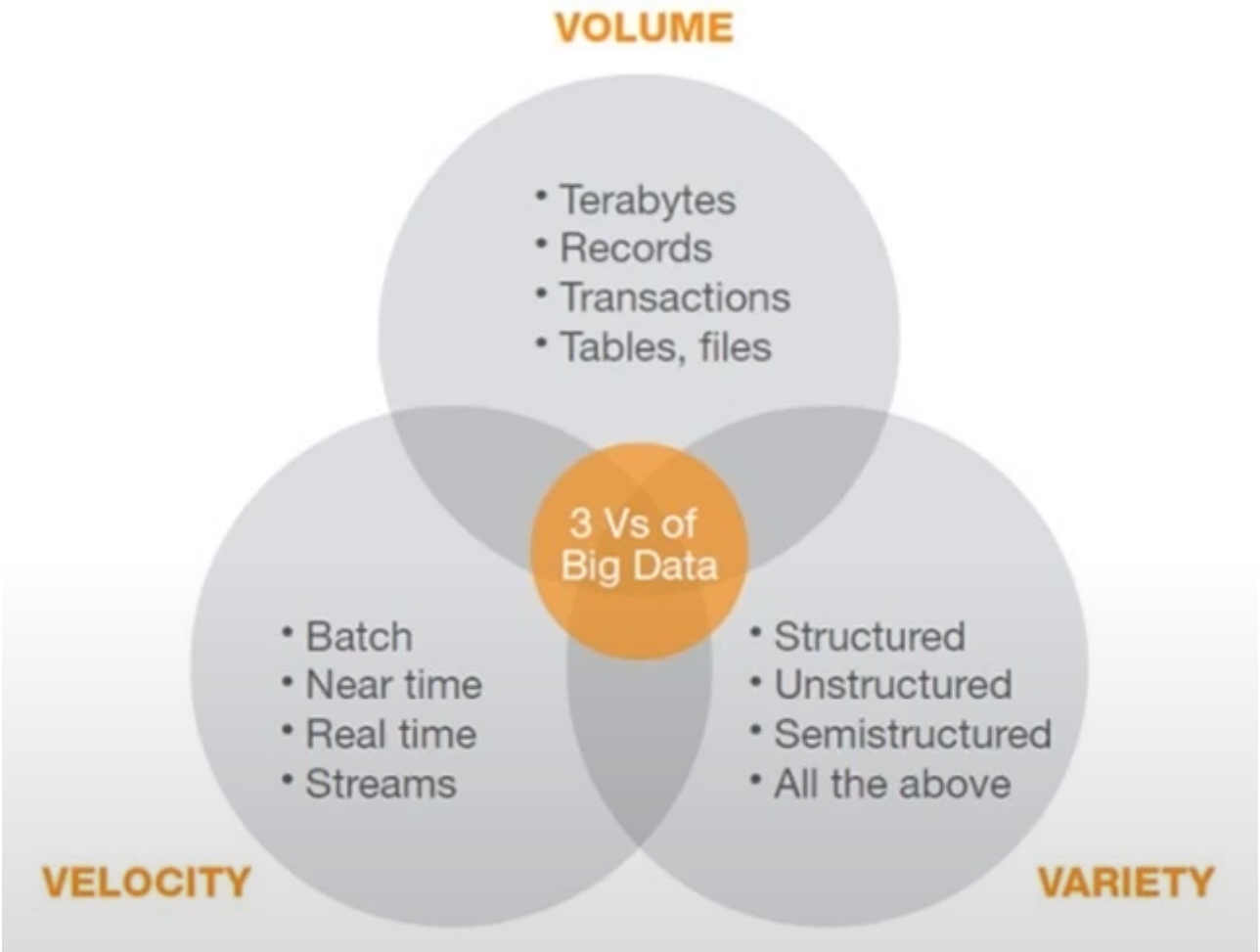
2. 빅데이터 처리

1. 빅데이터 처리 기술 개요

---

1. 빅데이터란 무엇인가
    - 크기가 크고 빠르게 증가하는 데이터 파일
    - 일반적으로 TB or PB
    - 비정형 데이터
    - 관계형 모델에 적합하지 않음
    - 사용자, 어플리케이션, 시스템, 센서 등에서 파생된 데이터



- 데이터 볼륨 (Lake)
- 데이터 속도 (Waterfall)
- 데이터 다양성 (Recreation)

- 빅데이터 전송/저장/특징(3V)/관리/처리 이슈

---

2. 빅데이터 처리 기술 아키텍처

- 하둡을 중심으로 생태계가 꾸려져 있음
- 하둡보다 더 알려지고 있는 제품들도 늘어나고 있음
- but, 근본인 하둡을 이해해야 다른 프로그램들도 편할 것임
- NoSQL (아파치, HBase)
- DataWarehouse (HIVE)
- 수집 및 전송 (Flume)
- 처리 및 워크플로우 (Oozie)
- 아파치 스파크 (메모리 기반, batch 처리라서 다소 느린 하둡을 보완)
- RDBMS (HDFS, 하둡)
    - RDBMS에서 하둡 분산 파일 시스템으로 (혹은 반대)로 파일을 옮기고 싶을 때의
    - 게이트 웨이 역할인 Sqoop(스쿱)이 있음

분야	솔루션
NoSQL	HBase, Cassandra, MongoDB, CouchDB, Couchbase, Cloudata, Riak, Neo4j
Cache	Redis, Memcached
RPC	Thrift, Avro, Protocol Buffer
Collect	Scribe, Flume, Chukwa, Logstash, Fluentd
Query	Hive, Pig, Hcatalog, Impala, Tajo, SparkSQL, BigQuery
Streaming	Akka, Storm, SparkStreaming, Esper, S4
Search	Elastic Search, Solr, Katta
File System	Hadoop, Swift, GlusterFS, Ceph
ETC	Machine Learning(Mahout), Distributed Coordinator(Zookeeper) Queue(Kafka), Data Integration(Sqoop), Statistics(R), Workflow(Oozie)

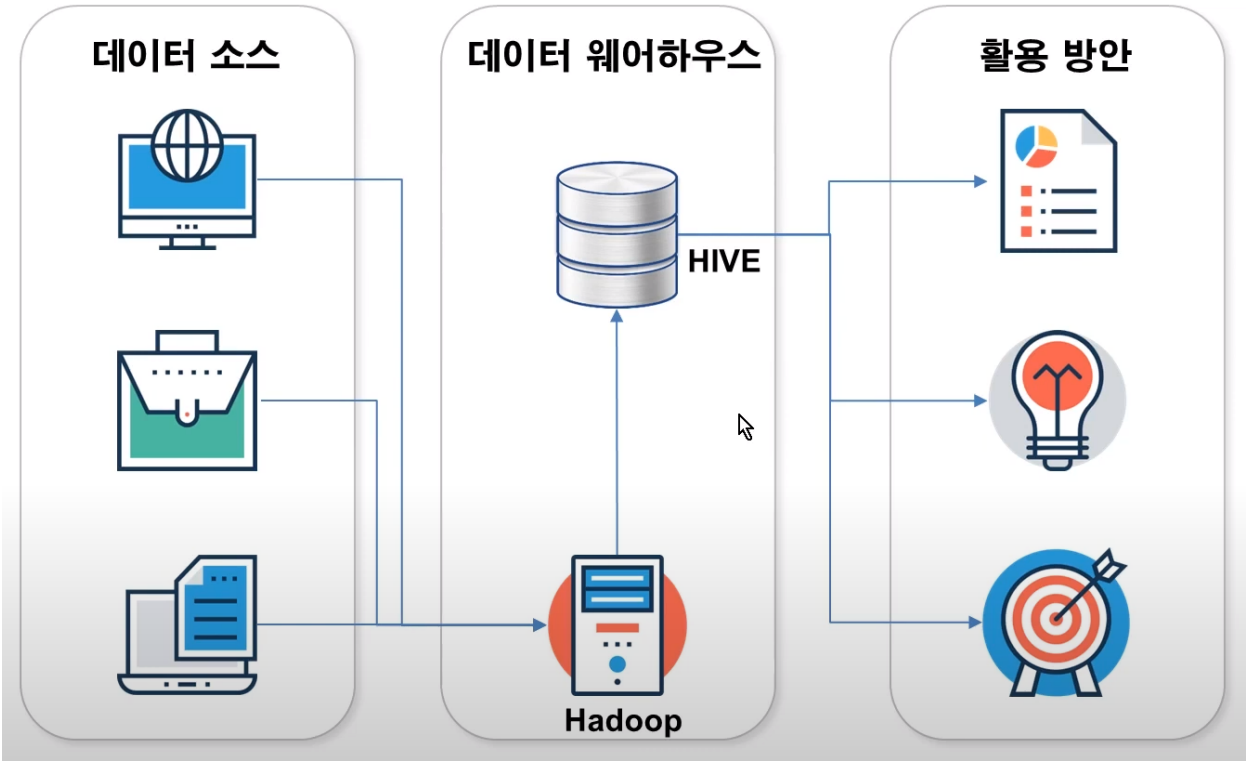
- 요즘에 많이 쓰이는 솔루션 - HBase, Cassandra, Neo4j / Redis, Memcached / ...

- 쿼리(질의)를 통해서 Raw데이터에서 인사이트를 얻을 수 있는 솔루션이 많음
  --〉 스톰, 스파크 스트리밍, 임팔라

- 비동기처리 및 동시성처리(Akka), 제조 및 보안(Elastic Search)

- 배치(batch) 처리 못지않게 실시간 처리가 중요해진 시대

---

3. 

- 과거) 고가의 서버를 이용해서 데이터웨어하우스를 구축

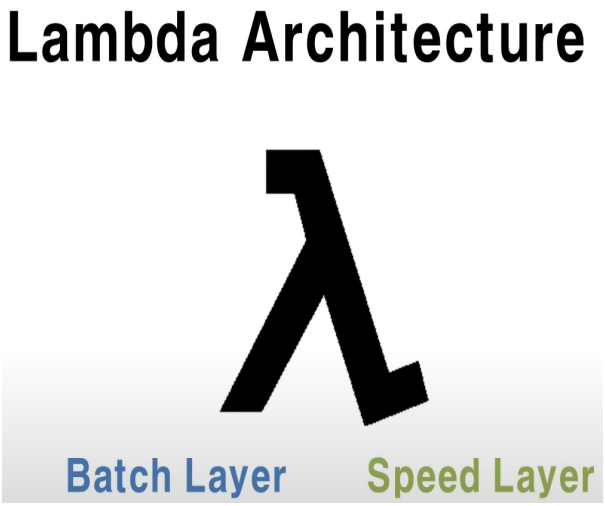
- 현재) 일반 범용 서버를 이용해서 하둡 구축 -〉데이터웨어하우스(DW) -〉SQL이용 대시보드/알람 등 활용

- 하둡 - 배치처리에 적합

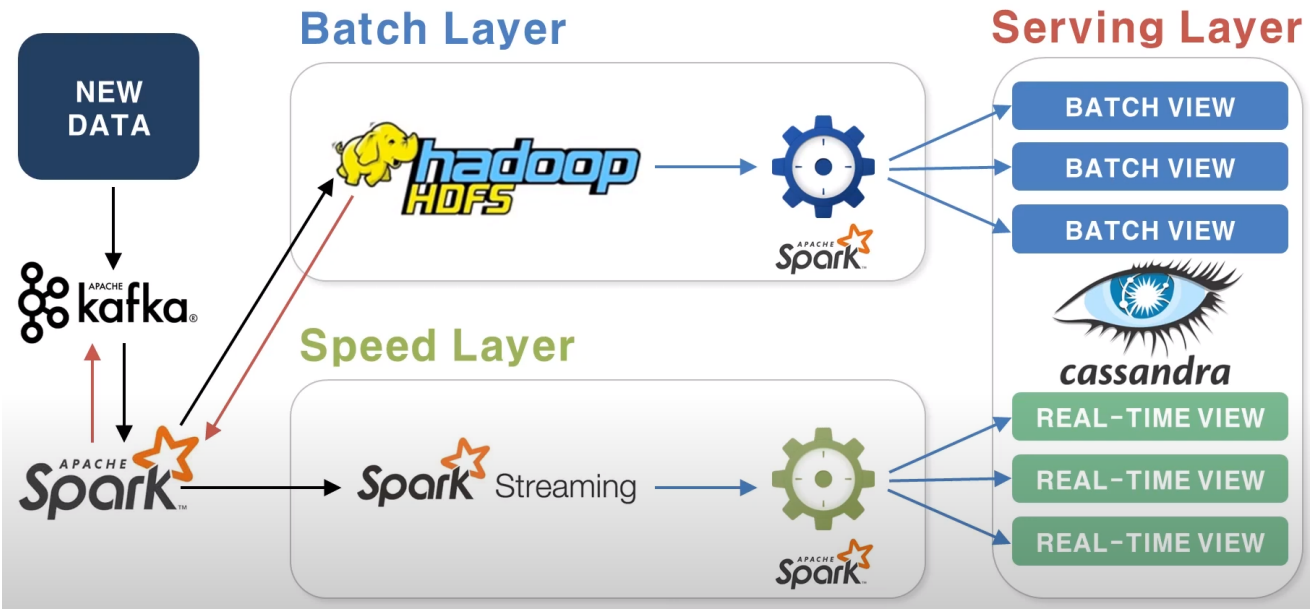
- 하이브 - 온라인에서 배치처리
  =〉 안정적이지만 느림

---

4. Lambda Architecture (람다 아키텍처) =〉 데이터 처리 프레임워크 아키텍처



- 람다 자체의 문자가 2가지로 이뤄짐 (Batch층, Speed층)



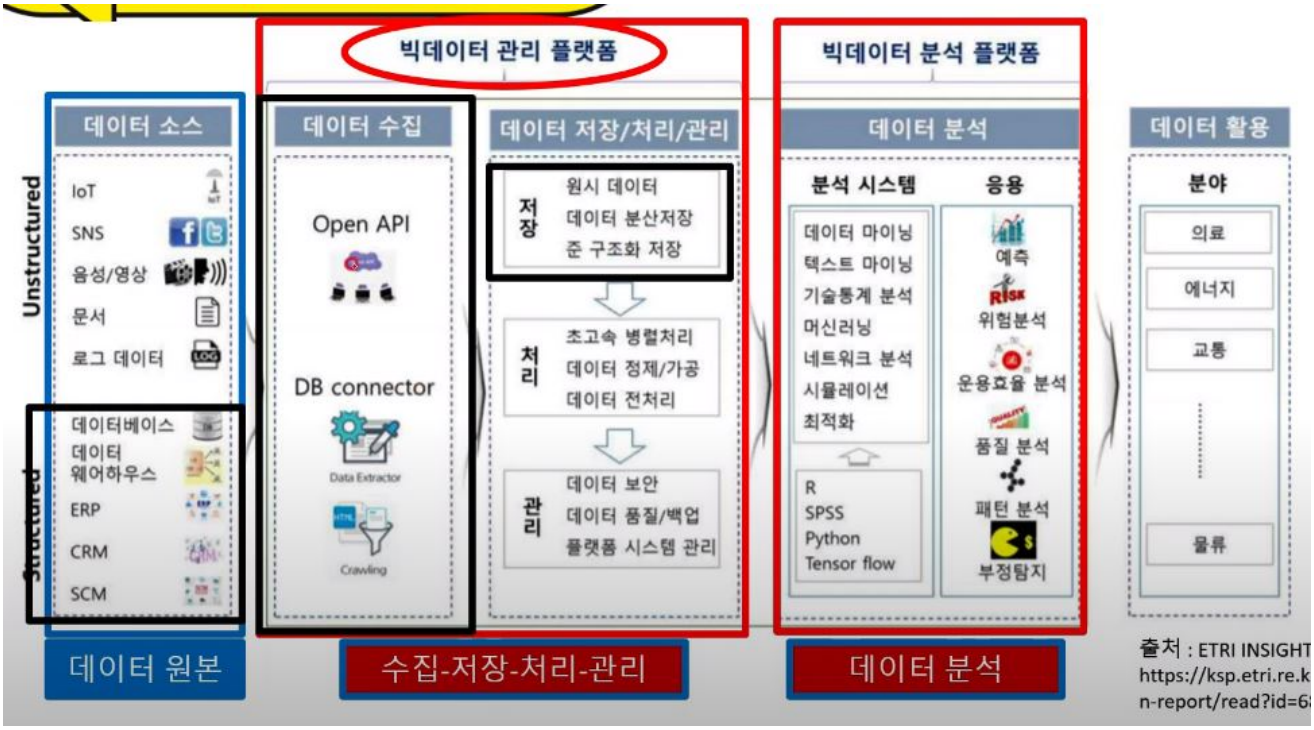
- 람다 아키텍처를 꾸릴때, Batch/Speed/Serving Layer까지 3가지를 구현함
- 새로운 데이터 빠르게 들어옴 -〉 Kafka(messageQ(?)역할)로 받음
  i) 아파치 스파크(배치 처리) -〉 하둡에 HDFS의 배치를 위해 Raw 데이터를 저장할 수도

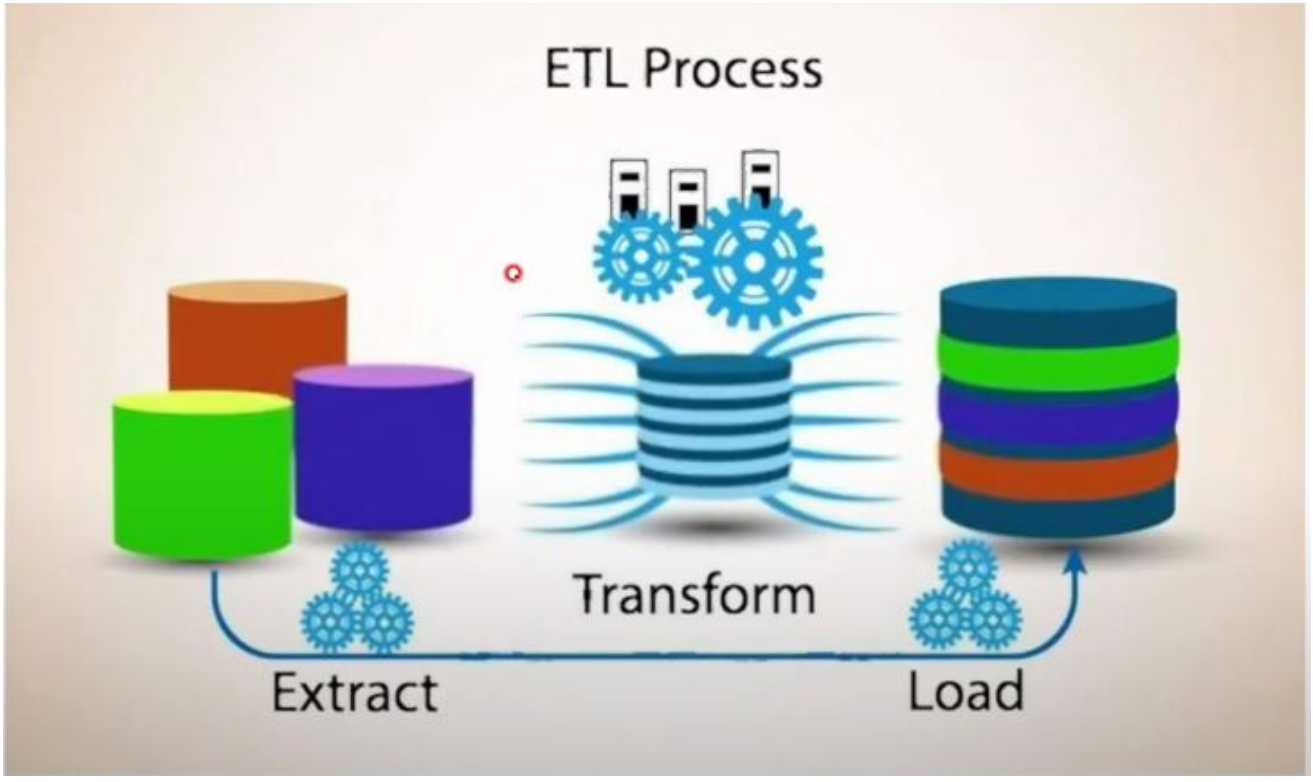
```
HDFS의 처리를 통해 -> 화면에 띄워주기/알람 등 -> 처리를 할 수 있는 솔루션(맵리듀스, 피그, Spark) (물론, 배치처리도 가능) -> Serving Layer로 보낼 수도
```

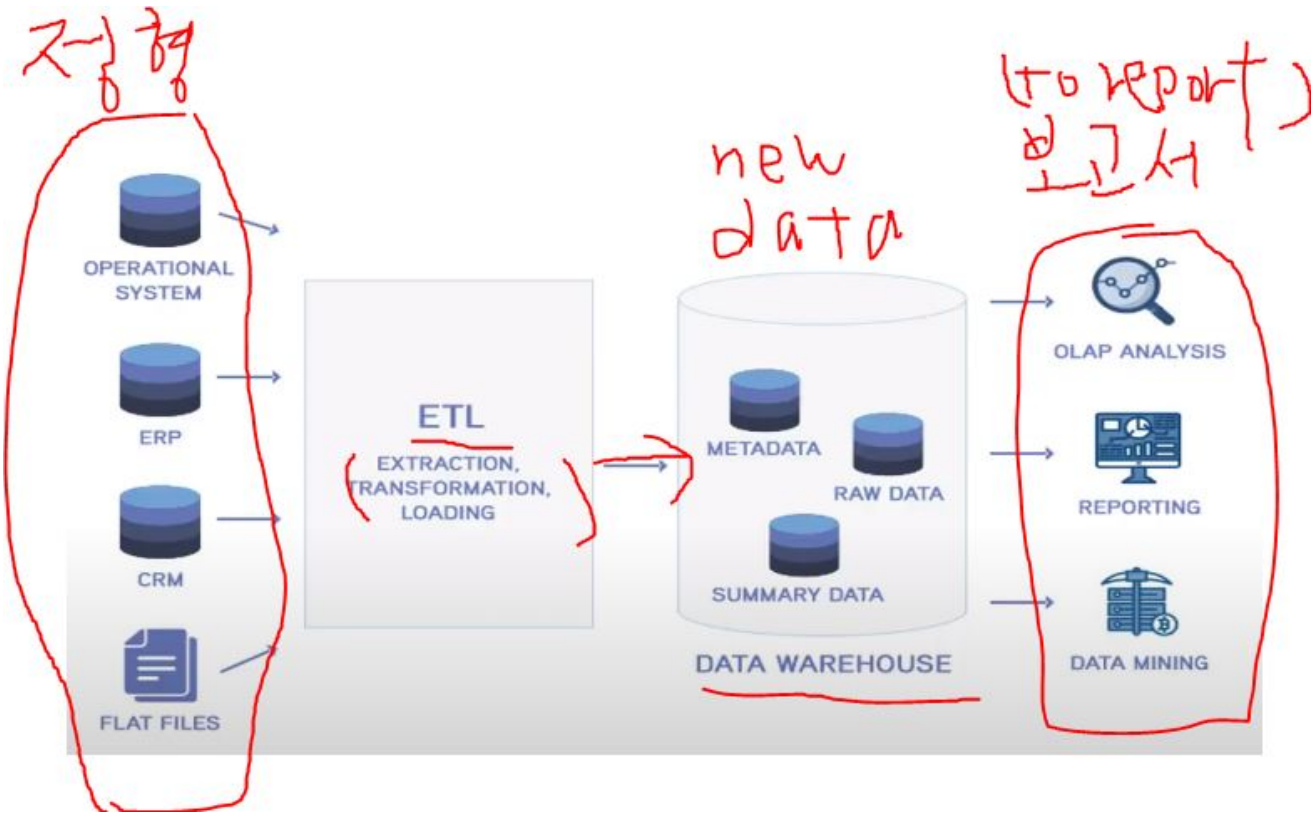
ii) 스파크 스트리밍(들어온 데이터를 실시간으로 처리) -〉 처리 결과를 Serving Layer로 보낼 수도

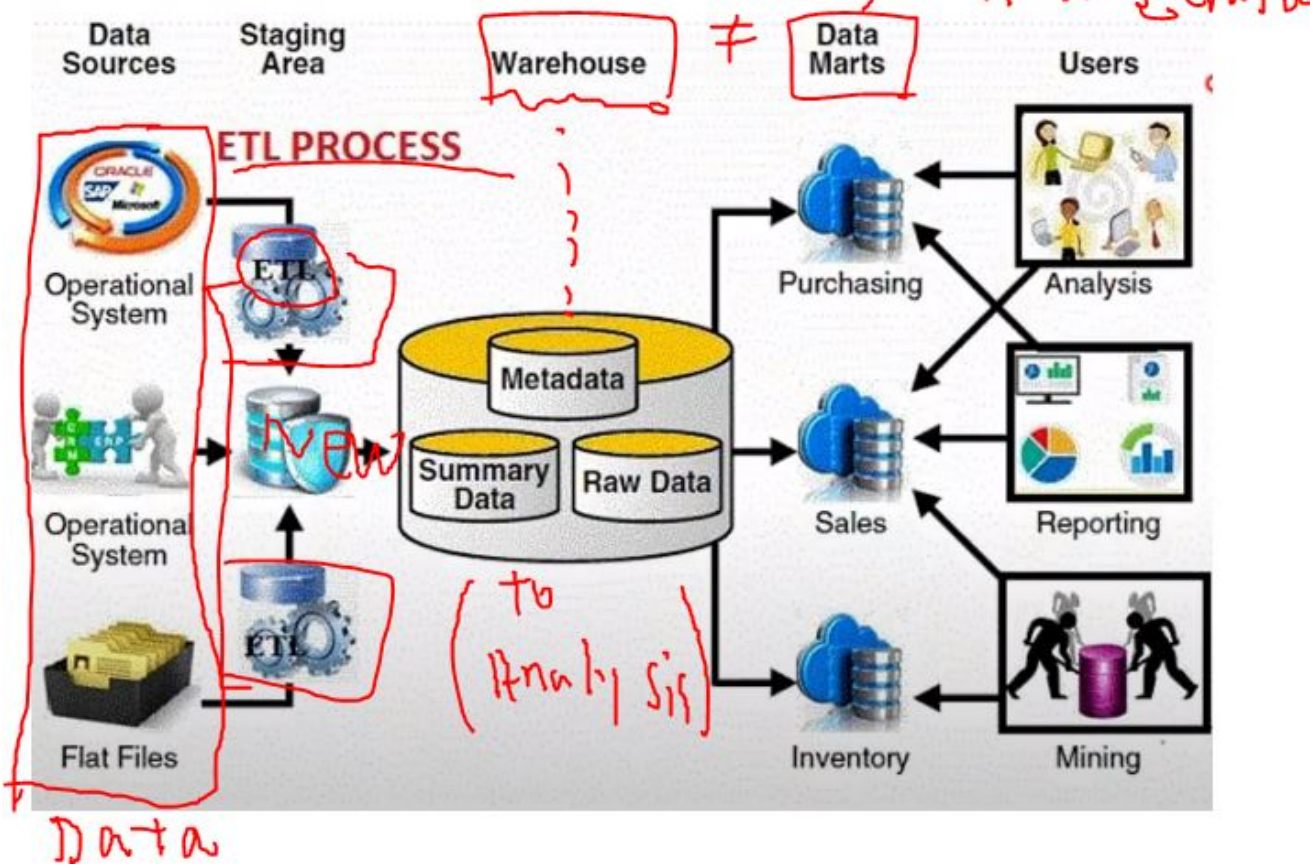
- NoSQL (몽고DB, 카산드라)
  --〉 Batch Layer(배치성으로 보여줄 View) / Speed Layer(실시간으로 처리해서 보여줄 Real-time View)
- Spark -〉 Lambda 꾸리는 경우가 많다고 함

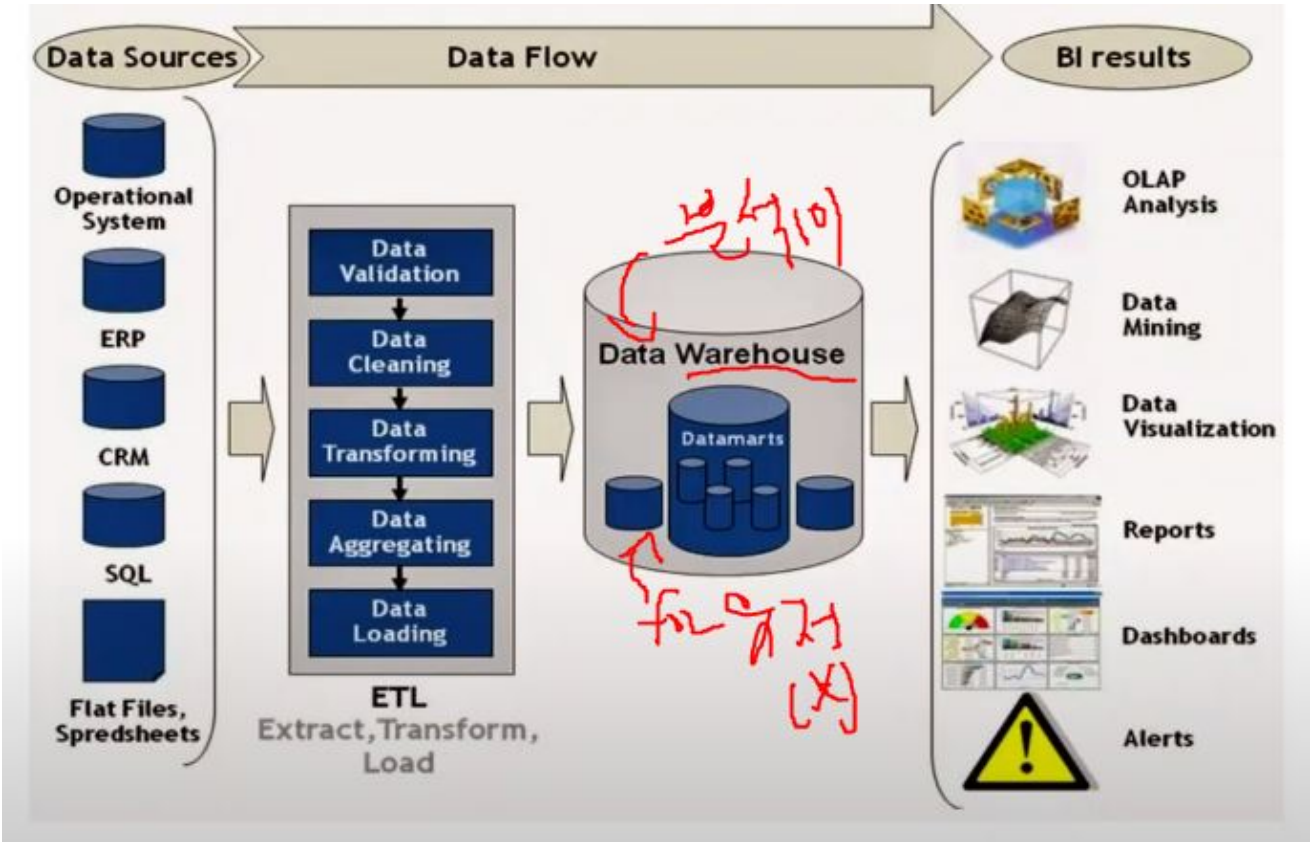
빅데이터 처리 프로세스

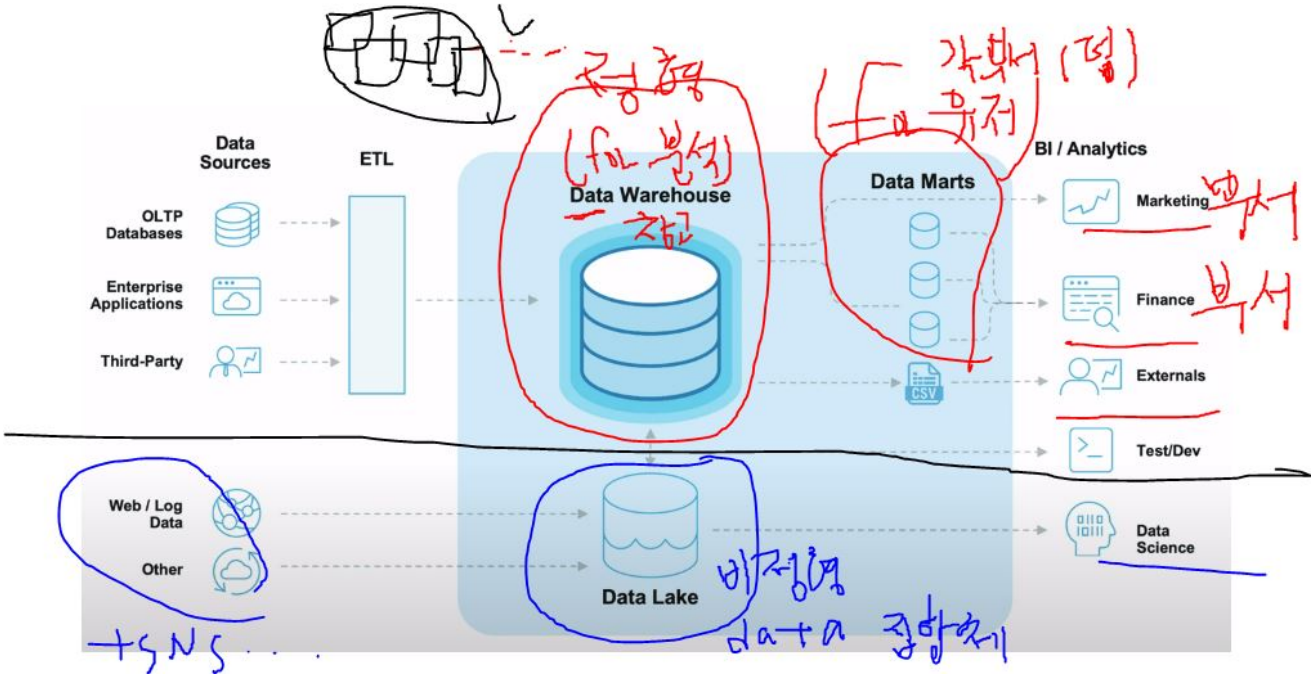

빅데이터 관리 플랫폼
빅데이터 분석 플랫폼
데이터 소스
데이터 수집
데이터 저장/처리/관리
데이터 분석
데이터 활용
Unstructured
Structured
IoT
SNS
음성/영상
문서
로그 데이터
데이터베이스
데이터 웨어하우스
ERP
CRM
SCM
Open API
DB connector
Data Extractor
Crawling
저장
원시 데이터
데이터 분산저장
준 구조화 저장
처리
초고속 병렬처리
데이터 정제/가공
데이터 전처리
관리
데이터 보안
데이터 품질/백업
플랫폼 시스템 관리
분석 시스템
응용
데이터 마이닝
텍스트 마이닝
기술통계 분석
머신러닝
네트워크 분석
시뮬레이션
최적화
R
SPSS
Python
Tensor flow
예측
위험분석
운용효율 분석
품질 분석
패턴 분석
부정탐지
분야
의료
에너지
교통
물류
데이터 원본
수집-저장-처리-관리
데이터 분석
출처 : ETRI INSIGHT
https://ksp.etri.re.k
n-report/read?id=6

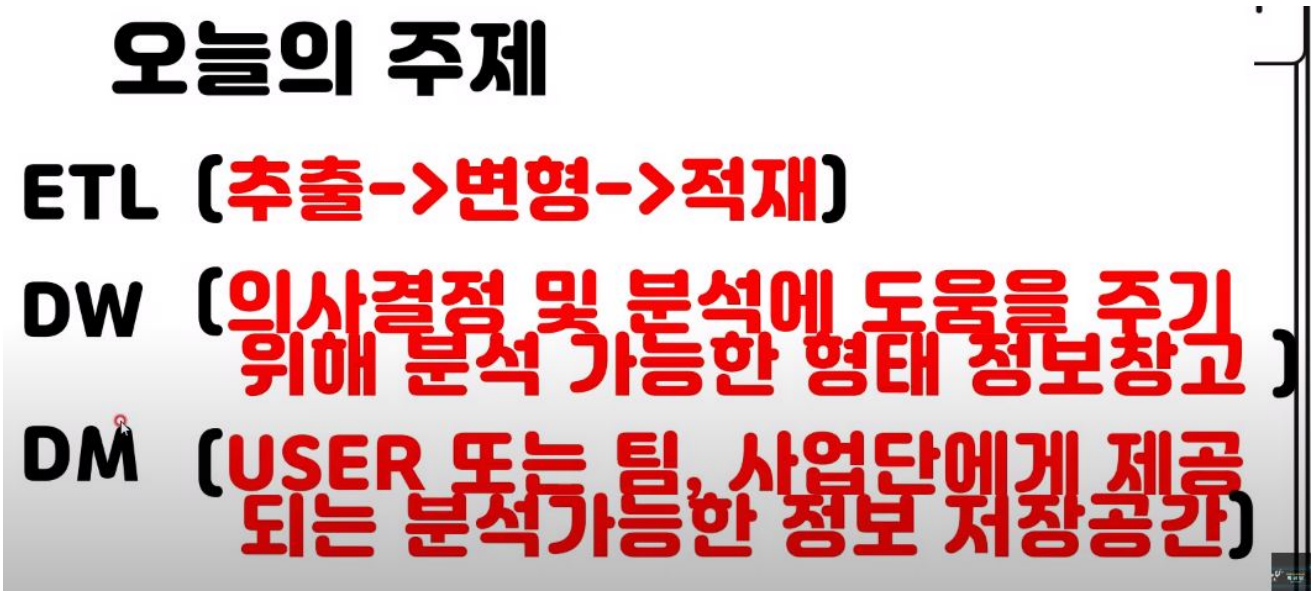

ETL Process
Transform
Extract
Load

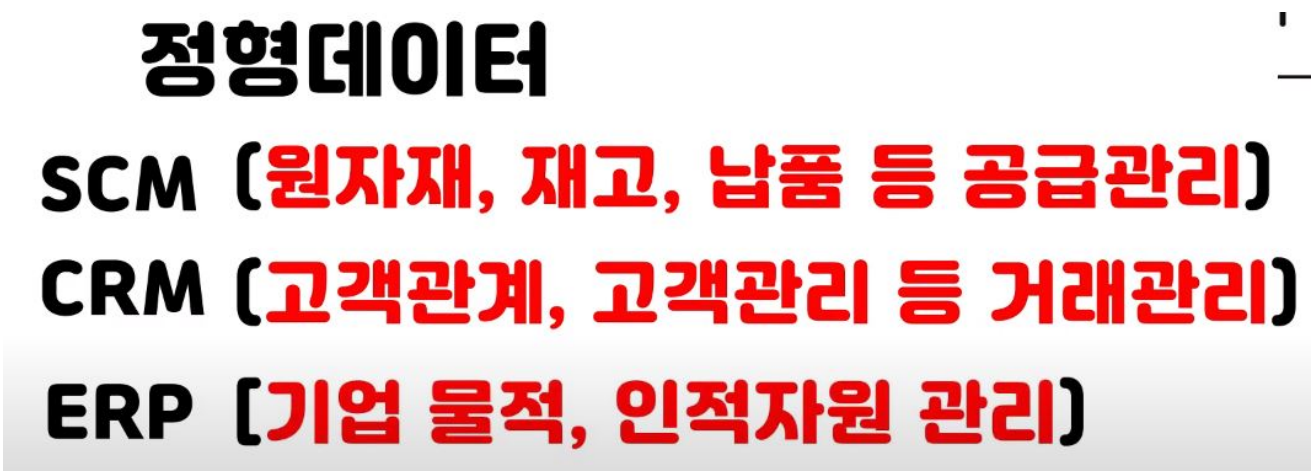

OPERATIONAL SYSTEM
ERP
CRM
FLAT FILES
ETL
EXTRACTION, TRANSFORMATION, LOADING
METADATA
RAW DATA
SUMMARY DATA
DATA WAREHOUSE
OLAP ANALYSIS
REPORTING
DATA MINING


Data Sources
Staging Area
Warehouse
Data Marts
Users
ETL PROCESS
Operational System
Operational System
Flat Files
ETL
ETL
Metadata
Summary Data
Raw Data
Purchasing
Sales
Inventory
Analysis
Reporting
Mining


Data Sources
Data Flow
BI results
Operational System
ERP
CRM
SQL
Flat Files, Spredsheets
Data Validation
Data Cleaning
Data Transforming
Data Aggregating
Data Loading
ETL
Extract, Transform, Load
Data Warehouse
Datamarts
OLAP Analysis
Data Mining
Data Visualization
Reports
Dashboards
Alerts


Data Sources
ETL
OLTP Databases
Enterprise Applications
Third-Party
Data Warehouse
Data Marts
BI / Analytics
Marketing
Finance
Externals
Test/Dev
Web / Log Data
Other
Data Lake
Data Science
CSV

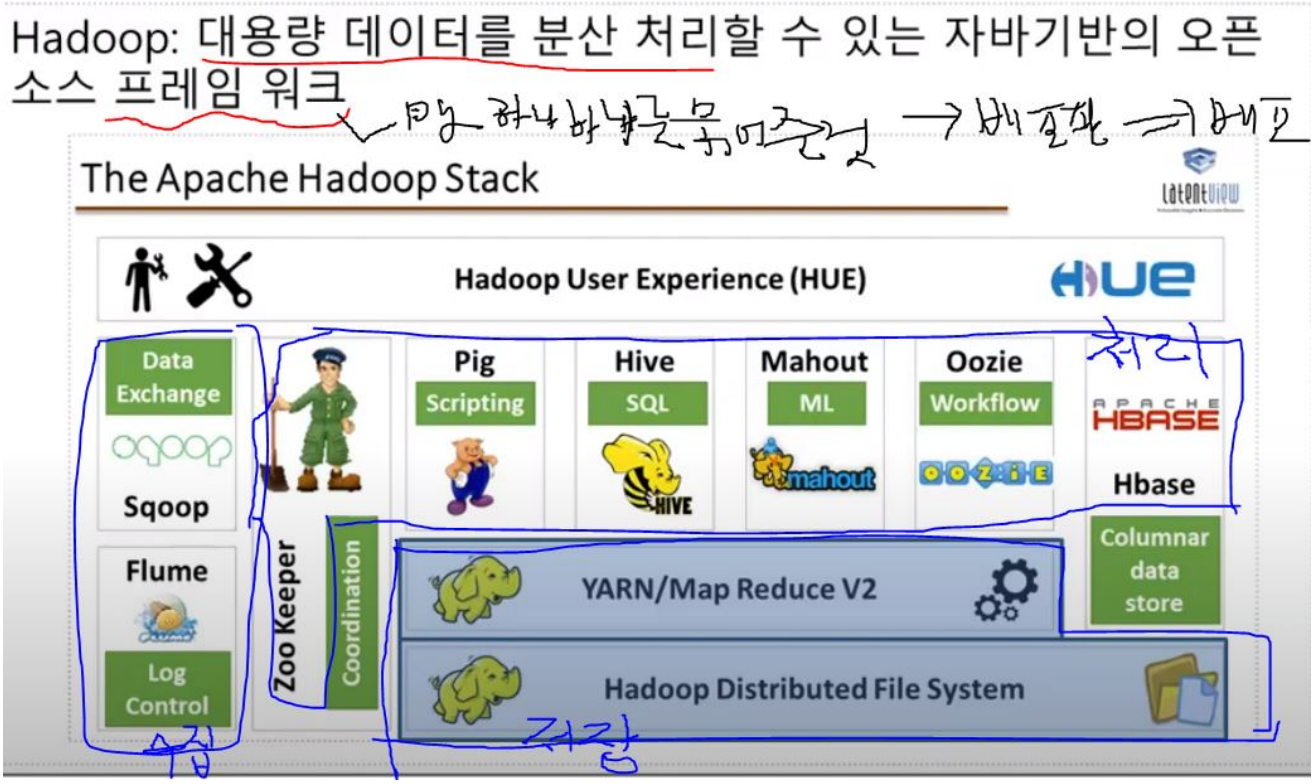




---

1. 빅데이터 처리 프로세스
2. 빅데이터 소스
3. 빅데이터 수집
4. 빅데이터 저장
5. 빅데이터 처리

6. 빅데이터 분석
7. 빅데이터 표현



---

2. 리눅스 기초 및 활용

---

3. 하둡 분산 파일 시스템