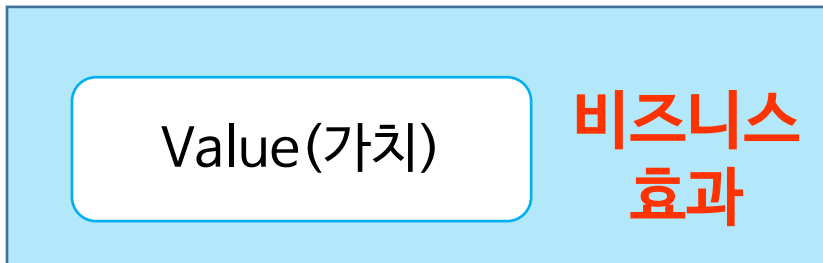


빅데이터 개요



빅데이터의 특징



■ Volume(크기)

- 데이터의 규모 및 양, 테라바이트, 엑사바이트, 기록, 거래, 사물인터넷 등
- 대용량 데이터를 저장, 처리, 관리하는 투자비용 필요

■ Variety(다양성)

- 데이터 종류/유형/형태, 정형, 반정형, 비정형 등
- 데이터 유형 또는 소스에 따라 내부/외부 및 소셜 데이터 등으로 구분

■ Velocity(속도)

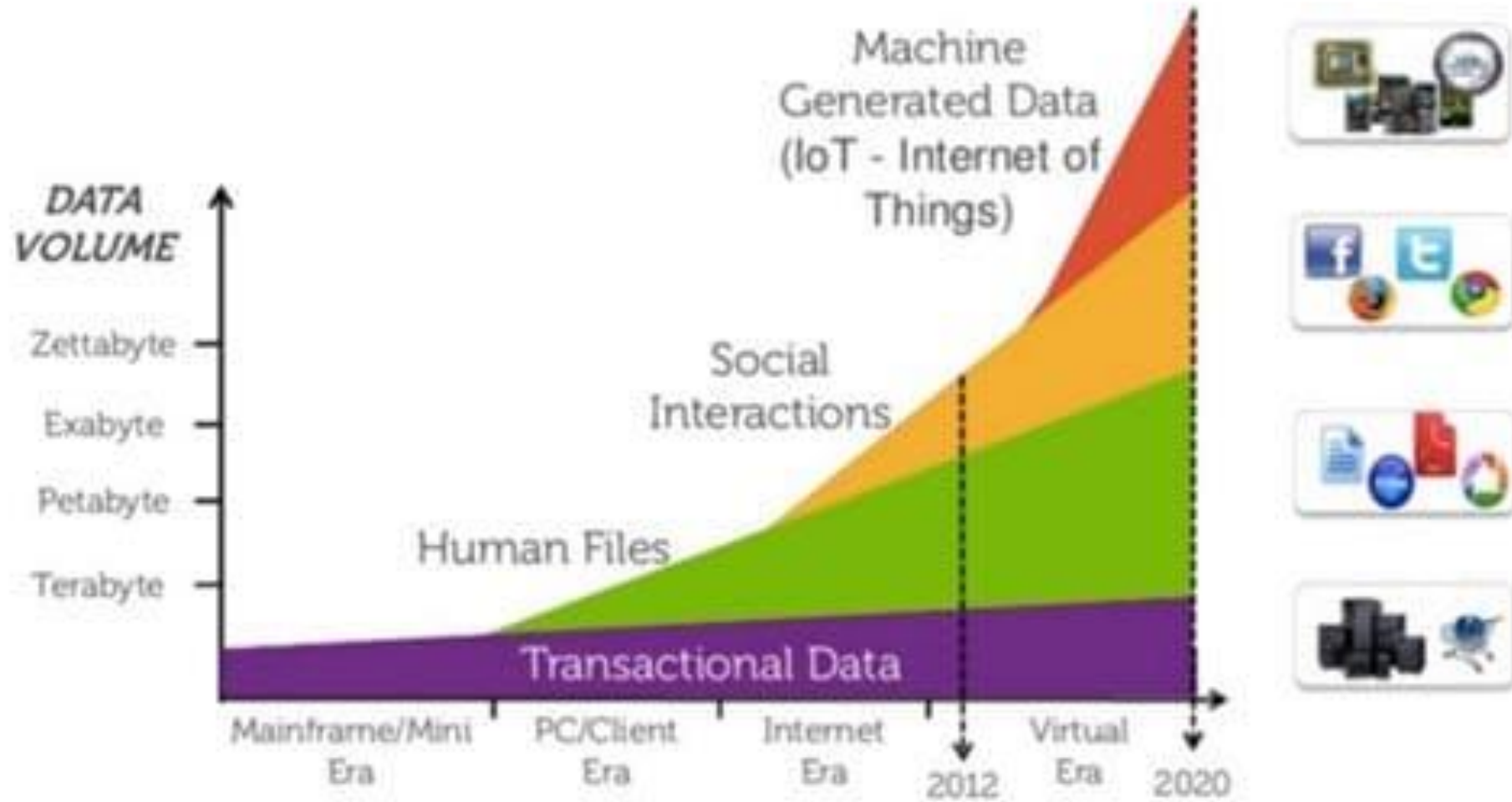
- 데이터 생성속도/처리속도, 실시간 처리, 배치처리, 모니터링, 스트리밍 등
- 시스템에 미치는 부하에 따라 기존 아키텍처/업무에 영향 발생, 투자비용 필요

■ Value(가치)

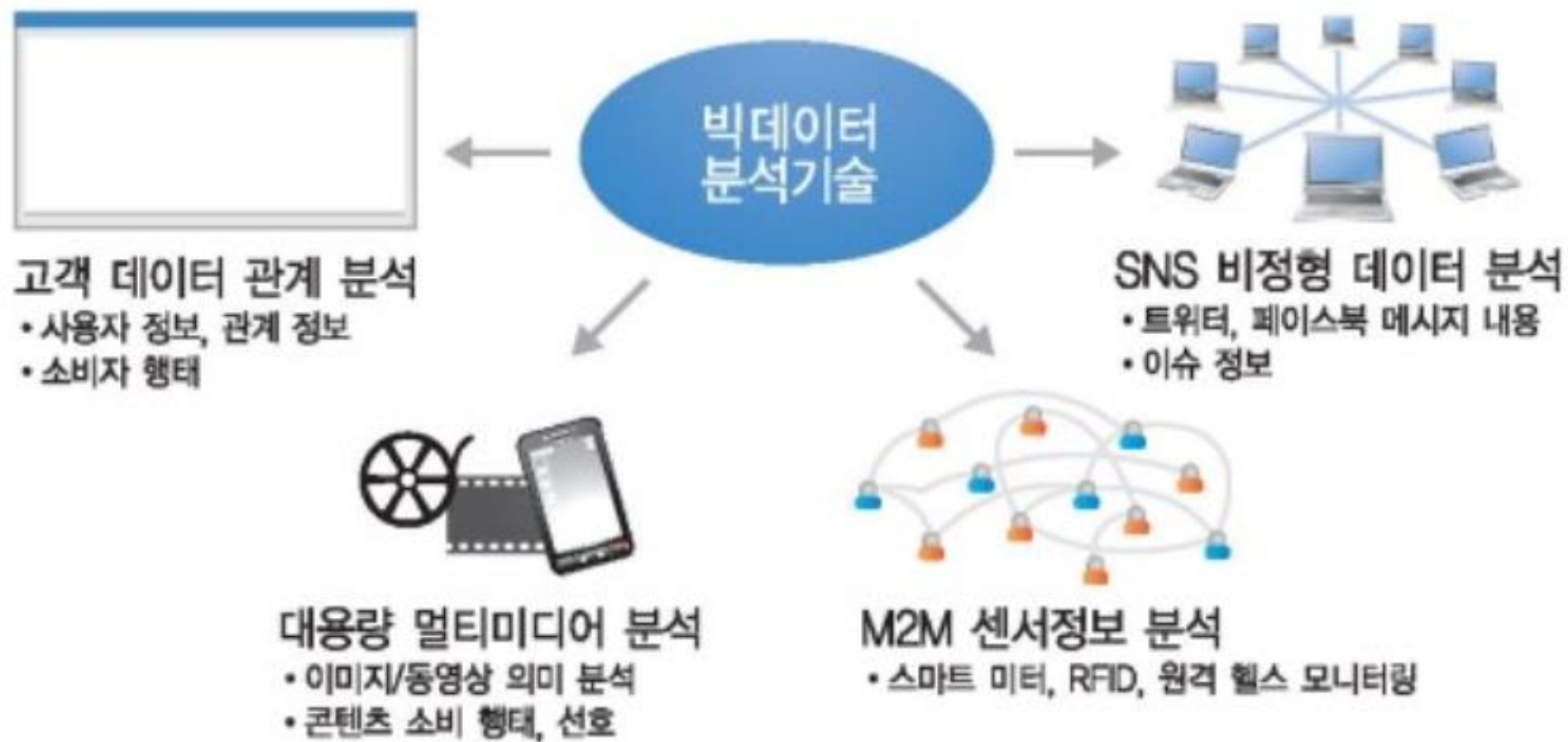
- 분석 결과 활용 및 실행을 통한 비즈니스 가치
- 기업이 데이터 분석을 통하여 추구 및 목표 가치를 정의할 수 있음

빅데이터 소스

The Explosion of Data



빅데이터 분석



빅데이터 분석 방법론

분석기획

분석하려는 비즈니스를 이해하고 도메인의 문제점을 파악하여 빅데이터 분석 프로젝트의 범위를 확정하는 단계

데이터 준비

비즈니스 요구사항을 데이터 차원에서 다시 파악하고 프로젝트별로 필요로 하는 데이터를 정의하여 전사차원의 데이터 스토어를 준비하는 단계

데이터 분석

데이터 준비 단계에서 확보된 데이터를 이용하여 수립된 프로젝트 목표를 달성하기 위하여 데이터 분석 프로세스를 진행함

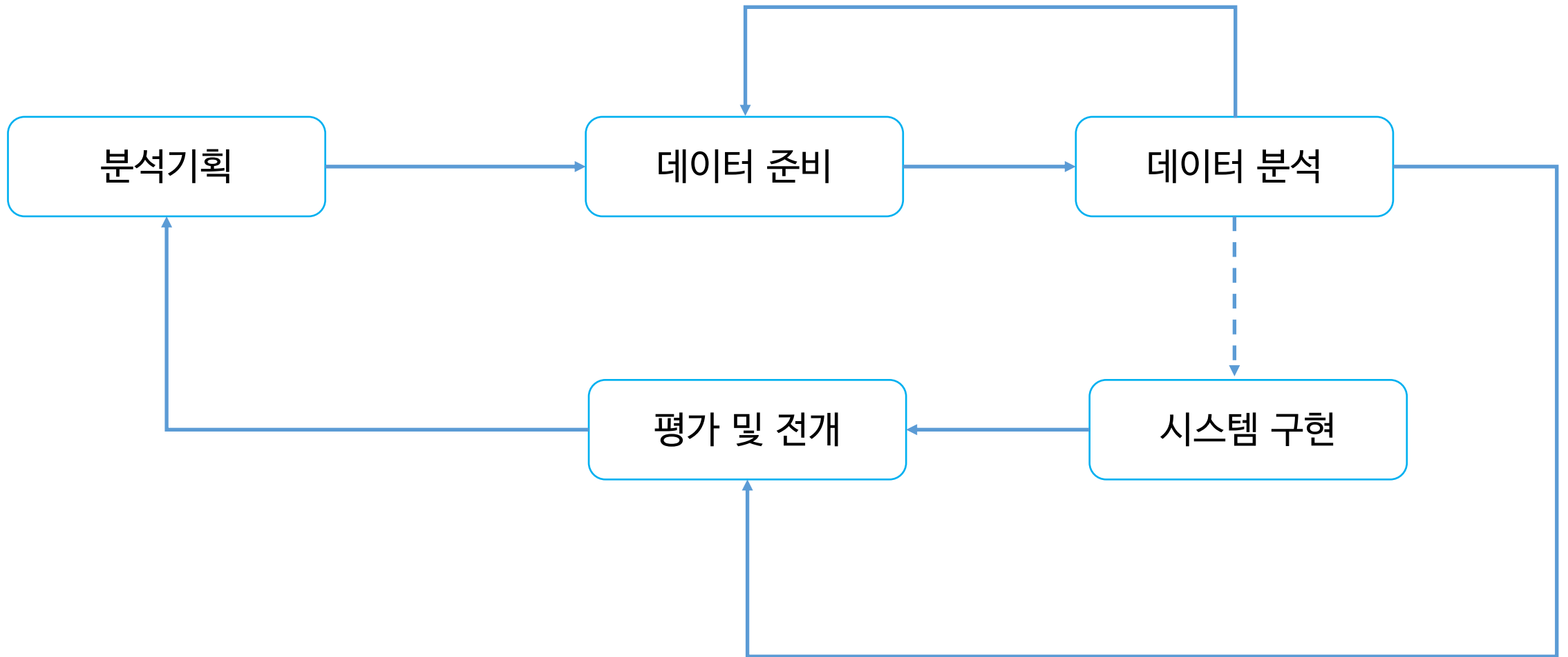
시스템 구현

분석 기획에 맞는 모델을 데이터 분석 단계를 진행하여 도출하고, 이를 운영중인 시스템에 적용하거나 프로토타입을 구현하고자 하는 경우 시스템 구현단계를 진행함

평가 및 전개

데이터 분석 단계와 시스템 구현 단계에서 구축된 모델의 발전계획을 수립 발생 된 모든 중간 산출물을 정하고 프로젝트 종료보고서를 작성/보고

빅데이터 분석 방법론



빅데이터 분석 방법론

비즈니스 문제 분석의 방법(How)

고객 이탈 증대

설비장애로 인한 판매량 감소

기존 판매정보 기반 영업사원이 판단시
재고관리 및 적정가격 판매 어려움

변 환

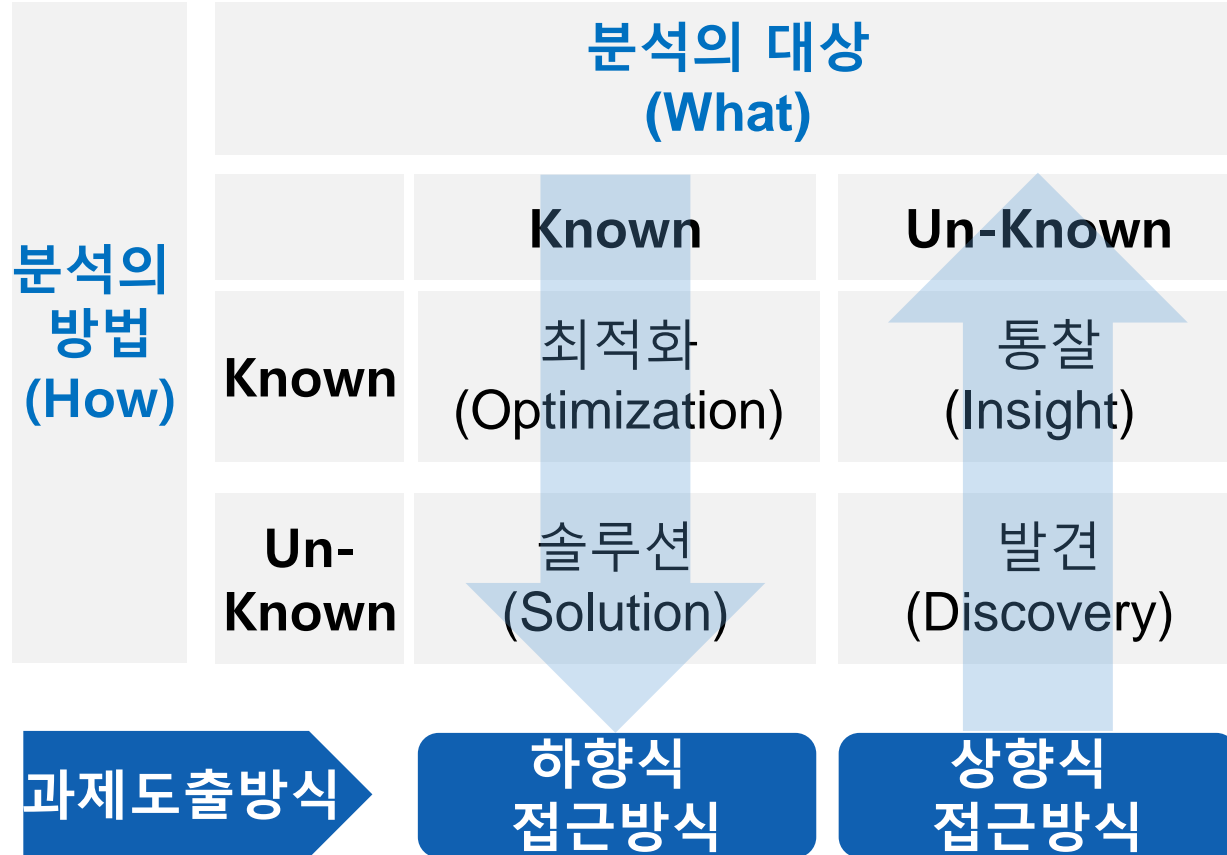
분석 문제 분석의 대상(What)

고객의 이탈에 영향을 미치는
요인을 식별하고, 이탈 가능성을 예측

설비 장애를 초래하는 신호를 감지하여
설비 장애 요인으로 식별하고
장애 발생 시점/가능성을 예측

내부 판매 정보외의 수요예측을
수행할 수 있는 인자의 추출 및
모델링을 통한 수요예측

분석 주제 유형



■ 최적화

분석의 대상과 방법을 알고 있는 경우에 개선을 통해 최적화의 형태로 분석이 수행됨

■ 솔루션

분석의 대상을 알며 방법을 모르는 경우

■ 통찰

분석의 대상이 명확하게 무엇인지 모르며, 분석의 방법을 아는 경우

■ 발견

분석의 대상 및 분석 방법을 알지 못하는 경우이며 이는 분석의 대상 자체를 새롭게 도출함

빅데이터 활용 테크닉

연관 규칙

예) 우유 구매자는 기저귀를 더 많이 구매하는가?
어떤 변인들간에 주목할 만한 상관관계가 있는지를 찾아내는 방법

유형 분석

예) 이 사용자는 어떤 특성을 가진 집단에 속하는가?
문서를 분류하거나 조직을 그룹으로 나눌 때. 특성에 따라 분류할 때 사용할 수 있음

유전 알고리즘

예) 최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?
최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 진화시키는 방법

기계 학습

예) 기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화 중에서 어떤 것으로 가장 보고 싶어할까?
훈련 데이터로부터 학습한 알려진 특성을 활용해 예측하는 일에 초점을 맞춤

회귀 분석

예) 구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?
분석가는 독립변수를 조작하며, 종속변수가 어떻게 변하는지를 보면 두 변인의 관계를 파악함

감정 분석

예) 새로운 환불정책에 대한 고객의 평가는 어떤가?
특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석함

소셜 네트워크 분석

예) 특정인과 다른 사람이 몇 촌 정도의 관계인가?
영향력 있는 사람을 찾아낼 수 있으며, 고객들 간 소셜 관계를 파악할 수 있음

데이터 형태

정형 데이터

- 형태가 있으며 연산 가능함. 주로 관계형 데이터베이스에 저장됨
- 데이터 수집 난이도가 낮음
- 내부 시스템인 경우가 대부분임
- 파일 형태의 스프레드시트라도 내부에 형식을 가지고 있어 처리가 쉬운 편임
- EX) 관계형 데이터베이스, 스프레드시트, CSV 등

반정형 데이터

- 형태(스키마, 메타데이터)가 있으며 연산이 불가능
- 주로 파일에 저장됨
- 데이터 수집 난이도가 중간
- 보통 API 형태로 제공되기 때문에 데이터 처리기술이 요구됨
- EX) XML, HTML, 로그형태(웹로그, 센서데이터), Machine Data 등

비정형 데이터

- 형태가 없으며, 연산이 불가능함
- 주로 NoSQL에 저장됨
- 데이터 수집 난이도가 높음
- 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어려움
- EX) 소셜데이터(트위터, 페이스북), 이메일, 보고서

빅데이터 활용

- 빅데이터가 가져다 주는 기회는 데이터의 크기에 있다가 보다는 음성, 텍스트, 로그, 이미지, 비디오 같은 새롭고 다양한 정보 원천의 활용에 있음
- 비즈니스의 핵심에 대해 보다 객관적이고 종합적인 통찰을 줄 수 있는 데이터를 찾는 것이 그 무엇보다 중요함

구분	과거	현재	미래
정보 (Information)	무슨 일이 일어 났는가? 리포팅	무슨 일이 일어나고있는가? 경고	무슨 일이 일어날 것인가? 추출
통찰력 (Insight)	어떻게, 왜 일어 났는가? 모델링, 실험 설계	차선 행동은 무엇인가? 권고	최악 또는 최선의 상황은 무엇인가? 예측, 최적화, 시뮬레이션

빅데이터 기술

■ 빅데이터 저장기술

- 다양하고 많은 양의 빅데이터를 저장하고 관리하는 기술이 필수
- 대표적인 빅데이터 저장기술 : 하둡(Hadoop), NoSQL(Not Only SQL)
- 하둡(Hadoop) : 대용량 데이터를 분산 처리할 수 있는 자바 기반의 오픈 소스 프레임워크
- NoSQL(Not Only SQL) : 관계형데이터베이스의 일관성 특징보다는 가용성과 확장성에 중점을 둔 데이터베이스

■ 빅데이터 분석기술

- 텍스트 마이닝, 오피니언 마이닝, 소셜 네트워크 분석, 패턴인식, 머신러닝, 딥러닝, 자연어 처리 기술 활용
- 텍스트 마이닝(text mining) : 텍스트 데이터에서 자연어 처리 기술을 기반해 가치 있는 정보를 추출하고 가공
- 오피니언 마이닝(opinion mining) : SNS, 블로그 게시물 등에서 사용자들의 의견을 수집하여 제품과 서비스에 대한 감성을 파악하거나 유용한 정보로 재가공하는 기술
- 소셜 네트워크 분석(social network analysis) : 소셜 네트워크상에서의 영향력인 사람/데이터 등 객체 간의 관계/특성을 분석하고 시각화하는 기법

빅데이터 플랫폼



빅데이터 플랫폼 - Data Foundational Services

Data Foundational Services

Coordination



Apache
Zookeeper

Resource Managers



PELTON

Engines



Messaging



Security



HELIX

Operability

빅데이터 플랫폼 - Data Storage Services

Data Storage Services

Filesystem



Alluxio



File Format



Table Format



빅데이터 플랫폼 - Data Management Services

Data Management Services

Data Query/Visualization



Integration



Orchestration



Metadata



빅데이터 플랫폼 - Data Processing Services

Data Processing Services

Batch



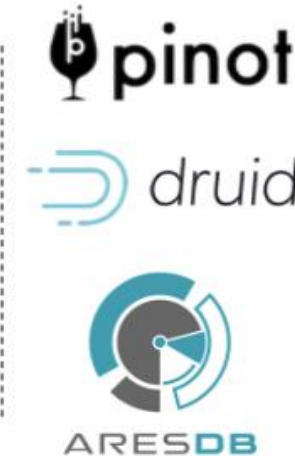
Streaming



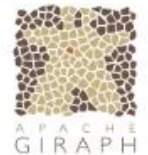
Interactive



Realtime



Iterative



빅데이터 플랫폼 - ML Services

ML Services

ML
Lifecycle



Processing
Frameworks



Feature
Stores



ML
Libraries



빅데이터 플랫폼 - Data Serving Systems

Data Serving Services

Column



Key-Value



Document



Graph



Search



Time Series



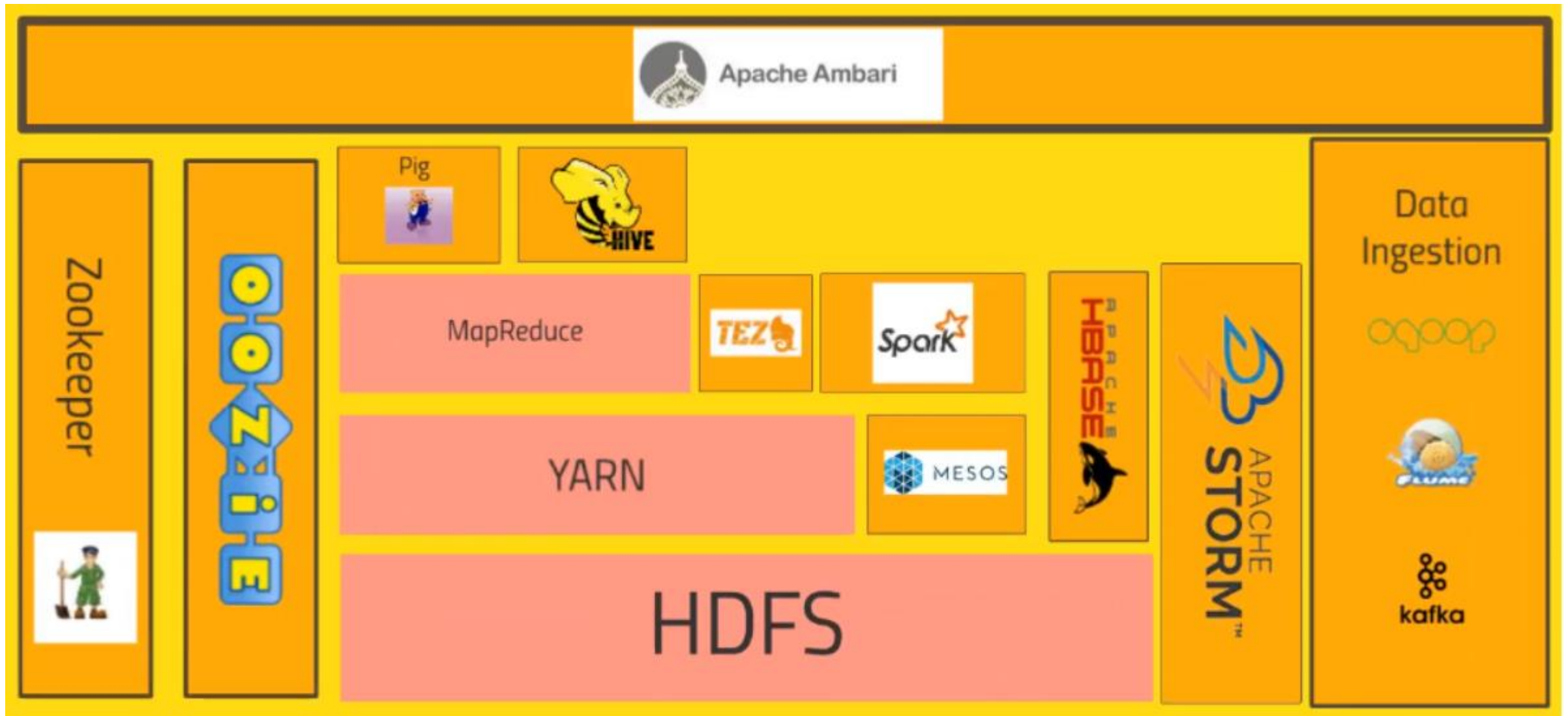
Relational



In-Memory



Hadoop Ecosystem



Hadoop Ecosystem

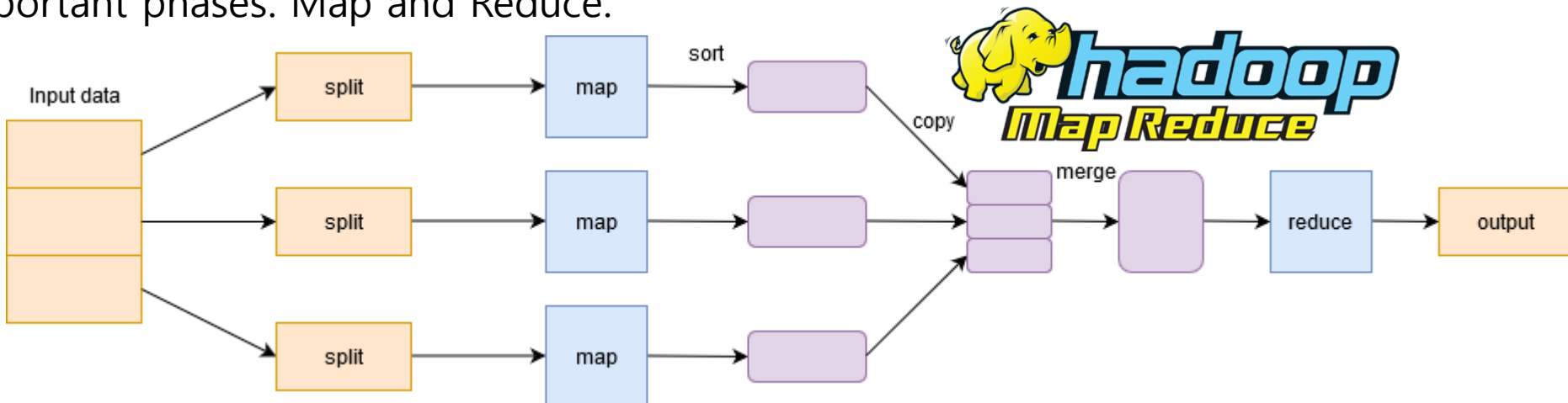
■ HDFS(Hadoop Distributed File System)

- It is the storage component of Hadoop that stores data in the form of files.
- Each file is divided into blocks of 128MB (configurable) and stores them on different machines in the cluster.
- It has a master-slave architecture with two main components: Name Node and Data Node.



■ MapReduce | [Hadoop Proper]

- To handle Big Data, Hadoop relies on the MapReduce algorithm introduced by Google and makes it easy to distribute a job and run it in parallel in a cluster.
- It essentially divides a single task into multiple tasks and processes them on different machines.
- It has two important phases: Map and Reduce.



Hadoop Ecosystem



■ YARN | [Hadoop Proper]

- YARN or Yet Another Resource Negotiator manages resources in the cluster and manages the applications over Hadoop.
- It allows data stored in HDFS to be processed and run by various data processing engines such as batch processing, stream processing, interactive processing, graph processing, and many more.
- This increases efficiency with the use of YARN.

■ Apache HBase

- HBase is a Column-based NoSQL database.
- It runs on top of HDFS and can handle any type of data.
- It allows for real-time processing and random read/write operations to be performed in the data.

Hadoop Ecosystem

■ Pig



Apache Pig

- Pig was developed for analyzing large datasets and overcomes the difficulty to write map and reduce functions.
- It consists of two components: Pig Latin and Pig Engine.
- Pig Latin is the Scripting Language that is similar to SQL.
- Pig Engine is the execution engine on which Pig Latin runs.
- Internally, the code written in Pig is converted to MapReduce functions and makes it very easy for programmers who aren't proficient in Java.

■ Hive

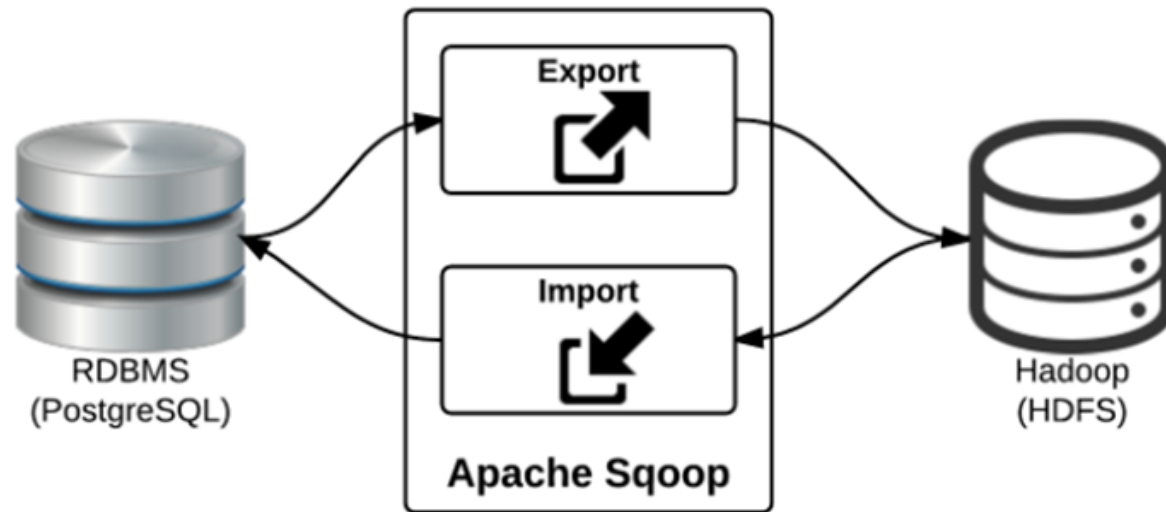


- Hive is a distributed data warehouse system developed by Facebook.
- It allows for easy reading, writing, and managing files on HDFS.
- It has its own querying language for the purpose known as Hive Querying Language (HQL) which is very similar to SQL.
- This makes it very easy for programmers to write MapReduce functions using simple HQL queries.

Hadoop Ecosystem

■ Sqoop

- A lot of applications still store data in RDBMS, thus making them a very important source of data.
- Therefore, Sqoop plays an important part in bringing data from Relational Databases into HDFS.
- The commands written in Sqoop internally convert into MapReduce tasks that are executed over HDFS.
- It works with almost all relational databases like MySQL, Postgres, SQLite, etc.
- It can also be used to export data from HDFS to RDBMS.

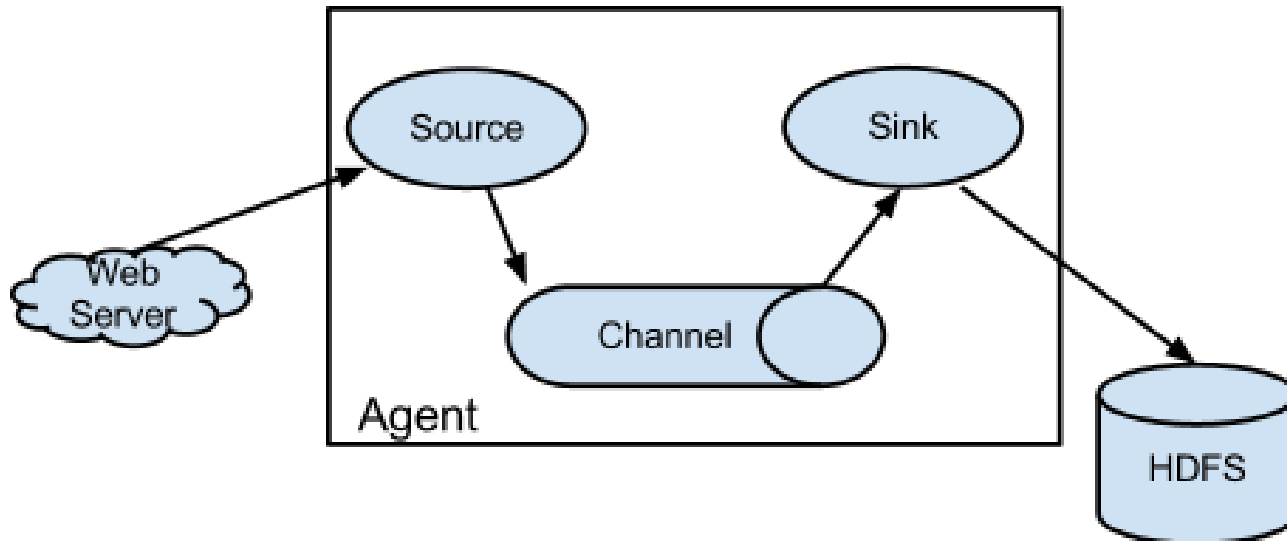


Hadoop Ecosystem

■ Flume



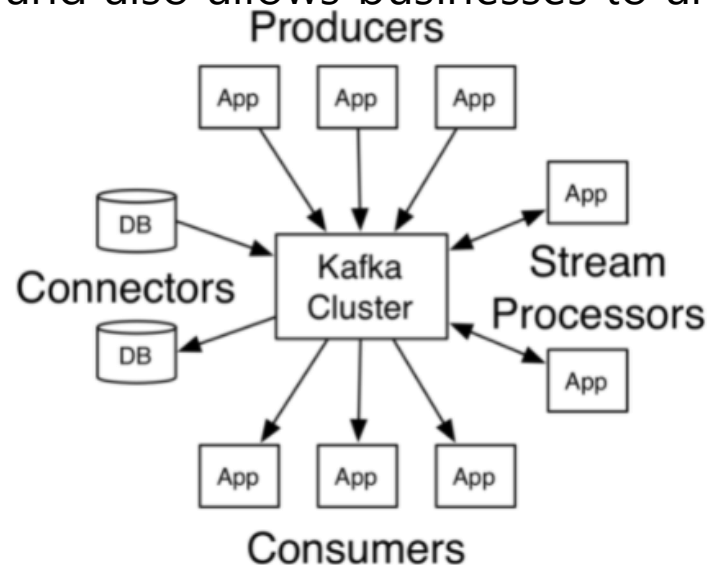
- Flume is an open-source, reliable, and available service used to efficiently collect, aggregate, and move large amounts of data from multiple data sources into HDFS.
- It can collect data in real-time as well as in batch mode.
- It has a flexible architecture and is fault-tolerant with multiple recovery mechanisms.



Hadoop Ecosystem



- There are a lot of applications generating data and a commensurate number of applications consuming that data.
- But connecting them individually is a tough task.
- That's where Kafka comes in. It sits between the applications generating data (Producers) and the applications consuming data (Consumers).
- Kafka is distributed and has in-built partitioning, replication, and fault-tolerance.
- It can handle streaming data and also allows businesses to analyze data in real-time.



Core Hadoop Ecosystem

■ Oozie



- Oozie is a workflow scheduler system that allows users to link jobs written on various platforms like MapReduce, Hive, Pig, etc. Using Oozie you can schedule a job in advance and can create a pipeline of individual jobs to be executed sequentially or in parallel to achieve a bigger task.
- For example, you can use Oozie to perform ETL operations on data and then save the output in HDFS..

■ Zookeeper

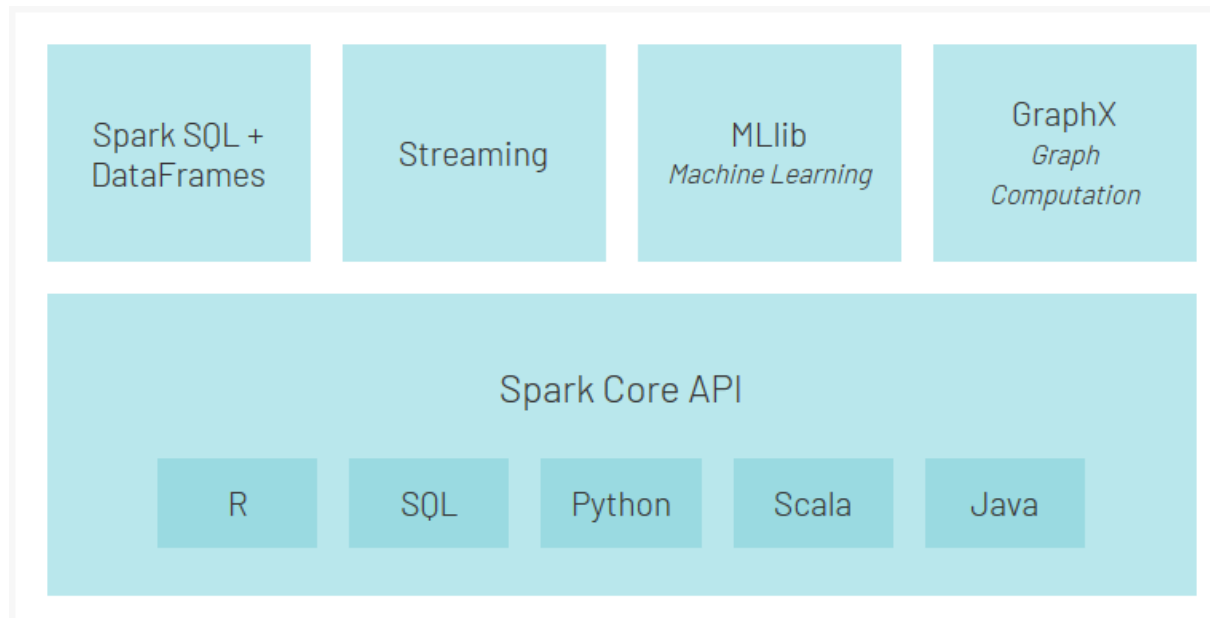


- In a Hadoop cluster, coordinating and synchronizing nodes can be a challenging task.
- Therefore, Zookeeper is the perfect tool for the problem.
- It is an open-source, distributed, and centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services across the cluster.

Core Hadoop Ecosystem

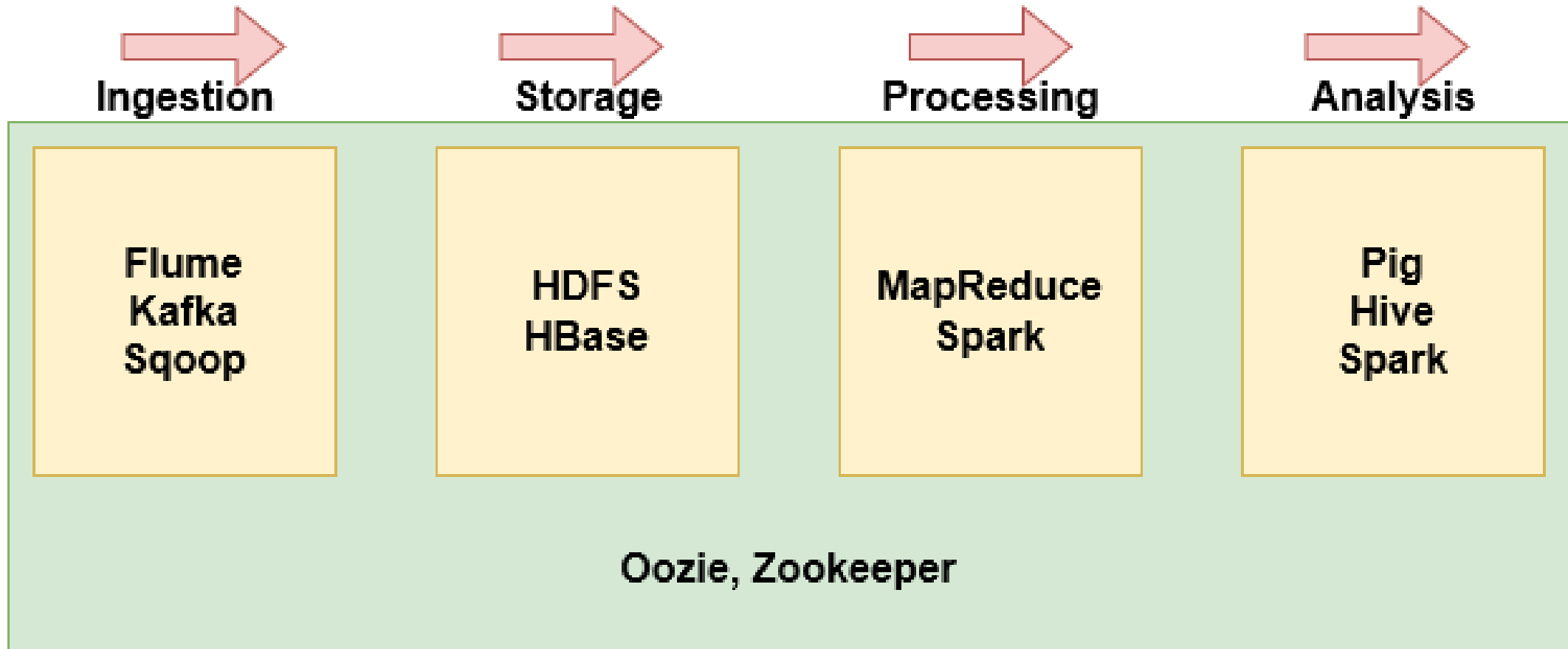


- Spark is an alternative framework to Hadoop built on Scala but supports varied applications written in Java, Python, etc.
- Compared to MapReduce it provides in-memory processing which accounts for faster processing.
- In addition to batch processing offered by Hadoop, it can also handle real-time processing.
- Further, Spark has its own ecosystem:



Core Hadoop Ecosystem

■ Stages of Big Data Processing

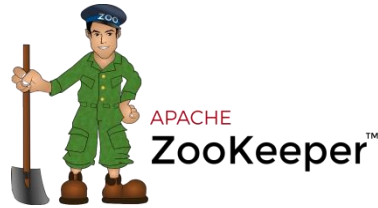


Core Hadoop Ecosystem

■ Apache Oozie

- Oozie is a workflow scheduler system to manage Apache Hadoop jobs on the cluster.
- However, Oozie is a scalable, reliable, and extensible system.
- So, it is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs.
- For example, it is ready for Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop, and Distcp.
- Also, it can perform system-specific jobs such as Java programs and shell scripts.

■ Apache ZooKeeper™



- Zookeeper also sits aside along with all of these technologies and enables highly reliable distributed coordination.
- So it is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.
- For example, Zookeeper coordinates which nodes are down, which are up. Also, keeps track of who is the master node, who are the slaves. etc.

빅데이터 분석툴



Hadoop은 클러스터 된 파일 시스템 및 빅 데이터 처리에 사용되는 소프트웨어 프레임워크
MapReduce 프로그래밍 모델을 통해 빅데이터의 데이터셋 처리



데이터 분석, 기계 학습 알고리즘 및 빠른 클러스터 컴퓨팅을 위한 오픈 소스 프레임워크



Apache HDFS/HBase의 데이터를 다룰 때 SQL쿼리로 쓸 수 있게 해주는 툴



Apache Hadoop, Apache Spark, Apache Impala 등을 포함하는 무료 플랫폼 배포를 제공
무제한 데이터를 수집, 처리, 관리, 발견, 모델링 및 배포 할 수 있음

빅데이터 분석툴



서버에 분산 된 방대한 양의 데이터를 관리하도록 구성된 오픈 소스 분산 NoSQL DBMS
CQL (Cassandra Structure Language)을 사용하여 데이터베이스와 상호 작용



문서 지향 데이터베이스

주요 기능: BSON(Binary JSON) 형식 사용, 샤딩, 인덱싱, 복제, 로드 밸런싱



Elasticsearch : 분산형 RESTful 검색 및 분석엔진

Logstash : 여러 소스에서 동시에 데이터를 수집하여 변환한 후 Elasticsearch 로

Kibana : 차트와 그래프를 이용해 데이터를 시각화



Tableau는 비즈니스 인텔리전스 및 분석을 위한 소프트웨어 솔루션

빅데이터 제공 사이트



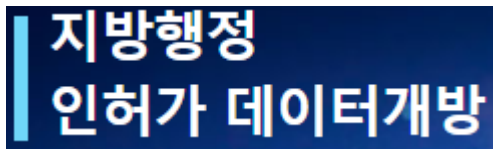
<https://www.data.go.kr>



<https://kostat.go.kr>



<https://opendata.hira.or.kr>



<https://www.localdata.go.kr>



<http://data.seoul.go.kr>



<https://data.gg.go.kr/>



<https://www.data.gov>



<https://open.fda.gov/>

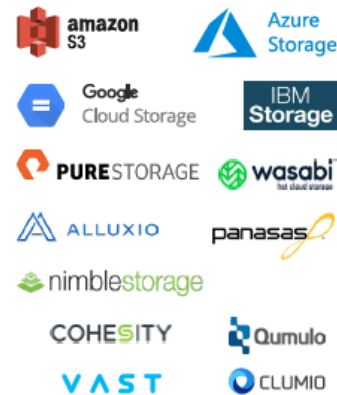
빅데이터 제공 사이트

- 통신 빅데이터 플랫폼: bdp.kt.co.kr
- 교통 빅데이터 플랫폼: <https://www.bigdata-transportation.kr/>
- 문화 빅데이터 플랫폼: culture.go.kr/bigdata
- 환경 빅데이터 플랫폼: <https://www.bigdata-environment.kr/user/main.do>
- 중소기업 빅데이터 플랫폼: datastore.wehago.com
- 지역경제 빅데이터 플랫폼: ggdata.kr
- 금융 빅데이터 플랫폼: fnbigdata.com
- 헬스케어 빅데이터 플랫폼: cancerportal.kr
- 유통소비 빅데이터 플랫폼: kdx.kr
- 산림 빅데이터 플랫폼: forestdata.kr
- 소방안전 빅데이터 플랫폼: <https://www.bigdata-119.kr/>
- 스마트치안 빅데이터 플랫폼: <https://www.bigdata-policing.kr/policy/main/index.do>
- 해양수산 빅데이터 플랫폼: <http://www.bigdata-sea.kr/>
- 농식품 빅데이터 플랫폼: <https://kadx.co.kr/>
- 라이프로그 빅데이터 플랫폼: <https://www.bigdata-lifelog.kr/portal>
- 디지털 산업혁신 빅데이터 플랫폼: <http://www.bigdata-dx.kr/>

Data and AI Landscape

INFRASTRUCTURE

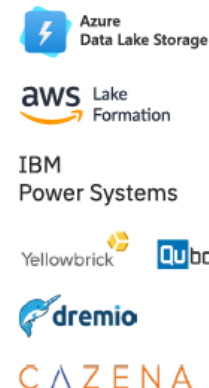
STORAGE



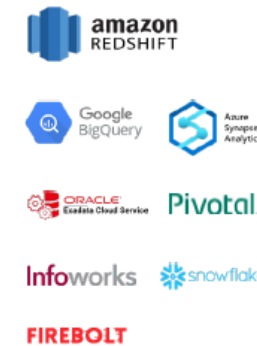
HADOOP



DATA LAKES



DATA WAREHOUSES



STREAMING / IN-MEMORY



NoSQL DATABASES



NewSQL DATABASES



GRAPH DBs



MPP DBs



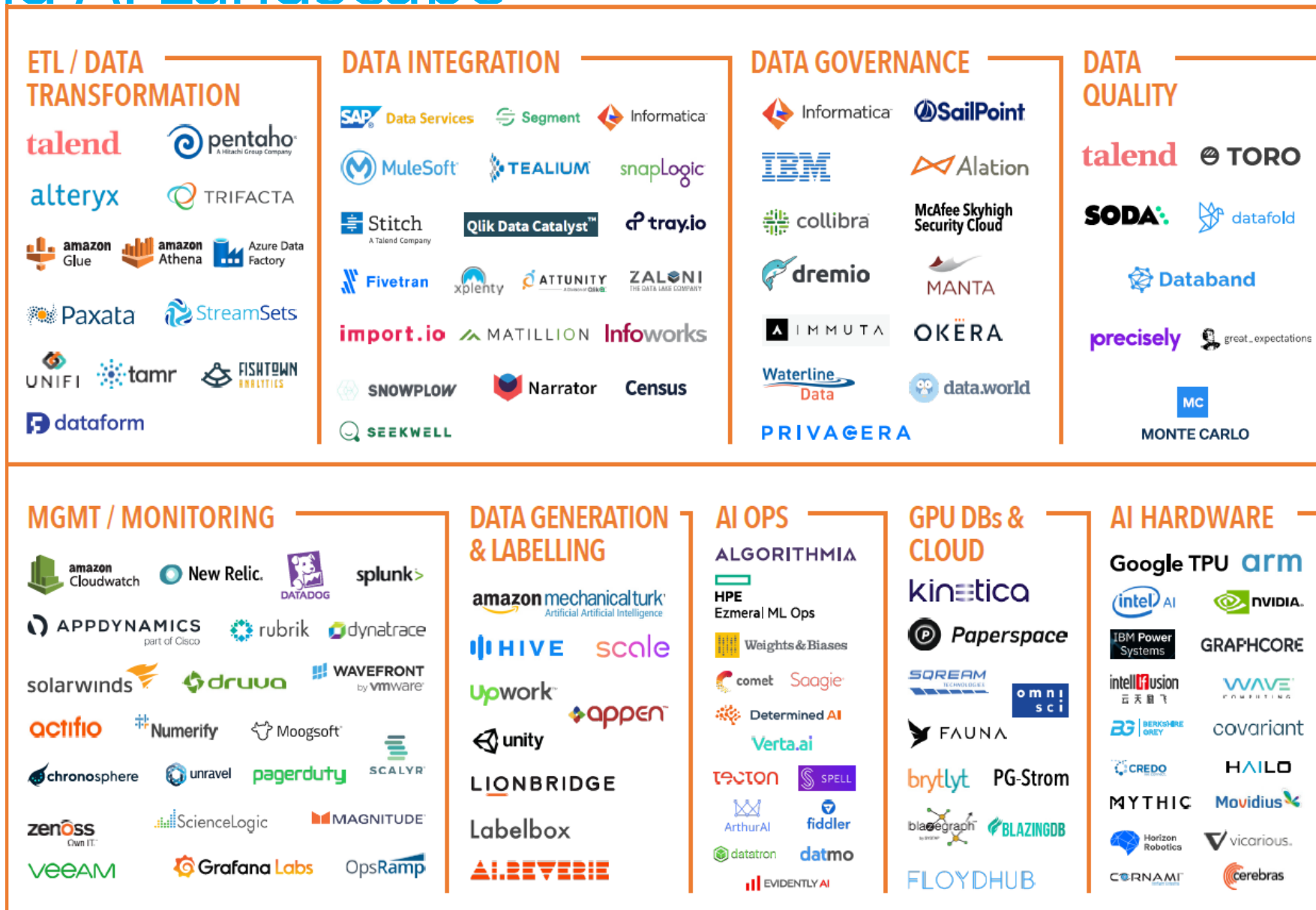
SERVER-LESS



CLUSTER SVCS



Data and AI Landscape



Data and AI Landscape

ANALYTICS & MACHINE INTELLIGENCE

BI PLATFORMS



VISUALIZATION



DATA ANALYST PLATFORMS



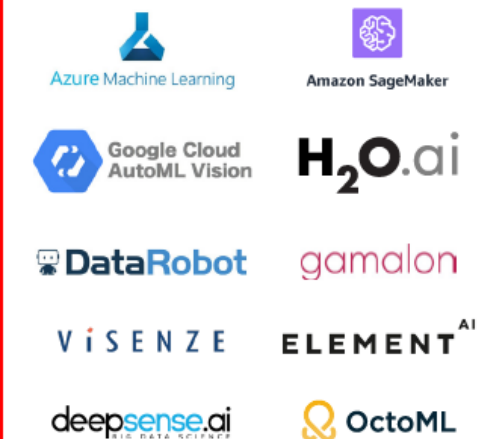
DATA SCIENCE NOTEBOOKS



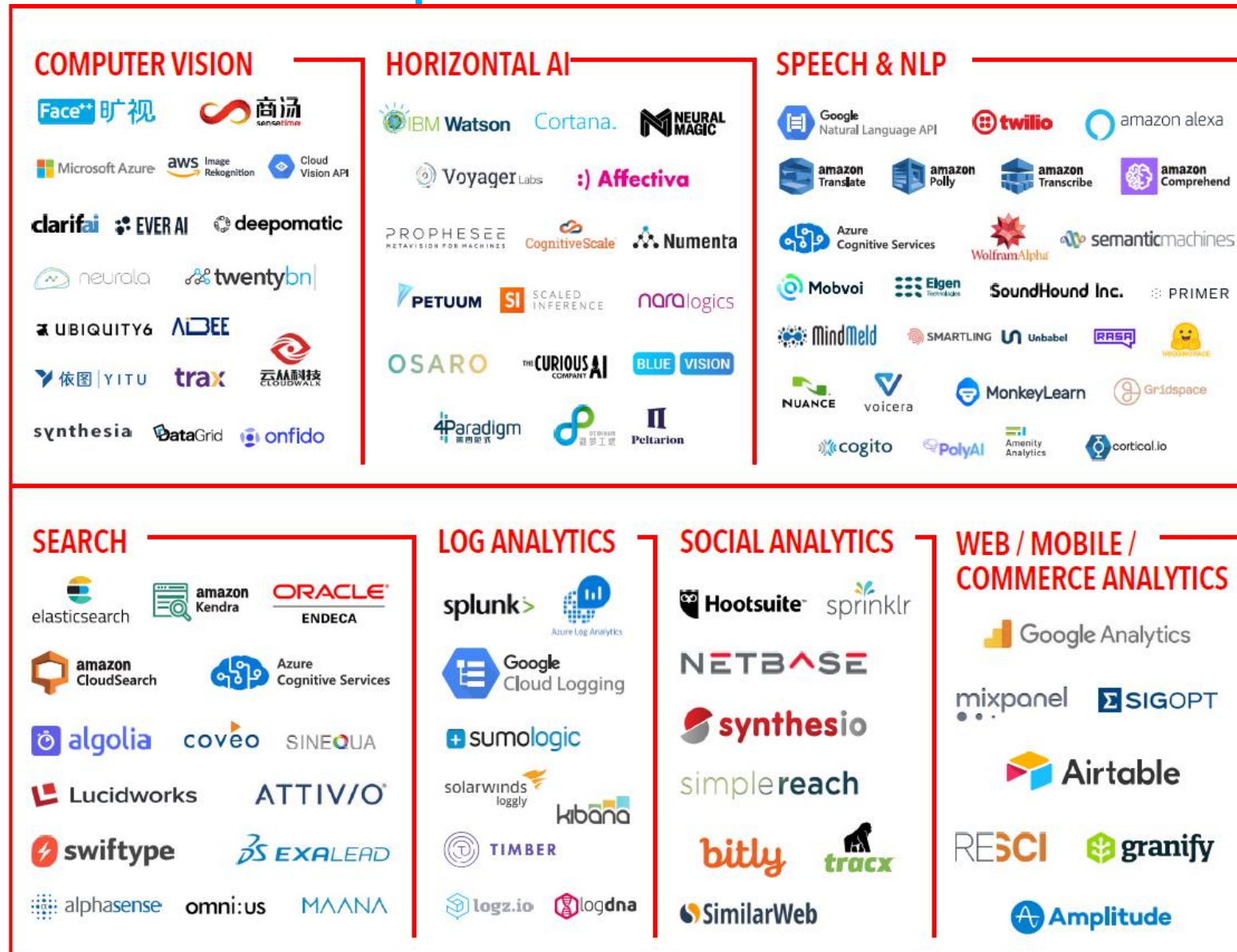
DATA SCIENCE PLATFORMS



MACHINE LEARNING

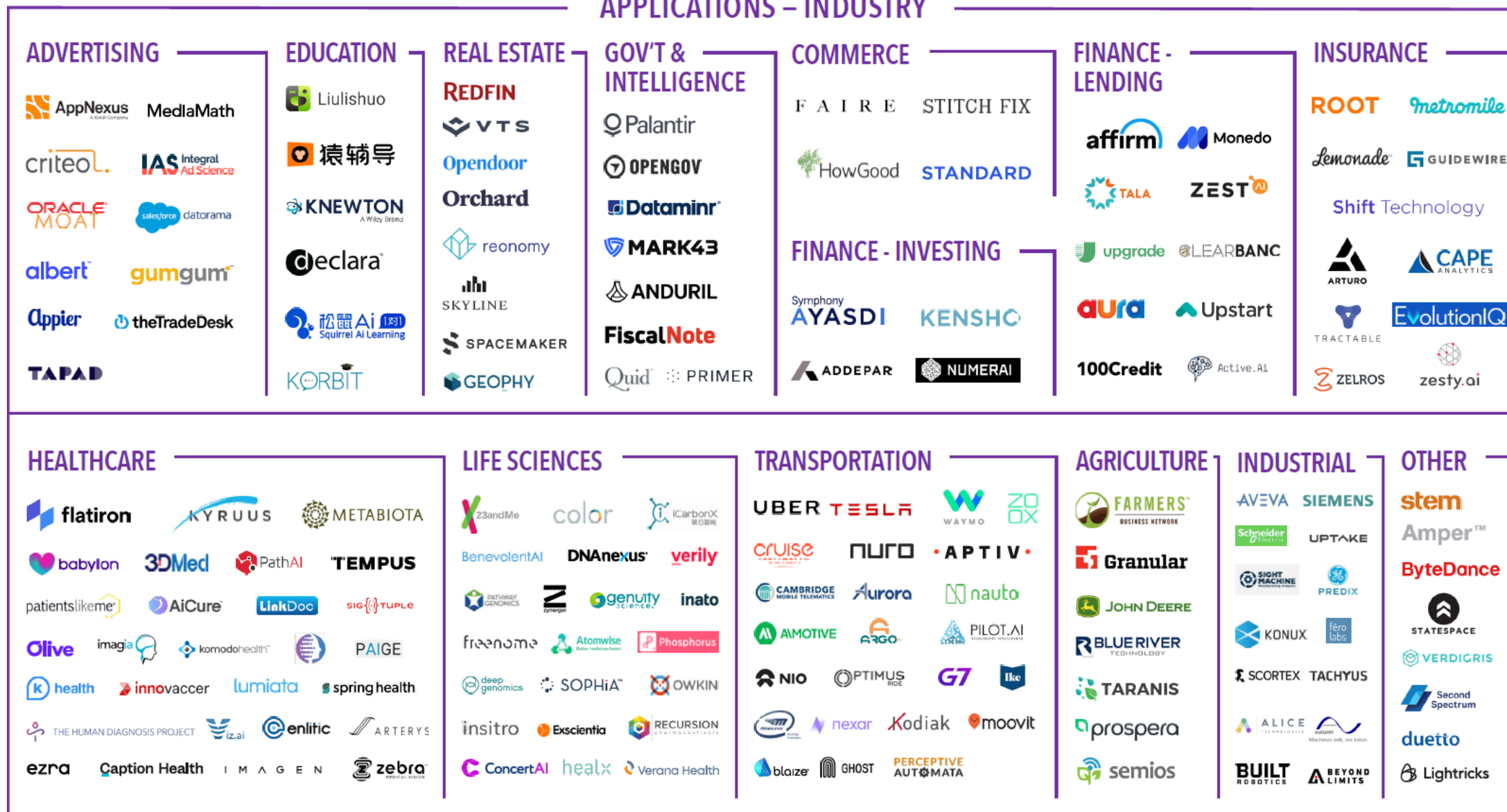


Data and AI Landscape

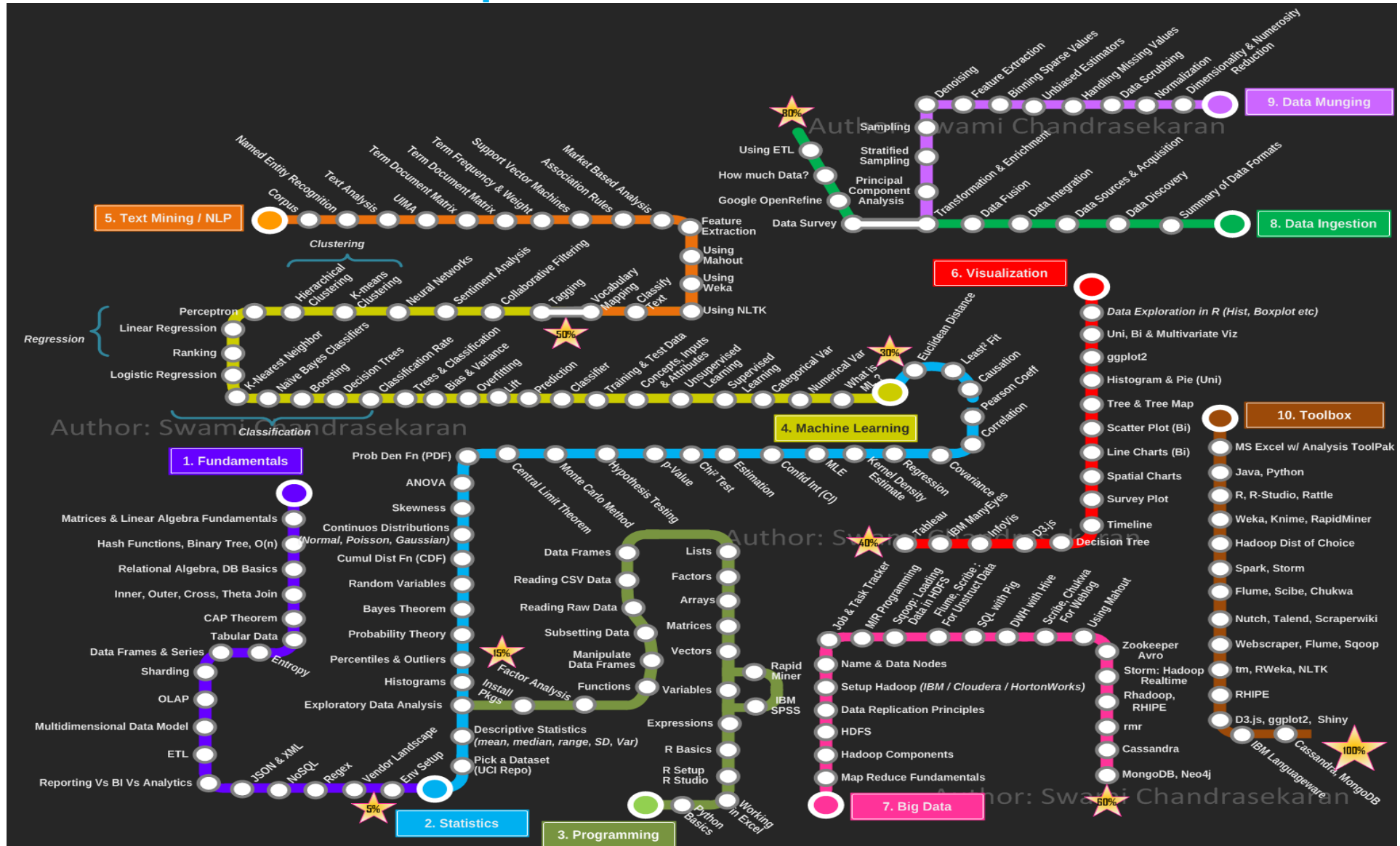


Data and AI Landscape

APPLICATIONS – INDUSTRY



Data Scientist Roadmap



Thank you