Data Platform

I. Issues and Trends

II. Architecture

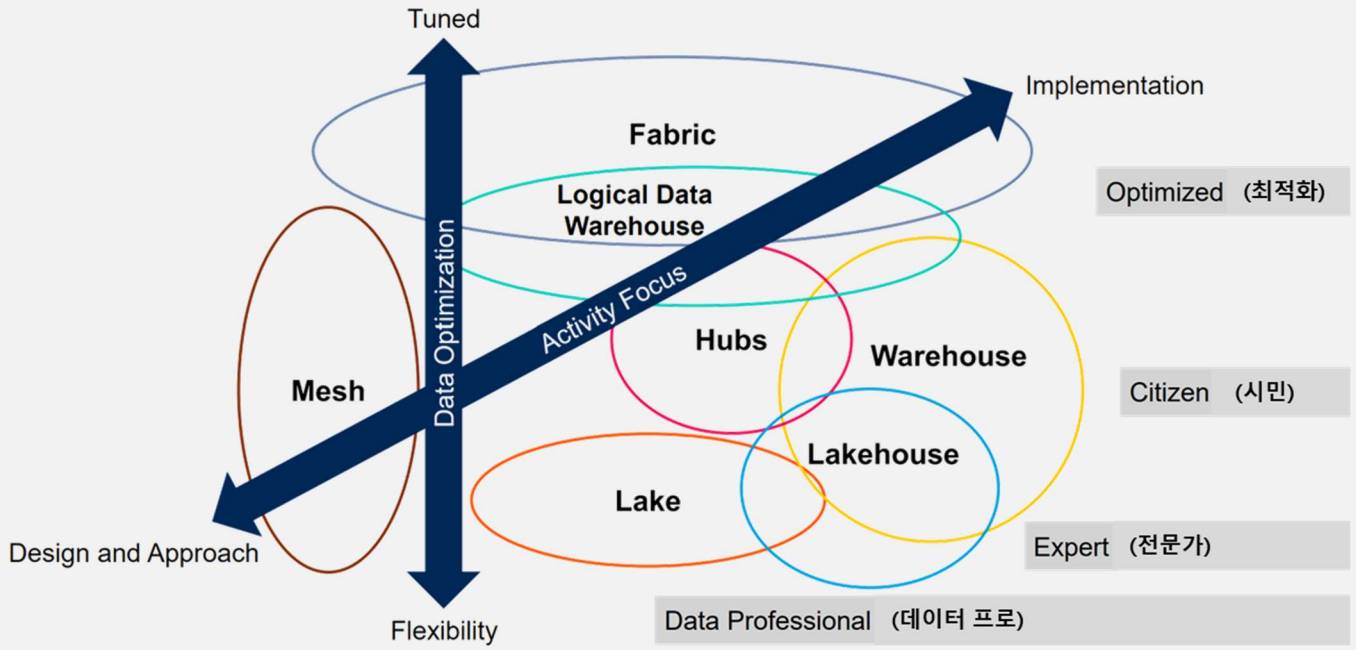
III. 오픈소스 기반 설계

2023년 1월 18일
안종석
james@jslab.kr

I. ISSUES AND TRENDS

❖ Data & Analytics Governance Platforms (Gartner 2022)

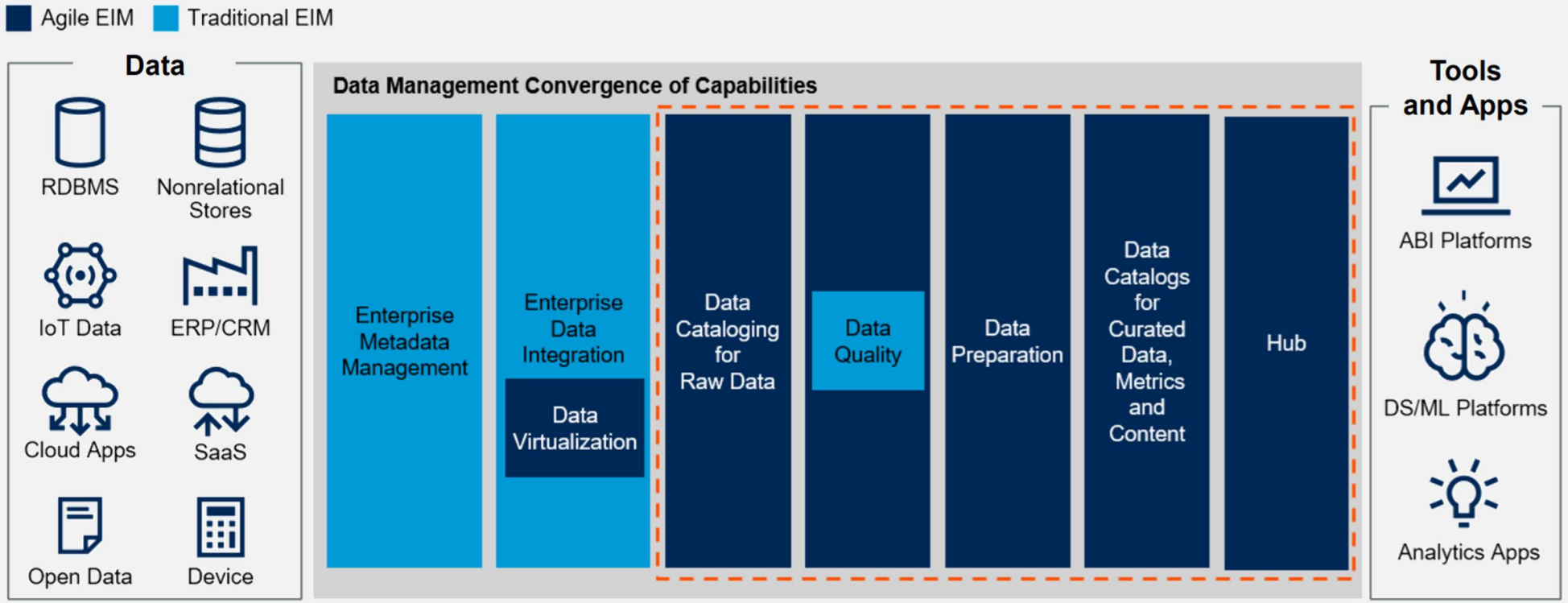
- **Data Optimization** (Flexibility←→Tuned)
- **Activity Focus** (Design/Approach←→Implementation)



Source: https://webinar.gartner.com/441976/agenda/session/1044039?login=ML

I. ISSUES AND TRENDS

❖ **Market Requires "a Convergence of Capabilities" in Data Management (Gartner 2022)**

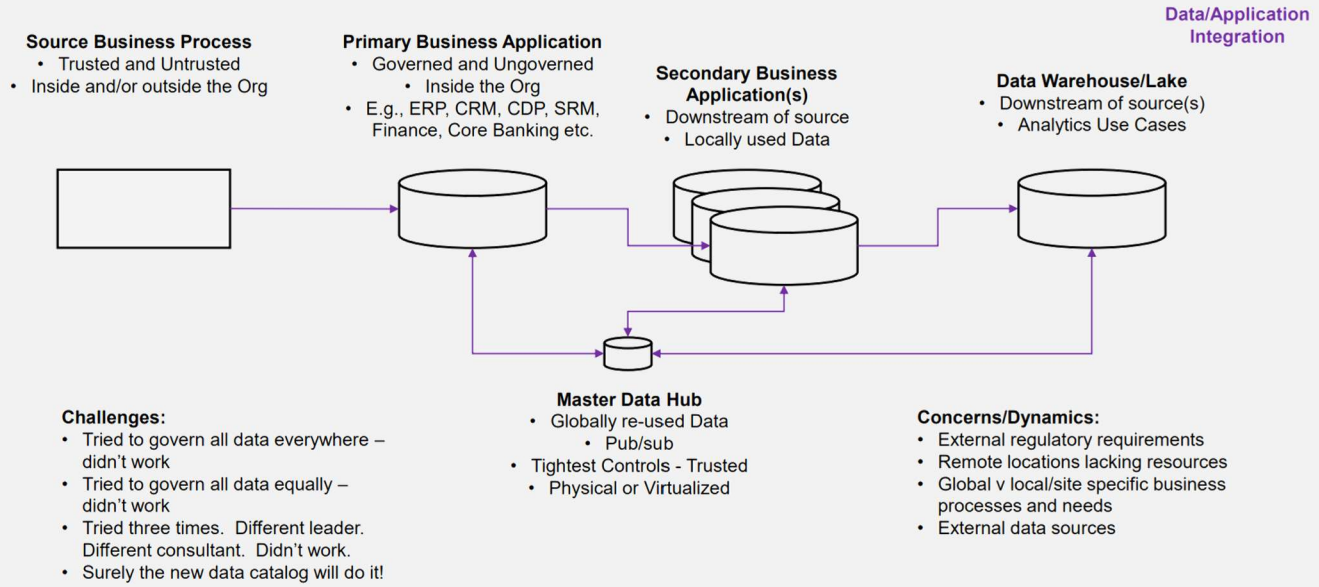


Source: https://webinar.gartner.com/441976/agenda/session/1044039?login=ML

I. ISSUES AND TRENDS

❖ Data & Analytics Governance in Operational Business Systems (Gartner 2022)

- **Challenges**
  - Tried to govern all data everywhere didn't work
  - Tried to govern all data equally didn't work
  - Tried three times. Different leader. Different consultant. Didn't work.
  - Surely the new data catalog will do it!



Source: https://webinar.gartner.com/441976/agenda/session/1044039?login=ML

II. ARCHITECTURE

❖ 데이터 플랫폼: 데이터 관리, 데이터 웨어하우징, 분석 및 기계 학습과 같은 추가 기능을 제공

- 데이터 플랫폼은 데이터 패브릭 위에 구축되며 데이터 관리, 데이터 웨어하우징, 분석 및 기계 학습을 위한 추가 도구 및 서비스도 포함하는 보다 포괄적인 시스템
- 데이터 거버넌스, 데이터 계보, 데이터 품질 및 데이터 보안과 같은 추가 기능을 제공

❖ 데이터 패브릭: 데이터 플랫폼의 백본으로 하위 구성 요소로 간주되며 서로 다른 소스 간의 데이터 이동 및 통합에 중점

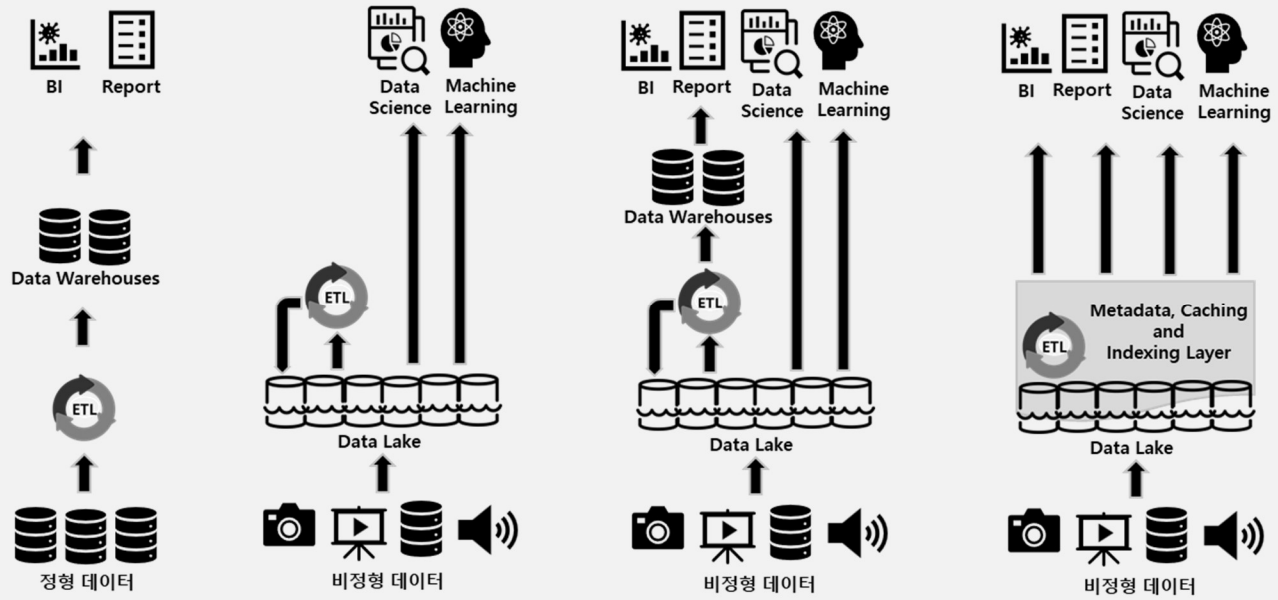
- 조직의 다양한 데이터 소스 및 시스템에서 데이터를 연결하고 관리하는 기술 및 서비스
- 데이터베이스, 파일 시스템 및 클라우드 서비스와 같은 서로 다른 시스템 간에 데이터가 원활하고 안전하게 흐르도록 하고 실시간 데이터 처리 및 분석이 가능
- 오픈소스 Apache Kafka 또는 Apache Nifi등으로 분산 시스템 상에 구축되어 데이터 이동 및 통합에 중점

❖ 데이터 가상화: 데이터가 단일 위치에 있는 것처럼 보이도록 추상화하여 데이터에 액세스하고 조작할 수 있도록 하는 것에 중점

II. ARCHITECTURE

❖ **Data architecture (데이터 아키텍처) or Data Storage Framesworks**

- **Data Warehouse**
- **Data Lake**
- **Data Lakehouse**

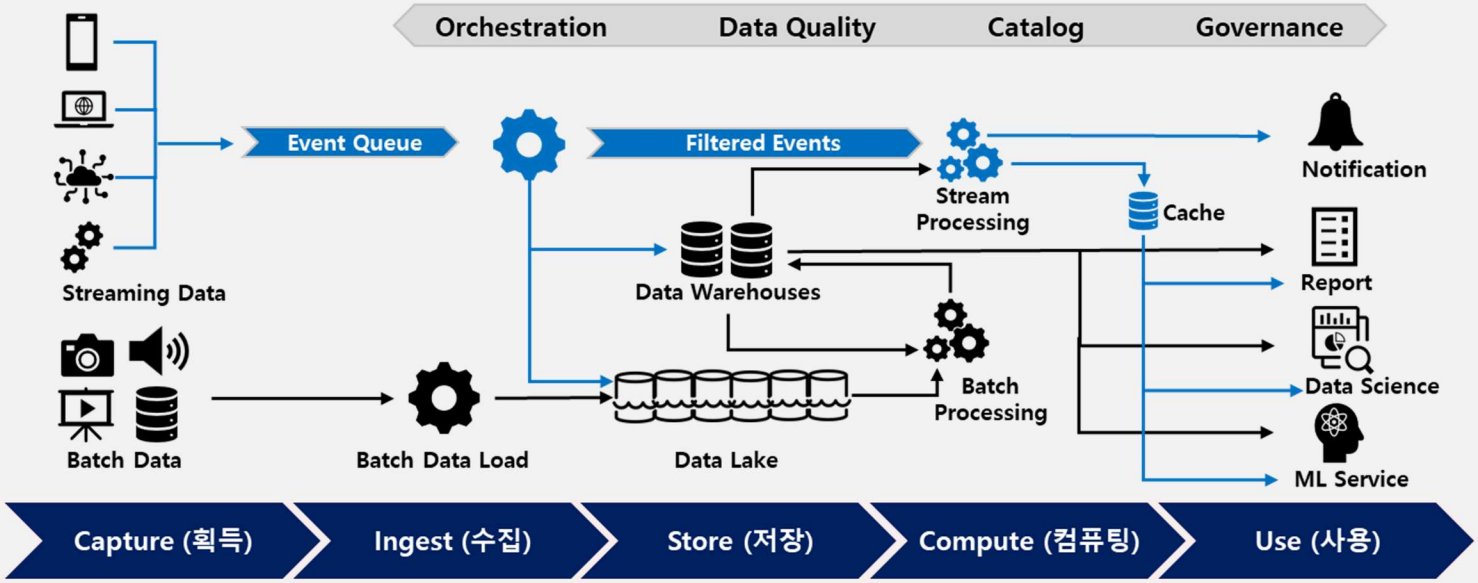


ETL: 추출(Extract), 변환(Transform), 로드(Load)

II. ARCHITECTURE

❖ 오픈소스로 구축 가능한 Data Pipeline Architecture (데이터 파이프라인 아키텍처)

- Capture (획득)
- Ingest (수집)
- Store (저장)
- Compute (컴퓨팅)
- Use (사용)



III. 오픈소스 기반 설계

❖ **오픈소스 사용 데이터 플랫폼 설계:**

- **오픈소스 커뮤니티 리소스를 활용 (Leveraging community resources)**
  - 오픈 소스 소프트웨어를 사용하여 데이터 플랫폼 설계에 대한 조언과 지침을 제공할 수 있는 포럼 및 사용자 그룹과 같은 온라인 커뮤니티
- **구축된 솔루션을 레퍼런스로 사용(Pre-built solution references)**
  - AWS, GCP, Azure와 같은 CSP는 데이터 인프라를 위한 사전 구축된 솔루션을 제공하여 필요에 맞는 데이터 인프라를 쉽고 빠르게 설정 가능
- **데이터 플랫폼의 요구 사항과 목표를 정의하여 요구 사항을 충족하는 방식으로 설계하며, 향후 확장성과 변화하는 요구 사항을 지속적으로 충족할 수 있도록 관리, 유지 관리 및 업데이트 가능**
- **오픈 소스 소프트웨어를 사용하는 설계 프로세스는 반복적이며, 새로운 이슈가 수집되고 프로젝트가 진행됨에 따라 업데이트 하여 최종 솔루션이 요구 사항을 충족 할 수 있도록 담당자와 협력하여 필요에 따라 조정**

III. 오픈소스 기반 설계

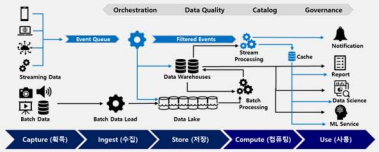
❖ 오픈소스 사용 데이터 플랫폼 설계:

1. **요구 사항 식별:**
   - 데이터 처리/저장/분석 측면에서 이슈와 요구 사항을 이해 (데이터 유형, 예상되는 데이터 양, 필요한 성능 및 확장성등)
2. **다양한 오픈 소스 소프트웨어 옵션 평가:**
   - 요구 사항에 적합한 것을 결정 (Hadoop, Spark, Hive, Flink, Storm, Samza, Druid, Airflow, Superset, 및 Cassandra 등등)
3. **아키텍처 설계:**
   - 요구 사항과 선택한 오픈 소스 소프트웨어를 기반으로 데이터 플랫폼의 아키텍처를 설계 (노드 수, 데이터 위치, 데이터 흐름 및 처리 파이프라인 등을 포함)
4. **오픈 소스 소프트웨어 구성:**
   - 아키텍처가 설계되면 선택한 노드에서 선택한 오픈 소스 소프트웨어를 구성
5. **보안 및 거버넌스 구현:**
   - 보안 및 거버넌스 사례를 레퍼런스로 데이터 프라이버시, 데이터 액세스 및 준수를 보장. (데이터가 조직의 요구 사항을 충족하는 방식으로 보호되고 사용되도록 인증, 권한 부여, 암호화 및 모니터링과 같은 보안 및 거버넌스 제어를 구현)
6. **테스트 및 배포(안):**
   - 데이터 플랫폼을 테스트하여 요구 사항을 충족하고 예상되는 데이터 양을 처리 확인 방안과 프로덕션에 배포(안)
7. **모니터링 및 유지 관리(안):**
   - 데이터 플랫폼 구성의 오픈 소스 소프트웨어를 정기적으로 유지 관리하고 필요에 따라 업데이트하여 조직의 요구 사항을 지속적으로 충족모니터링/실행/조정 (안)

III. 오픈소스 기반 설계

❖ 오픈소스 (예) @ Frameworks in data engineering (IEEE 2020)

- **Data Connect (연결):** Flume, kafka connect, nifi, FALCOR, Graph API
- **Data Ingest: (수집)** Spark streaming, RabbitMQ, HERON, kafka, PULSAR, STORM, FLINK, RockerMQ
- **Data Store (저장):** neo4j, redis, ClickHouse, GIRAPH, ceph, hadoop HDFS, HBASE
- **Data Analytics (분석)**
  - **Processing (프로세싱):** spark, Flink, samza, TEZ, hadoop Map Reduce, beam
  - **Log Analytics (로그분석):** splunk, logstash, Beats, fluentd, SENTRY, Prometeus
  - **Query (질문):** presto, DRILL, Pig, ARROW, pinot, HIVE
  - **Search (검색):** elasticsearch, Solr
  - **Analysis (분석):** DL4J, PyTorch, Big이, TensorFlow, Spark Mllib, rllib
- **Data Visualization (가시화):** kibana, Grafana, Superset, druid, graphite, Plotly Dash
- **Data Management (관리):** hadoop YARN, MESOS, Airflow, Kubernetes, Docker Swarm, LXC, YuniKorn



Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9743922&fbclid=IwAR08OlkAzEqezXobSwczEKsTII--MLOrvgZfLZlG8jJYFaq5zHgj4BHu4Ns

III. 오픈소스 기반 설계

❖ 데이터 레이크 (Data Lakes, IEEE 2022)

- **Data Lakes**
  - e.g. in distributed storage systems: Hadoop clusters HDFS, Ceph, NoSQL databases (such as MongoDB, Apache Cassandra (inspired by Amazon DynamoDB), CouchDB, Hbase, CosmosDB), NewSQL databases (such as MariaDB, MemSQL, VoltDB, InfluxDB, NuoDB).
- **Depending on the data models, NoSQL databases can also be divided into the following categories:**
  - Graph databases: Apache Spark's API GraphX
  - Key-value stores: Memcached and Redis
  - Column-oriented databases: Cassandra or Hbase
  - Document-oriented databases: MongoDB, Elasticsearch, Apache CouchDB
- **Data Lakes are designed to be much cheaper, easy to write, and store large amounts of data. However, it is difficult to access/read data from Data Lakes because the data may not be stored according to the analysis requirements and lacks consistency/isolation features. They also have complex system architectures due to multiple storage systems with different semantics. On the other hand, recent advances in ML libraries and data science ecosystem projects (e.g., TensorFlow) are starting to be integrated into data lakes after data preparation**

Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9743922&fbclid=IwAR1-UabuvnH7CiKINVJMcen3_TwSKcT5n__MVfSlIKmqPLwwISRE2Anjw3E

III. 오픈소스 기반 설계

❖ 데이터 모니터와 가시화 프레임워크 (IEEE 2022)

- **Notification services and alerts, monitoring dashboards using tools**
  - Grafana(https://grafana.com/) or
  - Kibana
- **Open graph visualization platform**
  - Gephi (https://gephi.org/)
- **Building web applications and data visualization apps**
  - Dash (https://dash.plotly.com/), a production Python framework using Flask, Plotly.js, and React.js
- **Apache Superset is an open-source visualization tool developed by Airbnb that is used to visualize analytics results (with chart types such as word counts, heat maps, boxplots, etc.).**
  - Apache Superset (https://superset.incubator.apache.org/)

Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9743922&fbclid=IwAR1-UabuvnH7CiKINVJMcen3_TwSKcT5n__MVfSlIKmqPLwwISRE2Anjw3E

III. 오픈소스 기반 설계

❖ 데이터 오케스트레이션과 관리 프레임워크 (IEEE 2022)

- SCALING
- PARALLELISM
- MICROSERVICES AND DEPLOYMENT
- CI/CD AND IaC
- AVAILABLE TOOLS

Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9743922&fbclid=IwAR08OlkAzEqezXobSwczEKsTII--MLOrvgZfLZlG8jJYFaq5zHgj4BHu4Ns


THANK
YOU