스파크 데이터 파이프라인

순천향대학교 컴퓨터공학과

이 상 정

학습 내용

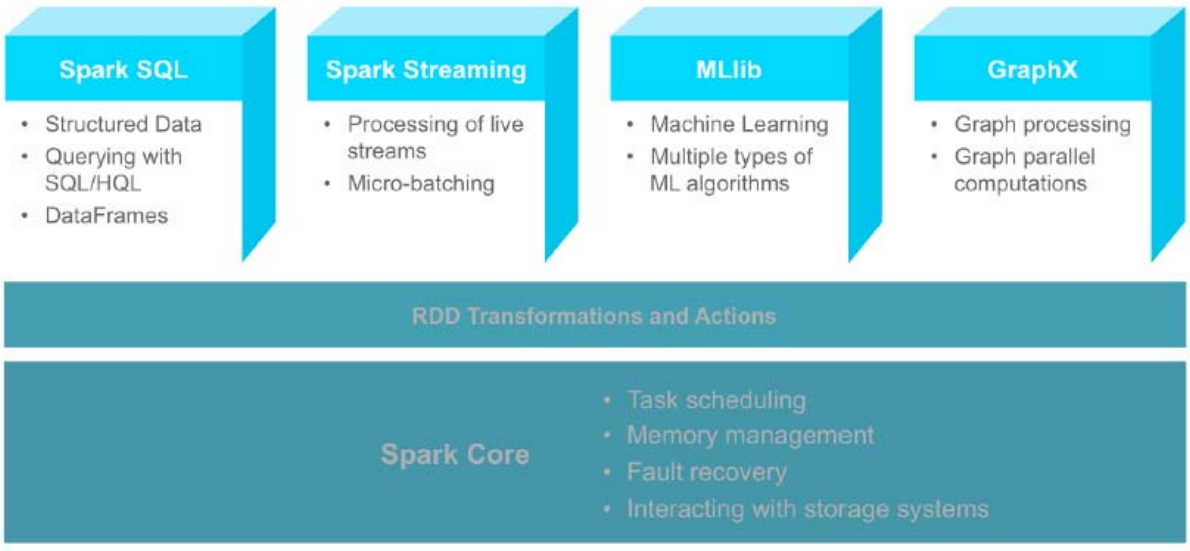
1. 스파크 통합 스택의 컴포넌트
2. 하둡 에코 시스템과 스파크
3. 스파크 데이터 파이프라인 활용 사례(Use Case)

1. 스파크 통합 스택의 컴포넌트

스파크 데이터 파이프라인

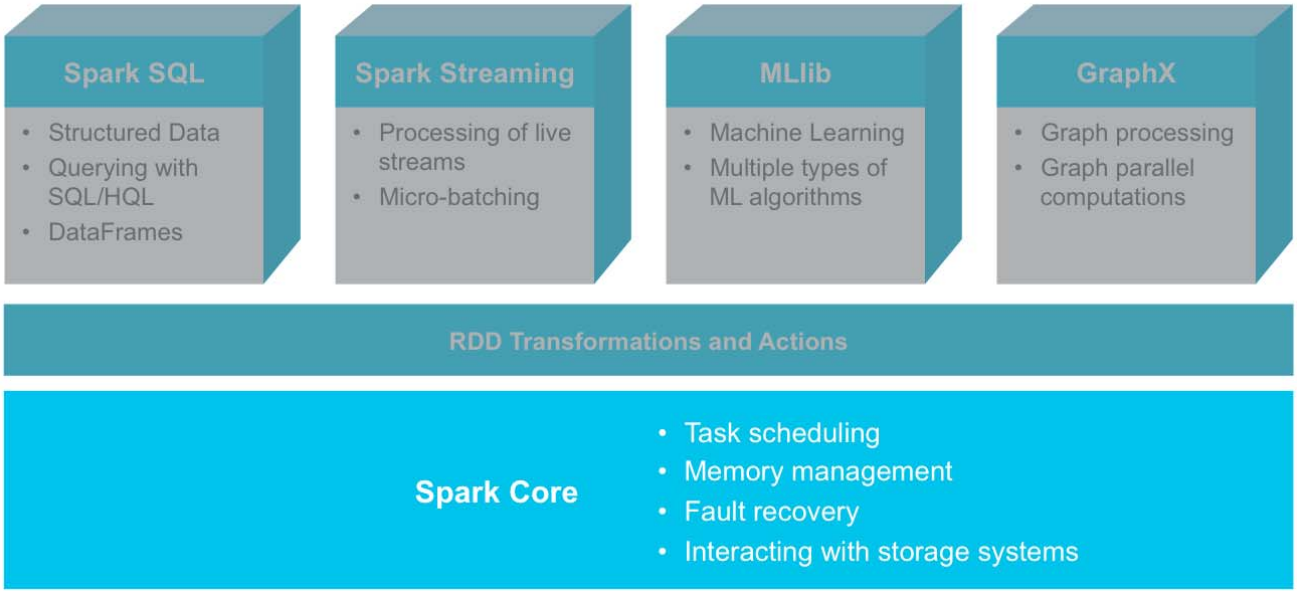
아파치 스파크 통합 스택 (Apache Spark Unified Stack)

- 스파크는 그래프 처리, 고급 질의, 스트림 처리, 기계 학습 등과 같은 고급 분석을 수행할 수 있는 통합 프레임워크
  - 같은 응용에서 단일 프로그래밍 언어를 사용하여 이들 라이브러리들을 결합 적용할 수 있음



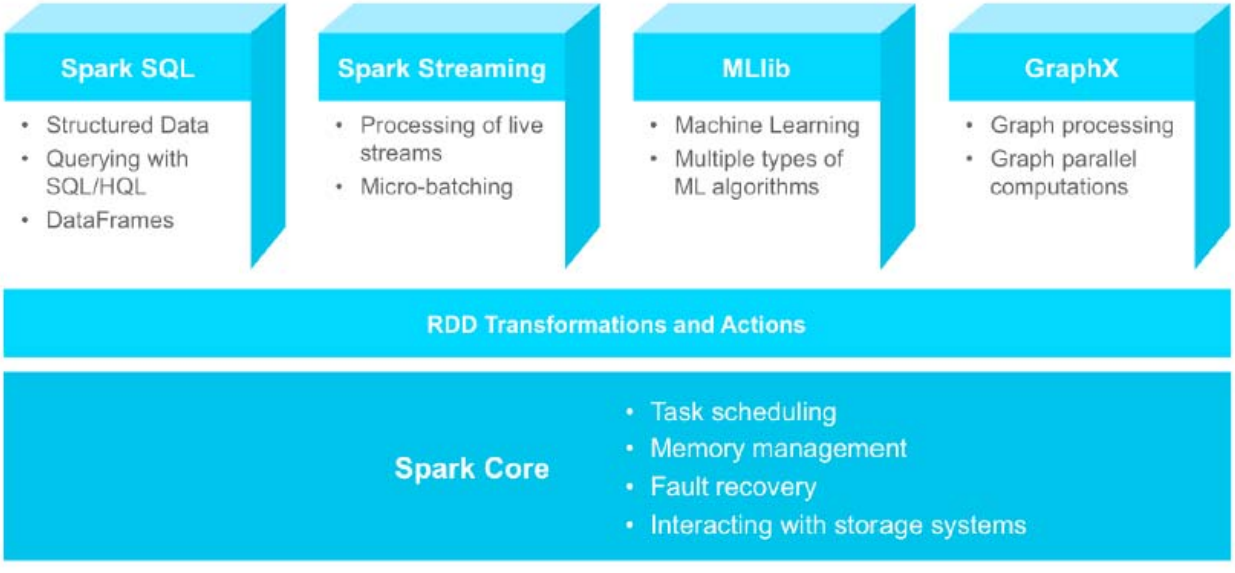
아파치 스파크 통합 스택 – 스파크 코어 (1)

- 스파크 코어는 태스크 스케줄링, 메모리 관리, 오류 회복, 저장 시스템과의 상호 작용을 책임지는 계산 엔진



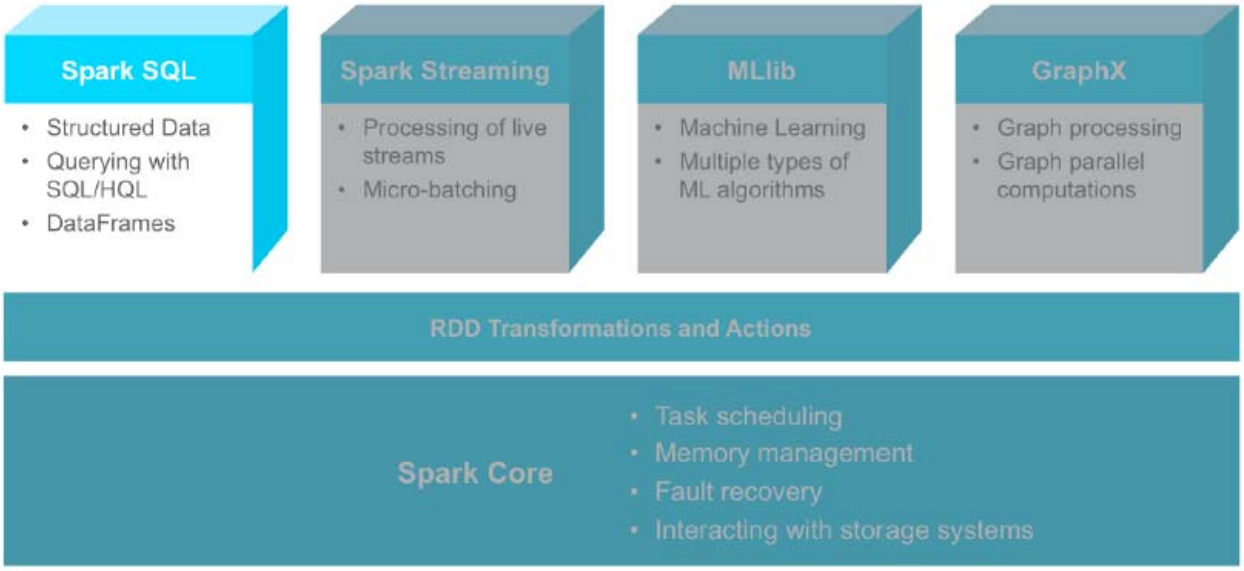
아파치 스파크 통합 스택 – 스파크 코어 (2)

- 다른 스파크 모듈들은 스파크 코어와 긴밀하게 결합



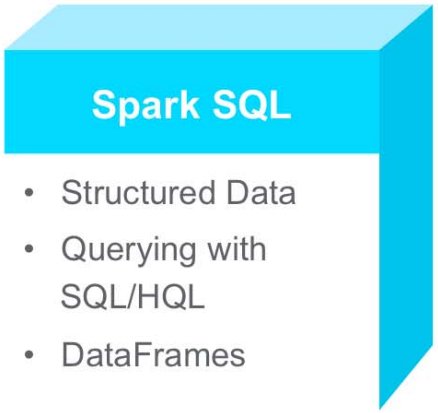
아파치 스파크 통합 스택 – 스파크 SQL (1)

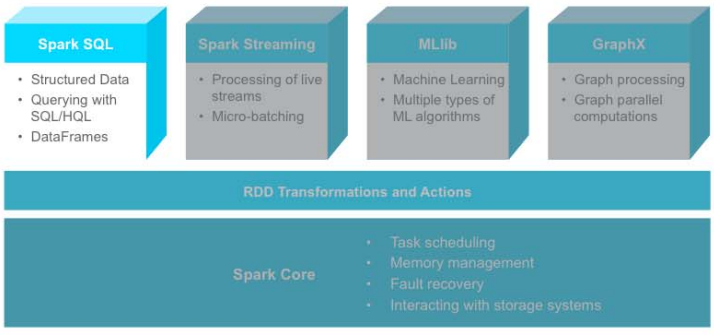
- 스파크 SQL은 구조화된 데이터를 처리



아파치 스파크 통합 스택 – 스파크 SQL (2)

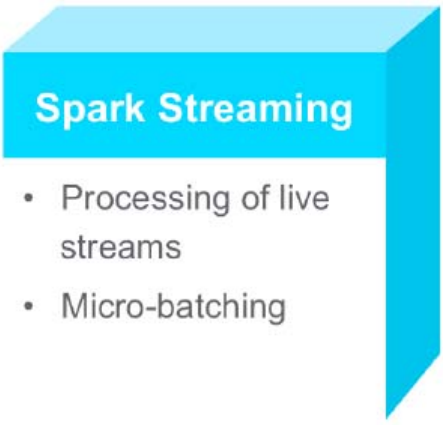
- 스파크는 SQL, HiveQL를 사용한 데이터 질의가 가능
  - 구조화된 Hive 테이블, 복잡한 JSON 데이터와 같은 다양한 데이터 소스 지원
  - SparkSQL의 기본적인 추상화는 데이터프레임

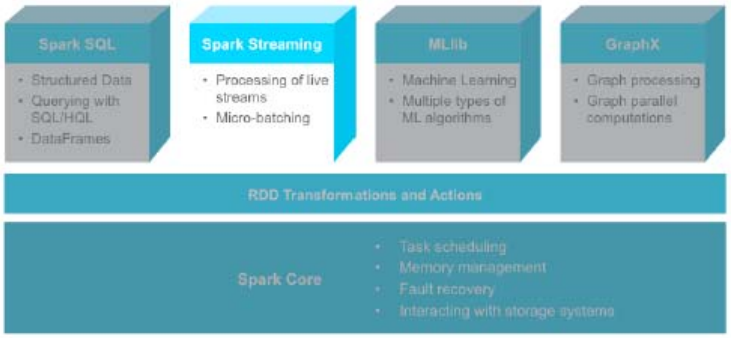




아파치 스파크 통합 스택 – 스파크 스트리밍 (1)

- 스파크 스트리밍(Spark streaming)은 데이터의 연속적인 스트림을 처리
  - 마이크로배치 실행 모델(microbatch execution model)을 적용하여 스트림 데이터 분석



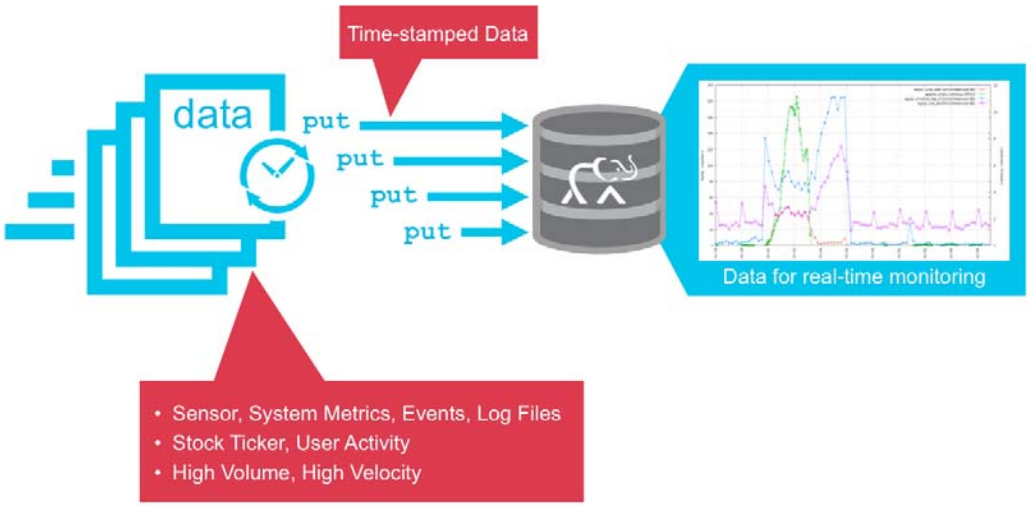


아파치 스파크 통합 스택 – 스파크 스트리밍 (2)

- 일반적으로 스트리밍 데이터 처리는 다음 요구사항을 만족해야 함
  - 실시간에 준하는(near-real-time) 처리 결과
  - 대규모 워크로드를 다룰 수 있어야 함
  - 수초 내의 지연 시간

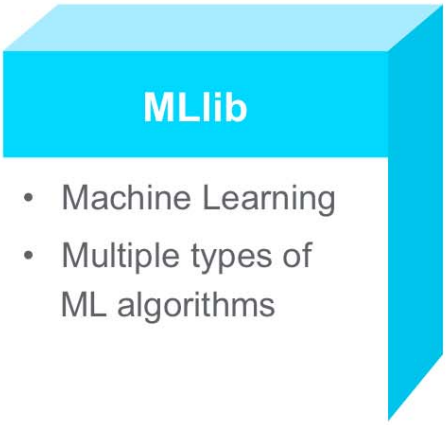
- 스트리밍 데이터 예로는 SNS 스트림 데이터, 모바일 응용, 타임스탬프된 로그 데이터, 트랜잭션 데이터, 센서 디바이스 네트워크의 이벤트 스트림 등이 있음
- 스트림 데이터의 실시간 처리 응용은 웹사이트 모니터링, 네트워크 모니터링, 사기 검출(fraud detection), 웹 클릭, 광고 등이 있음

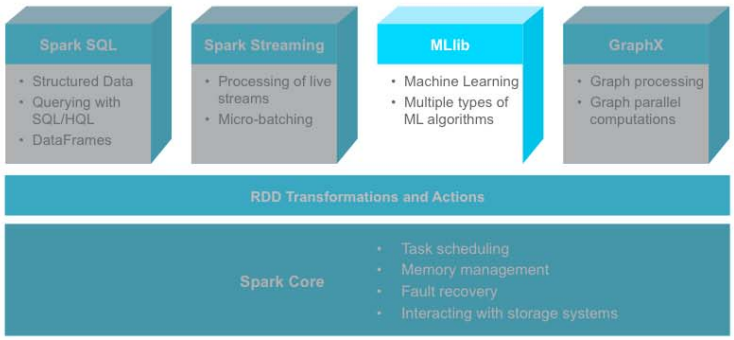
아파치 스파크 통합 스택 – 스파크 스트리밍 (3)



아파치 스파크 통합 스택 – 스파크 MLib (1)

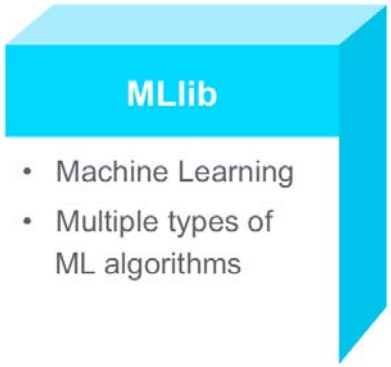
- MLib는 분류(classification), 회귀분석(regression), 클러스터링(clustering), 협업 필터링(collaborative filtering), 차원 축소(dimensionality reduction) 등 다양한 기계 학습 알고리즘을 제공하는 라이브러리





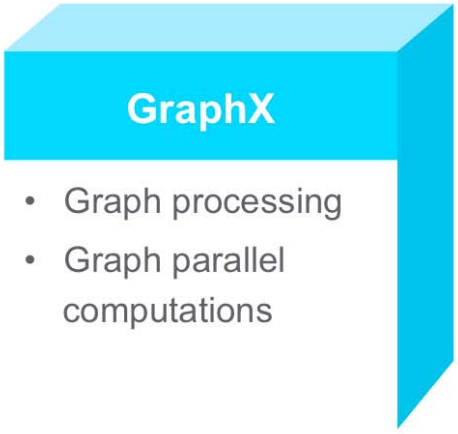
아파치 스파크 통합 스택 – 스파크 MLib (2)

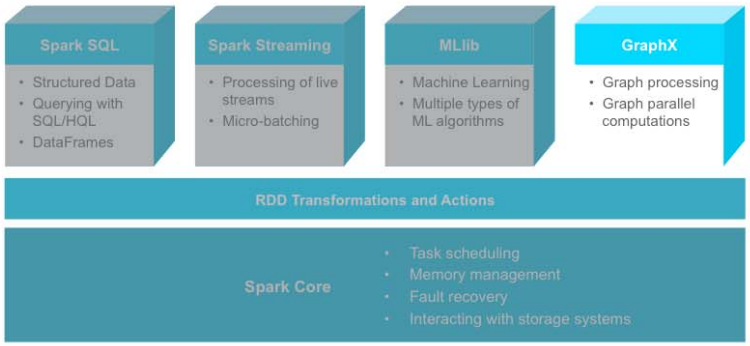
- 두 종류의 MLib 패키지
  - spark.mllib
    RDD 상에 적용되는 오리지널 API
  - spark.ml
    데이터프레임 상에 적용되는 상위 레벨 API로 스파크 2.0부터 기본 기계 학습 API가 됨



아파치 스파크 통합 스택 – 스파크 GraphX (1)

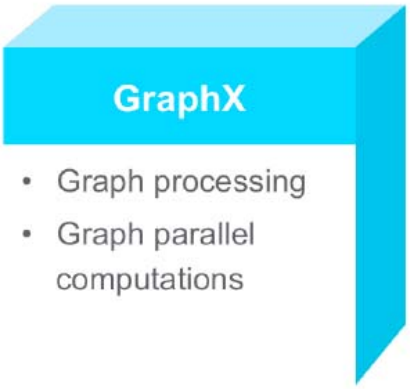
- GraphX는 그래프 처리와 그래프-병렬 계산을 수행하는 라이브러리

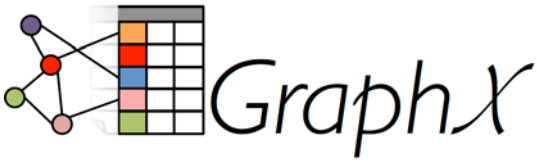




아파치 스파크 통합 스택 – 스파크 GraphX (2)

- GraphX는 스파크 RDD를 확장한 그래프 추상화
  - 최적화된 버전의 Pregel을 포함하여 그래프 계산을 많은 연산을 제공
    - Pregel은 Google에서 개발한 그래프 처리 아키텍처
  - 그래프 알고리즘과 그래프 분석을 위한 빌더를 제공
  - 데이터 이동이나 별도의 작업 없이 그래프와 컬렉션을 연동하여 작업

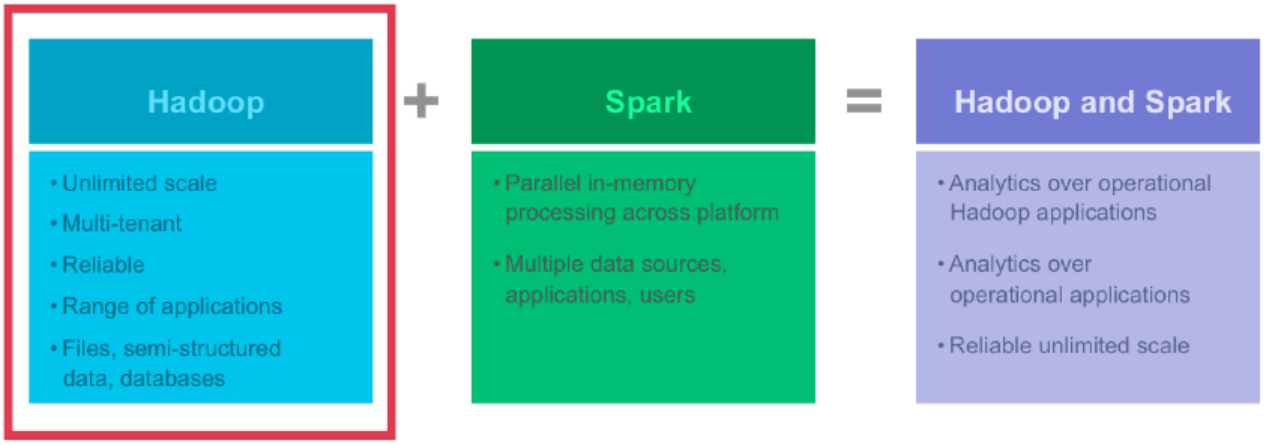




2. 하둡 에코 시스템과 스파크

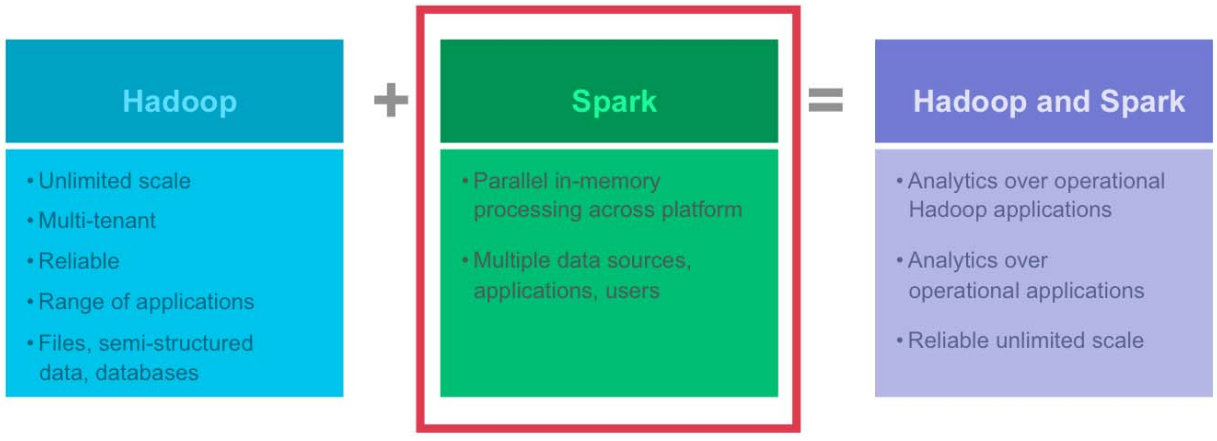
아파치 하둡 (Apache Hadoop)

- 하둡(Hadoop)은 무한 확장이 가능하고, 멀티-테넌트(multi-tenant, 다수의 사용자 서비스)와 신뢰성을 가면서 파일, 데이티베이스, 반-구조화된 데이터 등의 다양한 응용에 적용



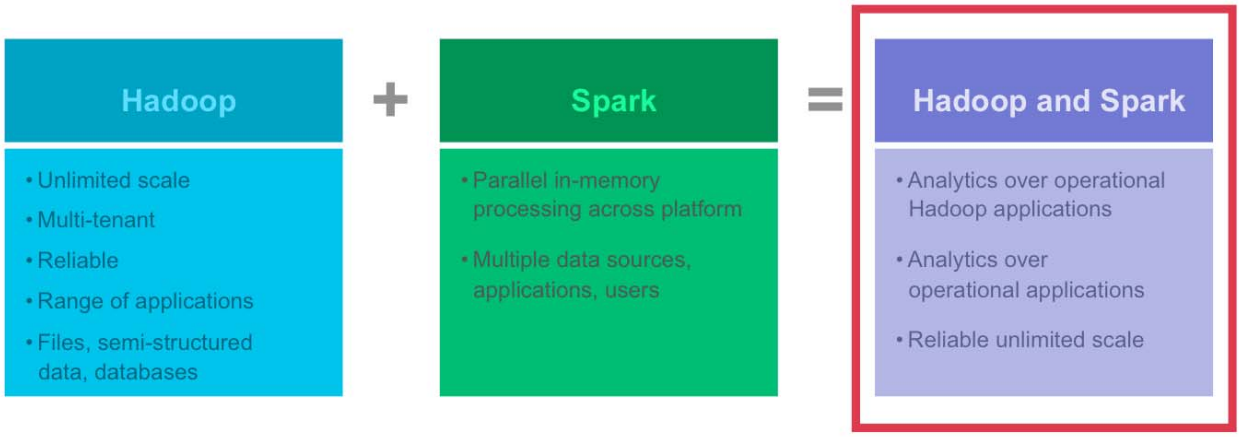
아파치 스파크 (Apache Spark)

- 스파크(Spark)는 하둡 플랫폼 상에서 메모리 적재 하에 병렬 처리를 제공하면서 다양한 데이터 소스, 응용, 사용자들을 수용

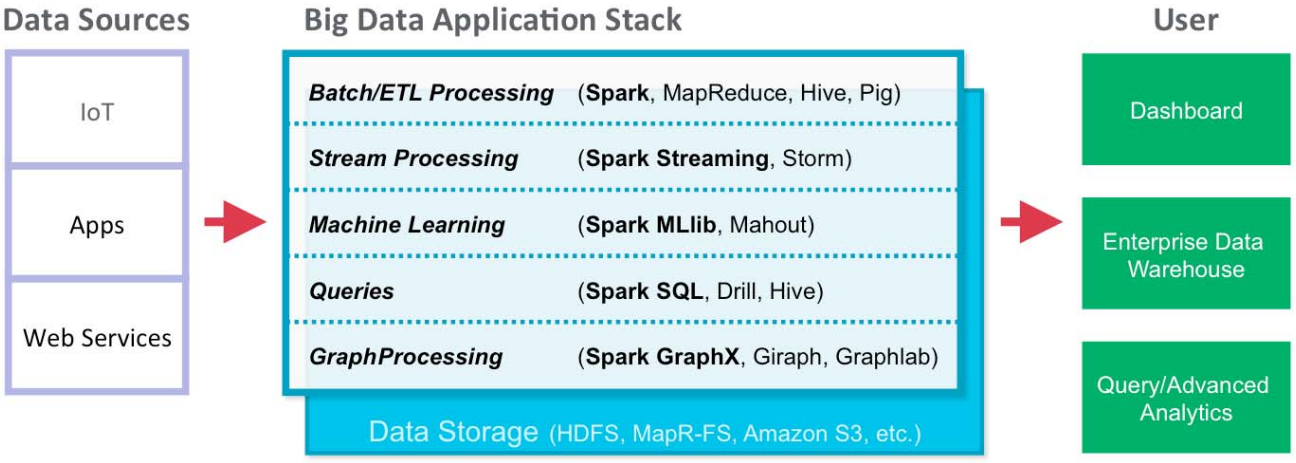


아파치 하둡과 스파크

- 스파크와 하둡을 결합하면 두 플랫폼의 장점을 활용
  - 신뢰성, 확장성, 빠른 메모리 상의 처리
  - 다양한 운영 응용 분석(operatinal application analytics)
    - 데이터를 수집, 분석, 보고 응용

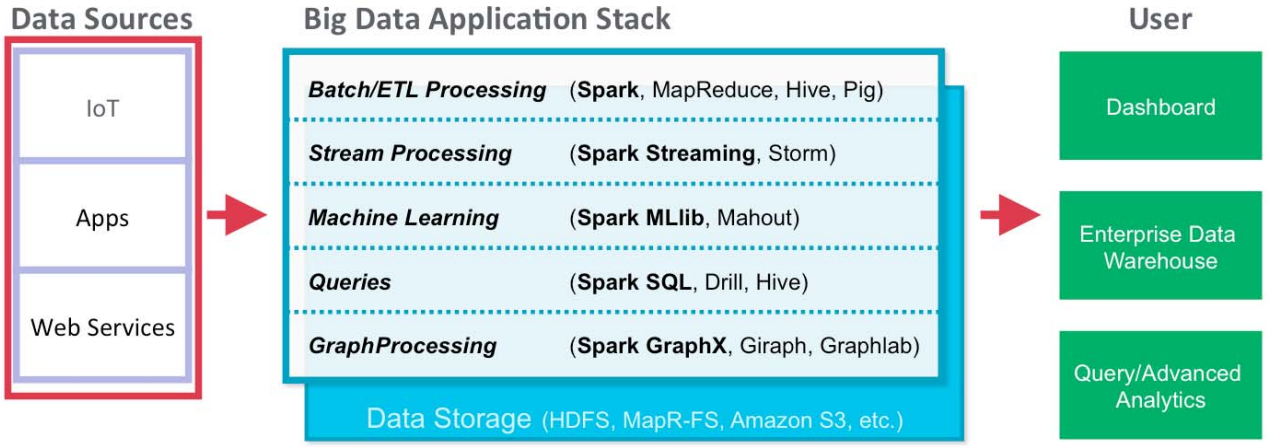


아파치 스파크와 빅데이터



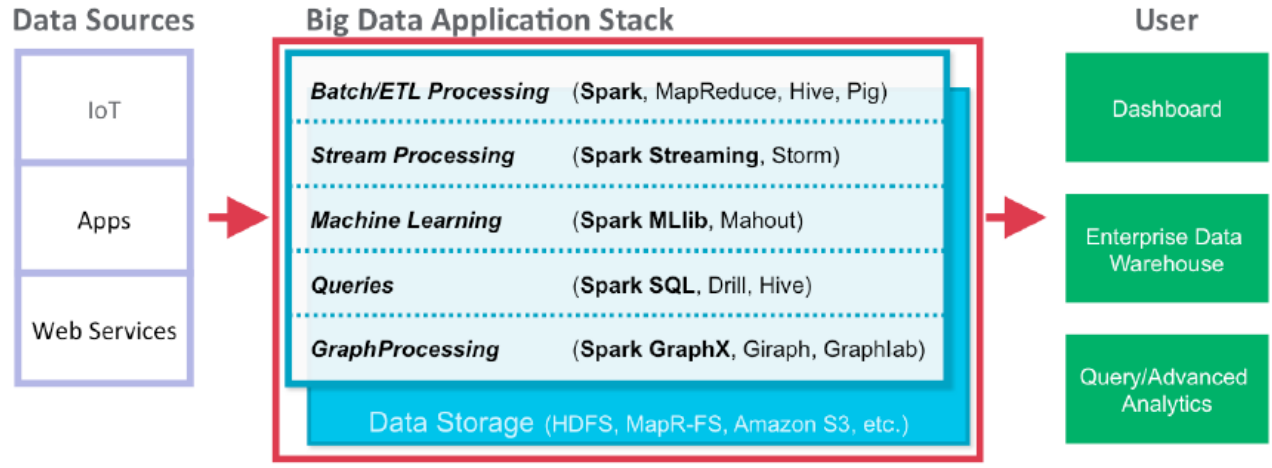
아파치 스파크와 빅데이터 – 데이터 수집

- 다양한 소스로 데이터 수집



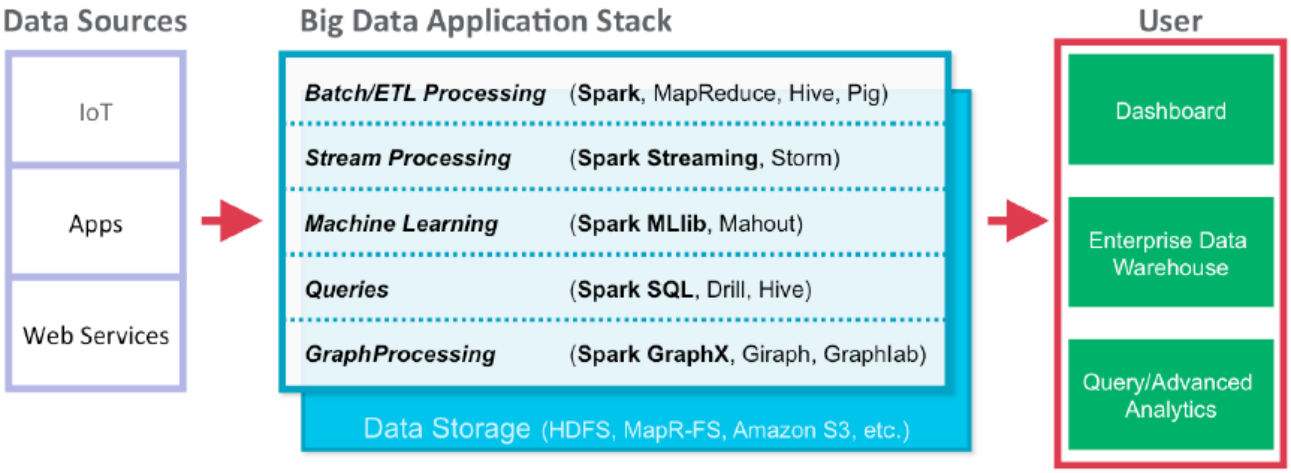
아파치 스파크와 빅데이터 – 데이터 처리

- 하둡은 다양한 툴과 프로그래밍 언어를 사용하여 작업흐름(workflow)를 결합하지만, 스파크를 사용하면 여러 작업 흐름의 통합 적용이 가능



아파치 스파크와 빅데이터 – 출력 리포트

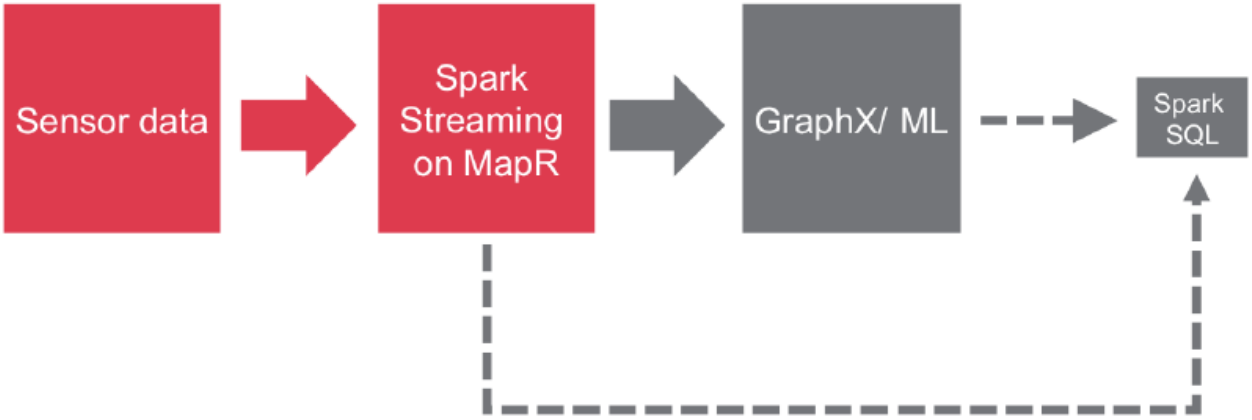
- 실시간 대시보드나 질의와 고급 분석을 위한 시스템을 생성하여 출력 리포트



3. 스파크 데이터 파이프라인 활용 사례 (Use Case)

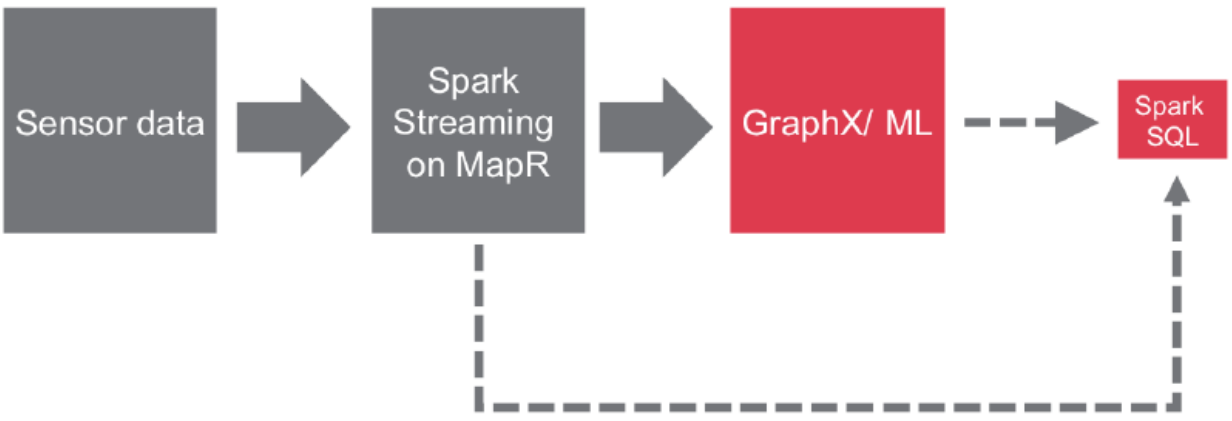
활용 사례 – 보안 서비스 제공자 (1)

- 스파크 스트리밍으로 센서 데이터를 수집하면서 알려진 보안 공격에 대해 조사



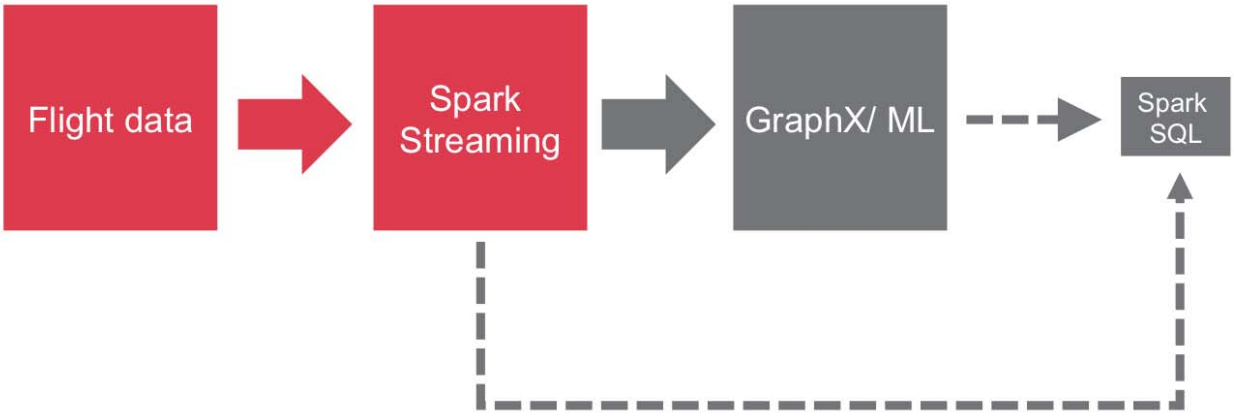
활용 사례 – 보안 서비스 제공자 (2)

- 수집된 데이터는 예측을 위해 그래프 처리와 기계 학습
  - SparkSQL을 사용하여 그래프 알고리즘의 결과, 예측 모델, 요약/집계 등과 같은 부가적인 질의를 수행



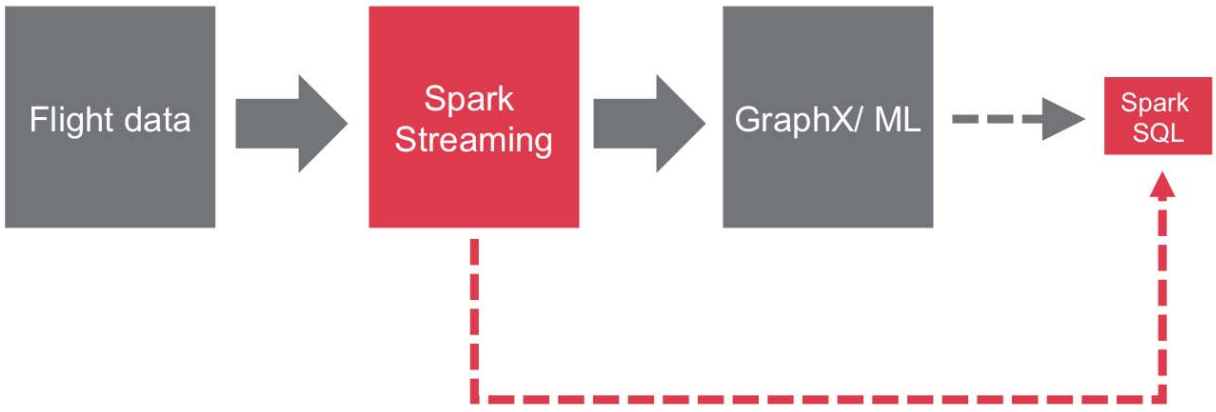
활용 사례 – 항공기 운항 최적화 (1)

- 연속적으로 발생하는 항공기 운항 정보가 스트리밍으로 수집



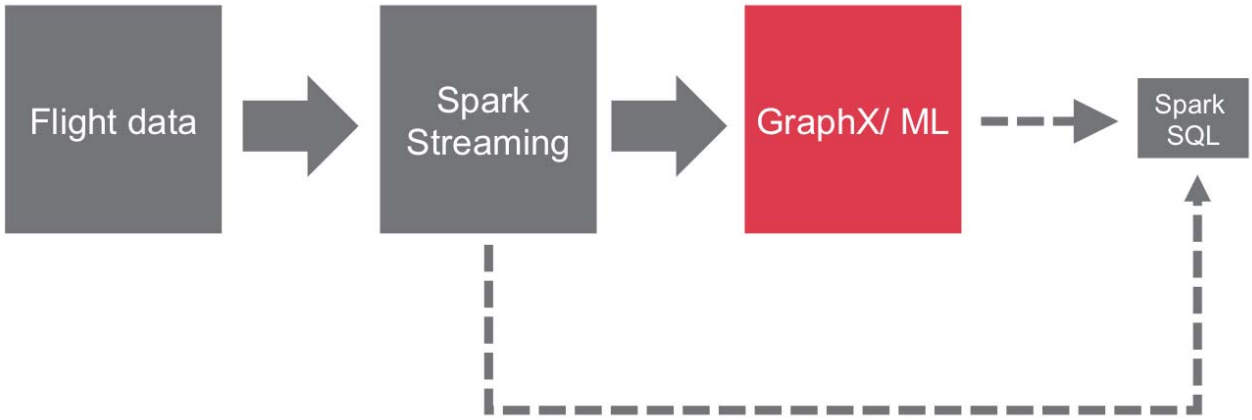
활용 사례 – 항공기 운항 최적화 (2)

- SparkSQL을 사용하여 스트리밍 데이터의 분석



활용 사례 – 항공기 운항 최적화 (3)

- GraphX를 사용하여 공항 및 비행 경로 분석
- 기계학습의 분류(classification) 또는 결정트리(decision tree) 알고리즘을 적용하여 항공기 지연을 예측



과제

- 국내외 아파치 스파크 활용 사례 조사
- 텀 프로젝트의 작업 흐름(workflow)을 도시하고, 스파크 데이터 파이프라인 적용 계획 작성

참고 자료

- MapR Academy DEV 362 Create Data Pipelines With Apache Stack
  - https://mapr.com/training/on-demand/dev-362/
  - Lesson 7: Create Data Pipelines With Apache Spark
- Spark Programming Guide
  - https://spark.apache.org/docs/latest/sql-programming-guide.html
  - https://spark.apache.org/docs/latest/streaming-programming-guide.html
  - https://spark.apache.org/docs/latest/ml-guide.html
  - https://spark.apache.org/docs/latest/graphx-programming-guide.html