8.1 분석환경 확장 개요

분산환경 분석 도구

VS.

외부환경 분석 도구

8.1 분석환경 확장 개요

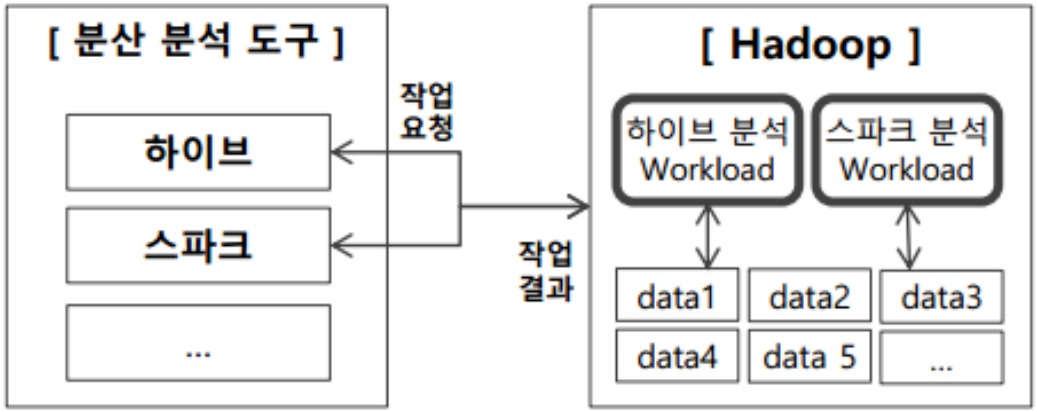


그림 8.1.1 분산 환경의 분석 도구

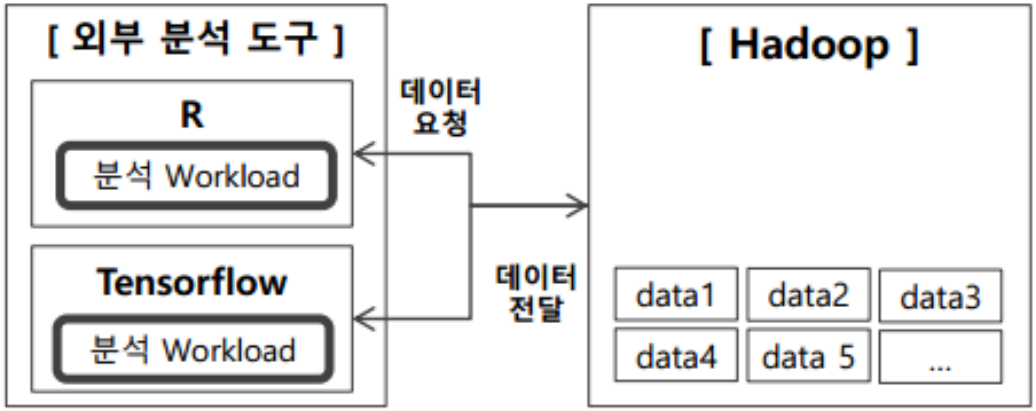
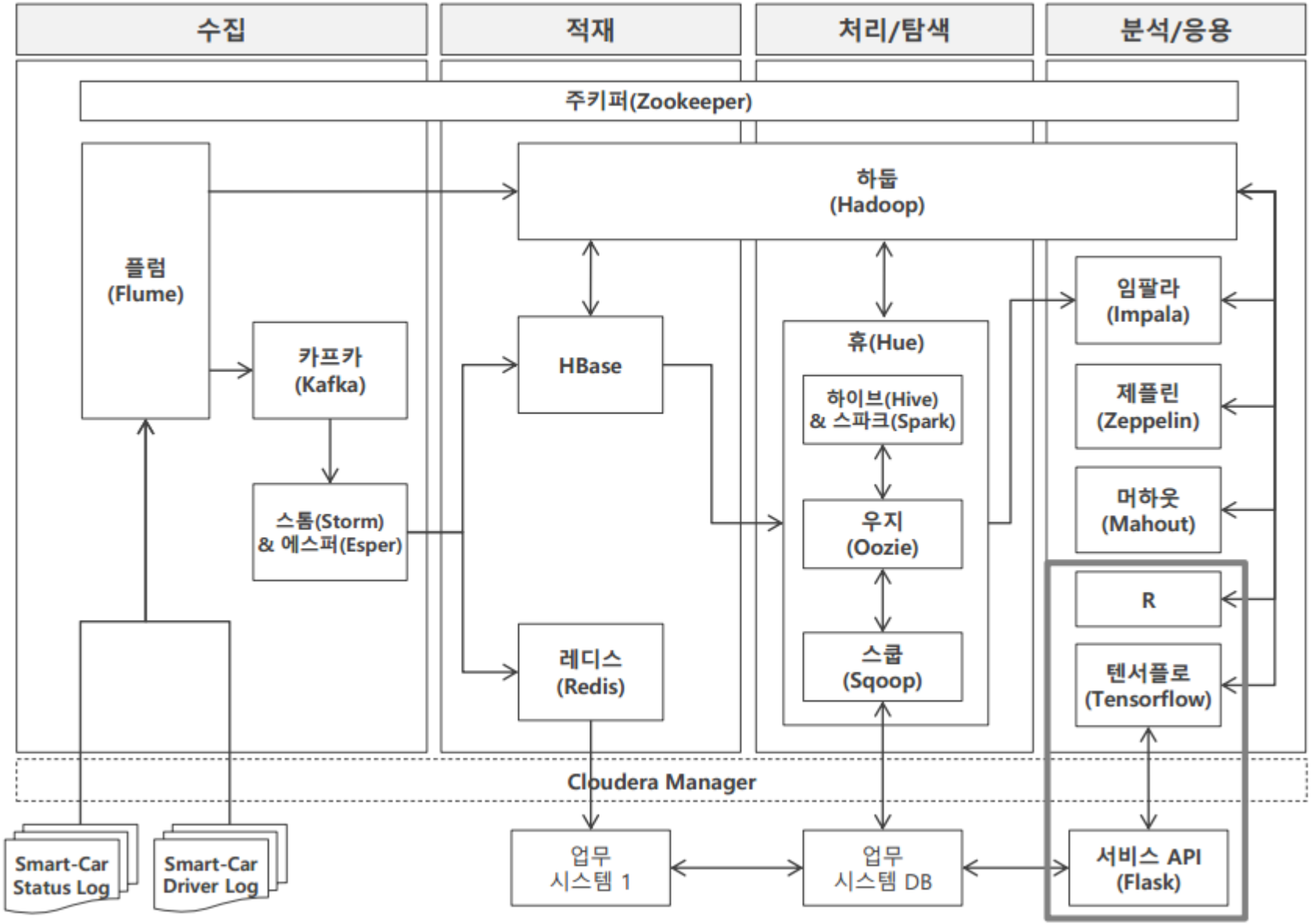


그림 8.1.2 외부 독립 분석 도구

8.1 분석환경 확장 개요


수집
적재
처리/탐색
분석/응용
주키퍼(Zookeeper)
하둡
(Hadoop)
플럼
(Flume)
카프카
(Kafka)
스톰(Storm)
& 에스퍼(Esper)
HBase
레디스
(Redis)
휴(Hue)
하이브(Hive)
& 스파크(Spark)
우지
(Oozie)
스쿱
(Sqoop)
임팔라
(Impala)
제플린
(Zeppelin)
머하웃
(Mahout)
R
텐서플로
(Tensorflow)
Cloudera Manager
Smart-Car
Status Log
Smart-Car
Driver Log
업무
시스템 1
업무
시스템 DB
서비스 API
(Flask)

Tip _ 분산 환경을 지원하는 분석 도구

외부 독립형 분석 도구의 단점은 대용량 데이터를 이용한 분석 및 학습에 과도한 시간이 소요된다는 것이다. 최근에는 GPU 파워를 이용해 빠른 연산은 가능해졌으나 대용량 데이터 로드에 발생하는 I/O 오버헤드와 자원(CPU/메모리) 증설 한계는 피할 수 없다. 이 같은 이유로 최근 들어 독립형 방식의 분석 도구도 대규모 분산 환경(스케일아웃 방식)을 지원하기 위한 다양한 시도를 하고 있다. 텐서플로는 지난 2016년 4월경 텐서플로 분산 버전을 공식적으로 출시했고, R은 오래전부터 RHive를 통해 하둡 분산 환경에서 분석 작업을 수행할 수 있도록 지원해 왔다.

하지만 분산 분석 환경의 경우 요청 작업을 여러 대의 서버에 나누어 실행해야 하고, 그 결과를 다시 하나로 모으는 맵/리듀스 메커니즘(그림 4.4 참고)이 기본적으로 요구된다. 이러한 이유로 분산 환경에서 기존 분석 알고리즘이 작동하기 위해서는 프로그램 수정과 복잡도가 증가하며 환경에 대한 호환성 문제를 고려해야 한다.

사용하려는 분석 도구와 라이브러리의 분산 컴퓨팅 지원 여부는 관련 소프트웨어의 공식 사이트에서 확인한 후 사용하기 바란다. 참고로 대표적 분산 환경인 아파치 스파크(http://spark.apache.org)에서 지원하는 주요 머신러닝 알고리즘(MLlib)은 다음 표와 같다.

Classification(분류)	Logistic regression
	Decision tree classifier
	Random forest classifier
	Gradient–boosted tree classifier
	Multilayer perceptron classifier
	Linear Support Vector Machine
	One–vs–Rest classifier (a.k.a. One–vs–All)
	Naive Bayes
Clustering(군집)	K–means
	Latent Dirichlet allocation (LDA)
	Bisecting k–means
	Gaussian Mixture Model (GMM)
Regression(회귀)	Linear regression
	Generalized linear regression
	Available families
	Decision tree regression
	Random forest regression
	Gradient–boosted tree regression
	Survival regression
	Isotonic regression

스파크 MLib에서 지원하는 알고리즘(2019년 8월 기준)

8.1 분석환경 확장 개요

- R을 이용한 회귀분석 – 스마트카 배기량에 따른 운전자의 연소득 예측
- 텐서플로 이용한 신경망 분석 – 주행 중 스마트카의 사고 위험 징후 판별
- 예측 모델 API 구성 – 딥러닝 모델을 웹서비스 REST API로 구성