# [S1-02] 1-7. 하둡 에코시스템(Hadoop EcoSystem)

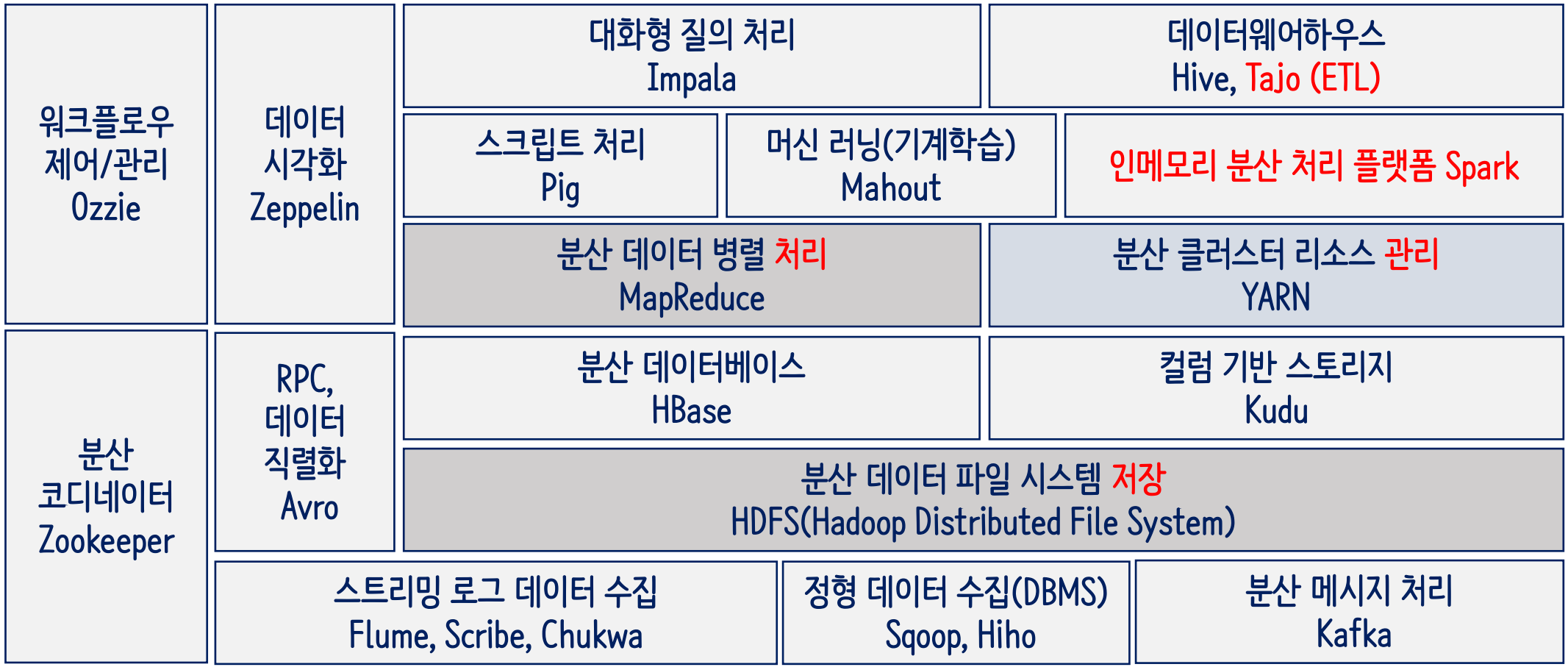
| | | | |
|---|---|---|---|
| Avro | ▪ RPC(Remote Procedure Call)과 데이터 직렬화를 지원하는 프레임워크 | | |
| 반정형/비정형 데이터 수집 | ▪ Flume, Scribe, Chukwa (로그 데이터)<br>▪ Kafka (분산 메시지 처리, 비정형) | 정형 데이터 수집 | ▪ Sqoop, Hiho |
| Yarn | ▪ 리소스(CPU, 메모리, 디스크 등) 할당 및 작업 스케쥴링 프레임워크 (리소스 관리) | | |
| Zeppelin | ▪ 분석 결과를 표 그래프로 제공하는 웹 기반 분석 도구 (시각화) | | |

- RPC : 원격 프로시저 호출, 현재 실행 중인 프로세스의 주소공간 내부가 아닌, 외부의 프로세스 또는 원격지의 프로세스와 상호작용 하기 위한 기능
- 데이터 직렬화 : 메모리에서 현재 실행하고 있는 프로세스의 특정 데이터를 서버로 통신하거나 디스크에 저장할 때 사용되는 기술 (메모리 → 데이터 직렬화 → 통신/디스크 저장)

- Sqoop : 관계형 데이터베이스 시스템에서 하둡 파일 시스템으로 데이터를 수집한다

- 정형 데이터 수집 Sqoop, Hiho (o 가 들어 있고)
- 반정형/비정형 데이터 수집은 Flume, Scribe, Chukwa, Kafka (o가 들어있지 않음)

# [S1-02] 1-8. 하둡 에코시스템(Hadoop Ecosystem)




- Flume, Scribe, Chukwa, Kafka : 반정형/비정형 데이터 수집
- **스트리밍 데이터** : 수천 개의 데이터 원본에서 연속적으로 생성되는 데이터(모바일/웹 애플리케이션을 사용하는 고객이 생성하는 로그 파일, 전자 상거래 구매, 주식 거래 등)
- **로그 데이터** : IT인프라에서 발생하는 모든 상황의 데이터로 사용자의 사용/행동 기반 데이터로 시간정보를 포함하기 때문에 시계열 데이터 개념에 포함됨

# [S3-03] 1-7 모수/비모수적 추론 방법

| 집단수 | 관계 | 비모수 연속형 | 비모수 명목척도 | 비모수 서열척도 | 모수 (정규분포) |
|---|---|---|---|---|---|
| 1 | | Kolmogorov-Smirnov test<br>(1 표본, 적합도 검정) | Run test<br>$\chi^2$ 적합도 검정 | Wilcoxon Signed Rank test<br>(1 표본) | One sample T test |
| 2 | 독립 | Kolmogorov-Smirnov test<br>(2 표본, 분포의 동질성 검정) | Crosstab<br>Fisher's test<br>$\chi^2$ 동질성, 독립성 검정 | Mann-Whitney U test<br>Wilcoxon Rank Sum test | Two sample T test |
| | 대응 | | McNemar test | Wilcoxon Signed Rank test(2 표본)<br>Sign Test | Paired T test |
| K<br>3이상 | 독립 | | $\chi^2$ 동질성, 독립성 검정 | Kruskal-Wallis test (중위수) | ANOVA test |
| | 대응 | | Cochran Q test | Friedman Rank-Sum test | |

적합도 검정(Goodness-of-Fit Test) : 데이터의 표본이 특정 분포를 가진 모집단에서 추출되었는지 여부를 검정하는 것
- 귀무가설 : 데이터가 특정 분포를 따른다, 대립가설 : 데이터가 특정 분포를 따르지 않는다

| | |
|---|---|
| Kolmogorov-Smirnov Test | ▪ 연속형 확률분포의 적합성 검정에 사용, 정규분포, 균일 분포, 로그정규분포, 지수분포 등 |
| Anderson Darling Test | ▪ 연속형 확률분포에 사용, K-S Test를 수정한 것으로 K-S보다 강력, 분포의 꼬리에 훨씬 더 민감 |
| Chi-Square Test | ▪ 이산적 확률분포의 적합성 검정에 사용, 이항 분포, 포아송 분포 등 |