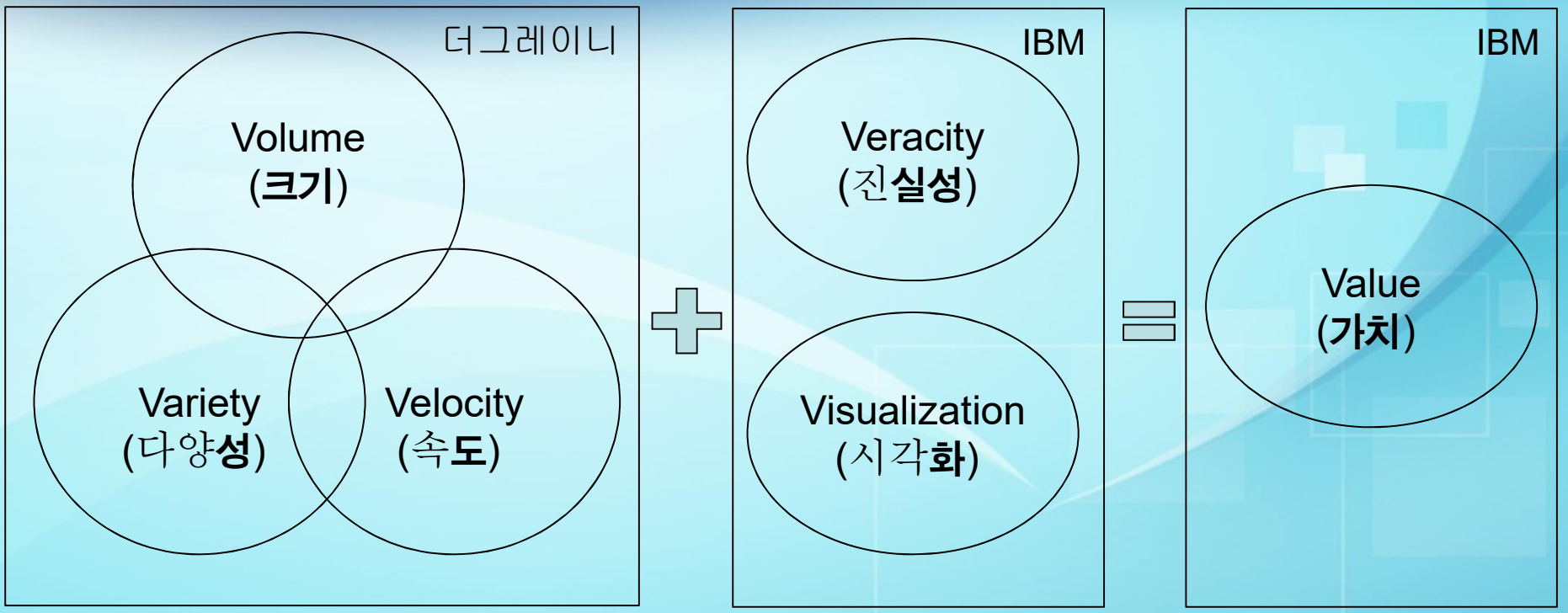
빅 데이터 처리 기술
(데이터수집, 적재, 처리, 분석)

강의 : 신 인 호 교수

1. 빅 데이터 실습을 위한 선수과목

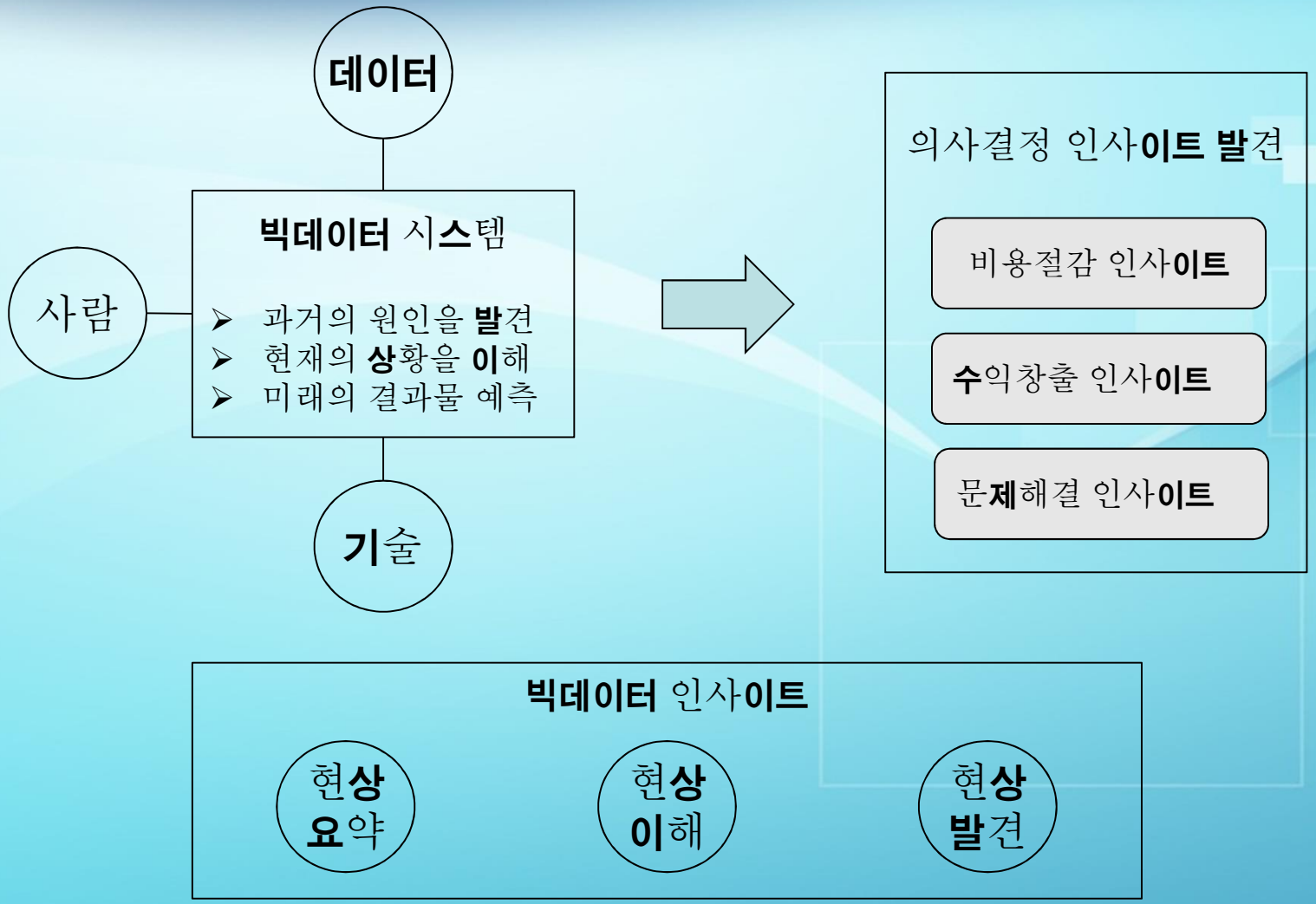
Unix OS	Unix 명령어, VI 편집기, Network 설정 등 기본적인 Unix명령을 사용할 수 있어야 한다.
SQL	DML(Create, Drop, Truncate), DDL(Select, Update, Delete, Insert)등 SQL문을 활용할 수 있어야 한다.
JAVA Programming	Java프로그램 및 소스코드를 읽고 해독할 수 있어야 한다.
DataBase ,Data	Oracle, Hbase, PostgreSQL데이터를 조회하고 생성하는것을 이해하고 구현할 수 있어야 한다. 정형, 비정형 데이터를 구분하고, 활용할 수 있어야 한다.
프로그램 개발에 관련 Tool 사용	Eclipse, FileZilla, PuTTY, **크롬**

2. 빅 데이터 정의 (6V)



크기(Volume)	**방대한 양의 데이터(테라, 페타 바이트이상)**
다양성(Variety)	정형, 비정형데이터
속도(Velocity)	실시간으로 생산되며, 빠른 속도로 데이터를 처리/분석
진실성(Veracity)	주요의사 결정을 위해 데이터의 품질과 신뢰성을 확보
시각화(Visualization)	복잡한 대규모 데이터를 시각적으로 표현
가치(Value)	비즈니스 효익을 실현하기 위해 궁극적인 가치를 창출

3. 빅 데이터 목적


데이터
사람
빅데이터 시스템
과거의 원인을 발견
현재의 상황을 이해
미래의 결과물 예측
기술
의사결정 인사이트 발견
비용절감 인사이트
수익창출 인사이트
문제해결 인사이트
빅데이터 인사이트
현상 요약
현상 이해
현상 발견

4. 빅 데이터 구현기술

단계	역할	활용기술	
수집	• 내.외부**데이터** 연동 및 통합	Crawing, FTP, Open Api,, Log Aggregation, Rss, DB Aggregation, Strieaming	**전**처**리**
적재	• 대용량/**실**시간 **데이터** 처**리** • 분**산 파일**시**스**템 저장	Distributed File, NO-SQL, Memory Cached, Message Queue	
처**리**	• **데이터** 선택, 변**환**, 통합, 축**소** • **데이터** 워**크플**로 및 **자**동**화**	Structured Processing, UnStructured Processing Workflow, Scheduler	**후**처**리**
탐색	• 대**화**형 **데이터** 질의 • 탐색적 Ad-Hoc 분석	SQL Like, Distributed Programming Exploration Visualization	
분석	• **빅데이터** 마**트구성** • 통계분석, **고**급분석	Data Mining, Machine Learning Analysis Visualization	
응용	• 보**고서** 및 시각**화** • 분석정보**제**공	Data Export / Import, Reporting Business Visualization	활용

5. 빅 데이터 구축단계(1/4)

6V	수집기술	중요도
Volume	대용량 데이터 수집 대규모 메시지 수집	상
Variety	정형/반정형/비정형 데이터 수집 (Log, Rss, Xml, 파일, DB, HTML, 음성, 사진, 동영상..)	상
Velocity	실시간 스트림 데이터 수집	상
Veracity	N/A	하
Visualization	N/A	하
Value	N/A	하

관련 솔루션	플롬(Flume), 스톰(Storm), 에스퍼(Esper)

5. 빅 데이터 구축단계(2/4)

6V	수집기술	중요도
Volume	대용량 데이터 수집 대규모 메시지 수집	상
Variety	정형/반정형/비정형 데이터 수집 (Log, Rss, Xml, 파일, DB, HTML, 음성, 사진, 동영상..)	중
Velocity	실시간 스트림 데이터 수집	상
Veracity	데이터의 품질과 신뢰성을 확보해 적재	상
Visualization	N/A	하
Value	N/A	하

관련 솔루션	분산 파일시스템 하둡, No SQL저장장치 Hbase, 분산 캐시저장소 레디스(Redis), 메시지 저장소 카프카(Kafka)

5. 빅 데이터 구축단계(3/4)

6V	수집기술	중요도
Volume	대용량 데이터 수집 대규모 메시지 수집	상
Variety	N/A	하
Velocity	N/A	하
Veracity	데이터의 품질과 신뢰성을 확보하기 위한 후처리 및 탐색	상
Visualization	후 처리된 데이터 셋을 시각화해서 탐색	상
Value	N/A	중

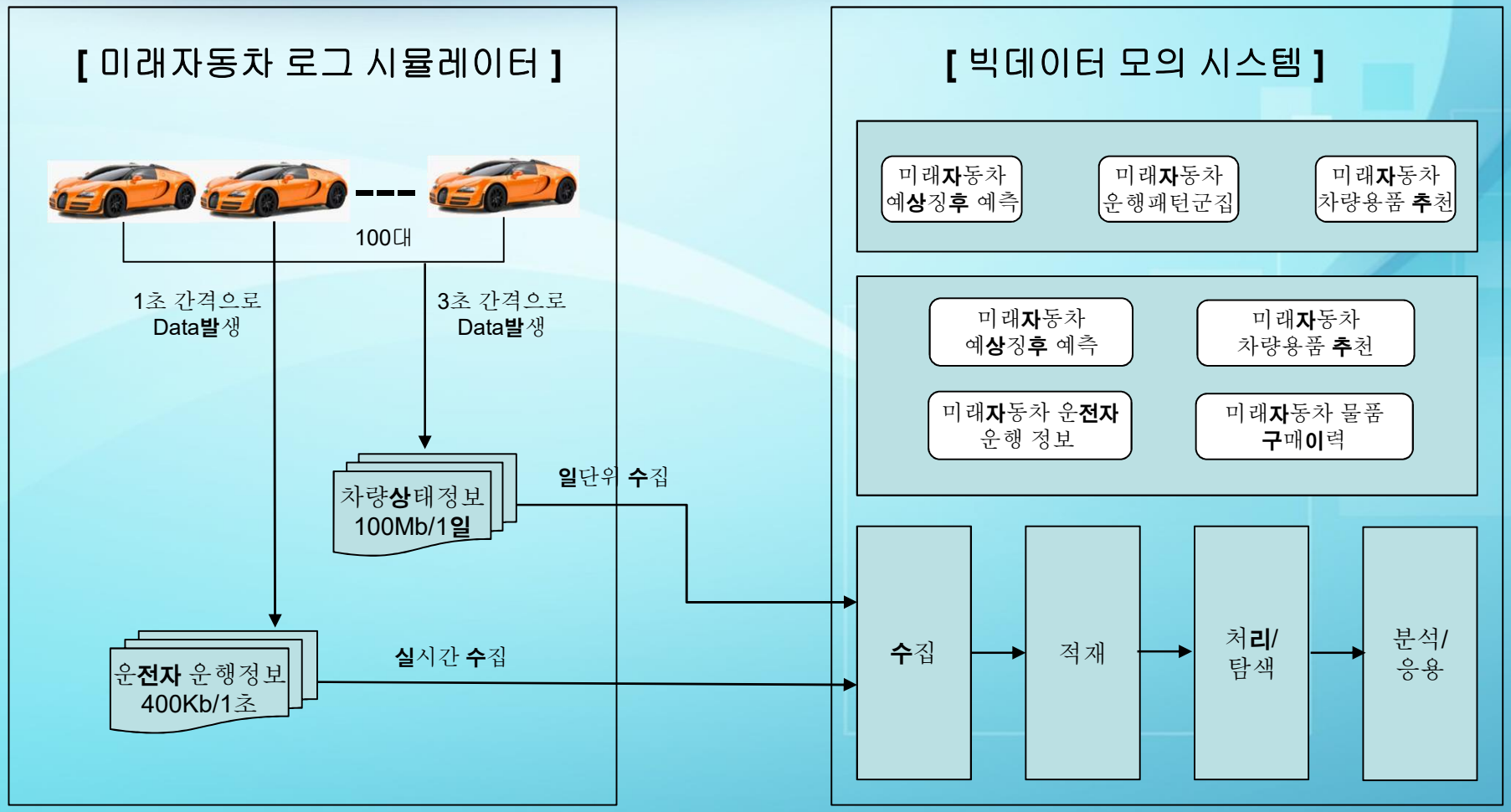
관련 솔루션	**휴**(HUe), **하이브**(Hive), **스파크**(Spark) SQL, **후처리** 작업을 **자동화하**는 워**크플**로에 **우지**(Oozie)

5. 빅 데이터 구축단계(4/4)

6V	수집기술	중요도
Volume	대용량 데이터 / 메시지 분석	상
Variety	정형/반정형/비정형 데이터 분석	상
Velocity	인메모리 기반으로 실시간 데이터분석	상
Veracity	신뢰도 높은 분석 결과를 비즈스에 적용	상
Visualization	분석 결과 및 찬출된 가치를 시각화	상
Value	분석된 결과를 비즈니스에 적용해 가치 창출	상

관련 솔루션	임팔라(Impala), 재플린(Zeppelin), 머하웃(Mahout), R, 텐서플로(TensorFlow), 스쿱을통한 DB 데이터 제공

6. 미래자동차 빅 데이터 요구사항



7. 빅 데이터 구축 (미래자동차)

1. 데이터 요구사항(1/2)

요구사항1	차량의 다양한 장치로 부터 발생하는 로그 파일을 수집해서 기능별 상태를 점검한다.
발생 데이터위치	100대의 시범 운행차량
발생 데이터종류	대용량 로그 파일
데이터 발생 주기	3초
수집 주기	24시간(1일 Batch)
수집 규모	1MB/1대 (1일 수집규모: 약 100Mb – 차량 100대 정보)
데이터형태	텍스트(UTF-8), 반정형
데이터분석주기	일/주/월/년
데이터 처리유형	배치처리
데이터구분자	콤마(,)
데이터스키마	발생일시, 차량번호, 앞타이어(좌), 앞타이어(우), 뒤타이어(좌), 뒤타이어(우)상태, 앞라이트(좌), 앞라이트(우), 뒤라이트(좌), 뒤라이트(우), 엔진상태, 브레이크, 배터리, 작업요청일

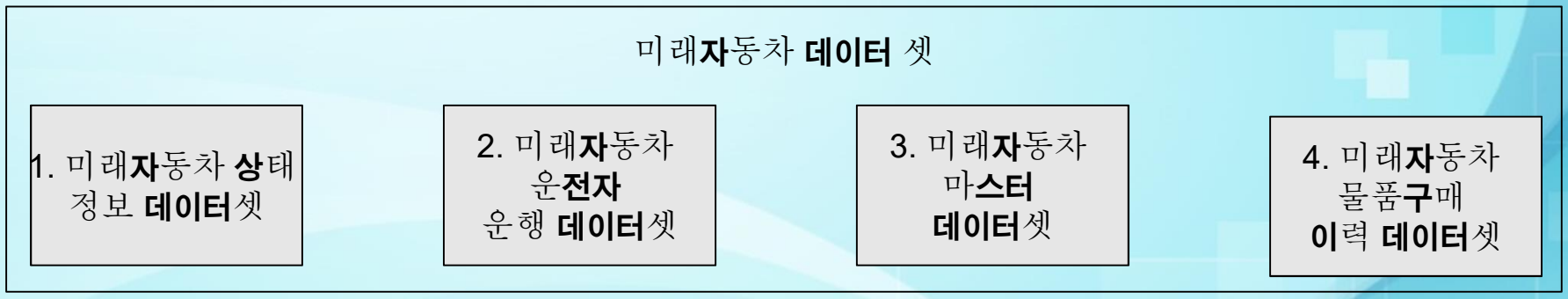
7. 빅 데이터 구축 (미래자동차)

1. 데이터 요구사항(2/2)

요구사항2	운전자의 운행 정보가 담긴 로그를 실시간으로 수집해서 주행패턴을 분석하다.
발생 데이터위치	100대의 시범 운행차량
발생 데이터종류	실시간 로그 파일
데이터 발생 주기	주행 관련 이벤트 발생시
수집 주기	1초
수집 규모	4KB/1대 (초당 수집규모: 약 400Kb – 차량 100대 정보)
데이터형태	텍스트(UTF-8), 반정형
데이터분석주기	실시간
데이터 처리유형	실시간
데이터구분자	콤마(,)
데이터스키마	발생일시, 차량번호, 가속페달, 브레이크페달,운전대 회전각, 방향지시등, 주행속도, 주행지역, 작업요청일

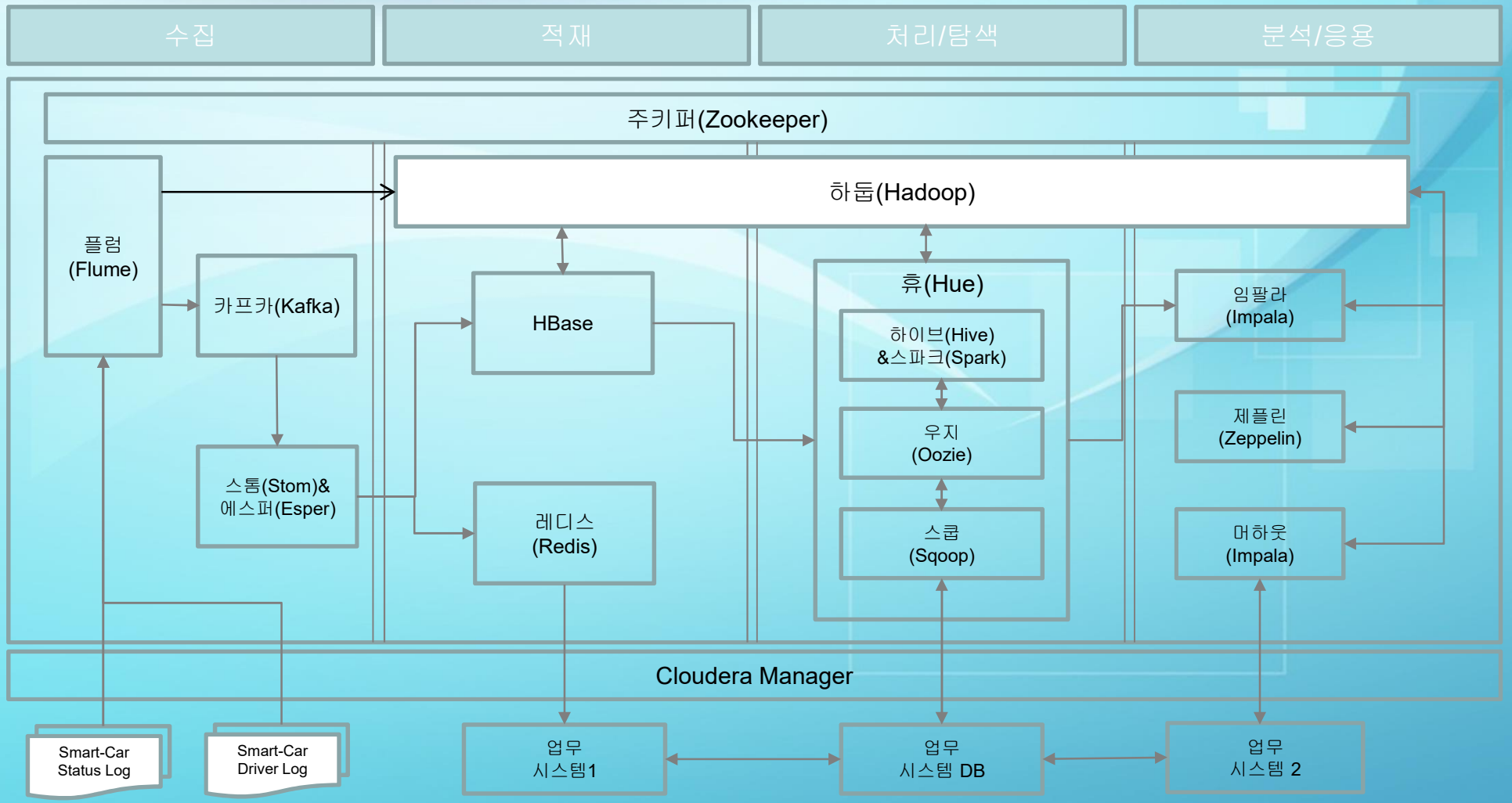
8. 미래자동차 빅 데이터 분석

1. 미래자동차 데이터 셋 분석

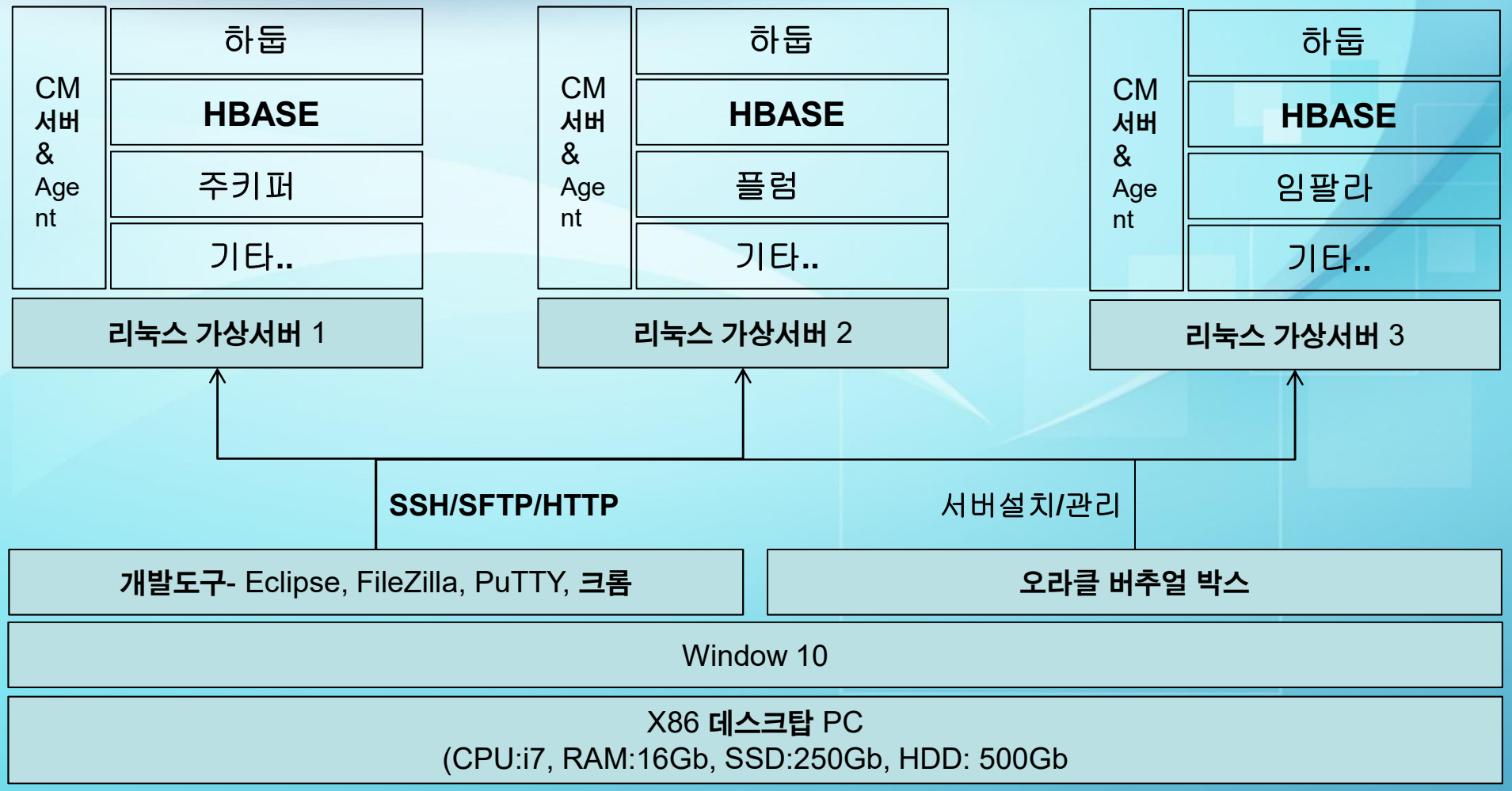


미래자동차 상태정보	미래자동차의 각종 센서로 부터 발생하는 차량의 정보 데이터 셋이다. 요구사항 1과 직접적인 관련이 있으며, 로그 시뮬레이션을 통하여 생성된다.
미래자동차 운전자 운행	미래자동차 운전자의 운전 패턴/운행정보가 담긴 데이터 셋이다. 요구사항 2와 직접적인 관련이 있으며, 로그 시뮬레이션을 통하여 생성된다
미래자동차 마스터	미래자동차 운전자의 프로필 정보가 담긴 데이터 셋이다. 요구사항 1,2과 관련된 분석 데이터 셋을 만들 때 활용한다. 이미 만들어진 샘플 파일을 이용한다.
미래자동차 물품 구매 이력	미래자동차 운전자가 차량내의 스마트 스크린을 통해 쇼핑몰에서 구입한 차량 물품 목록 데이터 셋인다. 요구사항 1,2과 관련되 분석 데이터 셋을 만들 때 활용한다. 이미 만들어진 샘플 파일을 이용한다.

9. 미래자동차 빅 데이터 소프트웨어 구성도(1/2)



9. 미래자동차 빅 데이터 소프트웨어 구성도(2/2)

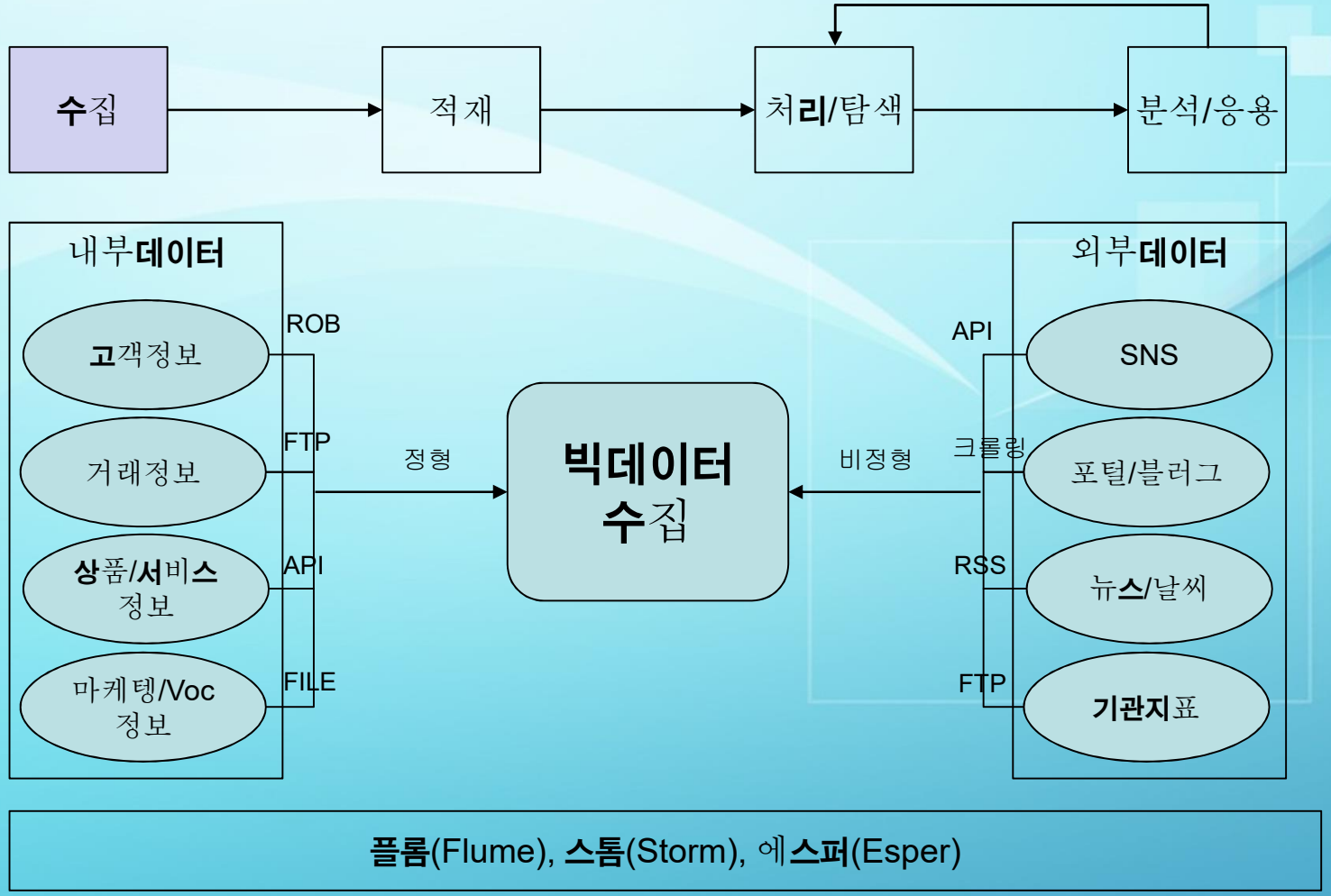


10. 미래자동차 빅 데이터 하드웨어 구성도

Server01	Server02	Server03
▪ Hadoop management Nodes ▪ Hadoop Data Nodes ▪ HBase management ▪ HBase Region Server ▪ PostgreSQL	▪ Hadoop Data Nodes ▪ HBase Region ▪ **플럼** ▪ **하이브 / 스파크** ▪ **휴** ▪ **카프카** ▪ **우지** ▪ **레디스** ▪ **스톰** ▪ **제플린** ▪ **주키퍼**	▪ Hadoop Data Nodes ▪ HBase Region ▪ Cloudera Management ▪ **임팔라** ▪ **스쿱**
리눅스 서버 1 ▪ Server01.hadoop.com ▪ 192.168.56.101 ▪ CentOS 6.10 ▪ **자바** 1.7환경	**리눅스 서버** 2 ▪ Server02.hadoop.com ▪ 192.168.56.102 ▪ CentOS 6.10 ▪ **자바** 1.7환경	**리눅스 서버** 3 ▪ Server03.hadoop.com ▪ 192.168.56.103 ▪ CentOS 6.10 ▪ **자바** 1.7환경
오라클 버추얼 박스		
Window 10		
X86 **데스크탑** PC (CPU:i7, RAM:16Gb, SSD:250Gb, HDD: 500Gb		

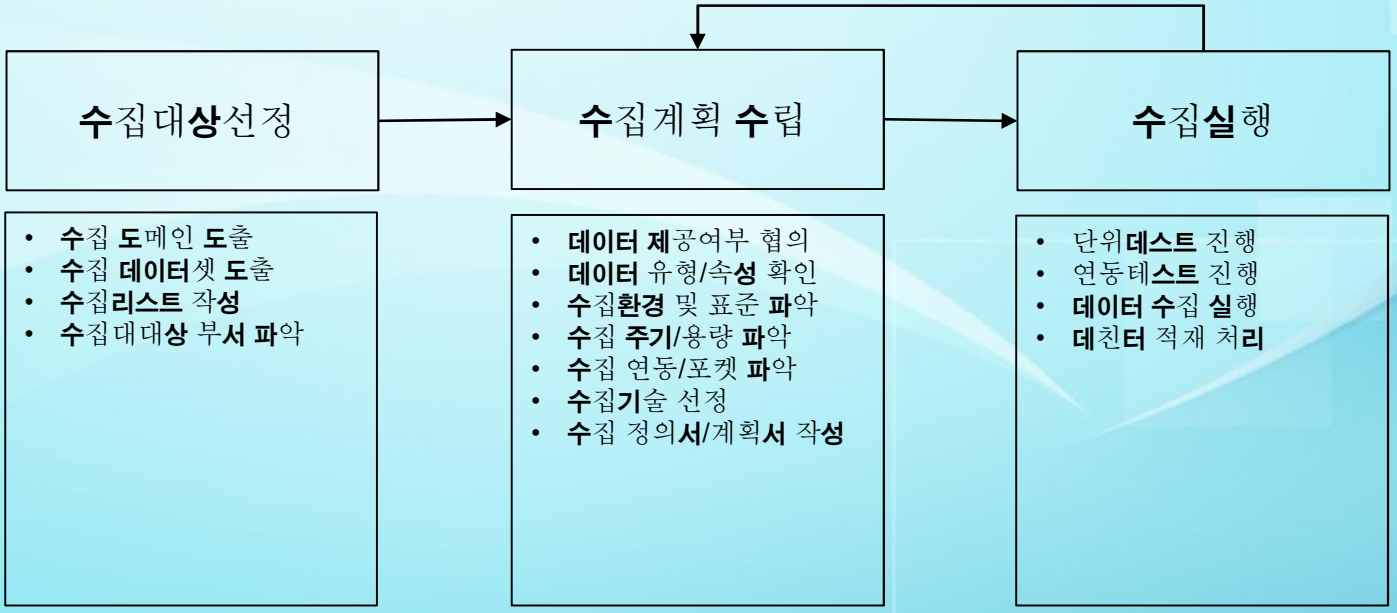
10. 데이터 수집 개요

빅데이터 시스템 구축은 수집에서부터 시작된다.
데이터 수집은 빅데이터 전체공정의 절반이상을 차지한다.



10. 데이터 수집 절차

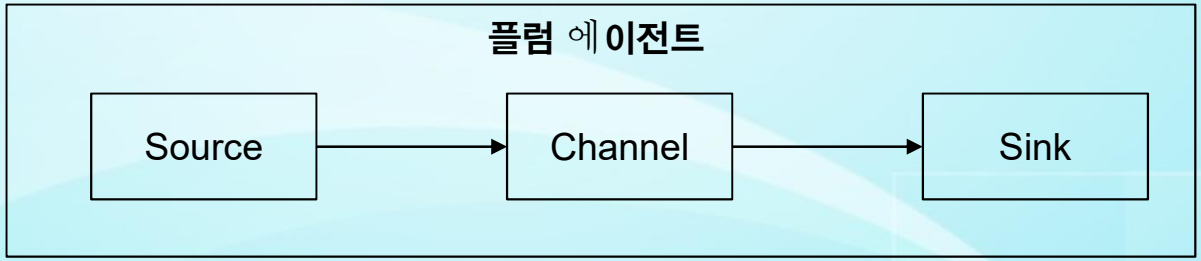
- 빅데이터 수집 실행 단계에서는 업무요건과 환경의 변화로, 이전단계인 수집계획 수립으로 다시 돌아가는 경우가 빈번하게 발생한다.



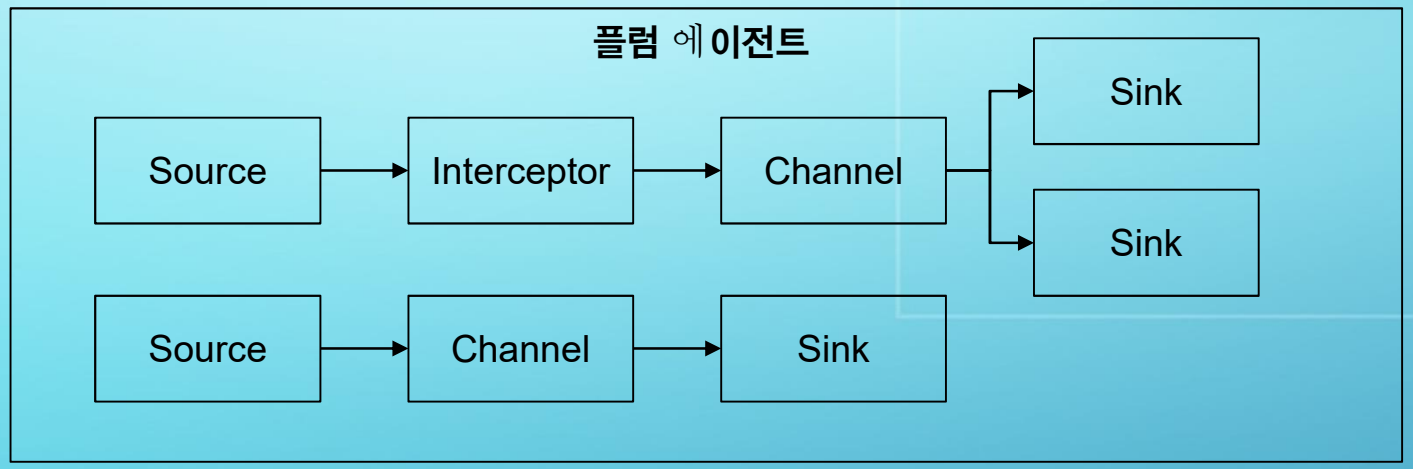
데이터 수집/적재, 데이터 분석 영역 도출 중 어느 것을 먼저 할 것인지는 빅 데이터 도입 주관 부서의 특성에 따라 달라질 수 있다.

10. 데이터 수집(플럼(Flume 1/3)

- 플럼(Flume)- 1/3
- 다양한 수집 요구사항으로 해결하기 위한 기능으로 구성된 소프트웨어
- 구성 유형1

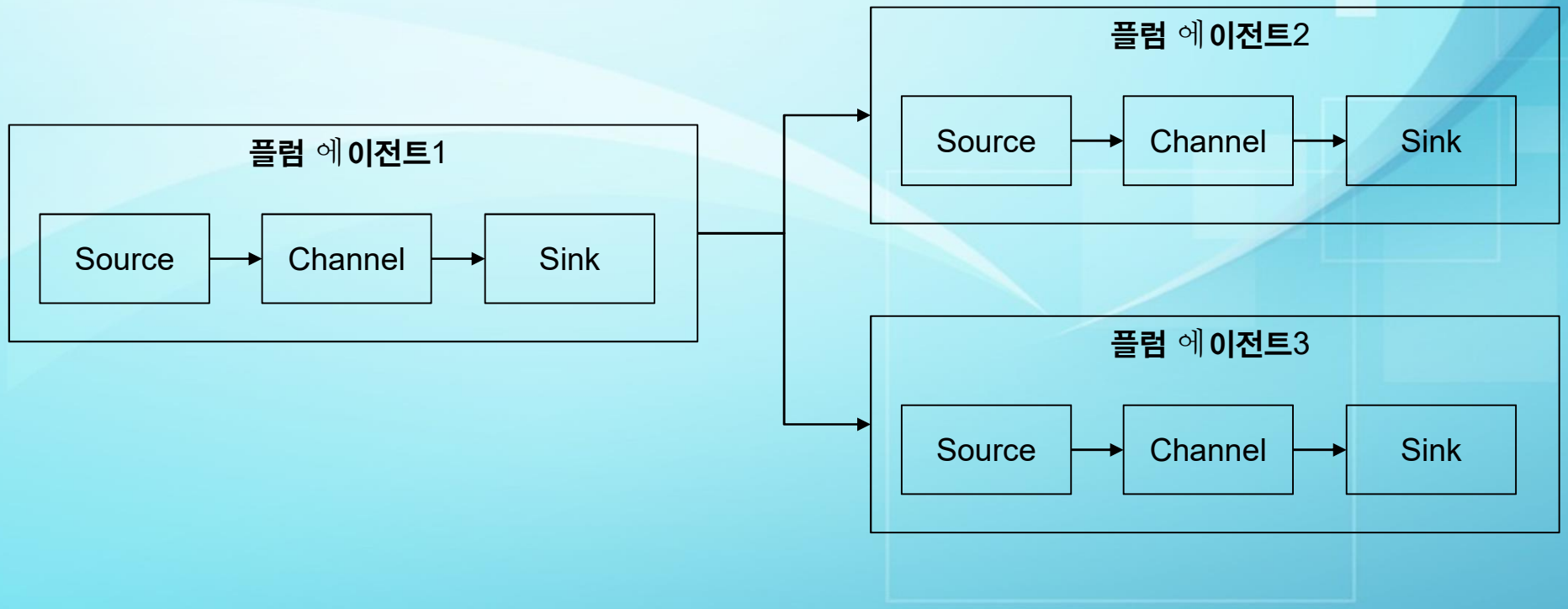


- 구성 유형2



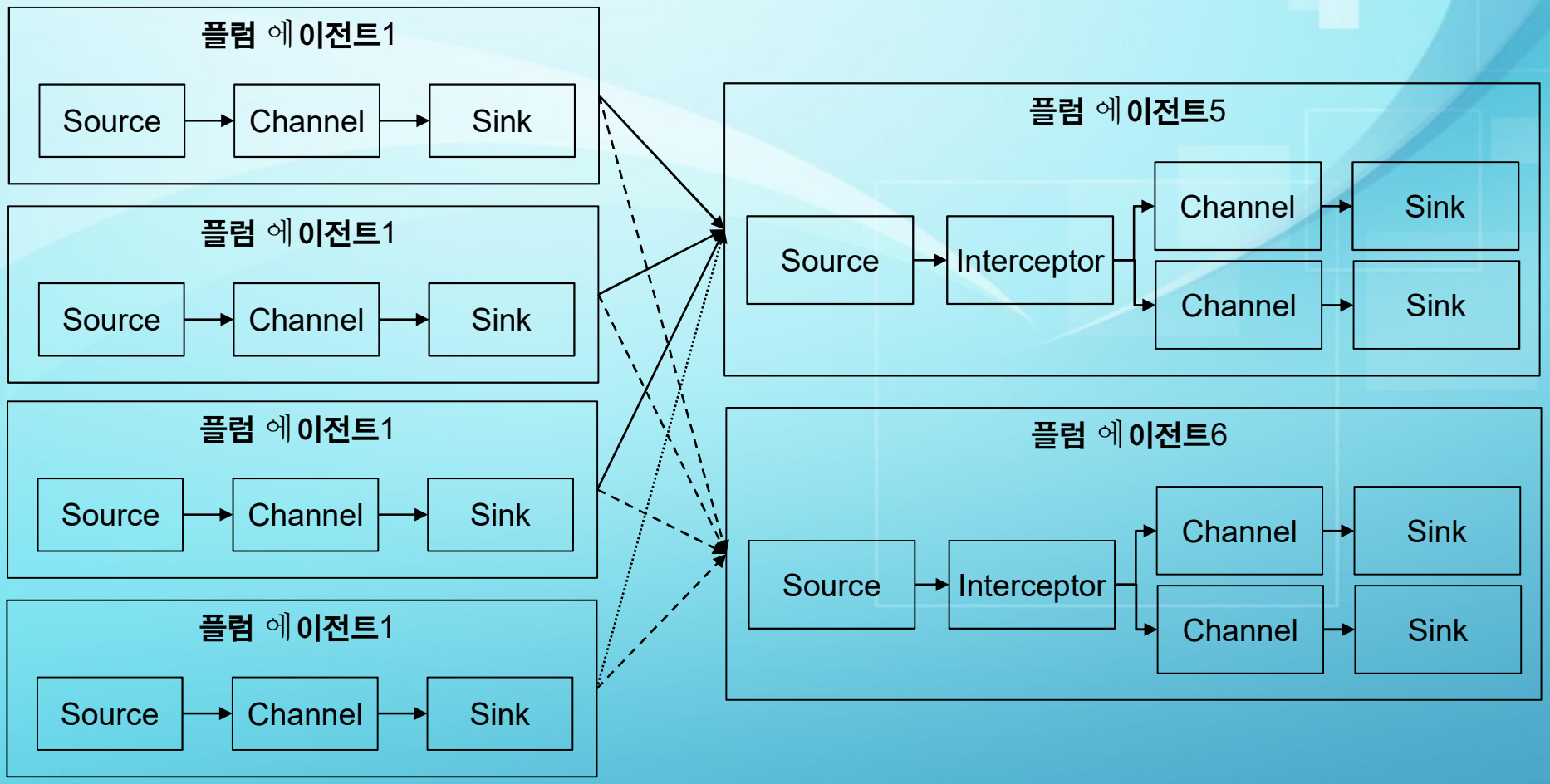
10. 데이터 수집(플럼(Flume 2/3)

- 플럼(Flume)- 2/3
- 구성 유형3



10. 데이터 수집(플럼(Flume 3/3)

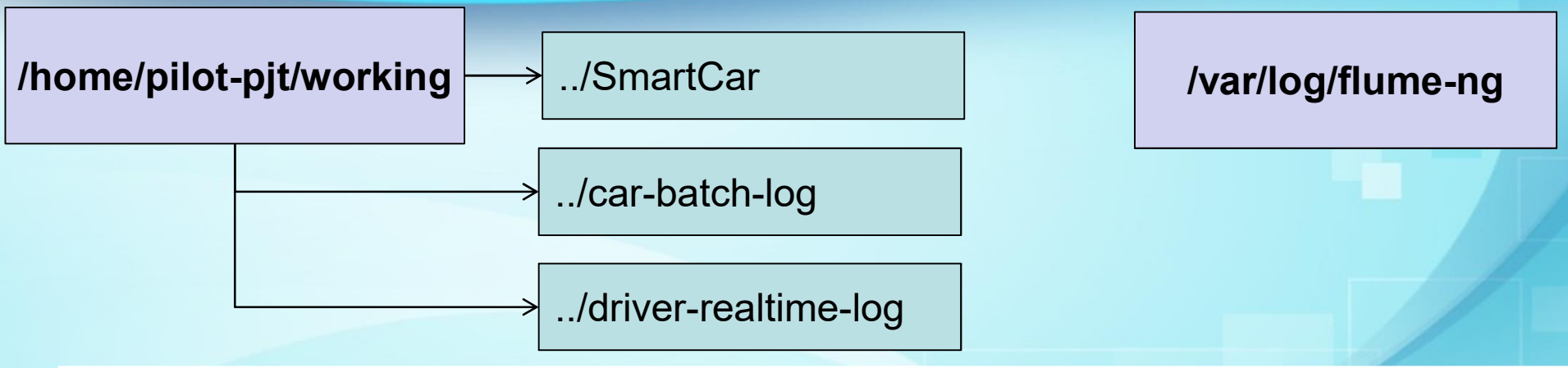
- 플럼(Flume)- 3/3
- 구성 유형4



11. 빅 데이터 생성 실습(1)

데이터 생성기	**bigdata.smartcar.loggen-1.0.jar.**
디렉토리 생성	mkdir -p /home/pilot-pjt/working/car-batch-log mkdir -p /home/pilot-pjt/working/driver-realtime-log
권한 부여	chmod 777 -R /home/pilot-pjt
실시간 빅데이터생성 **(400Kb/1초)**	java -cp bigdata.smartcar.loggen-1.0.jar com.wikibook.bigdata.smartcar.loggen.DriverLogMain 20201027 10
일 배치 빅데이터 생성 **(100Mb/일)**	java -cp bigdata.smartcar.loggen-1.0.jar com.wikibook.bigdata.smartcar.loggen.CarLogMain 20201027 10

11. 빅 데이터 생성 실습(2)



../working	빅데이터를 생성하는 자바프로그램이 존재하는 위치
../working/SmartCar	100Mb/일 배치 데이터 생성 위치
../working/car-batch-log	플럼이 처리 대기 중으로 지정된 Drectory (Drectory에 데이터파일을 있으면 플럼이 읽어 처리한 후 삭제함.)
../working/driver-realtime-log	400Kb/1초 실시간 생성 데이터 위치
/var/log/flume-ng	플롬이 처리한 Log파일 위치