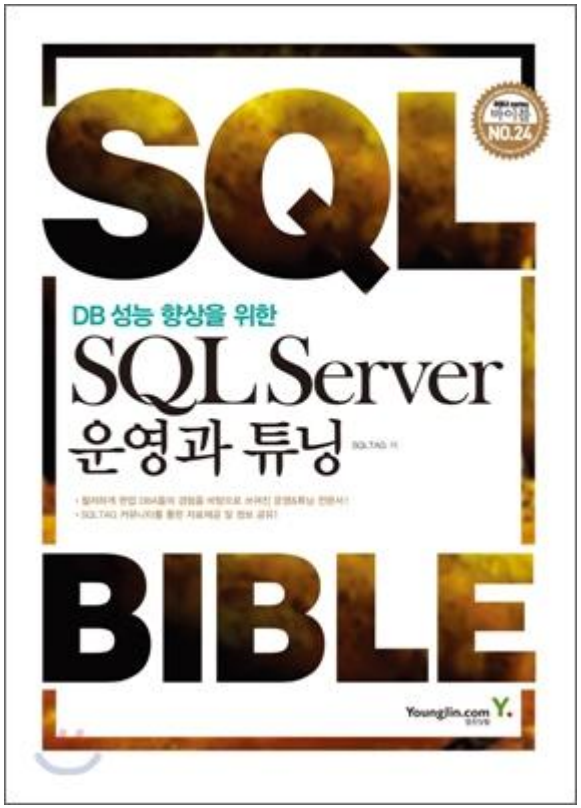
빅데이터 분석과 품질의 틀을 깬다

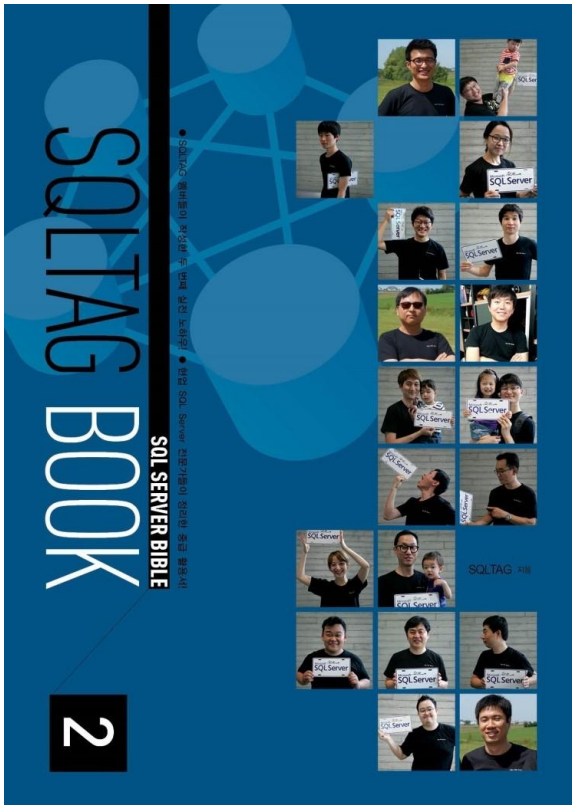
WISE OLAP BLU와 WISE Agile Analytics

2014 Data Grand Conference

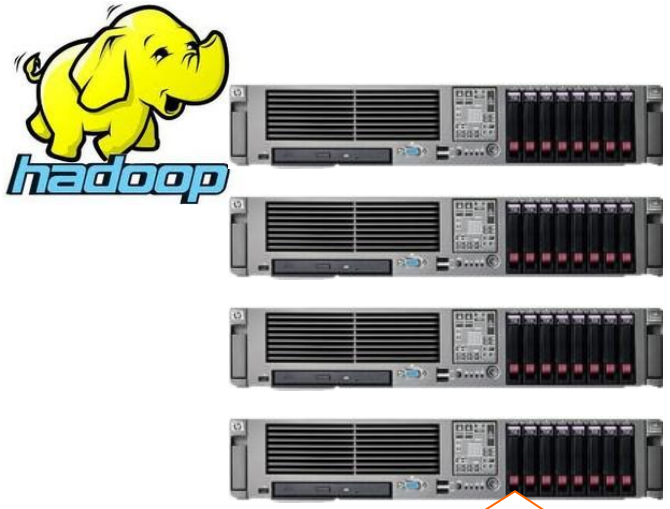
WISE I TECH

Best in Enterprise Intelligence


SQL
DB 성능 향상을 위한
SQL Server
운영과 튜닝
BIBLE
Youngjin.com


SQL SERVER BIBLE
SQLTAG BOOK
2
SQLTAG

빅데이터 분석이 아닌 빅데이터 저장에 초점

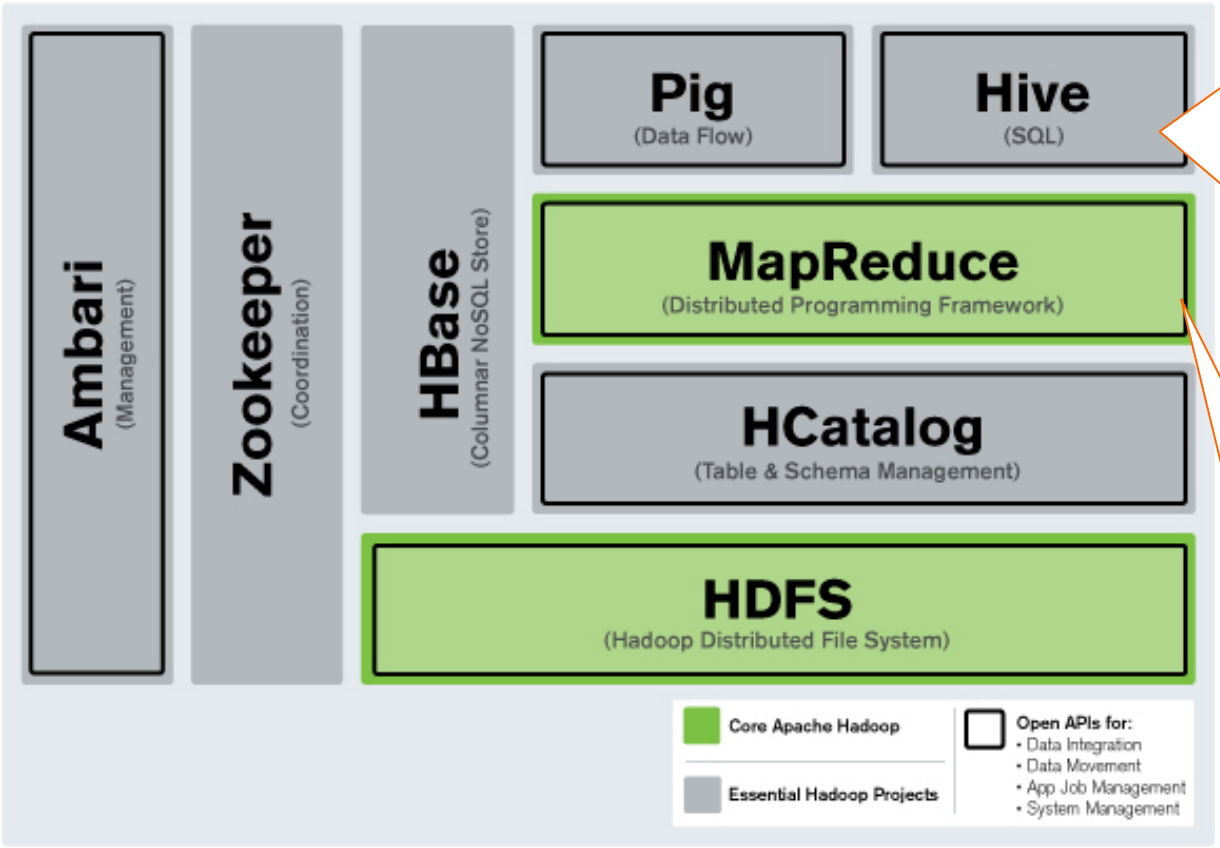


파일 저장, 대량의 로그, 센서, 텍스트 데이터 저장에 적합

단, 압축 없이 중복 분산 저장(x 3~5)하므로 저렴하나 많은 디스크 용량을 필요로 함

- ✓ 많은 데이터를 경제적으로 저장
  - Storage, RDB는 고비용
  - 최고의 확장성
- ✓ 특정한 분석과 서비스를 빠르게 수행
  - 단순 집계, 간단한 로직 처리는 아주 빠름
  - 예) 방문자수 카운트, 웹상의 간단한 추천

하둡으로는 빅데이터 분석에 한계



인텔, 클라우데라에 8천억 베팅...왜?

클라우데라는 2년전 '임팔라(Impala)'라는 기술을 통해 기존 업계 표준 데이터 처리문법인 SQL을 하둡 환경에도 쓸 수 있도록 만들었다

하둡을 사용하기 때문에 오히려 성능이 저하된다.

페이스북은 이러한 단점을 극복하기 위해서 최근에 맵리듀스 계층을 제거한 프레스토(Presto) 엔진을 도입

빅데이터 분석을 위한 NewSQL(1/2)

거대한 변화, NewSQL vs NoSQL - ZDNet Korea 2013.12.18 중에서

MIT 교수인 마이클 스톤브레이커는 관계형 데이터베이스 시스템에 대한 최고 권위자중 한 명이다. 그가 1973년에 개발한 인그레스(Ingres)라는 데이터베이스 시스템은 당시까지 추상적인 개념으로만 존재하던 관계형 데이터베이스를 최초로 구현한 사례였다.

최근에 업계에서는 NoSQL이 많은 주목을 받고 있는데, 그는 이러한 흐름에 대해서 설득력 있는 비판을 가하면서 전통적인 관계형 데이터베이스를 새롭게 구축하는 전략을 의미하는 NewSQL이라는 개념을 주창했다.

NoSQL에 대한 그의 비판은 'NoSQL'의 정확한 이름은 아마도 'NotYetSQL'이라고 봐야 할 것이라는 주장으로 압축된다. 스톤브레이커 교수가 보기에 NoSQL 시스템이 전통적인 관계형 데이터베이스의 한계를 통렬하게 지적하고 'SQL'을 부정하면서 출발했지만, 시간이 흐르면서 차츰 전통적인 관계형 데이터베이스의 틀 안으로 되돌아 올 수밖에 없다.

빅데이터 분석을 위한 NewSQL(2/2)

하둡만 해도 그렇다. 하둡은 가장 아래에 있는 파일시스템을 의미하는 HDFS, 중간에 존재하는 알고리즘인 맵리듀스(map reduce), 그리고 질의문을 구성하는 하이브(Hive)와 같은 테크놀로지, 이렇게 세 개 계층으로 구성된다.

하둡을 사용하는 대표적인 기업인 페이스북은 처음에 하이브를 이용해서 성공을 거두었지만, 얼마 지나지 않아서 HDFS와 맵리듀스가 기본적으로 배치처리를 중심으로 설계되었으며 그에 따르는 많은 오버헤드를 안고 있기 때문에 성능을 저하시킨다는 문제에 봉착했다.

스톤브레이커 교수는 페이스북이 발생시키는 트래픽의 95%가 SQL과 다를 것이 없는 하이브를 통해서 발생하는 트래픽이고 겨우 5% 정도만 맵리듀스를 통한 병렬처리의 덕을 볼 수 있는 트래픽이라고 지적한다. 이것은 곧 페이스북이 하둡이라는 묵직한 테크놀로지를 통해서 이득을 볼 수 있는 부분이 5%에 불과하다는 뜻이다.

나머지 95% 부분은 하둡을 사용하기 때문에 오히려 성능이 저하된다. 이런 의미에서 **하둡은 병렬처리가 가능한 최신의 관계형 데이터베이스의 성능에 미치지 못한다.**

지금까지 빅데이터 분석의 실질적인 대안은 Appliance

HW+OS+DBMS 일체형인 데이터 분석 전용 장비

HW/OS가 제외된 SW형 어플라이언스도 있음

수테라 ~ 수십테라를 고속으로 처리

But 수억 ~ 수십억원 가격

어플라이언스 제품 비교

	A	B	C	D	E
H/W 일체형	O	O	X	O	X
컬럼저장방식	X	O	O	X	X
Hadoop 지원 (브로셔 기준)	X	O	O	X	X
압축지원	O	O	O	O	O
튜닝요소	많음	많음	적음	적음	적음
확장용이	△	X	O	X	O
OLTP지원	O	O	X	O	O
OS	솔라리스	NT	Linux/Unix	Linux	Linux

SW형 어플라이언스가 늘어가고 있고

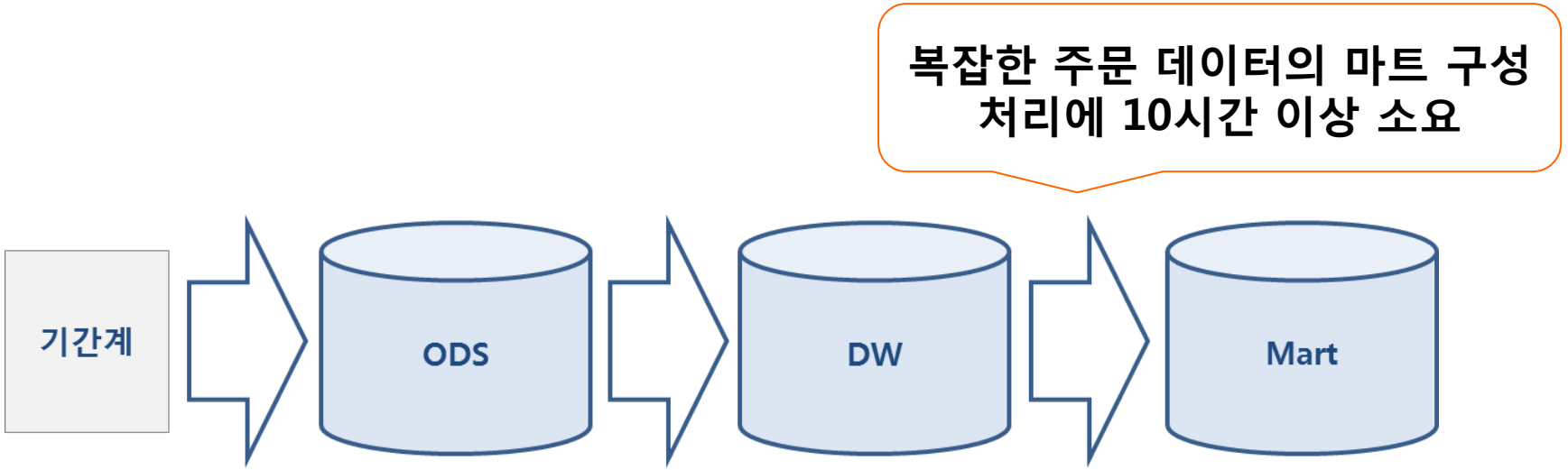
컬럼 저장과 압축 기술이 보편화

컬럼 저장 기술로 경제적인 빅데이터 분석 플랫폼 구성이 가능

기존 RDBMS에서도 컬럼, 압축, 메모리 캐싱 기능을
추가한 컬럼형 DB 기능이 추가되고 있음

컬럼 기반의 오픈소스 DBMS와 새로운 상용 DBMS 출현
InfiniDB, PetaSQL, ...

적용 사례(1/2)



- ✓ 저장된 데이터를 조회만 하지 않는다
- ✓ 다양한 분석을 위해서는 분석 모델에 맞게 데이터마트를 구성해야 하고
- ✓ 특정 테이블이 크거나 로직이 복잡하다면 조인, 업데이트 작업량은 엄청난 시간을 필요로 한다

적용 사례(2/2)

대용량의 데이터를 빠르게 분석하기 위해 많은 기업들이 수십억원에 이르는 어플라이언스를 이용하고 있는 반면, WISE OLAP은 컬럼 저장과 최적 메모리 캐싱을 지원하는 고성능의 경제적인 WISE OLAP BLU를 제공하고 있다.

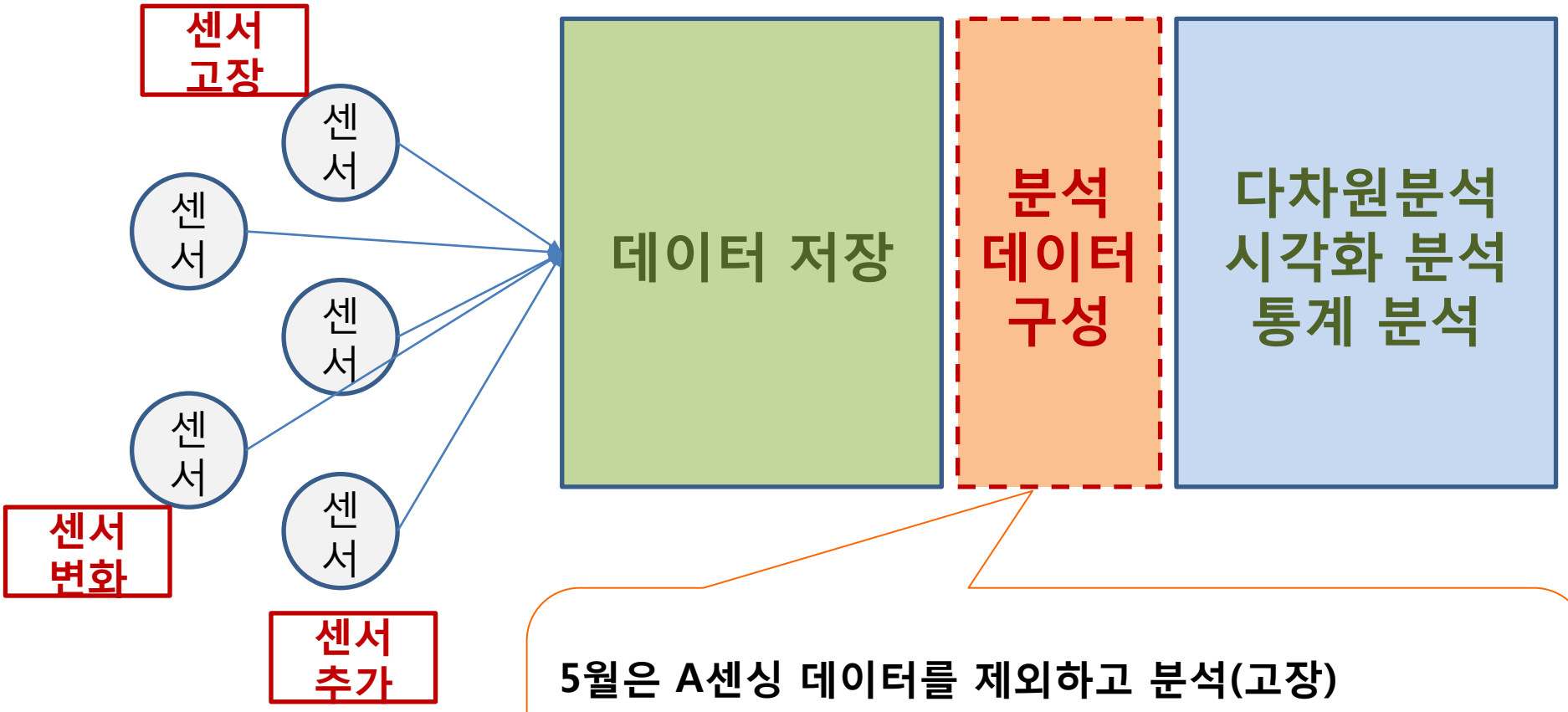
유통서비스 H사의 경우 기존의 BI 시스템을 WISE OLAP BLU로 업그레이드하면서, 분석 처리 속도는 10배 향상시키고, 비용은 어플라이언스의 10분의 1 이하라는 획기적인 성과를 얻었다

센싱 빅데이터는 새로운 분석 접근을 요구

IoT에서의 각종 센서, 웹/앱에서의 각종 센싱 데이터나 로그 데이터

- 빅데이터
- 해당 기기의 설정에 따라 데이터 항목과 형식이 변경
- 무엇을 분석해야 할지 불명확
- 매출액이나 기온과 같은 다른 데이터와의 결합 분석을 요구

센싱 빅데이터 이슈 예시(1/4)

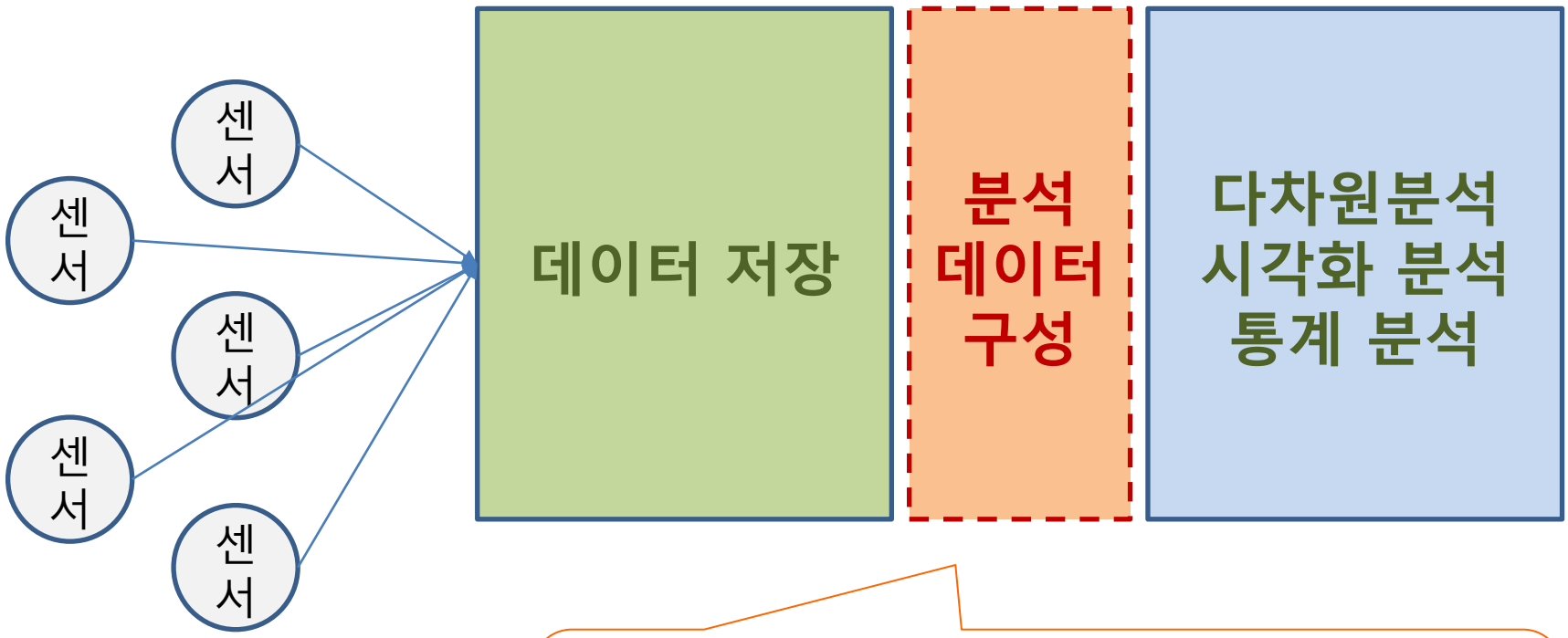


5월은 A센싱 데이터를 제외하고 분석(고장)

9월부터는 B센싱 데이터 추가한 후 평균하여 분석(추가)

C센싱 데이터는 +1보정하여 분석(센싱 감도 변화)

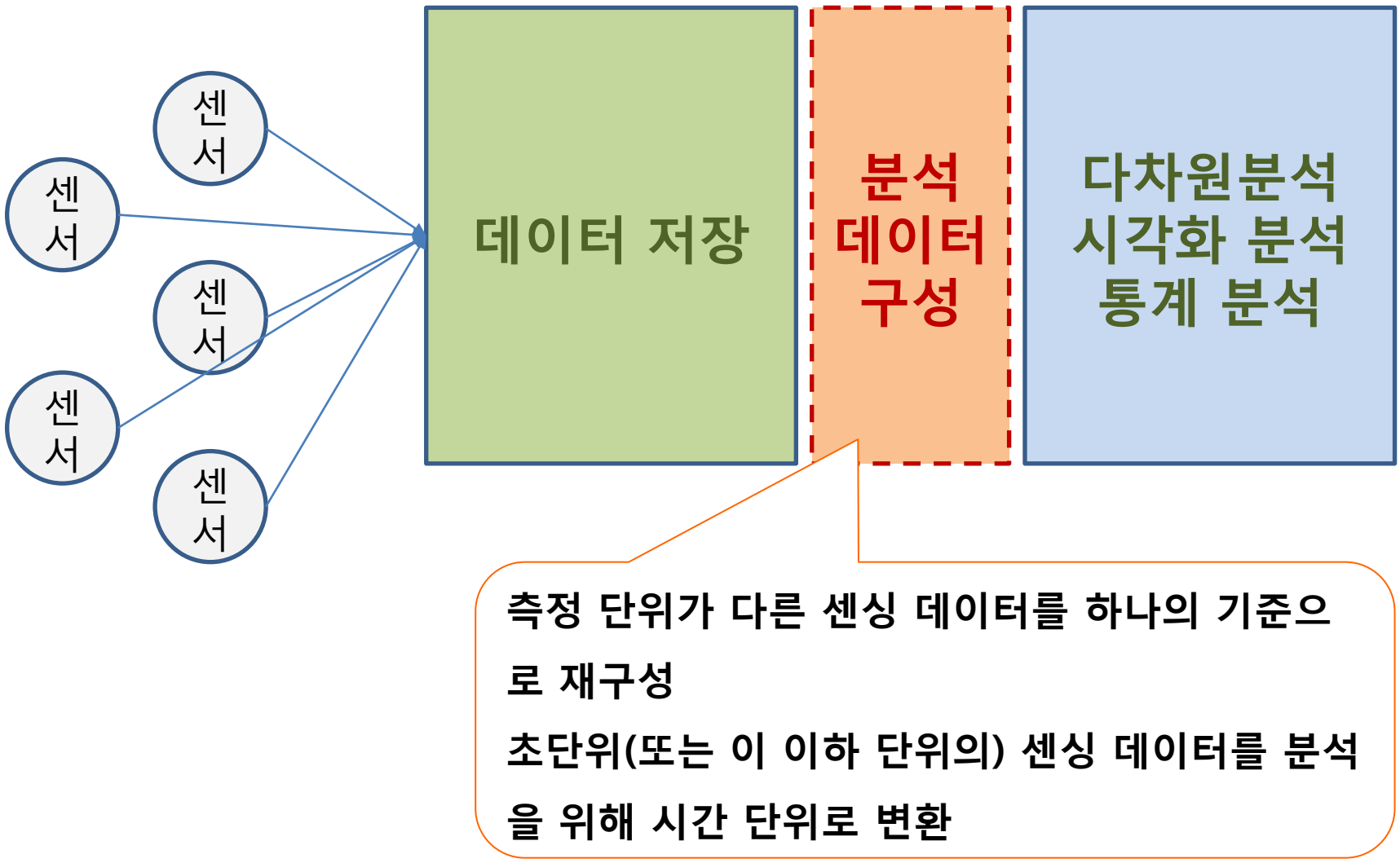
센싱 빅데이터 이슈 예시(2/4)



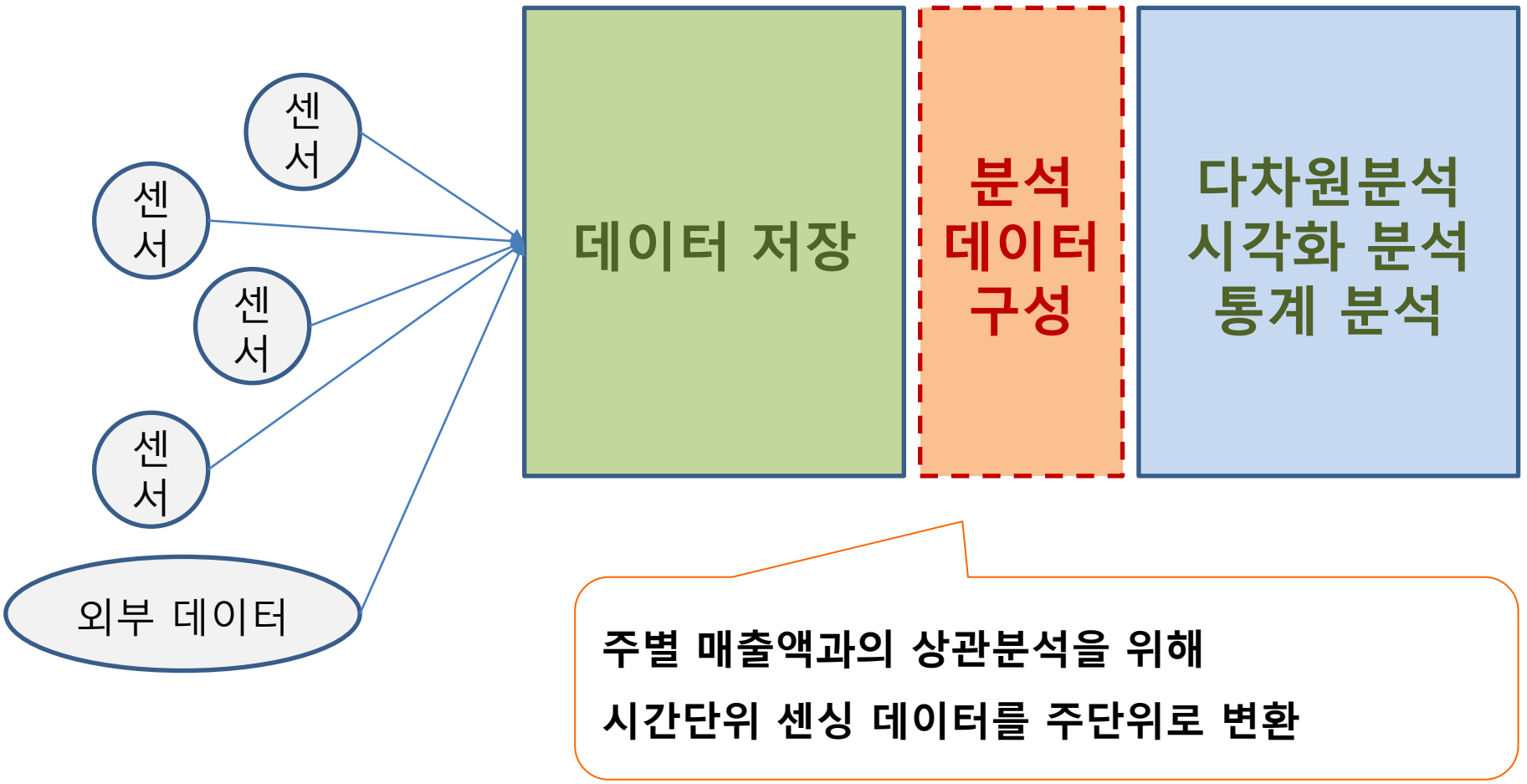
5개 센싱 데이터를 평균한 값으로 분석

최고값(또는 최저값)을 뽑은 후 분석

센싱 빅데이터 이슈 예시(3/4)



센싱 빅데이터 이슈 예시(4/4)



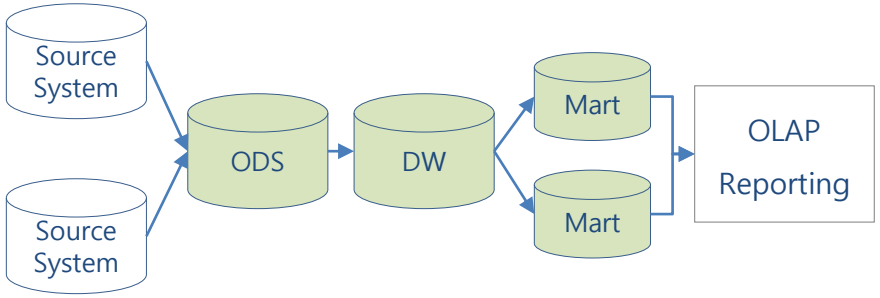
센싱 빅데이터 분석 데이터 구성은 사전에 정책을 정하기 어렵다

분석가 스스로 내가 무엇을 분석할 지 사전에 알기 어렵다. 분석 주제가 나올 때 마다 데이터를 구성해야 한다.

- ✓ 겨울철 이상 고온 분석을 할 경우 – 겨울철만 선별해서 최고/평균 등을 우선 계산(추출)
- ✓ 외부 데이터와 비교 분석을 할 경우 – 인구수와 비교할 경우에는 연도별로 데이터 구성, 매출액과 비교할 경우에는 월별로 구성

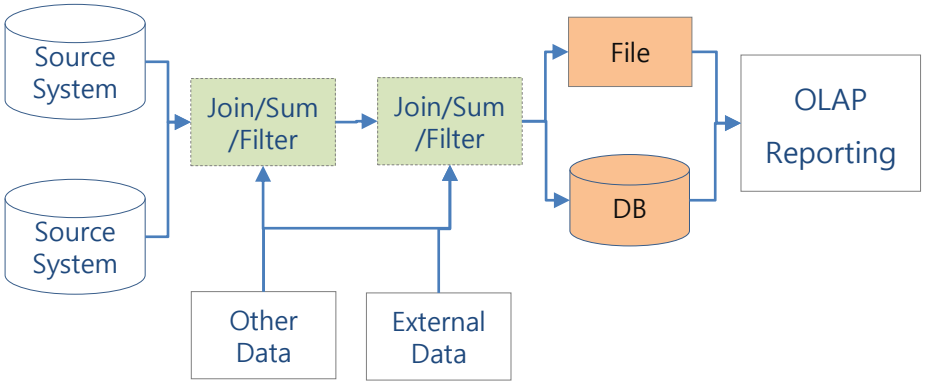
애자일 분석(Agile Analysis)

전통적인 데이터 분석



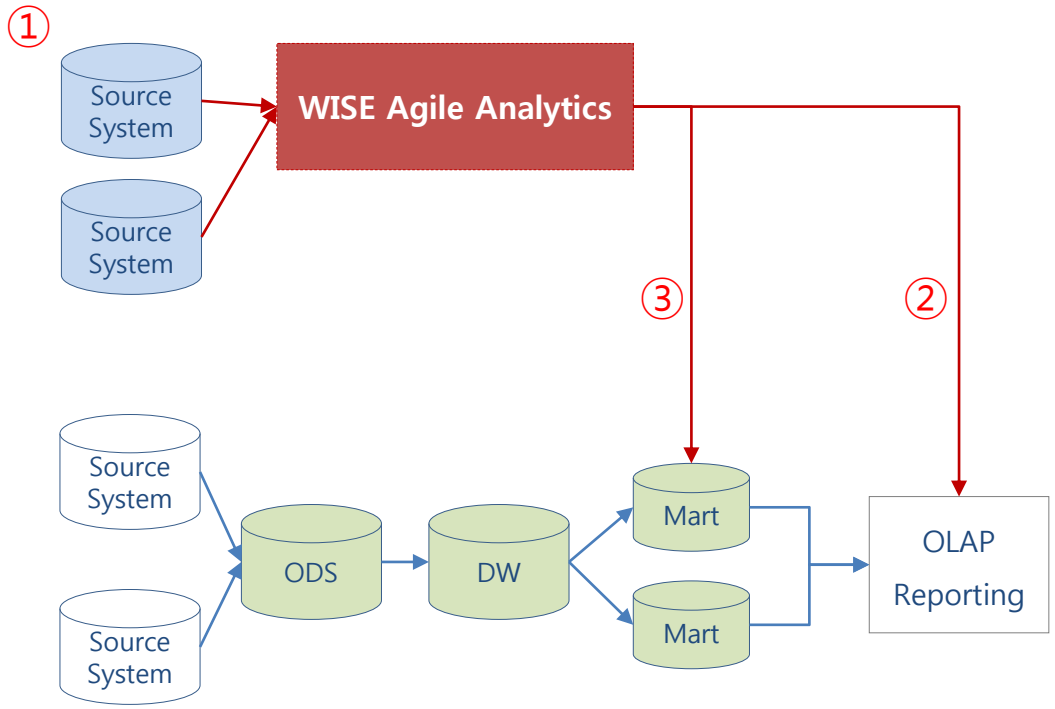
- 정해진 몇 개의 시스템에서 데이터 추출
- 분석, 설계를 거쳐 DW, Mart 구축
- 정형화 된 분석

Agile Analysis



- 다양한 데이터를 동적으로 조합
- 그때 그때 필요한 분석을 위해 빠르고 유연하게 분석 데이터를 생성

기존 DW를 보완하는 애자일 분석(Agile Analysis)

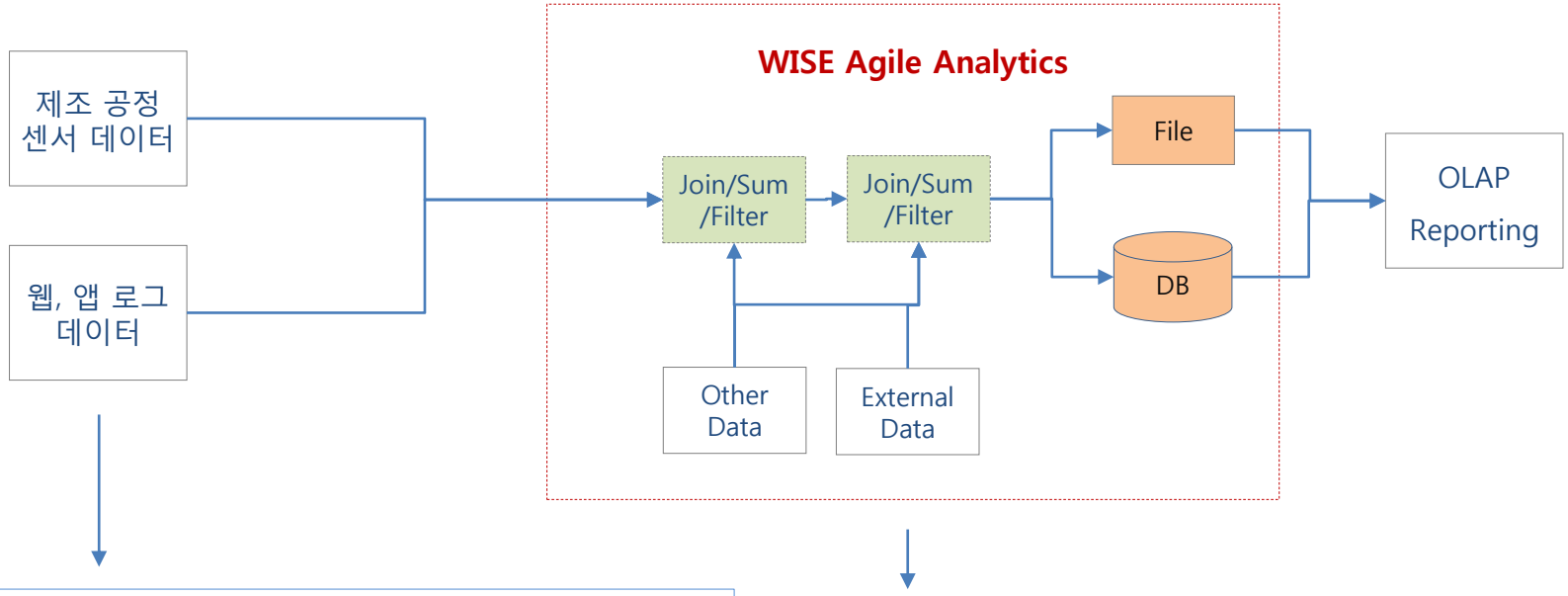


① 원천 데이터가 추가 되거나 변경되는 경우, 분석-설계-ETL-마트 구축 단계를 거치지 않고

② 빠르게 분석 데이터 집합을 생성하여 분석하거나

③ 지속적인 분석이 필요한 경우에는 정기적으로 기존 마트에 추가

WISE Agile Analytics in Sensing & Log data



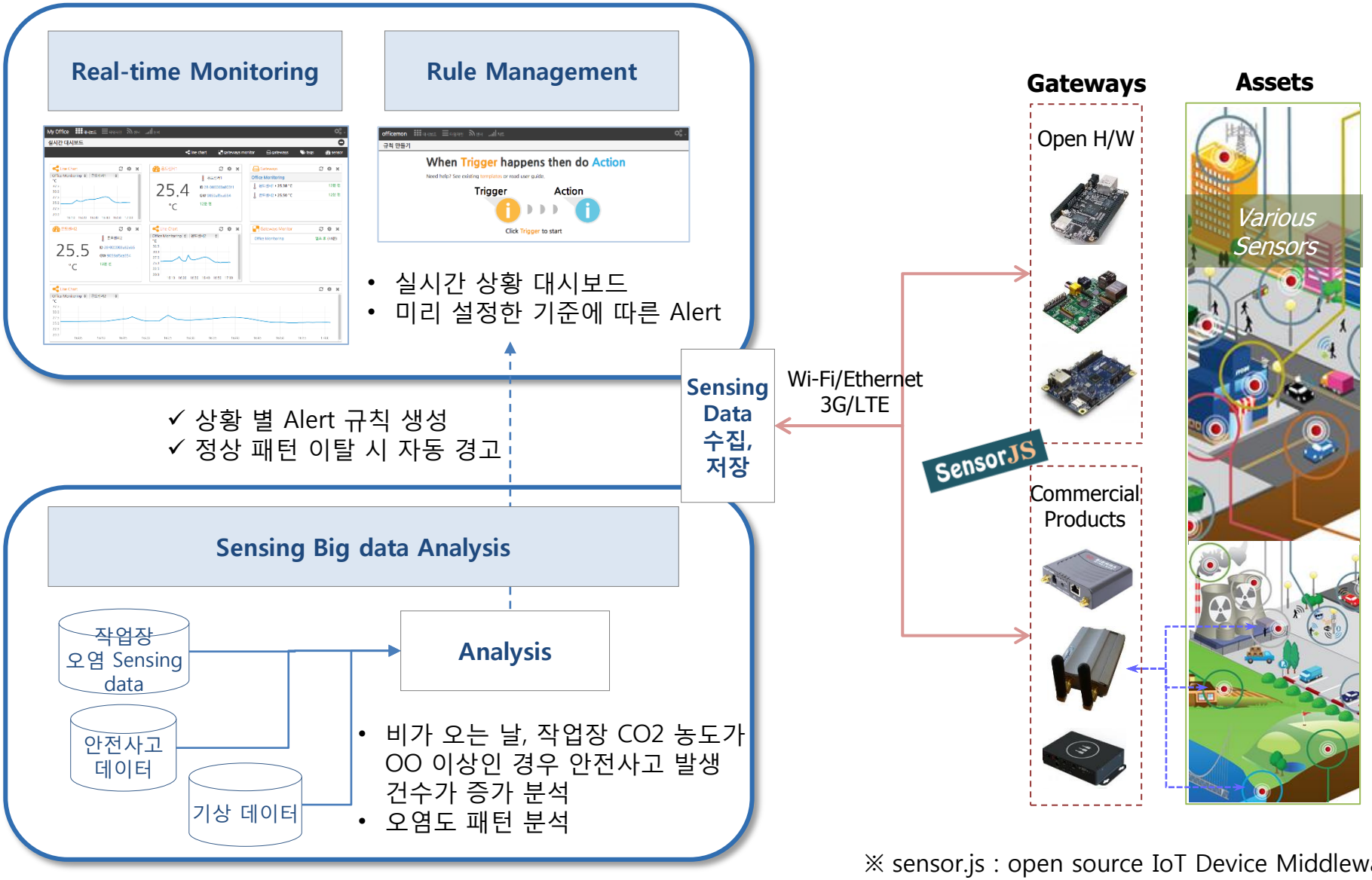
- Big Size
- 설정에 따라 데이터 항목, 형식 변경
- 제조 장비, 앱 메뉴 추가에 따라 데이터 변동
- 무엇을 분석해야 할지(분석 관점) 불명확
- 다른 데이터(예: 매출)와의 결합 분석 요구

✓분석 관점을 미리 정하지 않고 다양하게 접근

✓분석 가치가 있는 리포트가 정해졌을 경우, 주기적으로 자동 생성

✓다양한 센서 데이터간 조합 또는 외부 데이터를 조합하여 분석

WISE Agile Analytics in IoT

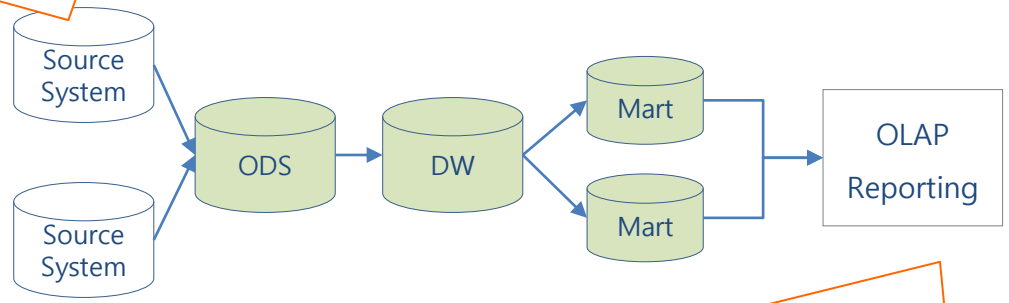


※ sensor.js : open source IoT Device Middleware

프로파일링 관점의 품질 분석

품질관리 DB를 대상으로

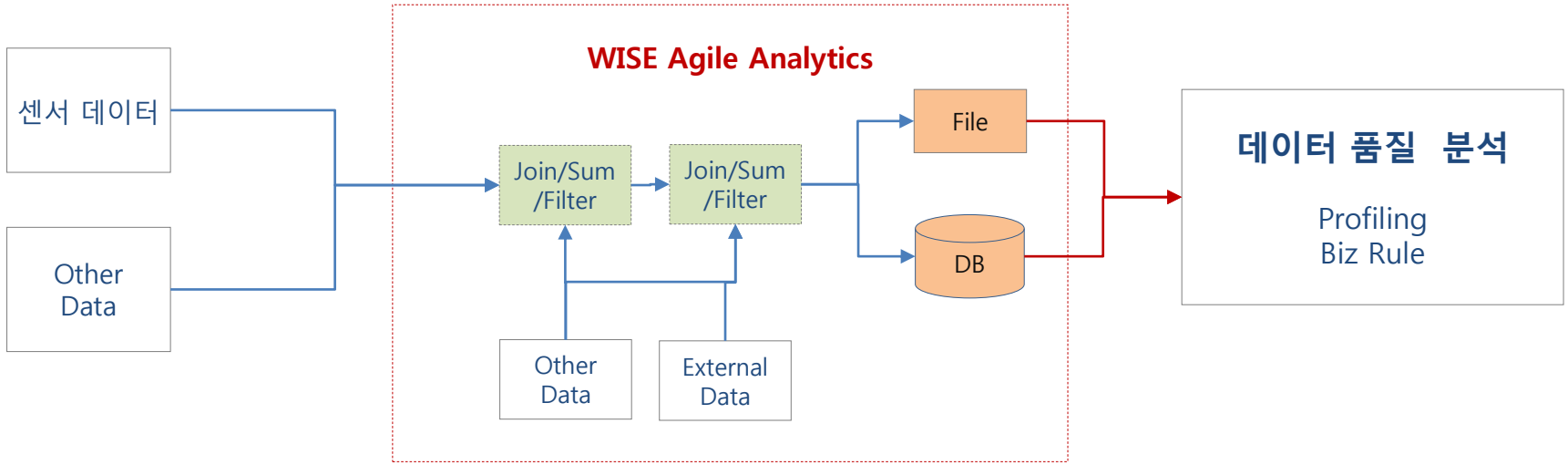
컬럼별 평균, 최대, 최소, Null 등의 분포 측정



리포트 오류 시 원인 파악 불분명
리포트 자체의 문제? DW의 문제? 원천의 문제?

리포트로부터 사후적인 품질 오류 추적

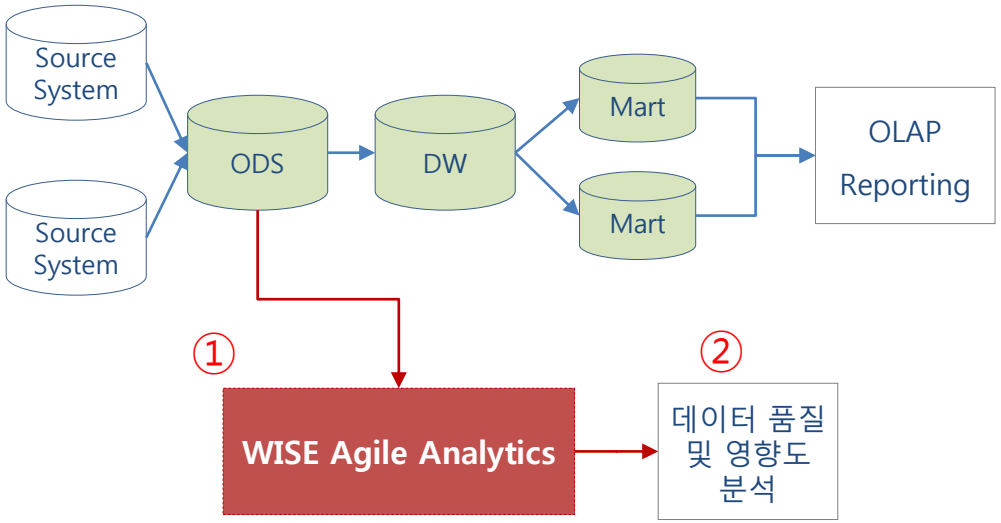
분석 관점의 데이터 품질 관리



① 분석을 위해 수집, 구성한 데이터를 대상으로 품질을 체크

② 조합한 데이터를 대상으로 Biz 관점의 품질 체크 용이 (압력 데이터 대비 30%를 넘는 온도 센싱 데이터는 비정상)

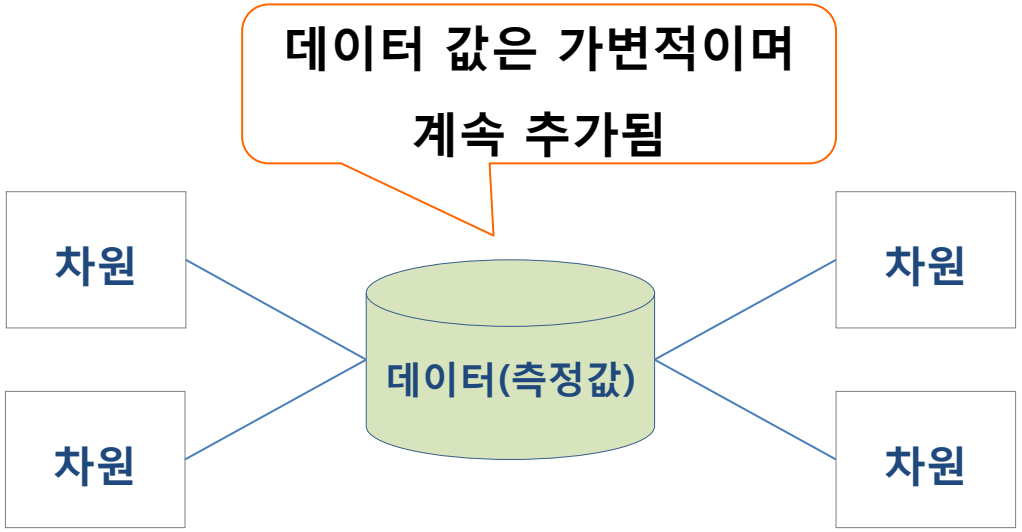
ODS 변경 탐지를 통한 품질 관리



① 분석을 위해 수집하는 원천 데이터의 변경 여부를 탐지하고 주요 항목에 대한 품질을 체크

② 원천 데이터 형식 변경 사항과 그로 인해 영향받는 리포트가 무엇인지를 제공하며, DW 주요 항목에 대한 품질 수준을 리포트

마트 표준 체계 관리

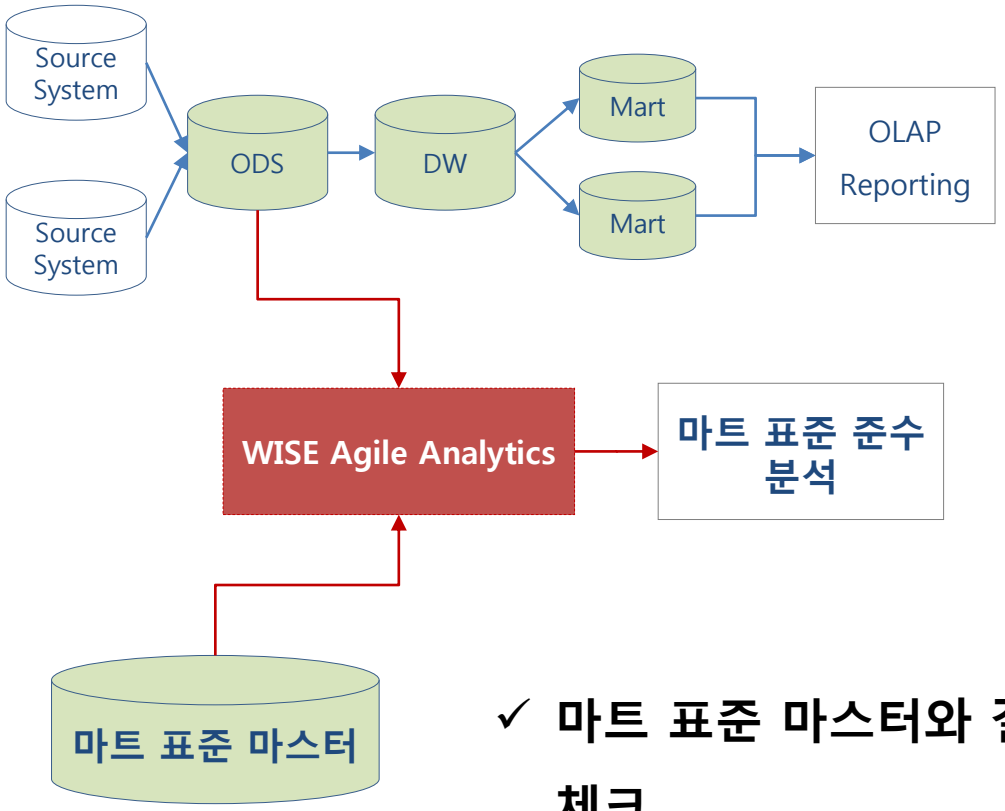


차원 멤버 코드는
표준 (차원 마스터)
에 맞춰서 관리

일자차원(주문일자, 배송일자, 반품일자, 결제일자 등)과 같이
기본적이고 반복적으로 사용되는 차원은 공통으로 관리

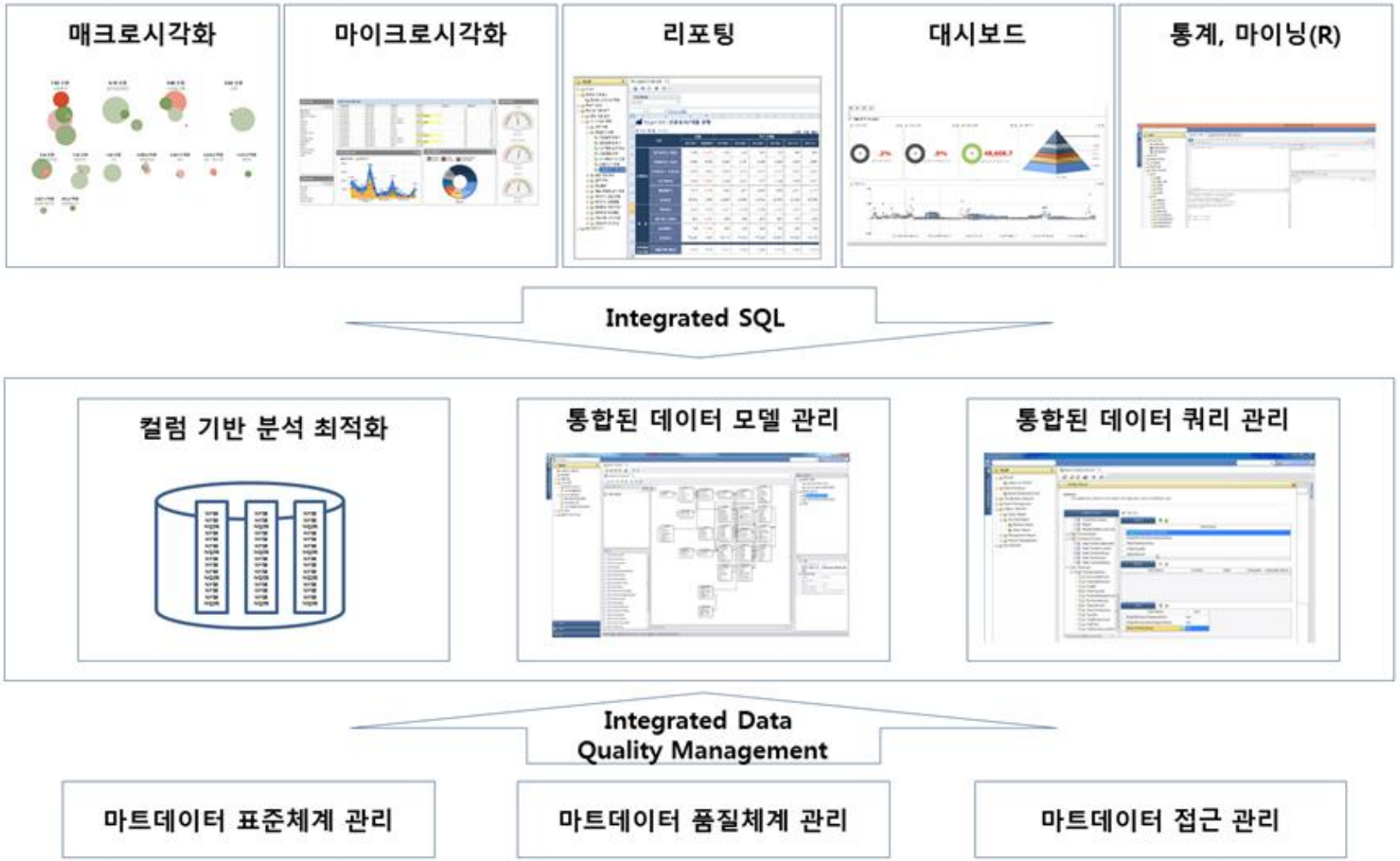
- 일자차원(연/반기/분기/월/일)
- 시간차원(시/분/초)
- 주차차원

마트 표준 준수 관리

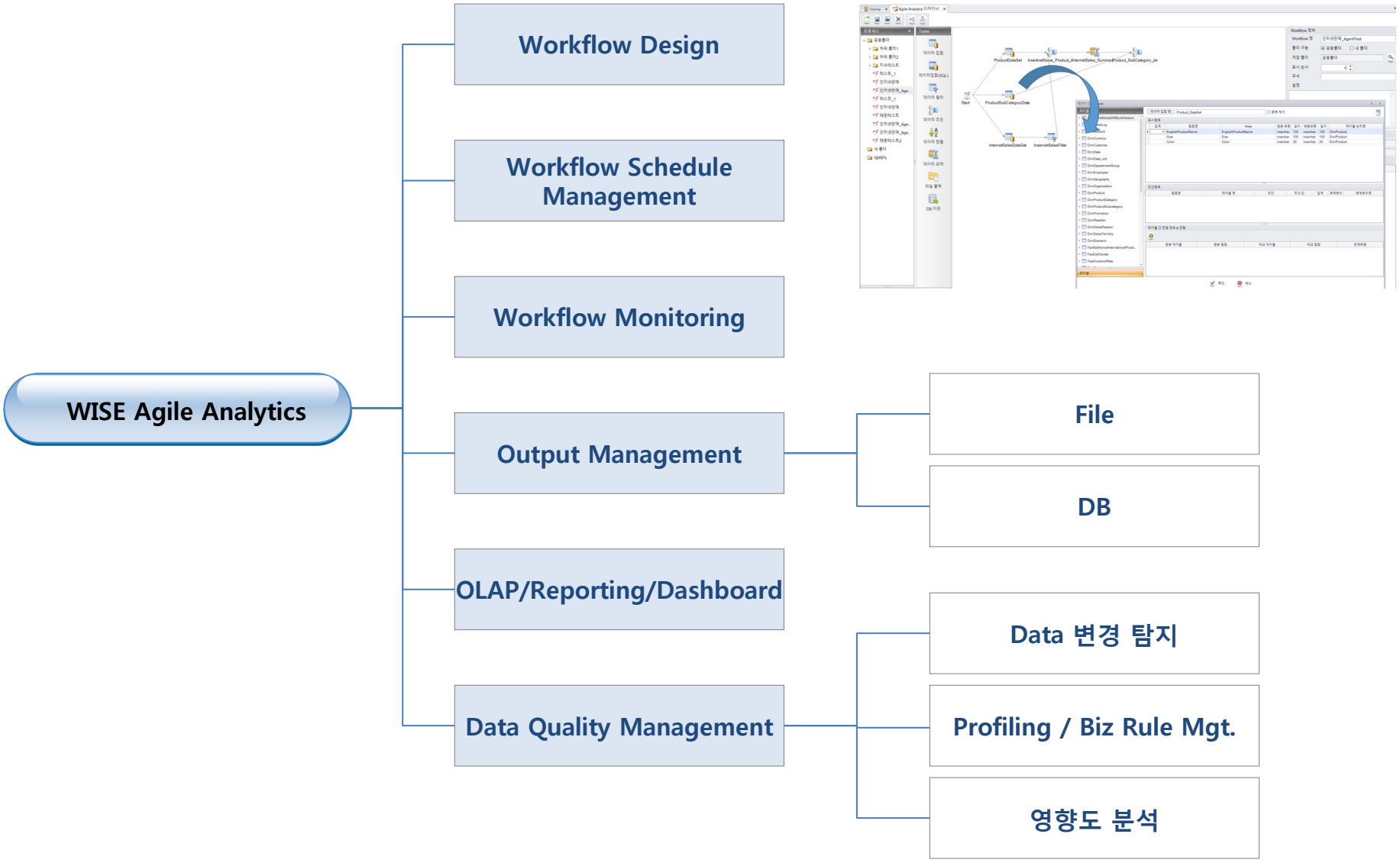


- ✓ 마트 표준 마스터와 결합하여 데이터 표준 준수 여부를 체크
- ✓ 수집되는 데이터에서 새로 발생되는 차원 멤버를 파악하여 마스터에 반영

빅데이터 분석 프레임워크 예



WISE Agile Analytics 기능 구성



감사합니다.

㈜위세아이텍
김 상 수
sskim@wise.co.kr