실무에 적용 가능한 Big Data 분석 개론

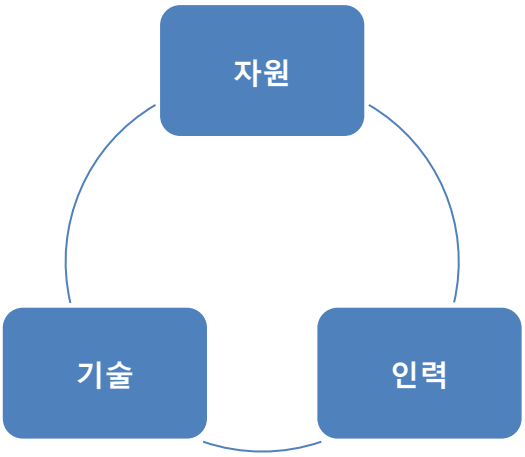
# 빅데이터 활용요소와 데이터 사이언티스트


한국기술교육대학교
온라인평생교육원

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 1. 빅데이터 활용을 위한 3대 요소

**빅데이터 활용을 위한 3대 요소**



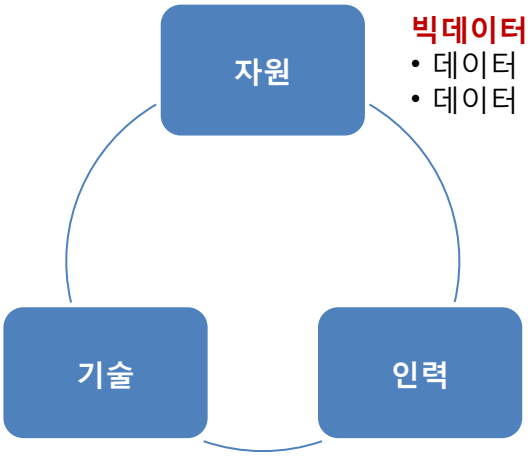

<출처 : 김현곤, 빅데이터 시대의 전망과 대응전략, 2012 >

- 데이터가 폭발적으로 증가하면서 빅데이터가 등장했지만, 방대한 양의 데이터 중에서 의미 있는 데이터는 소수에 불과함
- 빅데이터를 활용하기 위해서는 데이터를 자원화 해야 하고, 그 중 의미 있는 데이터를 찾아내기 위해 빅데이터를 효과적으로 처리할 수 있는 기술이 필요하며, 빅데이터의 의미를 통찰할 수 있는 인력이 있어야 함
- 성공적인 빅데이터 활용을 위해서는 자원, 기술, 인력 3가지 분야의 요소가 필수적임

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 1. 빅데이터 활용을 위한 3대 요소

**빅데이터 활용을 위한 3대 요소**




**빅데이터(Big Data)**
- 데이터 자원 확보
- 데이터 품질 관리

**빅데이터 플랫폼(Big Data Platform)**
- 데이터 저장, 관리 기술
- 대용량 데이터 처리
- 빅데이터 분석
- 시각화

**데이터 사이언티스트(Data Scientist)**
- 수학, 공학(IT기술과 엔지니어링) 능력
- 경제학, 통계학, 심리학 등 다문학적 이해
- 비판적 시각과 커뮤니케이션 능력
- 스토리텔링 등 시각화 능력

- 빅데이터 기술을 잘 사용할 수 있도록 준비된 환경
- 빅데이터 기술의 집합체
- NoSQL(Not only SQL)
- ETL
  - ➢ Extraction, Transformation, and Loading 즉, 데이터의 추출, 변환, 적재
- 하둡
- 맵리듀스
- 자연어처리
- 의미분석
- 데이터마이닝
- 빅데이터 표현

<출처 : 김현곤, 빅데이터 시대의 전망과 대응전략, 2012 >

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 2. 빅데이터 활용을 위한 자원 요소

**자원**

데이터는 유용한 정보 및 지식을
찾는 데 필요한 자원(Resource)

확보의 중요성!

통신, 포털, 금융, 의료, 공공 분야 및 대기업

중소기업

데이터 자체가 빅데이터 수준 **Vs.** 데이터 부족

정보 격차 발생 우려

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 2. 빅데이터 활용을 위한 자원 요소

**자원**


외부자원은 적극적으로 확보 **Vs.** 내부자원은 품질 향상 및 관리

기업의 투자 부담이 완화되고 있음

➡ 저렴한 가격으로 빅데이터 저장 공간과 분석 솔루션을 제공하는 기술 및 서비스 시장이 확대되고 있기 때문임

'자원' 측면의 전략수립이 필요함

➡ 활용할 수 있는 빅데이터를 발견 및 확보하고 빅데이터의 품질을 관리함

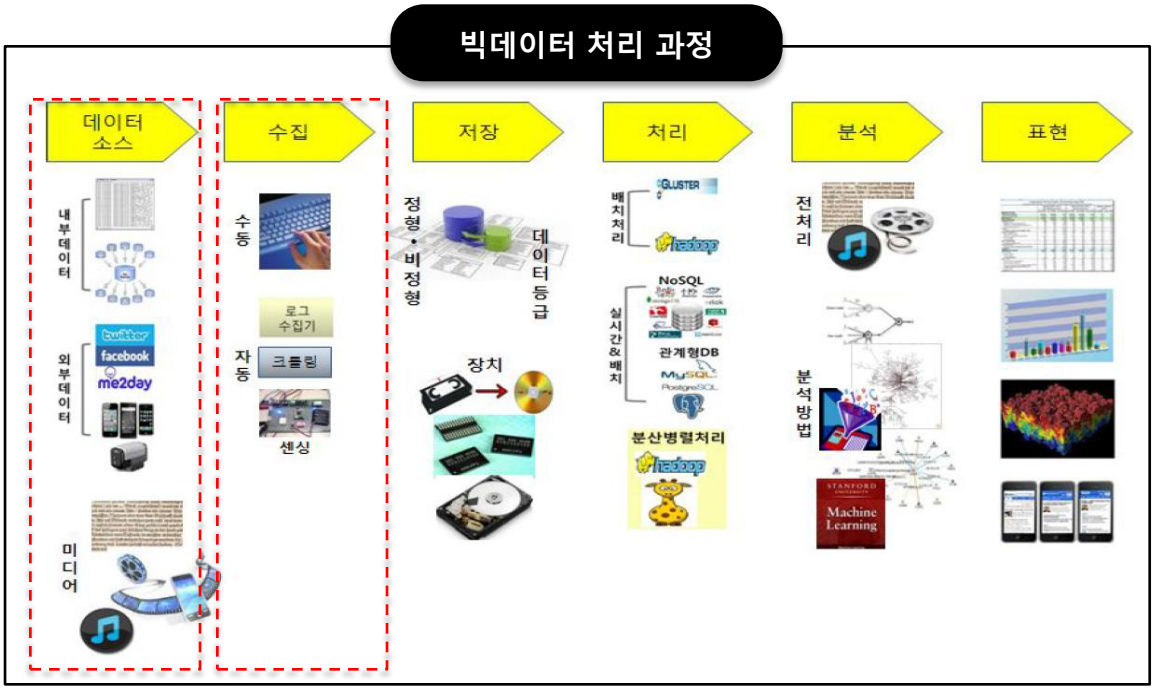
가트너의 빅데이터 자원을 확보하는 방안

1. 독자적으로 데이터를 생성, 저장하고 외부 데이터를 검색을 통해 수집함
2. 기업 데이터를 외부 기관들과 상호 교환함
3. 특정 활동이나 목적을 위해 모인 연합, 그룹들이 협력의 장을 형성하거나 표준화된 데이터 풀(Pool)을 연계하여 상호 이용함
4. 오픈 방식의 플랫폼을 통한 데이터 공유

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 3. 빅데이터 활용을 위한 기술 요소

**기술**




<출처 : 문혜정, 'Big Data 구축기술과 사례를 중심으로' 재구성, 2012 >

**▶ 데이터 소스**

- 데이터는 소스 위치에 따라 내부 데이터와 외부 데이터 그리고 미디어로 구분함
- 내부 데이터 소스 : 자체적으로 보유한 내부 파일 시스템이나 데이터베이스 관리 시스템, 센서 등
- 외부 데이터 소스 : 인터넷

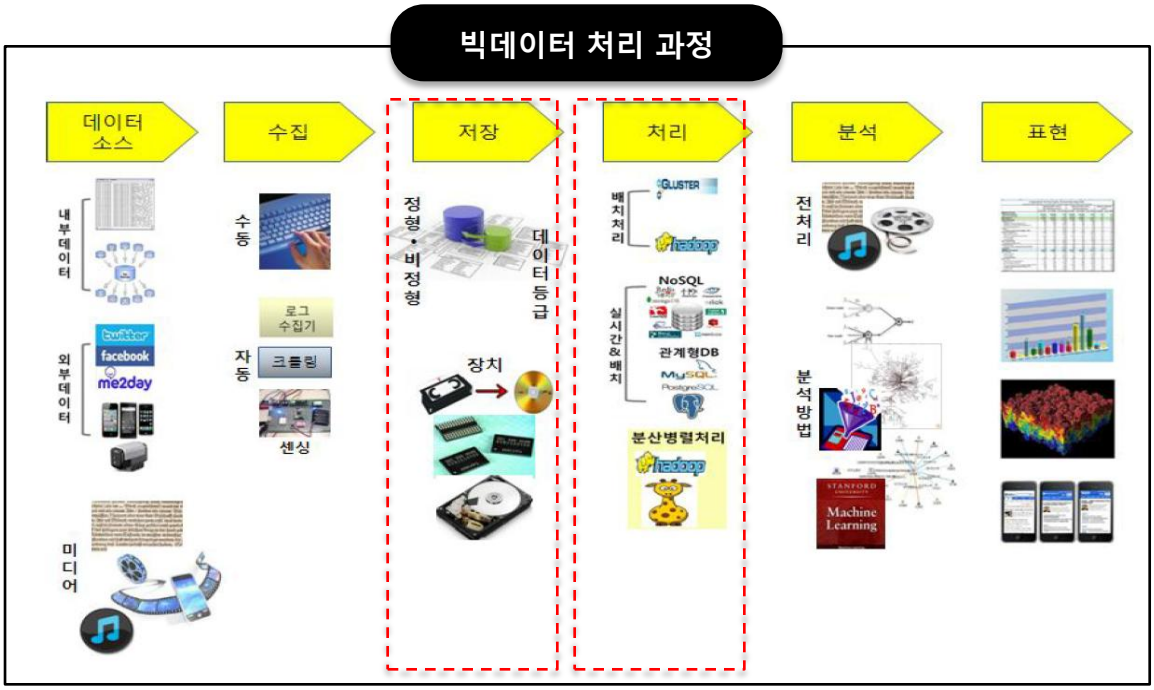
**▶ 수집**

- 데이터 수집은 소스 위치에 따라 구분함
- 내부 데이터는 정형 데이터로 수집하고, 외부 데이터는 비정형 데이터로 수집함
- 수집을 위한 기술은 수동과 자동 방식으로 나뉘는데 자동으로 데이터를 수집하는 기술에는 로그수집기, 크롤링, 센싱 등이 있음
- 로그수집기는 사용자가 웹이나 애플리케이션을 통해 어떤 서비스를 이용하게 되면 생성되는 로그데이터를 자동으로 수집함
- 크롤링 기술은 많은 컴퓨터에 분산 저장되어 있는 문서를 수집하여 검색 대상의 색인으로 포함시키는 기술
- 센싱은 각종 센서의 작동을 통해 물체, 소리, 빛, 압력, 온도 등을 탐지, 관측, 계측하여 데이터화하는 것

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 3. 빅데이터 활용을 위한 기술 요소

**기술**




<출처 : 문혜정, 'Big Data 구축기술과 사례를 중심으로' 재구성, 2012 >

#### ▶ 저장

- 수집한 데이터에서 의미 있는 정보를 추출하려면 효율적으로 저장 관리하는 기술이 필요한데 추후 사용할 수 있도록 데이터를 안전하고 효율적으로 저장하기 위함임
- 빅데이터는 '대용량, 비정형, 실시간성' 속성을 수용할 수 있는 저장 방식이 필요함

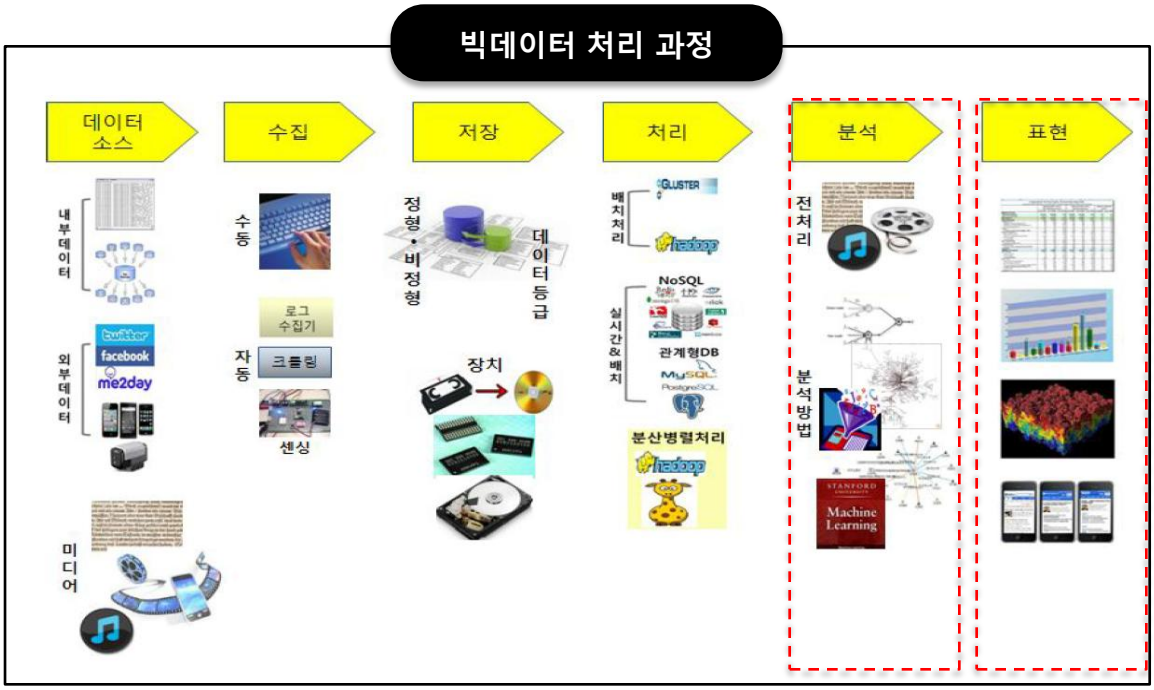
#### ▶ 처리

- 저장된 빅데이터는 방대한 양의 데이터와 데이터 생성 속도, 데이터 종류의 다양성을 통합적으로 고려할 수 있는 기술이 필요함
- 빅데이터의 처리를 수행하는데 활용되는 기술에는 하둡, NoSQL, 관계형 데이터베이스, 분산 병렬처리 등이 있음
- 하둡은 저가의 서버와 하드디스크를 이용하여 빅데이터를 상대적으로 쉽게 활용, 처리할 수 있는 분산파일 시스템 기술을 의미하며 NoSQL은 Not-only SQL을 줄인 말로 전통적인 관계형 데이터베이스 보다 덜 제한적인 일관성 모델을 이용하는 데이터의 저장 및 검색을 위한 매커니즘을 제공하는 데이터베이스를 의미하며 관계형 데이터베이스는 일련의 정형화된 테이블로 구성된 데이터 항목들의 집합체임
- 분산병렬처리 기술은 여러 대의 컴퓨터에 하나의 작업을 나누어 병렬로 처리하여 그 내용이나 결과가 통신망을 통해 상호 교환되도록 연결되도록 해주는 기술을 의미함

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 3. 빅데이터 활용을 위한 기술 요소

**기술**




<출처 : 문혜정, 'Big Data 구축기술과 사례를 중심으로' 재구성, 2012 >

**▶ 분석**

- 빅데이터 분석에 사용하는 기술은 대부분 통계학과 전산학, 특히 기계 학습과 데이터 마이닝 분야에서 이미 사용한 것들로, 이 분석 기술들의 알고리즘을 대규모 데이터 처리에 맞게 개선하여 빅데이터 처리에 적용시키고 있음
- 분석을 위해 데이터 전처리 과정을 거친 정형 및 비정형 데이터를 통계, 데이터 마이닝, 머신 러닝 기술 등을 적용하여 분석함
- 데이터 마이닝 기술은 데이터 베이스에 저장된 데이터로부터 의미있는 정보나 지식을 추출하는 기술이고, 머신 러닝 기술은 데이터를 기반으로 분석한 내용을 기계가 학습하고 미래를 예측할 수 있는 기술

**▶ 표현**

- 데이터 분석 결과를 한눈에 쉽게 이해할 수 있도록 간단한 도표나 3D 이미지 등으로 표현하는 정보 표현 기술이 발전하고 있음

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 4. 빅데이터 활용을 위한 인력 요소

**인력**

**데이터 사이언티스트(Data Scientist)**

대규모 데이터 속에서 숨겨진 정보를
찾아내 제품이나 서비스를 개선하는 최고의 인재

빅데이터 경쟁 시대에서 데이터를 관리하고 분석할 수 있는 인력의
중요성이 높아짐

정부와 기업들은 경쟁적으로 '데이터 사이언티스트 모시기'에 나서고 있음

## ▣ 빅데이터 활용을 위한 자원, 기술, 인력 요소

### 4. 빅데이터 활용을 위한 인력 요소

**인력**

**글로벌 IT 업체와 데이터 사이언티스트**

| 데이터 사이언티스트 확보 | 내부 역량 강화 |
|---|---|

· 고객 데이터를 분석하고 가공하는 일을 맡은 직원만 5,000명에 이름

· '애널리틱스' 부서 운영

➢ 경제학, 통계학, 심리학 등을 전공한 박사급 인재들인 데이터 사이언티스트로 구성됨

· 사내에 200명 이상의 수학자들 보유

➢ '분석학(Analytics)'을 집중연구하고 있음

➢ 500개 이상의 관련 특허를 취득하면서 미래 사업을 준비하고 있음

**데이터 사이언티스트의 현황과 미래**

현재 데이터 사이언티스트의 역량을 갖춘 인재는 매우 부족함

➢ 미국에서 2018년까지 14만 명에서 19만 명의 전문가와 150만 명 정도의 데이터 관리자와 분석인력이 부족할 것이라고 전망함(2011, 매킨지)

데이터 사이언티스트는 21세기 유망직업

➢ 수요 급증이 예상되고, 기업 내에서도 중요한 역할을 담당할 것임

➢ 데이터 처리와 분석능력을 갖춘 인력은 IT분야 뿐만 아니라 대부분의 기업과 조직에서 필수적으로 확보해야 할 핵심인력으로 분류하고 있음

실무에 적용 가능한 Big Data 분석 개론

# 빅데이터 활용요소와 데이터 사이언티스트


한국기술교육대학교
온라인평생교육원

## ▣ 데이터 사이언티스트의 역량과 조건

### 1. 여러 업계에서 보는 데이터 사이언티스트의 필요 역량

**존 라우저가 제시하는 데이터 사이언티스트 자질**

- 데이터 사이언티스트는 승리를 좌우할 핵심 인재로 평가 받고 있음
- 여러 기업과 정부는 역량을 갖춘 데이터 사이언티스트를 찾는 데 많은 노력을 기울이고 있음

데이터 사이언티스트가 가져야 할 자질로는 6가지가 있습니다.


존 라우저(아마존 수석 엔지니어)

1. 수학
2. 공학능력
3. 데이터 분석을 위한 가설수립 및 검증에 필요한 비판적 시각
4. 이를 잘 작성할 수 있는 글쓰기 능력
5. 다른 사람에게 잘 전달할 수 있는 대화능력
6. 호기심과 개인의 행복

## ▣ 데이터 사이언티스트의 역량과 조건

### 1. 여러 업계에서 보는 데이터 사이언티스트의 필요 역량

**한국정보화진흥원이 제시하는 데이터 사이언티스트 역량**

NIA 한국정보화진흥원

**시각화 역량도 중요하다!**

| | |
|---|---|
| 시각화 | 데이터 분석 결과를 전달하는 마지막 단계로서, 데이터의 문맥화를 통한 해석 작업에 해당함 |

정교한 모형 및 시각화 도구를 활용하면
더 큰 비즈니스 가치와 통찰력을 제공할 수 있음

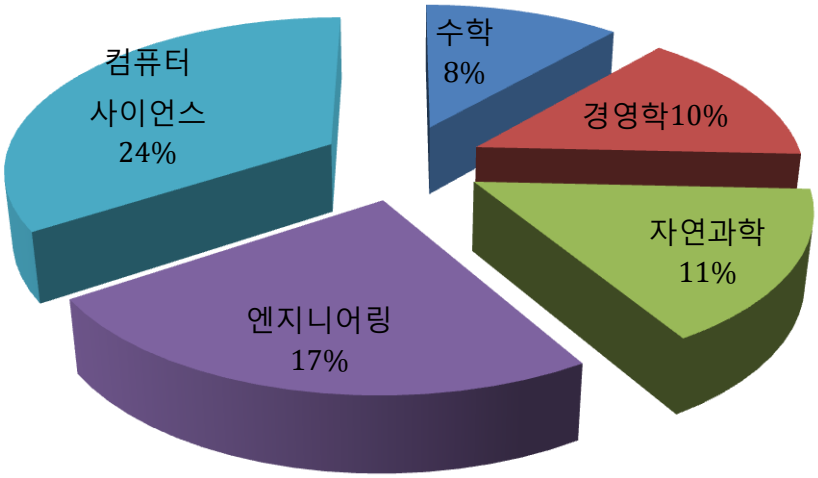
## ▣ 데이터 사이언티스트의 역량과 조건

### 1. 여러 업계에서 보는 데이터 사이언티스트의 필요 역량

**EMC 컨퍼런스가 조사한 데이터 사이언티스트 전공**

EMC² where information lives®

**정보관리와 저장 소프트웨어 및 시스템 생산업체로**
**데이터 사이언티스트에 대한 높은 관심**



## ▣ 데이터 사이언티스트의 역량과 조건

### 1. 여러 업계에서 보는 데이터 사이언티스트의 필요 역량

**MIT Slon**


**업계에서 활동하는 경영자, 관리자, 분석가를 대상으로**
**조직에게 가치를 제공하는 최고의 분석기법 조사**

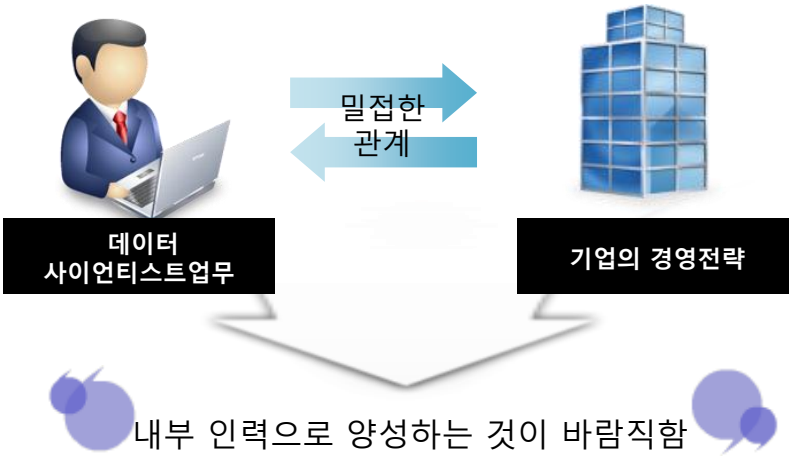
현재는 과거 추세 분석 및 예측 기능이 중요함

향후에는 데이터를 **시각화**하는 역량의 가치가 상승함

- 한눈에 표현할 수 있는 정보 표현 역량을 갖추어야 함

## ▣ 데이터 사이언티스트의 역량과 조건

### 2. 데이터 사이언티스트 양성




기업의 사업 기회를 찾아내고, 전략적인 통찰력을 발휘하기 위해서 조직과 비즈니스에 대한 충분한 이해가 필수적임

보안상 중요한 조직의 데이터나 비공개 데이터를 분석하기 위해서는 내부 인력을 활용하는 것이 적합함