# 데이터 관리 및 시각화

## 화공 데이터의 관리

화공 산업 현장에서 발생 및 수집된 데이터는 곧바로 제어 · 운영 · 설계 등 공정에 관한 의사결정 모델에 직접 사용되지 못한다. 전형적인 복잡계 문제인 화공 산업의 다양한 의사결정 문제를 해결하기 위해서는 수집된 데이터로부터 핵심적인 특징추출 (Feature extraction) 과정이 필요하다. 즉, 데이터 근원으로부터 수집된 원시 데이터 (Raw data) 는 사용처에 특화된 인공지능 모델 개발을 위해 데이터 선별과 전처리 과정 등 다양한 데이터 추출 단계를 필수적으로 수반한다. 본 장에서는 수집된 데이터를 사용 단계에 적합한 수준으로 가공 및 가시화하는 과정을 학습하며 더불어 효과적인 저장 방법을 소개한다.

### 데이터 큐레이션 (Data curation)

실험 및 공정 운전 등에서 확보된 원데이터 (raw data)을 문제 해결을 위한 정보원료로 사용하기 위해서는 원데이터의 전처리가 반드시 필요하다. 특히, 수많은 정제되는 않은 원데이터로부터 무제 해결에 적합한 데이터를 선별하는 이른바 데이터 큐레이션 작업은 빅데이터 가치 사슬 (Big data value chain)에서 선행되는 단계이다. 데이터 큐레이션은 데이터의 수집 · 분석 · 저장 등 은 데이터 이용을 위해 데이터의 가치 발견을 위한 과정으로, 데이터로부터 정보를 부각시키는 역할을 한다 (그림 1). 이를 위해 데이터의 수집 · 조직화 · 인덱싱 · 그룹화가 수행되며, 용도에 따라 원시 데이터는 데이터 사용을 위해 소그룹으로 구성 및 통합된다. 데이터 큐레이션을 거친 데이터는 연계성과 연관성을 지니므로 사용자들의 접근성을 향상시킨다.

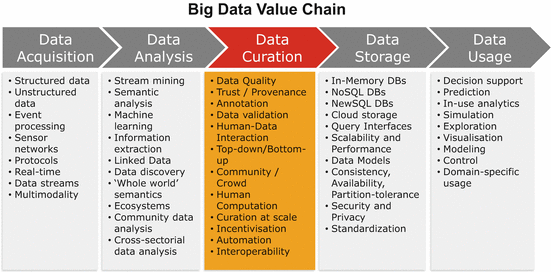


그림 1. 데이터 확보, 처리 및 저장의 순서도.

### 데이터 베이스 (Database)의 구축

데이터 베이스는 데이터 저장소를 의미하며, 특정한 목적에 맞게 정돈된 데이터로부터 정보를 추출하기 위해 필요하다. 데이터는 보통 가공되지 않은 원시적인 자료를 의미하고, 특정한 수집 의도가 존재하지 않는다. 따라서, 정보 추출을 위해 체계적이지 않는 자료들의 연계성과 연관성을 찾아 그룹화할 필요가 있다. 이를 위해, 데이터는 위와 같은 분류 · 선별 · 저장 단계를 거쳐 관리 · 이용된다. 각 단계에 대한 설명은 아래와 같다.



그림 2. 효과적 데이터 저장을 위한 데이트베이스 구조.

* 데이터 근원 (Data source)

화공산업에서의 데이터는 센서 네트워크 및 웹 스크래핑으로부터 수집된 물리량/정보량 (온도, 압력, 농도 등) 이다. 데이터가 잠재되어 있는 추상적인 장소를 데이터 근원이라 하며, 보통 데이터는 비정형화된 형태에서 변환과정을 거쳐 정형화된 형태로 수집된다. 데이터 근원으로부터 추출된 데이터는 곧 바로 사용되는 것이 아닌 데이터 웨어하우스에 축적된다. 이때 데이터 축적을 위해 추출 · 운송 · 축적 (Extraction, Transportation, Load; ETL) 기술을 사용한다. 데이터 축적을 위해 데이터 간의 관계성 (ex 독립 및 종속 관계) 을 규명하는데, 이는 데이터 스키마 (Schema) 로 불린다.

* 데이터 웨어하우스 (Data warehouse)

데이터 웨어하우스는 데이터 근원으로부터 추출 · 운송 · 축적 과정을 거친 데이터 저장매체이다. 사용자의 의사결정을 돕기 위해, 데이터는 최초로 데이터 웨어하우스에서 분석 가능한 형태를 갖는다. 데이터 웨어하우스에는 크게 4가지 특징으로 주제지향성, 통합성, 시계열성, 그리고 비휘발성이 있다. 먼저, 주제지향성은 데이터 웨어하우스가 직접적인 사용보다는 데이터의 관계성 파악에 더 중점적이라는 것에서 기인한다. 따라서, 특정 주제 (반응기/분리기, 온도 변화/압력 변화 등)에 따라 데이터를 표현하는 것이 데이터 웨어하우스에서 중점적으로 이루어진다. 통합성은 데이터 웨어하우스의 스키마가 일관성을 유지한다는 것으로, 데이터 근원으로부터의 데이터가 모두 전사적으로 통합되어 있다는 의미이다. 시계열성은 데이터 웨어하우스의 수집된 데이터가 시간에 따른 갱신없이 보관한다는 의미이다. 마지막으로, 비휘발성은 데이터 웨어하우스에 한번 적재된 데이터는 삭제 혹은 변경 작업이 발생하지 않는다는 의미이다. 데이터 웨어하우스의 데이터는 이후 목적에 맞게 더 작은 규모의 저장소인 데이터 마트로 이동한다.

* 데이터 마트 (Data mart)

데이터 마트는 특정 목적을 위해 정돈된 작은 규모의 저장소이다. 보통 데이터 마트 단계에서 데이터 분석에 용이하게 데이터가 재구성되어 있으며, 유연성과 접근성이 뛰어난 다차원 구조를 갖는다. 데이터 마트는 데이터를 용도 혹은 역할로 세분화함으로써 유저의 정보 접근성을 향상시킬 수 있다.

* 운용 데이터 베이스 (Operating database)

운용 데이터 베이스는 직접적인 인공지능 기반 예측/분류 모델 개발을 위해 사용되는 데이터 저장소이다. 사용자에 가장 가깝게 존재하며, 정보와 서비스 생성을 위해 데이터 마트로부터 필요한 데이터의 추출을 통해 생성된다.

### 데이터 저장 매체

데이터 베이스 관리 시스템 (Data Base Management System; DBMS)은 데이터 베이스를 체계적으로 관리하는 시스템이다. DBMS는 사용자의 요구에 맞춰 정보를 생성하고 데이터 베이스를 편리하게 관리하는 소프트웨어의 종합을 의미한다. 데이터의 종속성과 중복성 문제들을 해결하기위해 제안되었고, 물리적으로 데이터를 저장하는 장소이다. DBMS의 예시는 다음과 같다.

- Amazon Aurora

- Azure Synapse Analytics

- Google Cloud

- Microsoft SQL Server

- Mongo DB

- PostgreSQL

- Presto

- Hadoop ecosystem



그림 3. 주요 대용량 데이터 자장 플랫폼.

## 화공 데이터의 가시화

화공 산업에서의 데이터는 다양한 환경 변수들이 존재하여 복잡한 고차원의 형태를 갖는다. 사람이 인식할 수 있는 차원은 한정되어 있고, 고차원의 독립 변수-종속 변수에 대한 관계 · 추세 파악은 단순한 선형적 구조로 표현하기 어렵다. 따라서, 데이터로부터 인사이트를 얻기 위해서는 창의적이고 다채로운 데이터 시각화 방법이 필요하다. 이를 통해 데이터의 수많은 패턴을 파악함으로써 직관적이고 명확한 공정 제어 · 운영 · 설계 등의 의사결정을 도울 수 있다. ‘화공 데이터의 가시화’는 화공산업에서 자주 발생하는 데이터들의 효과적인 시각화 예시를 소개한다.

### 데이터 시각화의 절차



그림 4. 데이터 시각화 시 주요 확인 사항.

한국전자통신연구원 (Electronic telecommunications research institute; ETRI)은 데이터 시각화를 위한 그림 4의 세가지 주요 절차를 제안했다. 많은 양의 데이터로부터 효율적인 정보 제공에 시각화는 필수적이다. 전통적인 데이터 시각화는 단순히 독립 변수와 종속 변수 간의 일차원적인 관계성만을 보여줬다. 하지만, 화학공정은 다양한 독립 변수-종속 변수 관계성 이 존재하므로 더 높은 수준의 데이터 시각화가 필요하다. 정보 구조화-정보 시각화-정보 시각표현으로 이루어지는 데이터 시각화의 절차는 더 높은 차원의 데이터 표현을 위한 고찰 과정이다.

1) 정보 구조화

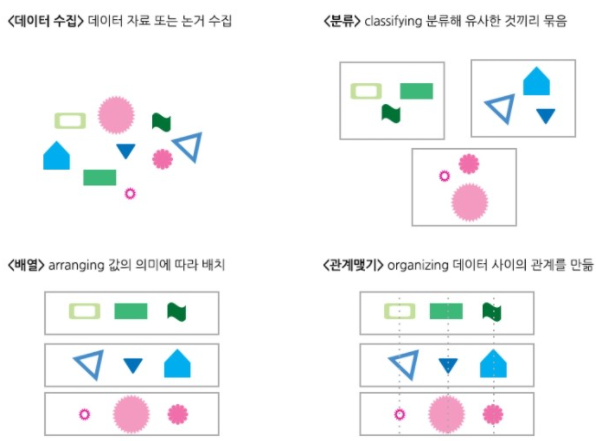


그림 5. 데이터 및 정보 구조화 과정.

정보 구조화는 사용자가 정보를 쉽게 인식시키도록 수집된 데이터의 조직적이고 체계화된 자료 분류 · 배열 과정이다. 위의 그림 5와 같이 무작위 · 무질서한 원시적인 데이터 자료들을 효과적으로 구조화할 수 있다.

- 데이터 수집 (Acquisition): 데이터 근원으로부터 데이터 자료들을 수집한다.

- 분류 (Classification): 사용자의 판단에 따라 유사한 자료 (색깔, 형태, 종류 등) 들을 묶어 정리한다.

- 배열 (Arrangement): 분류된 자료들의 의미 파악을 통해 배치 순서 · 종속 관계를 파악하여 재배열한다.

- 관계맺기 (Organization): 특정 데이터의 획득 및 자료 그룹 사이의 관계와 분류된 자료 사이의 관계를 각각 규명한다.

2) 정보 시각화

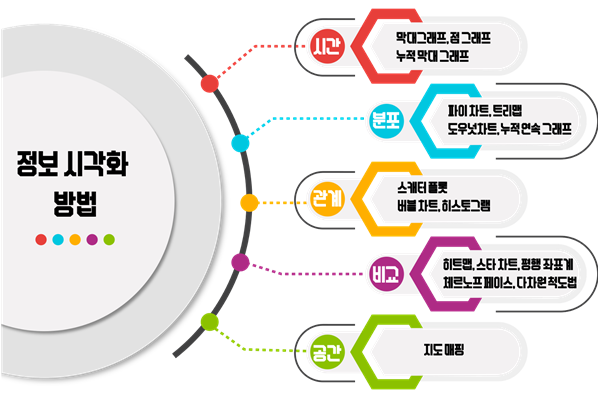


그림 6. 시간, 관계, 공간 등 데이터 및 정보 시각화 방법.

정보 시각화는 시간 · 공간 · 수량 · 비수량을 색체 · 형태 · 통계 · 이미지 등을 통해 표현한다. 분석을 위해 도표나 그래프를 활용하는데, 적절한 시각화 방법을 위해 정보의 특성 파악이 필요하다. 시각화 방법으로는 시간 시각화, 분포 시각화, 관계 시각화, 비교 시각화, 그리고 공간 시각화가 있다. 적절한 시각화 방법은 빠른 의사결정을 돕고 의사결정자들 사이의 커뮤니케이션을 효과적으로 만든다. 아래와 같은 차트들은 전통적으로 사용되는 데이터 시각화 형태로, 단독적으로도 많은 정보를 설명할 수 있지만, 효과적인 전달을 위해선 추가적인 시각 표현이 필요하다.



그림 7. 전통적인 데이터 및 정보 시각화 그래프.

3) 정보의 시각화 표현

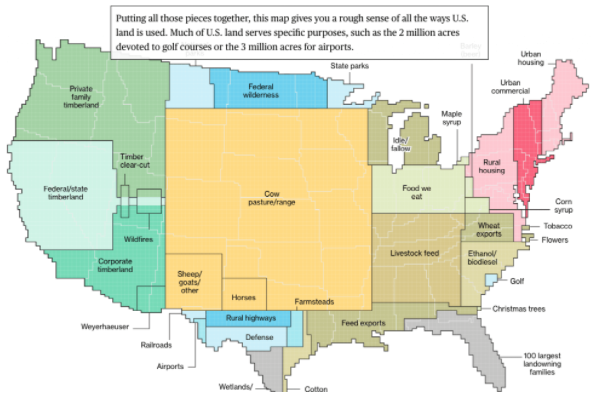


그림 8. 공간 구조 표현을 위한 데이터 및 정보 시각화: 예, 미국의 주요 토지 사용처

정보의 시각화 표현은 단순한 정보 시각화를 넘어 별도의 그래픽적인 요소를 추가해 정보를 완성시키는 과정이다. 정보 시각화는 통계량과 차트를 통해 데이터를 설명할 수 있지만, 사용자의 입장을 충분히 더 고려한다면 실제 영향을 주는 대상과 통계량을 결부시켜 표현할 수 있을 것이다. 이를테면 미국에서 국토 (7,356,027 km2) 중 단지 5% 미만이 미국인을 위한 농업 지역 (312,822 km2) 이라는 점은 매우 흥미로운 정보이다. 하지만, 이는 단지 수치적일 뿐이며 말하고자 하는 내용을 위해 ‘농업 부지를 늘려야한다.’ 혹은 ‘많은 국토가 낭비된다.’같은 추가적인 해석이 더 필요하다. <미국의 주요 토지 사용처>는 훌륭한 정보의 시각화 표현 예시이다. 목축업지 (Cow pastor/range)를 미국 지도 중앙에 위치하여 효과적으로 강조했다. 또한, 토지의 사용처에 따라 비슷한 색깔을 사용해 효율적인 정보 전달이 가능하다. 결국, 효과적인 정보 시각화 표현을 통해 정보를 읽는 사용자들의 판단 및 의사결정을 가속화했다.

* Benjamin Fry의 데이터 시각화 단계

비슷하게, 효과적인 데이터 시각화를 위해 MIT 미디어랩의 Benjamin Fry는 7가지의 데이터 시각화 단계를 제시했다. 그는 데이터로부터 중요한 정보를 추출하기 위해선 근본적인 질문 (데이터 수집 목적, 주요 관심사, 예상 결과 등)을 구성해야 한다고 주장했다. 데이터 분석을 위한 질문들은 데이터 시각화를 위한 줄거리로 작용하여, 정보 전달과 흥미로움을 갖게 할 수 있다는 까닭이다.

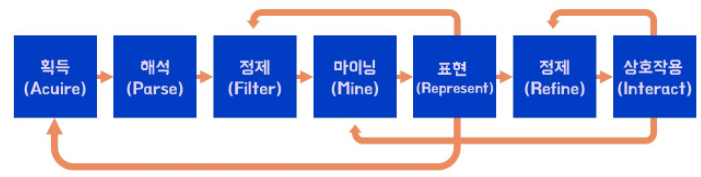


그림 9. 효과적 데이터 및 정보 시각화를 위한 단계.

- 획득 (Acquire): 특정 데이터의 획득

- 해석 (Parse): 데이터의 분류 및 구조화

- 여과 (Filter): 주요 관심사가 인덱싱된 데이터만을 추출

- 마이닝 (Mine): 데이터 분석 (통계적, 머신러닝 등)을 통한 정보 획득

- 표현 (Represent): 차트 혹은 그래픽을 통한 정보 표현

- 정제 (Refine): 정보의 가독성 증가를 위한 그래픽화

- 상호작용 (Interact): 사용자의 관심사와 필요에 따른 데이터 편집 가능성

해당 데이터 시각화 단계는 모든 과정에서 순차적이지 않고 명확한 정보 전달을 위해 반복적인 과정을 거치는 것이 중요하다.

### 화공 산업: 정보의 시각화 표현 예시

화공 산업에서의 데이터 시각화는 재료 · 기술 · 운송 · 공정 · 수익 등의 공정의 조작 변수와 결과물을 표현하는 형태이다. 그래픽 개요도와 인포그래픽은 대표적인 형태의 데이터 시각화의 예시이다.

* 그래픽 개요도 (Graphical abstract)



그림 10. 인포그래픽를 이용한 데이터 및 정보 시각화: 예, 바이오매스를 이용한 연료 생산 전략 최적화 프레임워크 설계 개요도 (DOI: https://doi.org/10.1039/C3EE24243A)

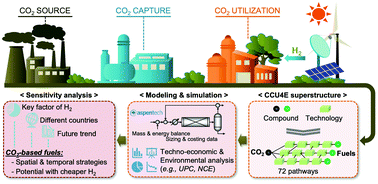


그림 11. 인포그래픽를 이용한 데이터 및 정보 시각화: 예, CO2를 이용한 액체연료 생산 경로 분석 설계 개요도 (DOI: <https://doi.org/10.1039/D1EE01444G>)

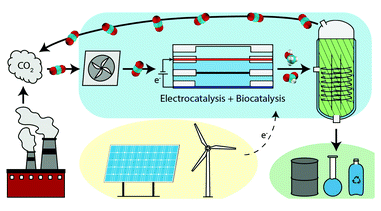


그림 12. 인포그래픽를 이용한 데이터 및 정보 시각화: 예 CO2 저감 및 활용을 위한 전기적 촉매 및 바이오 촉매 제작 개요도 (DOI: https://doi.org/10.1039/D1EE03753F)

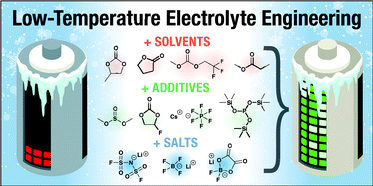


그림 13. 인포그래픽를 이용한 데이터 및 정보 시각화: 예, 리튬이온 배터리의 저온 활용을 위한 액체 전해질 개발 개요도 (DOI: https://doi.org/10.1039/D1EE01789F)

* 인포그래픽 (Infographics)

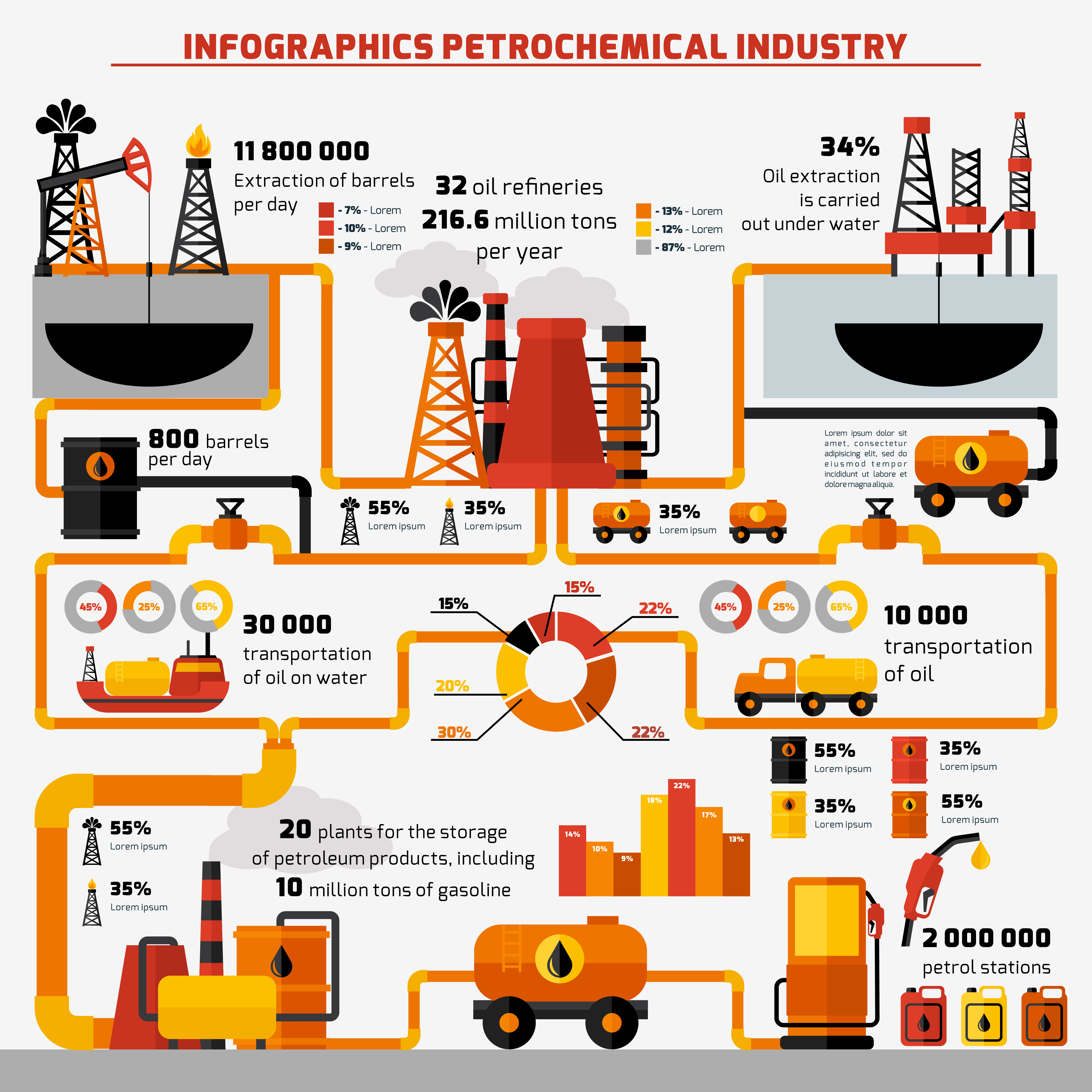


그림 14. 인포그래픽를 이용한 데이터 및 정보 시각화: 예, 석유 시추와 활용에 대한 통계적인 수치 표현 (“Microvector” from Freepik.com).

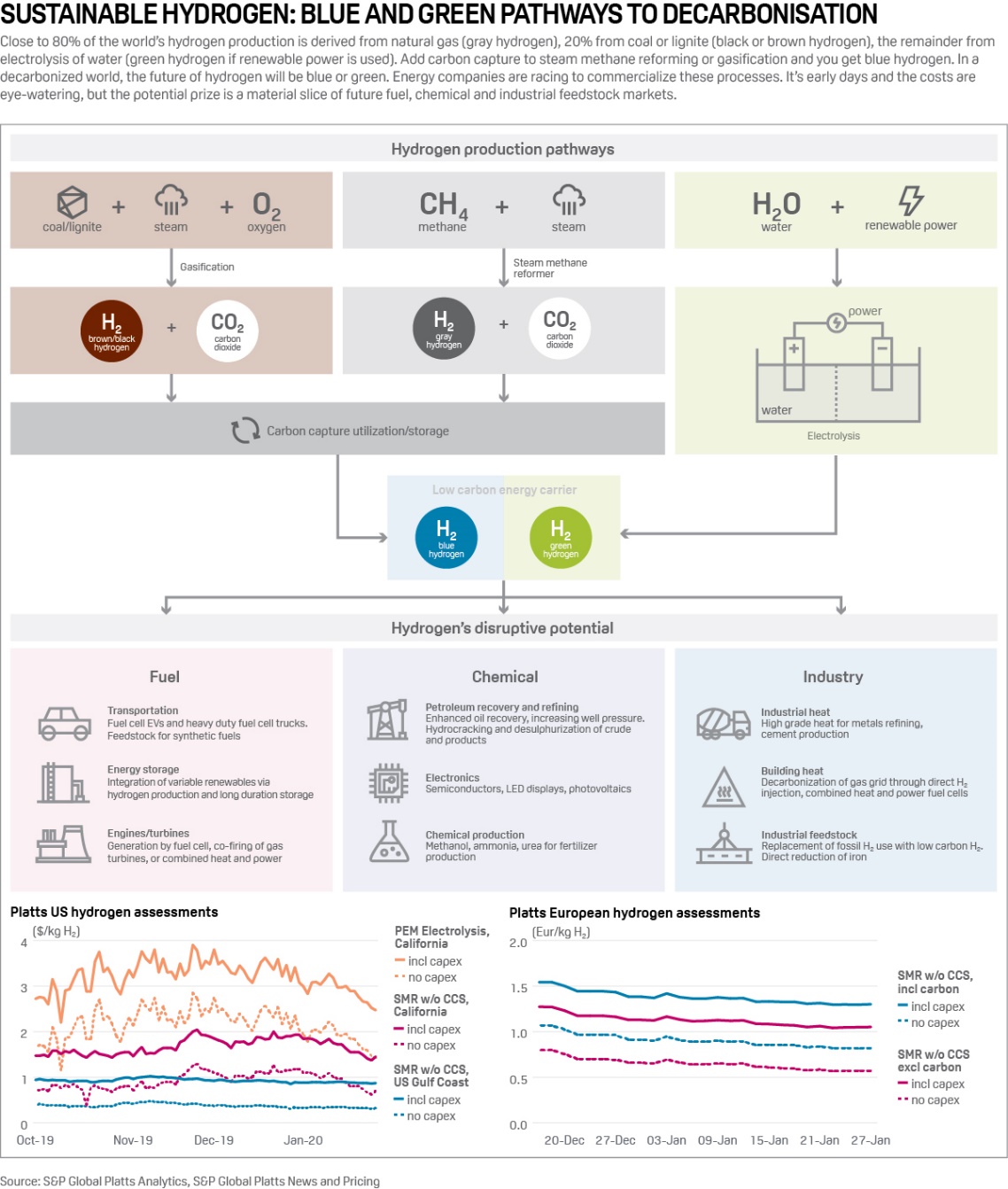


그림 15. 인포그래픽를 이용한 데이터 및 정보 시각화: 예, 수소 생산 및 활용에 관한 기술과 사용처 및 예상 생산 단가 수치 표현 (“Infographics: Sustainable hydrogen: blue and green pathway to decarbonization” from Spgobal.com).

## 학습 결과

* 학습 내용

화공 데이터 관리 및 가시화를 위한 데이터 저장 매체와 시각화 방법론

* 학습 결과 확인하기

대표적 복잡계인 화공 데이터의 효과적 관리를 위한 시각화 절차와 방법을 익히기.

* 학습 결과 응용하기

본 장의 학습내용에 기반해 화공 산업에서 발생하는 데이터를 적절히 매체를 사용해 처리 · 저장 역량을 확보함으로, 실제 화공 산업 현장에서 발생되는 다양한 빅데이터를 효과적으로 수집, 관리, 시각화에 응용