* 주의해야 할 데이터 이슈
* 데이터의 활용이 크게 늘어나고 있음
* 그에 따라 다양한 데이터 관련 이슈가 함께 발생함
* 1. 무분별한 개인 정보 확산

개인정보가 한번 Table에 노출되면 그 Table 기반으로 생성된 다른 Table로 전파,

불필요한 악순환 발생

해법) 개인정보가 무엇인지 명확하게 정의(PII)하고,

어떤 데이터가 ETL되어 시스템에 들어오는 단계부터 개인정보를 Tagging해야 함

개인 정보에 접근할 수 있는 권한을 제어하고 누가 언제 접근했는지 기록을 남겨야(logging)

특정 개인 정보 추출과 삭제를 자동화해야 함

* 2. 같은 데이터에 대한 여러 다른 해석의 발생

지표를 정확히 정의할 필요가 있음.

데이터 사전/데이터 용어집을 통해 사람들마다 지표를 자의적으로 해석하는 것을 방지

Data Source, Filtering 등의 조건을 명확화

Source of Truth를 기록: 각 지표 계산에서 바탕이 되는 데이터를 규정

Being Consistent is more important than being correct

* 3. Dashboard의 범람

Data Democratization이 정보 과잉으로 이어져

Data Discovery 이슈: 어느 테이블을 봐야 원하는 데이터를 볼 수 있나요?

Data Infra 비용 증가: 비슷한 정보 반복처리가 빅 데이터 규모에서 엄청난 비용 증가로

해법) Data Catalog의 도입

전사 내 어떤 데이터가 어디서 발생하고 어디서 소비되는지 한눈에 볼 수 있도록

자동화를 통해 Meta Data 관리, Dataset마다 Owner 지정, 주기적으로 Dashboard Cleanup 수행

* 4. Meta Data의 부족

특정 Dataset의 경우 누가 주인인지, 어떻게 생성된 데이터인지 알 수 없음 (Data Lineage의 부재)

모든 데이터의 Meta Data를 관리할 필요는 없지만, 중요한 데이터들의 Meta Data를 잘 관리할 필요가 있다.

* 5. 불분명한 데이터 소유권(Data Ownership)

데이터 양이 증가하며 품질 이슈가 발생: Dataset의 불분명한 소유권과 밀접하게 관련

Meta Data 문제와 연관 있음.

* 6. Data Silo

데이터를 중복 수집, 처리나 같은 데이터를 분석했는데 전혀 다른 수치 도출하는 일이 발생, 데이터 독점의 권력화 등 앞서 말한 이슈들이 증폭되며 Data Silo 현상 발생

해법) 조직 전반에 걸친 Meta Data 관리/유지 및 Data Mesh 도입

* 개인 정보 보호 및 관련법
* 개인 정보란? 개인을 식별할 수 있는 정보(PII, Personally Identifiable Information)

1) 식별자: 개인을 식별할 수 있는 정보. (성명, E-mail, 전화번호, 주소, 신용카드 정보 등)

2) 준식별자: 몇 가지 조합을 통해 개인을 식별할 수 있는 정보 (나이, 거주지, 직장 등)

* 한국 개인정보보호법, 정보통신망법

**살아 있는 개인**에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통해 **개인을 알아볼 수 있는 정보**(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 **다른 정보와 쉽게 결합하여 알아볼 수 있는 경우도 포함**)를 말한다 (개인정보보호법 제2조 1항)

생존하는 개인에 관한 정보로서 성명, 주민등록번호 등에 의하여 특정한 개인을 알아볼 수 있는 부호, 문자, 음성, 음향 및 영상 등의 정보(해당 정보만으로는 특정 개인을 알아볼 수 없어도 다른 정보와 쉽게 결합하여 알아볼 수 있는 경우 포함)를 말한다 (정보통신망법 제2조 6항)

* 미국 캘리포니아주 CCPA → CPRA (California Privacy Rights Act)

Information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a **particular consumer or household**.

인터넷 검색 기록까지도 개인 정보에 포함하고 있다.

2020년 1월 1일부터 효력이 발효되었고, 개인뿐만 아니라 가구까지 확대, 포괄.

이로 인해 미국 회사들의 Compliance 관련 Data Governance 활동이 증가했다.

2023년 1월 CPRA로 개정되며 GDPR과 비슷하게 개인의 정보주체권을 강조한다.

* 유럽연합 GDPR (General Data Protection Regulation)

‘Personal Data’ shall mean any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

EU 사용자가 이용하는 모든 서비스를 대상으로 하는 법률

GDPR 위반시 벌칙: 동 법령 위반시 과징금 등 행정처분 대상이 됨

Lv1: 일반적 위반사항 (전 세계 매출 2% 또는 1천만 유로 중 높은 금액을 과징금 처분)

미성년자 대리인 미지정 위반, 개인정보 유출 내역 통지 위반,

개인정보 처리활동 기록 위반 등

Lv2: 중요한 위반사항 (전 세계 매출 4% 또는 2천만 유로 중 높은 금액을 과징금 처분)

국외이전 규정 위반, 개인정보 처리 기본원칙 위반,

정보주체의 권리 보장 의무 위반 등

* 개인 정보 보호

개인의 정보를 적절한 동의 없이 저장하거나 사용하지 않는 것

개인의 정보를 적절한 동의 없이 노출하거나 배포하지 않는 것

국제적으로 개인 정보 보호를 위한 다양한 법률이 입법되고 있으며, 해당 법률을 준수하기 위해 Data Catalog와 Data Governance가 도입되고 있음. 또한 개인의 정보주체권 – Data as a property right – 이 중요시되고 있음

개인 정보 노출: 내부 유출 vs 외부 공격

내부자들의 실수 등으로 인한 Data Leak

외부 사이버 범죄조직이나 국가(러시아, 북한 등) 해커들의 해킹, 랜섬웨어 협박

**<개인 정보 보호 Checklist)**

보호가 필요한 중요 정보가 무엇인지,

가지고 있는 정보가 정말 필요한 정보인지,

이 정보에 대한 접근이 정말 필요한 사람이 누구인지,

이 사람들이 정말로 해당 정보가 필요할 때만 접근하는지

이러한 접근들이 모두 기록되는지, 그리고 그러한 기록들이 주기적으로 감사를 받는지

* Excel의 구성요소
* SUM, COUNTA, AVERAGE, MIN, MAX
* RANGE
* IF
* COUNTIF
* SUMIF, AVERAGEIF
* SUMIFS
* 논리검사
* 데이터 필터링
* 틀고정
* OR
* ABS
* AND
* OR + OR / AND + OR