* 데이터는 우리 생활 모든 곳에 존재
* 시작은 Digitization부터. 온라인 환경에서 데이터 수집이 용이
* 웹과 모바일 사용 증가로 데이터 크기의 폭발적 성장: Big Data, Cloud, 분산처리
* Big Data Processing
* Google: Big Data 기술 발전에 지대한 공헌
* Web 사용자 검색어와 클릭 정보 데이터 마이닝으로 개인화 및 별도 서비스 개발
* Web 자체를 NLP[[1]](#footnote-1) 거대 모델 개발의 훈련 데이터로 사용
* 데이터 팀의 역할(1)
* 데이터 조직의 미션은 **신뢰**할 수 있는 데이터를 바탕으로 **부가가치**를 생성하는 것

Data is the new oil, but it can be a liability. 데이터 오용으로 인한 위험 줄여야 함

* 고품질 데이터 바탕으로 의사 결정권자에게 데이터 고려(data-informed)/데이터 기반(data-driven) 결정을 가능하게 해 줌.

데이터 기반 지표 정의, 대시보드/리포트 생성

* Data Literacy
* 데이터 팀의 역할(2)
* 고품질 데이터를 기반으로 사용자 서비스 경험 개선, 프로세스 최적화(Product Science)
* ML 등 데이터 기반 알고리즘을 통해 고객 서비스 경험 개선, 공정 과정 최적화
* 데이터의 흐름
* 데이터 발생 – 데이터 인프라에 저장 – 데이터 분석 – 데이터 과학 적용
* 데이터 팀의 발전 단계
* Data Infrastructure 구축: Data Engineer가 수행.

Data Analyst/Scientist가 존재하지 않는 상황에서 Data를 ETL(추출/전송/적재)

다루는 데이터의 크기가 커지면 Big Data Processing 기술 습득 필요. Apache Spark

* Data Warehouse: 기업이 필요한 모든 데이터를 모아놓은 중앙 SQL RDB

일반적으로 Cloud 기반으로 구축되며, 데이터 분석용 DB가 아니라 기업의 서비스용 DB다. 온갖 종 류의 정보를 모두 담고 있다. 보존 기한이 있으며 Structured Data 위주로 저장, 처리한다.

Cloud 기반: AWS Redshift, Google Cloud BigQuery, Snowflake 등

Open-source 기반: Hadoop 기반으로 만들어진 Hive, Presto 등

* Data Lake

Structured Data뿐만 아니라 Semi-Structured Data, Unstructured Data까지 모든 데이터를 원형 그대로 보존. Data Warehouse보다 대규모. 보존 기한 없음.

* Data Analysis: Data Analyst들이 지표 정의, 시각화, 리포트를 통해 의사 결정권자들을 지원.

Data Engineer들이 ETL을 통해 Table의 형태로 Data Warehouse에 적재한 Data들을 Analyst들 이 다시 조합해서 사용하기 쉽게 가공(ELT)하여 새로운 정보를 생성해 지표 정의, 대시보드 생성/관 리, 데이터 기반 리포트 작성에 사용한다.

* Key Performance Indicators (KPI)
* Visualization / Dashboard

중요한 지표를 시간의 흐름과 함께 보여줌. 3A(Accessible, Actionable, Auditable)

Salesforce Tableau, Microsoft Power BI, Google Cloud Looker

* Data Science 적용

Data Scientist가 ML 모델을 만들어 가치 생성(예. 사용자 행동 분석을 통해 추천이나 검색 개인화)

* Cloud 기술
* Cloud: Computing Resource를 Network을 통해 서비스 형태로 사용하는 것

No provisioning 준비할 필요가 없다 / Pay As You Go 사용한 만큼 지불한다

자원을 필요한 만큼 실시간으로 할당 받아 사용하고 그만큼 지불하는 탄력적인 시스템

* Cloud Computing 덕분에 비용 절감 효과를 누릴 수 있다.

초기 투자 비용 감소, 리소스 준비 대기시간 대폭 감소, 글로벌 확장 용이, 개발 시간 단축

* Amazon Web Services
* AWS는 가장 큰 Cloud Computing 서비스 업체.

현재 100여개의 서비스를 전세계 15개 지역에서 제공(서울 포함)

대부분의 서비스는 오픈 소스 프로젝트를 기반으로 하며 최근 ML/AI 서비스도 추가됨

* Compute, Storage, Database, Network/Content Delivery, Developer Tools, Management Tools, Security, Analytics, AI, IoT, Game, Mobile, API 등 다양한 서비스를 제공하고 있다.

EC2(AWS의 서버 호스팅 서비스, 다양한 종류의 서버/가상서버를 제공)

S3(Simple Storage Service, AWS가 제공하는 대용량 Cloud Storage 서비스)

그 외 RDS, DynamoDB, RedShift, MongoDB, Elastic Search 등등

AWS의 AI, ML 관련 서비스:

SageMaker(DL & ML end-to-end framework),

Lex(Chatbot), Polly, Rekognition, Comprehend(NLP → LLM Service) 등

* 데이터 조직 구성
* 조직에 따라 한 사람이 몇 개의 역할을 동시 수행

Data Engineer, Data Analyst, Data Scientist, Machine Learning Engineer, MLOps Engineer, Privacy Engineer, …

**기본 Skill**: Python, SQL

**Data Analyst에게 요구되는 Skill**: 지표 정의, 통계 지식, **A/B 테스트**, 데이터 모델링, 비즈니스 도메인, BI Tool(Tableau, Looker 등)

**Data Scientist에게 요구되는 Skill**: 가설 기반 접근, ML, Data 청소, EDA, Feature Engineering, A/B 테스트 디자인, 알고리즘, Hyper Parameter, 모델 관리, ML 개발 Framework(SageMaker, MLflow, Kubeflow 등)

**Data Engineer에게 요구되는 Skill**: ETL, Cloud, Big Data 분산처리기술(Spark, Hadoop), NoSQL 등

* Data Engineer

기본적으로 SW 개발자로, 데이터 인프라 구축 및 관리, Data Warehouse에 요청에 따라 Data Source 추가(ETL)

* Data Analyst

DW의 데이터를 기반으로 지표를 만들고 시각화, 대시보드 생성. dbt 등의 도구를 사용해 ELT 수행, 새로운 정보 만들어 기업 내부의 데이터 관련 질문에 응답.

* Data Scientist

과거 데이터를 기반으로 미래를 예측하는 ML 모델을 생성.

데이터 수집에 왜곡이 없는지, 개인정보가 사용되고 있는지, 모델 동작을 설명할 수 있는지 알아야 하며, 모델 개발 및 배포 과정이 자동화되어야 한다.(MLOps)

* Machine Learning Modelling Cycle

Training Data 수집 > Model 빌드 & 테스트 > Model Distribution > A/B Test > 테스트 결과 분석

결과가 좋으면 배포, 그렇지 못하면 다시 사이클 투입

* A/B Test

실제 사용자를 두 개의 그룹으로 나눠서 기존 버전과 새 버전을 온라인으로 테스트

* 데이터 문해력(Data Literacy)
* 데이터를 이해하고 활용할 수 있는 능력
* 데이터 기반 의사 결정과 제품 개선, 생산성 증대
* 데이터/IT 조직에서 시작해 회사 전체로 문해력 전파

데이터는 자산이라는 인식이 필요.

특정 조직이 아니라 기업 내 모든 조직이 데이터 활용, 점점 더 많은 인력이 역량 갖춤

* 생성형 AI와 같은 도구를 사용해 생산성 증대: 개발 뿐만 아니라 다양한 업무
* 데이터 관리

잘못 관리된 데이터는 커다란 위험 요소. 개인정보 보호 법안(GDPR, CCPA, HIPAA) 준수 필요.

* 데이터 교훈
* 데이터를 통해 매출이 생겨야 한다.
* 데이터 인프라 없이는 제대로 된 일을 하기 힘들다.
* 데이터의 품질이 중요하다: GIGO

Data Scientist들은 Data Cleanup에 가장 많은 시간을 쏟는다.

중요한 데이터의 경우 보다 품질 유지에 노력이 필요

* 항상 지표부터 생각해야 한다.
* 가능하면 간단한 솔루션에서부터 시작: 모든 문제를 DL로 해결할 필요 없다

1. NLP(Natural Language Processing), 자연어처리는 컴퓨터가 인간의 언어를 이해, 생성, 조작할 수 있게 해주는 AI의 한 분야입니다. <https://www.oracle.com/kr/artificial-intelligence/what-is-natural-language-processing/> [↑](#footnote-ref-1)