2과목 데이터 분석 기획

작성자 : 윤소영

ADsP 헬프 데스크

구분	링크
종합 정보 안내	https://eduatoz.notion.site/f55a34175dd14f42b14161140a27c1d2?v=6be282fd277c43e2a2bf014db0514fb0
1과목 QnA 정리문서	https://colab.research.google.com/drive/1cCk43pbnapr0mxOSiC8ssN3ya5TJwSqG
2과목 QnA	https://colab.research.google.com/drive/1J7Ul9W-bVobaAob0wQKIsho3Uo8HkeVq
3과목 R – QnA	https://colab.research.google.com/drive/1VmixW_RYpn8_XycGAjgqSCZ00sR-_8Ke
3과목 통계분석 – QnA	https://colab.research.google.com/drive/1QDuCKk86lKTP8oxOTwOo1D2935Tu-5-e
3과목 정형분석 – QnA	https://colab.research.google.com/drive/1_NOLfLHlYrmAxBXcpcqNV1pwHme9C4IO
NoSQL 읽기자료	https://meetup.toast.com/posts/274
과목별 요약 강의 듣기 (R은 시험대비만 가능)	https://youtube.com/playlist?list=PLnp1rUgG4UVZO4ndD_HITLiBb8GlrtUOI
추가 강의 듣기	https://youtube.com/playlist?list=PLnp1rUgG4UVaHL5KKWkJxpT02X7Fh6ggv

링크는 변경이 없으며, 내용은 매 시험 때마다 계속 추가 됩니다.

교재 하단의 페이지는 영상 강의 페이지와 맞춘 것입니다.
일련번호가 아님에 유의하세요!
영상과 일치하지 않는 페이지가 있으면 imbgirl@naver.com 으로 문의 주세요 ^^!

합격을 기원합니다!

02-01

데이터 분석 기획 - 데이터분석 기획의 이해

2-01. 분석 기획

분석 기획이란?

- 실제 분석을 수행에 앞서 분석을 수행할 과제의 정의 및 의도했던 결과를 도출할 수 있도록 이를 적절하게 관리할 수 있는 방안을 사전에 계획하는 일련의 작업
- 어떤 목표(what)를 달성하기 위해 어떤 데이터를 가지고 어떤 방식(how)을 수행할지에 대한 일련의 계획을 수립하는 작업
- 성공적인 분석 결과 도출을 위한 중요 사전 작업
- 해당 문제 영역에 대한 전문성 역량 및 통계학적 지식을 활용한 분석 역량과 분석 도구인 데이터 및 프로그래밍 기술 역량에 대한 균형 잡힌 시각을 가지고 방향성 및 계획을 수립해야 함

2-02. 분석 주제 유형 4가지

- 분석의 대상(what), 분석의 방법(how)에 따라 4가지로 구분한다

분석방법 (how) \ 분석대상 (what)	Known	Un-Known
Known	최적화(Optimization)	통찰(Insight)
Un-Known	솔루션(Solution)	발견(Discovery)

- Optimization : 분석 대상 및 분석 방법을 이해하고 현 문제를 최적화의 형태로 수행함
- Solution : 분석 과제는 수행되고, 분석 방법을 알지 못하는 경우 솔루션을 찾는 방식으로 분석 과제를 수행함
- Insight : 분석 대상이 불분명하고, 분석 방법을 알고 있는 경우 인사이트 도출
- Discovery : 분석 대상, 방법을 모른다면 발견을 통해 분석 대상 자체를 새롭게 도출함

분석 대상을 알면 (OS) 모르면 (ID)!

2-03. 목표 시점 별 분석기획 방안

- 과제 중심적인 접근방식의 단기방안, 마스터플랜 단위의 중장기 방안으로 구분

	과제 단위 당면한 분석 주제의 해결	마스터플랜 단위 지속적 분석 문화 내재화
1차 목표	Speed & Test	Accuracy & Deploy
과제의 유형	Quick - Win	Long Term View
접근 방식	Problem Solving	Problem Definition

→ 두가지를 융합적으로 적용하는 것이 바람직함

Quick - Win : 즉각적인 실행을 통한 성과 도출
프로세스 진행 과정에서 일반적인 상식과 경험으로 원인이 명백한 경우 바로 개선함으로써 과제를 단기로 달성하고, 추진하는 과정

2-04. 분석 기획 시 고려 사항

- 가용한 데이터, 적절한 유스케이스 탐색, 장애요소들에 대한 사전 계획 수립

가용한 데이터 (available data)	▪ 분석을 위한 데이터 확보 ▪ 데이터 유형에 따라 적용 가능한 Solution 및 분석 방법이 다름 ▪ 데이터의 유형 분석이 선행적으로 이루어져야 함 (정형, 비정형, 반정형)
적절한 유스케이스 탐색 (Proper Use-Case)	▪ 유사분석 시나리오 및 솔루션이 있다면 이것을 최대한 활용함
장애요소들에 대한 사전 계획 수립 (Low Barrier of Execution)	▪ 장애요소들에 대한 사전 계획 수립 필요 ▪ 일회성 분석으로 그치지 않고 조직 역량을 내재화 하기 위해서는 충분하고 계속적인 교육 및 활용방안 등의 변화관리가 고려되야 함

2-05. 데이터 유형, 저장 방식

- 데이터를 유형으로 분류하면 정형, 비정형, 반정형 데이터로 분류할 수 있다

정형 데이터	▪ ERP, CRM Transaction data, Demand Forecast
반정형 데이터	▪ Competitor Pricing, Sensor, machine data
비정형 데이터	▪ email, SNS, voice, IoT, 보고서, news

- 데이터 저장 방식

RDB	▪ 관계형 데이터를 저장, 수정, 관리할 수 있게 해주는 데이터 베이스, Oracle, MSSQL, MySQL 등
NoSQL	▪ 비관계형 데이터 저장소 ▪ MongoDB, Cassandra, Hbase, Redis
분산파일시스템	▪ 분산된 서버의 디스크에 파일 저장, HDFS

2-06. 분석 방법론 개요

20 22

분석 방법론의 필요

- 데이터 분석을 효과적으로 기업에 정착하기 위해
  데이터 분석을 체계화하는 절차와 방법이 정리된 데이터 분석 방법론 수립이 필요

분석방법론의 구성요소

상세한 절차, 방법, 도구와 기법, 템플릿과 산출물

기업의 합리적 의사결정 장애요소

고정관념, 편향된 생각, 프레이밍 효과(Framing Effect)

Framing Effect : 동일한 사건이나 상황임에도 불구하고 사람들의 선택이나 판단이 달라지는 현상으로, 특정 사안을 어떤 시각으로 바라보느냐에 따라 해석이 달라진다는 이론

2-07. 분석 방법론의 모델 3가지

17 18 19 23

폭포수 모델

- 단계를 순차적으로 진행하는 방법
- 이전 단계가 완료되어야 다음 단계로 순차 진행하는 하향식 진행
- 문제점이 발견되면 전단계로 돌아가는 피드백 수행

나선형 모델

- 반복을 통해 점증적으로 개발
- 반복에 대한 관리 체계가 효과적으로 갖춰지지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있음

프로토타입 모델

- 사용자 요구사항이나 데이터를 정확히 규정하기 어렵고 데이터 소스도 명확히 파악하기 어려운 상황에서 사용
- 일단 분석을 시도해보고 그 결과를 확인해가면서 반복적으로 개선해 나가는 방법
- 신속하게 해결책 모형제시, 상향식 접근방법에 활용

폭포-하향! 프로토-상향!!

2-07. 분석 방법론의 모델 3가지

프로토타입 모델

- '사용자들이 이렇게 만들면 편하게 사용할거야' 라는 가설을 생성하게 됨
- 프로토타입을 보고 완성시킨 결과물을 통해 가설을 확인할 수 있음
- 특정 가설을 갖고 서비스를 설계하고 디자인에 대한 실험도 실행함
- 시제품이 나오기 전의 제품의 원형으로 개발 검증과 양산 검증을 거쳐야 시제품이 될 수 있음
- '정보시스템의 미완성 버전 또는 중요한 기능들이 포함되어 있는 시스템의 초기 모델'

2-07. 분석 방법론의 모델 3가지 - 문제 4

4. 다음은 어떤 모델에 관한 설명인가?

반복을 통하여 점증적으로 개발, 처음 시도하는 프로젝트에 적용이 용이하지만, 반복에 대한 관리체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다

나선형 모델

- 반복을 통해 점증적으로 개발
- 반복에 대한 관리 체계가 효과적으로 갖춰지지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있음

KDD(Knowledge Discovery in Database) 분석 방법론

데이터베이스에서 의미 있는 지식을 탐색하는 데이터 마이닝 프로세스

분석 대상의 비즈니스 도메인에 대한 이해와 프로젝트 목표를 정확하게 설정

데이터셋 선택	
데이터 전처리	▪ 데이터셋에 포함되어 있는 잡음(Noise), 이상값(Outlier), 결측치 (Missing Value)를 식별하고 필요시 제거
데이터 변환	▪ 분석 목적에 맞는 변수 선택, 데이터의 차원 축소 ▪ 데이터 마이닝을 효율적으로 적용할 수 있도록 데이터셋 변경 작업
데이터 마이닝	▪ 분석 목적에 맞는 데이터 마이닝 기법 및 알고리즘 선택 ▪ 데이터의 패턴을 찾거나 분류 또는 예측 등의 마이닝 작업 시행
데이터 마이닝 결과 평가	▪ Interpretation/Evaluation, 분석 결과에 대한 해석과 평가, 활용

2-09. CRISP-DM 분석 방법론 - 1/3

16 18 21 *2

- 6단계로 구성, 일방향으로 구성되어 있지 않고 단계간 피드백을 통하여 단계별 완성도를 높이게 구성됨
  - Cross-Industry Standard Process for Data Mining

6단계 : 업무 이해 - 데이터 이해 - 데이터 준비 - 모델링 - 평가 - 전개

업무 이해 Business Understanding	▪ 비즈니스 관점 프로젝트의 목적과 요구사항을 이해하기 위한 단계 ▪ 도메인 지식을 데이터 분석을 위한 문제 정의로 변경하고 초기 프로젝트 계획을 수립하는 단계 ▪ 업무 목적 파악 -> 상황 파악 ->데이터 마이닝 목표 설정-> 프로젝트 계획 수립
데이터 이해 Data Understanding	▪ 분석을 위한 데이터 수집, 데이터 속성 이해를 위한 과정 ▪ 데이터 품질에 대한 문제점 식별 및 숨겨져 있는 인사이트를 발견하는 단계 ▪ 초기 데이터 수집, 데이터 기술 분석, 데이터 탐색, 데이터 품질 확인

KDD	CRISP-DM
데이터셋 선택 데이터 전처리	데이터 이해

2-09. CRISP-DM 분석 방법론 - 2/3

- 6단계로 구성, 일방향으로 구성되어 있지 않고 단계간 피드백을 통하여 단계별 완성도를 높이게 구성됨

6단계 : 업무 이해 - 데이터 이해 - 데이터 준비 - 모델링 - 평가 - 전개

데이터 준비 Data Preparation

- KDD의 Transformation == CRISP-DM 분석 방법론의 데이터 준비
- 분석을 위해 수집된 데이터에서 분석 기법에 적합한 데이터셋을 편성하는 단계
- 많은 시간이 소요될 수 있음
- 분석용 데이터셋 선택, 데이터 정제, 데이터 통합, 데이터 포맷팅

모델링 Modeling

- 다양한 모델링 기법과 알고리즘을 선택
- 모델링 과정에서 사용되는 파라미터를 최적화해 나가는 단계
- 모델링 단계를 통해 찾아낸 모델은 테스트용 프로세스와 데이터셋으로 평가하여 모델 과적합(Overfitting)등의 문제를 발견하고 대응 방안 마련
- 데이터 분석 방법론, 머신러닝을 이용한 수행 모델을 만들거나 데이터를 분할하는 부분
- 모델링 기법 선택, 모델링 작성, 모델 평가

KDD	CRISP-DM
데이터 변환	데이터 준비

2-09. CRISP-DM 분석 방법론 - 3/3

- 6단계로 구성, 일방향으로 구성되어 있지 않고 단계간 피드백을 통하여 단계별 완성도를 높이게 구성됨

6단계 : 업무 이해 - 데이터 이해 - 데이터 준비 - 모델링 - 평가 - 전개

평가 Evaluation	▪ 모델링 단계에서 얻은 모델이 프로젝트의 목적에 부합하는지 평가 ▪ 데이터 마이닝 결과를 수용할 것인지 최종적으로 판단하는 과정 ▪ 분석 결과 평가, 모델링 과정 평가, 모델 적용성 평가
전개 Deployment	▪ 완성된 모델을 실제 업무에 적용하기 위한 계획 수립 ▪ 전개 계획 수립, 모니터링과 유지보수 계획 수립, 프로젝트 종료 보고서 작성, 프로젝트 리뷰

- 모델 평가는 '모델링'단계
- 모델링 과정 평가와 모델 적용성 평가는 '평가'단계에서!!

2-10. 빅데이터 분석 방법론

분석 기획	데이터 준비	데이터 분석	시스템 구현	평가 및 전개
비즈니스 이해 및 범위 설정	필요 데이터 정의	분석용 데이터 준비	설계 및 구현	모델 발전 계획
프로젝트 정의 및 계획수립	데이터 스토어 설계	텍스트 분석	시스템 테스트 및 운영	프로젝트 평가 보고
프로젝트 위험 계획 수립	데이터 수집 및 적합성 점검	탐색적 분석		평가 및 전개
		모델링		
		모델 평가 및 검증		

2-10-1. 분석 기획(Planning) 단계의 Task - 1/3

비즈니스 이해 및 범위 설정 | 프로젝트 정의 및 계획 수립 | 프로젝트 위험 계획 수립

비즈니스 이해 및 범위 설정

비즈니스 이해	분석 대상인 업무 도메인을 이해하기 위해 내부 업무 매뉴얼과 관련 자료, 외부의 관련 비즈니스 자료 조사 및 프로젝트 진행을 위한 방향 설정
프로젝트 범위 설정	▪ 프로젝트 목적에 부합하는 범위를 명확히 설정 함 ▪ 프로젝트에 참여하는 관계자들의 이해를 일치시키기 위하여 구조화된 프로젝트 범위 정의서 SOW(Statement of Work)를 작성

2-10-1. 분석 기획(Planning) 단계의 Task - 2/3

비즈니스 이해 및 범위 설정 | 프로젝트 정의 및 계획 수립 | 프로젝트 위험 계획 수립

프로젝트 정의 및 계획 수립

데이터 분석 프로젝트 정의	상세 프로젝트 정의서 작성, 프로젝트의 목표를 명확화 하기 위해 모델 이미지 및 평가 기준 설정
프로젝트 수행 계획 수립	▪ 프로젝트 수행 계획서 작성, 프로젝트의 목적, 배경, 기대효과, 수행방법 일정 및 추진 조직 WBS 작성 ▪ WBS : Work Breakdown Structure, 작업 분할 구조도

WBS : 전체 업무를 분류하여 구성 요소로 만든 후 각 요소를 평가하고 일정별로 계획하며 그것을 완수할 수 있는 사람에게 할당해주는 역할

비즈니스 이해 및 범위 설정 | 프로젝트 정의 및 계획 수립 | 프로젝트 위험 계획 수립

프로젝트 위험 계획 수립

- 데이터 분석 위험 식별
- 계획 수립 단계에서 빅데이터 분석 프로젝트를 진행하면서 발생 가능한 모든 위험을 식별함
- 위험에 대한 대응 방법 : 회피(Avoid), 전이(Transfer), 완화(Mitigate), 수용(Accept)

2-10-2. 데이터 준비(Preparing) 단계

필요 데이터 정의 | 데이터 스토어 설계 | 데이터 수집 및 정합성 점검

필요 데이터 정의

데이터 정의	▪ 정형, 비정형, 반정형 등의 모든 내/외부 데이터를 포함하고 데이터의 속성, 데이터 오너, 데이터 관련 시스템 담당자 등을 포함하는 데이터 정의서 작성 ▪ 예) 메타데이터 정의서, ERD(Entity Relationship Diagram) 포함
데이터 획득 방안 수립	▪ 내부 데이터 : 부서 간 업무 협조와 개인정보보호 및 정보 보안과 관련한 문제점을 사전에 점검 ▪ 외부 데이터 : 시스템 간 다양한 인터페이스 및 법적인 문제점을 고려하여 상세한 계획 수립

2-10-2. 데이터 준비(Preparing) 단계

필요 데이터 정의 | 데이터 스토어 설계 | 데이터 수집 및 정합성 점검

데이터 스토어 설계

정형 데이터 스토어 설계	▪ 관계형 데이터베이스(RDBMS)를 사용하고, 데이터의 효율적 저장과 활용을 위해 데이터 스토어의 논리적 물리적 설계를 구분하여 설계함
비정형 데이터 스토어 설계	▪ 하둡, NoSQL 등을 이용하여 비정형 또는 반정형 데이터를 저장하기 위한 논리, 물리적 데이터 스토어 설계

2-10-2. 데이터 준비(Preparing) 단계

필요 데이터 정의 | 데이터 스토어 설계 | 데이터 수집 및 정합성 점검

데이터 수집 및 정합성 점검

데이터 수집 및 저장

- 크롤링 등의 데이터 수집을 위한 ETL 등의 다양한 도구와 API, 스크립트 프로그램 등으로 데이터를 수집
- 수집된 데이터를 설계된 데이터 스토어에 저장함

데이터 정합성(무결성) 점검

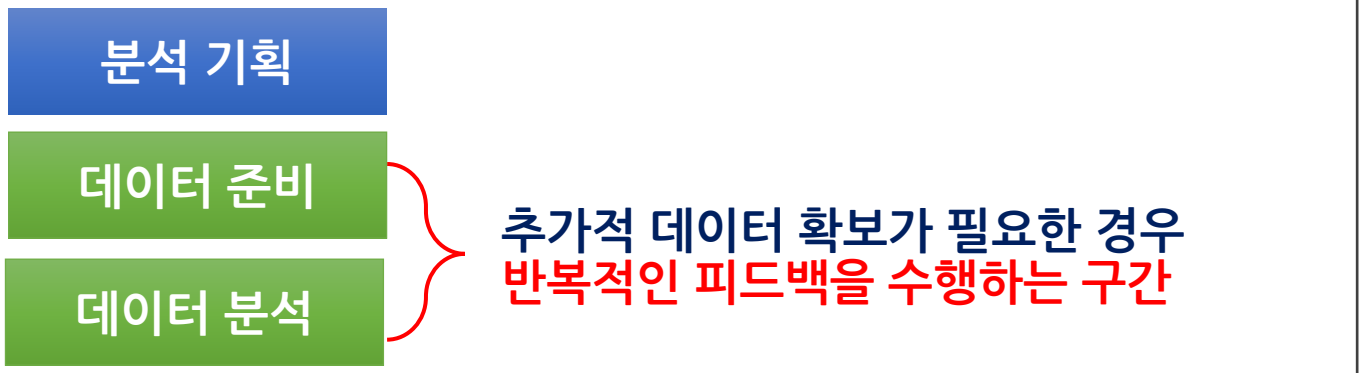
- 데이터 스토어의 품질 점검을 통해 데이터의 정합성 확보
- 데이터 품질개선이 필요한 부분에 대해 보완 작업 진행

ETL(Extract Transformation Loading)
다양한 데이터를 취합해 데이터를 추출하고 하나의 공통된 포맷으로 변환해 데이터 웨어 하우스나 데이터 마트 등에 적재하는 과정을 지원하는 도구

API(Application Programming Interface)
라이브러리에 접근하기 위한 규칙들을 정의한 것

데이터 분석

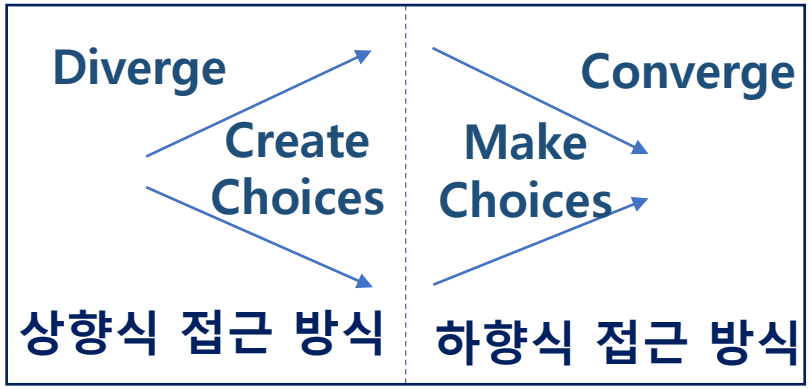
분석 기획 단계에서 수립된 프로젝트 목표를 달성하기 위해 데이터 준비 단계에서 확보된 정형,비정형 데이터를 이용하여 데이터 분석 프로세스를 진행함



모델링 : 분석용 데이터를 이용한 가설 설정을 통해 통계 모델을 만들거나 기계학습을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 과정

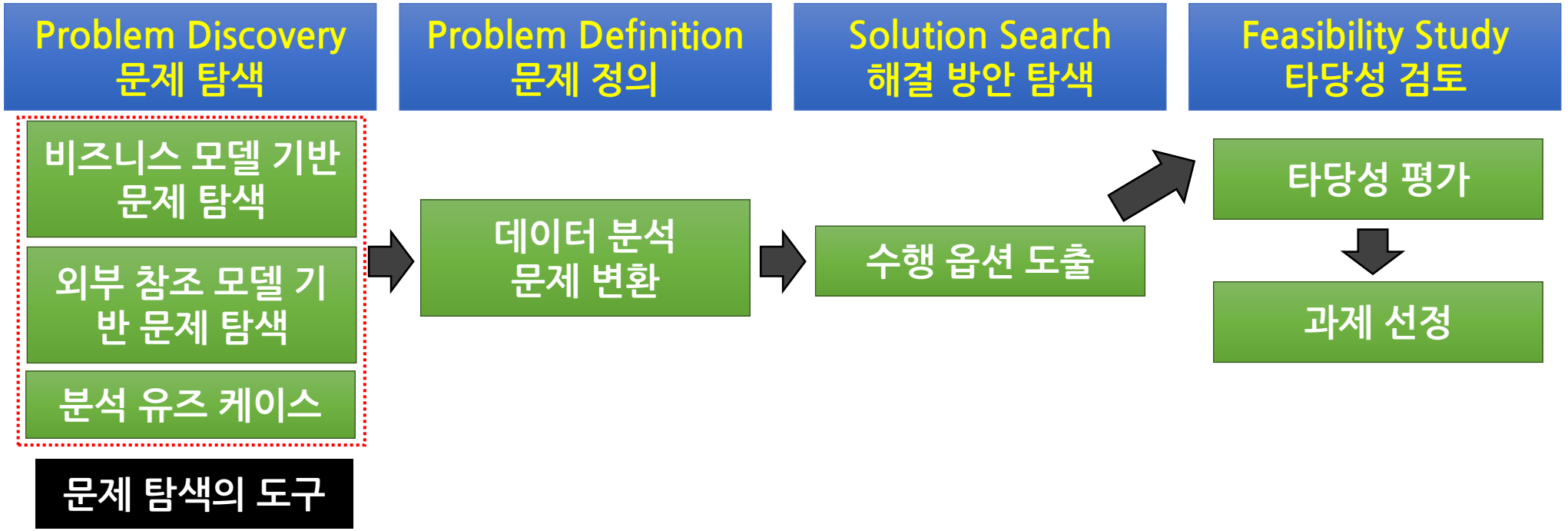
2-11. 분석 과제 도출 방법

하향식 접근 방법	▪ 문제가 확실할 때 사용함, 문제가 주어지고 해법을 찾기 위해 사용함
상향식 접근 방법	▪ 문제의 정의 자체가 어려운 경우 사용함
디자인 싱킹 (Design Thinking)	▪ 중요한 의사결정시 상향식과 하향식을 반복적으로 사용 ▪ 기존의 논리적인 단계별 접근법에 기반한 문제해결 방식은 최근 복잡하고 다양한 환경에서 발생하는 문제에 적합하지 않을 수 있음 ▪ "디자인 사고" 접근법을 통해 전통적인 분석적 사고를 극복하려 함 ▪ 상향식 방식의 발산(Diverge)단계와 도출된 옵션을 분석하고 검증하는 하향식 접근 방식의 수렴(Converge)단계를 반복하여 과제를 발굴함



2-12. 하향식 접근 방식

- 하향식 접근 방식(Top-Down Approach)의 데이터 분석기획 단계



2-12-1. 하향식 접근 방식 - 문제 탐색 단계

16 17 20 22 23

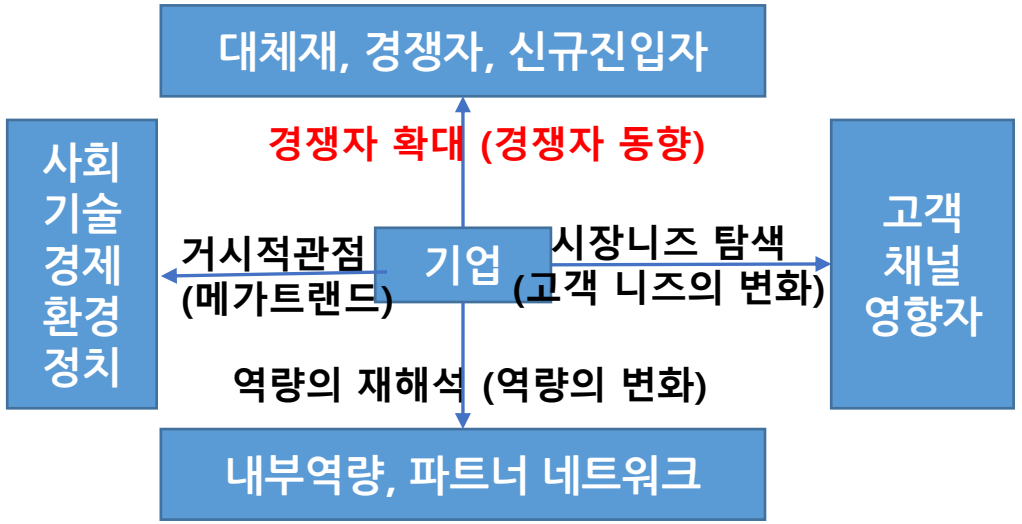
Problem Discovery 문제 탐색 | Problem Definition 문제 정의 | Solution Search 해결 방안 탐색 | Feasibility Study 타당성 검토

비즈니스 모델 기반 문제 탐색

- 비즈니스 모델 캔버스를 활용하여 가치가 창출될 문제를 누락없이 도출할 수 있음
- 기업의 사업 모델을 도식화한 비즈니스 모델 캔버스의 9가지 블록을 단순화하여 업무, 제품, 고객 단위로 문제를 발굴 (뒷 페이지 그림 참조)
- 이를 관리하는 지원 인프라, 규제와 감사 영역에 대한 기회를 추가로 도출하는 작업 수행
- 5가지 영역: 업무, 제품, 고객, 지원 인프라, 규제와 감사

분석 기회 발굴의 범위(영역) 확장

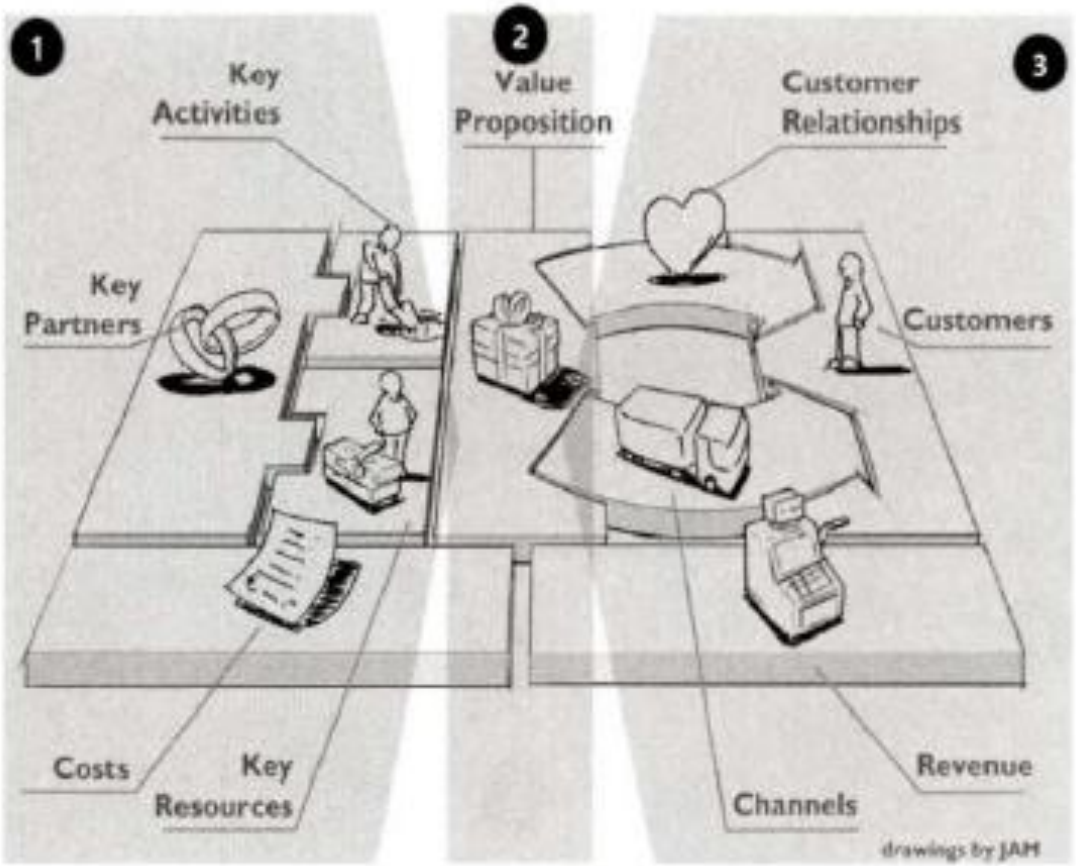
- 거시적 관점의 메가트랜드 : STEEP-사회, 기술, 경제, 환경, 정치
- 경쟁자 확대 관점: 대체재, 경쟁자, 신규진입자
- 시장의 니즈 탐색: 고객(소비자), 채널, 영향자들
- 역량의 재해석 관점: 내부역량, 파트너 네트워크



2-12-1. 하향식 접근 방식 - 문제 탐색 단계

16 17 20 22 23

Problem Discovery 문제 탐색 | Problem Definition 문제 정의 | Solution Search 해결 방안 탐색 | Feasibility Study 타당성 검토



비즈니스 모델 캔버스

빈 칸 채우기로 종종 출제됩니다. 각 영역의 위치를 기억해 두시기 바랍니다. **(출처:한국데이터진흥원-데이터분석전문가가이드)**

2-12-1. 하향식 접근 방식 - 문제 탐색 단계

16 17 20 22 23

Problem Discovery 문제 탐색 | Problem Definition 문제 정의 | Solution Search 해결 방안 탐색 | Feasibility Study 타당성 검토

외부 참조 모델 기반 문제 탐색

- 유사/동종 사례 벤치마킹을 통한 분석 기회 발굴
- 제공되는 산업별, 업무 서비스별 분석 테마 후보 그룹을 통해 Quick & Easy 방식으로 필요한 분석 기회가 무엇인지에 대한 아이디어를 얻고 기업에 적용할 분석 테마 후보 목록을 빠르게 도출

분석 유즈 케이스

- 풀어야 할 문제에 대한 상세 설명 및 해당 문제를 해결했을 때 발생하는 효과를 명시
- 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용하도록 함

2-12-1. 하향식 접근 방식 - 문제 탐색 단계

Problem Discovery
문제 탐색

Problem Definition
문제 정의

Solution Search
해결 방안 탐색

Feasibility Study
타당성 검토

분석 유즈 케이스의 예

업무	분석 유즈 케이스	설명	효과
재무	자금 시재 예측	일별로 예정된 자금지출과 입금 추정	자금 과부족 현상 예방, 자금 운용 효율화
	구매 최적화	구매 유형과 구매자별로 과거 실적과 구매 조건을 비교/분석하여 구매 방안 도출	구매 비용 절감

2-12-2. 문제 정의(Problem Definition) 단계

- 식별된 비즈니스 문제를 데이터의 (분석) 문제로 변환하여 정의하는 단계
  - 문제 탐색 단계 - 무엇(What)을 어떤 목적으로(Why) 수행해야 하는지 관점
  - 문제 정의 단계 - 달성을 위해 필요한 데이터 및 기법(How)을 정의하기 위한 데이터 분석 문제로 변환을 수행

비즈니스 문제		데이터 분석 문제
예상치 않은 설비 장애로 인한 판매량 감소	변환	설비의 장애를 이끄는 신호를 감지하여 설비장애 요인으로 식별하고 장애 발생 시점 및 가능성을 예측
기존판매 정보 기반 영업사원의 판단 시 재고 관리 및 적정 가격 판매 어려움		내부 판매 정보 외의 수요예측을 수행할 수 있는 인자의 추출 및 모델링을 통한 수요 예측

출처: 한국데이터베이스진흥원

2-12-3. 해결 방안 탐색, 타당성 검토 단계

- 해결 방안 탐색 : 어떤 데이터 또는 분석 시스템을 사용할 것인지 검토하는 단계
  - 데이터 및 분석 시스템에 따라 소요되는 예산 및 활용 가능 도구가 다름

분석 역량 (who) / 분석 기법 및 시스템	확보	미확보
기존 시스템	기존 시스템 개선 활용	교육 및 채용을 통한 역량 확보
신규 도입	시스템 고도화	전문업체(Sourcing)

- 타당성 검토 단계
  - 경제적 타당도 : 비용 대비 편익 분석 관점의 접근
  - 데이터 및 기술적 타당도 : 데이터 존재 여부, 분석 시스템 환경, 분석 역량

상향식 접근 방식(Bottom Up Approach)

- 문제의 정의 자체가 어려운 경우 상향식 접근 방식 사용
- 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식
- 상향식 접근 방식의 데이터 분석은 비지도학습(Unsupervised Learning) 방법에 의해 수행됨
- 디자인 싱킹(Design Thinking)의 발산 단계에 해당함
- 인사이트 도출 후 반복적인 시행착오를 통해 수정하며 문제를 도출하는 일련의 과정

2-14. 지도학습 vs 비지도학습

지도 학습 (Supervised Learning)

- 명확한 input, output이 존재함
- 예측(Regression) : 데이터를 대표하는 선형모델 등을 만들고 그 모델을 통해 미래의 사건을 예측하는 것
- 분류(Classification) : 이전까지 학습된 데이터를 근거로 새로운 데이터가 기존에 학습된 데이터에 분류 여부

비지도 학습 (Unsupervised Learning)

- 컴퓨터가 알아서 분류를 하고, 의미 있는 값을 보여줌
- 데이터가 어떻게 구성되어 있는지 밝히는 용도로 사용함
- 군집화(Clustering)

분석 프로젝트의 특징

- 분석 프로젝트는 다른 프로젝트 유형처럼 범위, 일정, 품질, 리스크, 의사소통 등 영역별 관리가 수행되어야 한다
- 다양한 데이터에 기반한 분석 기법을 적용하는 특성 때문에 5가지 주요 특성을 고려하여 추가적 관리가 필요하다
- 분석 과제 주요 특성에는 Data Size, Data Complexity, Speed, Analytic Complexity, Accuracy & Precision 등이 있다
- 분석 프로젝트는 도출된 결과의 재해석을 통한 지속적인 반복 및 정규화가 수행되기도 한다
- 분석 과제 정의서를 기반으로 분석 프로젝트를 진행하게 된다

분석 과제 정의서

- 다양한 분석 과제 도출 방법을 통해 도출된 분석 과제를 분석 과제 정의서로 정리함
- 필요한 소스 데이터, 분석 방법, 데이터 입수 난이도, 데이터 입수 사유, 분석 수행주기, 분석결과에 대한 검증, 분석 과정 상세 등을 작성함
- 프로젝트 수행 계획의 입력물로 사용됨
- 이해관계자가 프로젝트의 방향을 설정하고, 성공 여부를 판별할 수 있는 중요한 자료로 명확하게 작성해야 함

2-16. 분석 프로젝트의 특성 관리 영역

- 분석 과제의 주요 5가지 특성 관리 영역
  - Data Size, Data Complexity, Speed, Analytic Complexity, Accuracy & Precision

Data Size	▪ 분석하고자 하는 데이터의 양을 고려하는 관리방안 수립 필요
Data Complexity	▪ 비정형데이터 및 다양한 시스템에 산재되어 있는 데이터들을 통합해서 분석 프로젝트를 진행할 때는 해당 데이터에 잘 적용될 수 있는 분석 모델 선정에 대한 고려 필요
Speed	▪ 분석 결과 도출 후, 활용하는 시나리오 측면에서 일, 주 단위 실적은 배치 형태 작업, 사기 탐지, 서비스 추천은 실시간 수행되어야 함 ▪ 분석 모델의 성능 및 속도를 고려한 개발 및 테스트가 수행되어야 함

2-16. 분석 프로젝트의 특성 관리 영역

17 18 19 21 *2

- 분석 과제의 주요 5가지 특성 관리 영역
  - Data Size, Data Complexity, Speed, Analytic Complexity, Accuracy & Precision

Analytic Complexity	▪ 정확도(Accuracy)와 복잡도(Complexity)는 트레이드 오프 관계가 존재 ▪ 분석 모델이 복잡할수록 정확도는 올라가지만 해석이 어려워 짐 ▪ 기준점을 사전에 정의해 두어야 함
Accuracy & Precision	▪ Accuracy : 분석의 활용적인 측면 (모델과 실제 값의 차이) ▪ Precision : 분석의 안정성 측면 (모델을 반복했을 때의 편차) ▪ Accuracy, Precision은 트레이드 오프인 경우가 많음 ▪ 모델의 해석 및 적용 시 사전에 고려해야 함

2-17. 10 개 주제별 프로젝트 관리 체계 - 1/2

- 분석 프로젝트의 경우 관리 영역에서 일반 프로젝트와 다르게 유의해야 할 요소 존재
  - 시간, 범위, 품질, 통합, 이해관계자, 자원, 원가, 리스크, 조달, 의사소통

시간	▪ 프로젝트 활동의 일정을 수립, 일정 통제의 진척 상황 관찰
범위	▪ 작업과 인도물을 식별하고 정의하는데 요구되는 프로세스
품질	▪ 품질보증과 품질통제를 계획하고 확립하는 데 요구되는 프로세스
통합	▪ 프로젝트와 관련된 다양한 활동과 프로세스를 도출, 정의, 결합, 단일화, 조정, 통제, 종료에 필요한 프로세스
이해관계자	▪ 프로젝트 스폰서, 고객사, 기타 이해관계자 식별,관리에 필요한 프로세스

2-17. 10 개 주제별 프로젝트 관리 체계 - 2/2

- 시간, 범위, 품질, 통합, 이해관계자, 자원, 원가, 리스크, 조달, 의사소통

자원	▪ 인력, 시설, 장비, 자재, 기반 시설, 도구와 같은 적절한 프로젝트 자원을 식별하고 확보하는 데 필요한 프로세스
원가	▪ 개발 예산과 원가통제의 진척 상황을 관찰하는데 요구되는 프로세스
리스크	▪ 위험과 기회를 식별하고 관리하는 프로세스
조달	▪ 계획에 요구된 프로세스를 포함하며, 제품 및 서비스 또는 인도물을 인수하고 공급자와의 관계를 관리하는데 요구되는 프로세스
의사소통	▪ 프로젝트와 관련된 정보를 계획, 관리, 배포하는 데 요구되는 프로세스

02-02

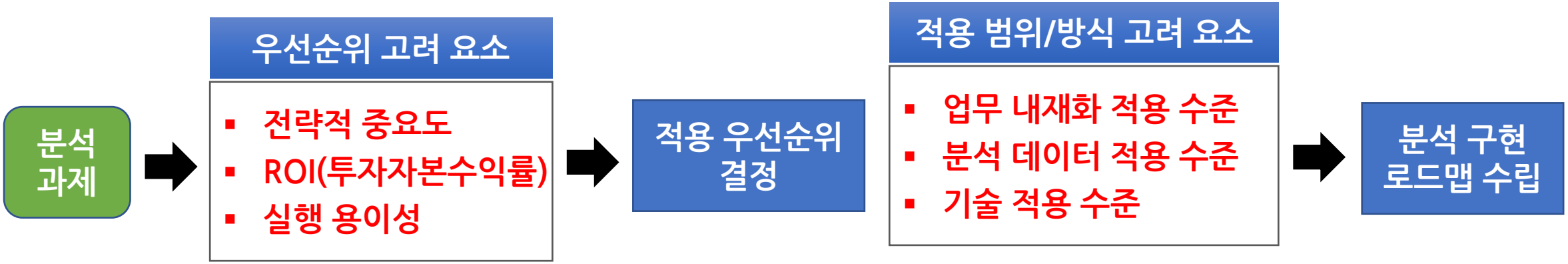
데이터 분석 기획 - 분석 마스터 플랜

분석 마스터플랜

1. 분석 마스터플랜 수립 프레임워크
2. 수행 과제 도출 및 우선순위 평가
3. 이행계획 수립

2-18. 분석 마스터플랜 수립

분석 마스터 플랜 수립 프레임워크



- 중장기적 마스터 플랜 수립을 위해서는 분석 과제를 대상으로 다양한 기준을 고려해 적용할 우선순위를 설정할 필요가 있다
- 분석 과제 수행의 선/후행 관계를 고려하여 우선순위를 조정해 나간다
- 분석 과제의 적용 범위 및 방식에 대해서도 종합적으로 고려하여 결정한다

2-19. 수행 과제 도출 및 우선순위 평가

우선순위 평가 방법 및 절차

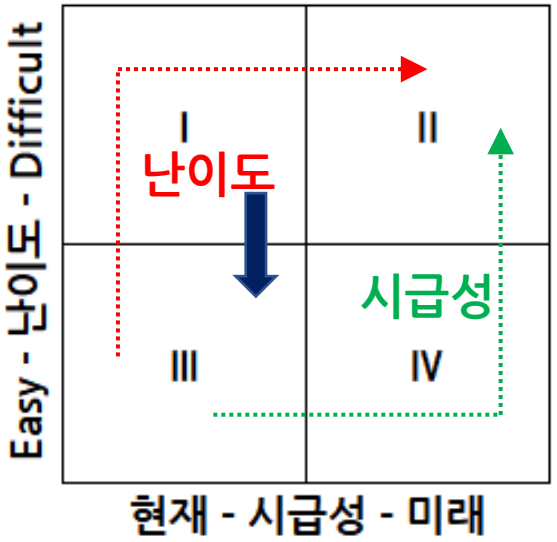
분석 과제 도출	→	우선순위 평가	→	우선순위 정련
		과제 우선순위 기준 수립		분석 과제 수행의 선/후 관계 분석을 통해 순위 조정

ROI(Return On Investment) 관점에서의 빅데이터 4V

Volume	데이터의 크기/양	투자비용 요소 (Investment)
Variety	데이터 종류/유형	
Velocity	데이터의 생성/처리 속도	
Value	분석 결과 활용 및 실행을 통한 비즈니스 가치	비즈니스 효과 요소 (Return)

2-20. 분석 과제 우선순위 선정 기법

포트폴리오 사분면을 통한 과제 우선순위 선정



- 3사 분면 : 일반적으로 가장 먼저 하는 것
- 우선순위를 '시급성'에 둔다면 Ⅲ - Ⅳ - Ⅱ순서 진행
- 우선순위를 '난이도'에 둔다면 Ⅲ - Ⅰ - Ⅱ순서 진행

- **시급성** 판단 기준: **전략적 중요도** 및 목표가치

- **난이도**는 현시점에서 과제를 추진하는 것이 **분석 비용과 적용 범위 측면**에서 쉬운(Easy) 것인지 어려운(Difficulty)것인지에 대한 판단 기준
- **시급성이 높고 난이도가 높은 영역(1사분면)**은 경영진 또는 실무 담당자의 의사결정에 따라 적용 우선순위를 조정할 수 있음

2-21. 이행계획 수립

1. 로드맵 수립

- 결정된 과제의 우선순위를 토대로 분석 과제별 적용 범위 및 방식을 고려하여 최종적인 실행 우선순위를 결정 후 단계적 구현 로드맵 수립

2. 세부 이행계획 수립

- 반복적인 정렬 과정을 통해 프로젝트의 완성도를 높이는 방식을 주로 사용
- 모든 단계 반복보다 데이터 수집 및 확보와 분석 데이터를 준비하는 단계를 순차적 진행하고 모델링 단계는 반복적으로 수행하는 혼합형을 많이 적용함

분석 거버넌스 체계 수립

1. 거버넌스 체계 개요
   - 거버넌스, 분석 거버넌스, 데이터 거버넌스, 분석 거버넌스 체계 구성 요소
2. 데이터 분석 준비도
3. 분석 성숙도 모델
4. 분석 수준 진단 결과
5. 분석 지원 인프라 방안 수립
6. 데이터 거버넌스 체계 수립
7. 데이터 조직 및 인력 방안 수립
8. 분석 과제 관리 프로세스 수립
9. 분석 교육 및 변화 관리

2-22. 거버넌스 체계 개요

16 22 23

거버넌스 (Governance)	▪ Government와 같은 어원 ▪ 더 폭넓은 의미로 진화하여 기업, 비영리 기관 등에서 규칙, 규범 및 행동이 구조화, 유지, 규제되고 책임을 지는 방식 및 프로세스를 지칭함
분석 거버넌스	기업에서 데이터가 어떻게 관리, 유지, 규제되는지에 대한 내부적인 관리 방식이나 프로세스
데이터 거버넌스	▪ 데이터의 품질보장, 프라이버시 보호, 데이터 수명 관리, 전담조직과 규정정립, 데이터 소유권과 관리권 명확화 등을 통해 데이터가 적시에 필요한 사람에게 제공되도록 체계를 확립하는 것 ▪ 데이터 거버넌스가 확립되지 못하면 빅브라더의 우려가 현실화될 가능성이 높음 ▪ 빅브라더 : 정보의 독점으로 사회를 통제하는 관리 권력 혹은 그러한 사회체계

2-23. 분석 거버넌스 체계 구성 요소

- 분석 거버넌스 체계 구성 요소 (**분석 비용 및 예산 없음에 주의**!)

Process	과제 기획/운영 프로세스
Organization	분석 기획/관리 및 추진 조직
System	IT기술/프로그램
Human Resource	분석 교육
Data	데이터 거버넌스

2-24. 데이터 분석 수준 진단

- 데이터 분석 수준 진단
  - 데이터 분석 기법을 구현하기 위해 무엇을 준비하고 보완해야 하는지 알 수 있음
  - 분석의 유형 및 분석의 방향성 결정에 도움
  - 분석 준비도와 분석 성숙도를 함께 평가함으로써 수행될 수 있음
- 분석 준비도
  - 기업의 데이터 분석 도입의 수준을 파악하기 위한 진단방법, 6가지 영역을 대상으로 현 수준을 파악함

> 분석 업무 파악, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라

- 분석 성숙도
  - 시스템 개발 업무능력과 조직의 성숙도 파악을 위해 CMMI 모델을 기반으로 분석 성숙도를 평가함

> 비즈니스 부문, 조직/역량 부문, IT 부문을 대상으로
> 성숙도 수준에 따라 도입단계, 활용 단계, 확산 단계, 최적화 단계로 구분해 살펴 볼 수 있음

능력 성숙도 통합 모델(Capability Maturity Model Integration, CMMI)
- 소프트웨어 개발 및 전산장비 운영 업체들의 업무 능력 및 조직의 성숙도를 평가하기 위한 모델을 말함

2-25. 분석 준비도

데이터 분석 준비도 프레임워크

분석 업무 파악	인력 및 조직	분석 기법	분석 데이터	분석 문화	IT 인프라
발생한 사실 분석 업무 예측 분석 업무 시뮬레이션 분석 업무 최적화 분석 업무 분석 업무 정기적 개선	분석 전문가 직무 존재 분석 전문가 교육 훈련 프로그램 관리자의 기본 분석 능력 전사 분석 업무 총괄 조직 존재 경영진 분석 업무 이해 능력	업무별 적합한 분석 기법 사용 분석 업무 도입 방법론 분석 기법 라이브러리 분석 기법 효과성 평가 분석 기법 정기적 개선	분석 업무를 위한 데이터 충분성 및 신뢰성 적시성 비구조적 데이터 관리 외부 데이터 활용 체계 기준 데이터 관리	사실에 근거한 의사결정 관리자의 데이터 중시 회의 등에서 데이터 활용 경영진의 직관보다 데이터의 활용 데이터 공유 및 협업 문화	운영 시스템 데이터 통합 EAI, ETL 등 데이터 유통체계 분석 전용 서버 및 스토리지 빅데이터 분석 환경 비주얼 분석 환경

IT인프라 = 분석인프라

2-25. 분석 성숙도 모델

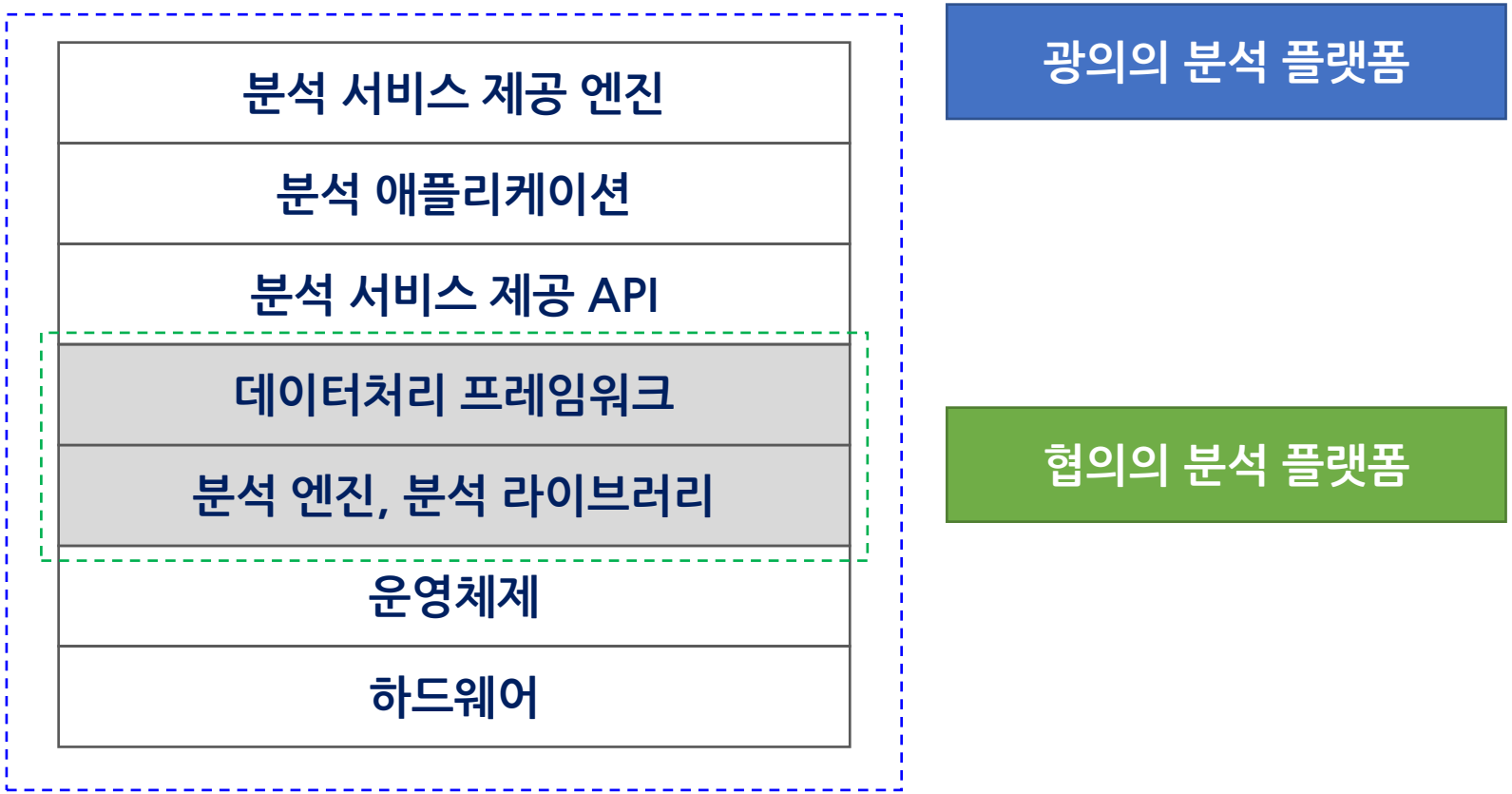
단계	도입 단계	활용 단계	확산 단계	최적화 단계
설명	분석을 시작하여 환경과 시스템 구축	분석 결과를 실제 업무에 적용	전사 차원에서 분석을 관리하고 공유	분석을 진화 시켜 혁신 및 성과 향상에 기여
비즈니스 부문	실적분석 및 통계 정기보고 수행 운영 데이터 기반	미래 결과 예측 시뮬레이션 운영 데이터 기반	전사 성과 실시간 분석 프로세스혁신 3.0 분석 규칙 관리, 이벤트 관리	외부환경 분석 활용 최적화 업무 적용, 실시간 분석 비즈니스 모델 진화
조직 역량 부문	일부 부서에서 수행 담당자 역량에 의존	전문 담당부서에서 수행 분석 기법 도입 관리자가 분석 수행	전사 모든 부서 수행 분석 CoE 조직 운영 데이터 사이언티스트 확보	데이터 사이언스 그룹 경영진 분석 활용 전략 연계
IT 부문	데이터 웨어하우스, 데이터 마트, ETL/EAI, OLAP	실시간 대시보드 통계분석 환경	빅데이터 관리 환경 시뮬레이션/최적화 비주얼 분석, 분석 전용 서버	분석 협업 환경,분석 Sandbox 프로세스 내재화 빅데이터 분석

2-26. 분석 수준 진단 결과

- 사분면 분석
  - 분석 수준 진단 결과를 구분하여 향후 고려해야 하는 데이터 분석 수준에 대한 목표 방향을 정의하고 유형별 특성에 따라 개선 방안을 수립할 수 있음

2-27. 분석 지원 인프라 방안 수립

- 장기적, 안정적으로 활용할 수 있는 확장성을 고려한 플랫폼 구조를 도입하는 것이 적절함



2-28. 데이터 거버넌스 체계 수립

16 18 19 20

데이터 거버넌스 체계요소

데이터 표준화	▪ 데이터 표준용어 설정, 명명규칙 수립, 메타 데이터 구축, 데이터 사전 구축
데이터 관리체계	▪ 메타데이터와 데이터 사전(Data Dictionary)의 관리 원칙 수립
데이터 저장소관리	▪ 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성
표준화 활동	▪ 데이터 거버넌스 체계 구축 후, 표준 준수 여부를 주기적으로 점검, 모니터링

데이터 거버넌스의 데이터 저장소 관리

- 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성
- 저장소는 데이터 관리 체계 지원을 위한 워크프로우 및 관리용 응용 소프트웨어를 지원하고 관리 대상 시스템과의 인터페이스를 통한 통제가 이루어져야 한다.
- 데이터 구조 변경에 따른 사전영향평가도 수행되어야 효율적인 활용이 가능하다.

2-28. 데이터 거버넌스 체계 수립

16 18 19 20

데이터 거버넌스 구성 요소

원칙	데이터를 유지 관리하기 위한 지침과 가이드 및 보안, 품질 기준, 변경 관리
조직	데이터를 관리할 조직의 역할과 책임 및 데이터 관리자, 데이터 아키텍트
프로세스	데이터 관리를 위한 활동과 체계 및 작업 절차, 모니터링 활동

2-29. 데이터 분석을 위한 조직 구조

집중형 조직 구조

- 조직내에 별도의 독립적인 분석 전담 조직 구성
- 분석 전담조직에서 회사의 모든 분석 업무를 담당함
- 일부 협업 부서와 분석 업무가 중복 또는 이원화될 가능성이 있음

기능중심 조직 구조

- 별도로 분석 조직을 구성하지 않고 각 해당 업무부서에서 직접 분석하는 형태
- 일반적인 분석 수행구조, 전사적 핵심 분석이 어려움

분산 조직 구조

- 분석 조직의 인력들이 협업부서에 배치 되어 신속한 업무에 적합
- 전사 차원의 우선순위 수행, 부서 분석 업무와 역할 분담 명확히 해야 함

2-30. 분석 과제 관리 프로세스, 분석 교육 및 변화 관리

분석 과제 관리 프로세스

과제 발굴	분석 아이디어 발굴	분석 과제 후보 제안	분석 과제 확정	
과제 수행	팀 구성	분석 과제 실행	분석 과제 진행 관리	결과 공유/개선

분석 교육 및 변화 관리

- 예전에는 기업 내 데이터 분석가가 담당했던 일 → 모든 구성원이 데이터를 분석하고 이를 바로 업무에 활용할 수 있도록 조직 전반에 분석 문화를 정착 시키고 변화시키려는 시도

- 분석 조직 및 인력에 대한 지속적인 교육과 훈련이 필요함

2-31. 빅데이터 거버넌스의 특징

빅데이터 거버넌스 특징

- 기업이 가진 과거 및 현재의 모든 데이터를 분석하여 비즈니스 인사이트를 찾는 노력은 비용면에서 효율적이지 못함 ⋯› 분석 대상 및 목적을 명확히 정의하고, 필요한 데이터를 수집, 분석하여 점진적으로 확대해 나가는 것이 좋음
- 빅데이터 분석에서 품질관리도 중요하지만, 데이터 수명주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리 비용 증대 문제에 직면할 수 있음
- ERD는 운영 중인 데이터베이스와 일치하기 위해 계속해서 변경사항을 관리하여야 함
- 산업 분야별, 데이터 유형별, 정보 거버넌스 요소별로 구분하여 작성함
- 적합한 분석 업무를 도출하고 가치를 높여줄 수 있도록 분석 조직 및 인력에 대해 지속적인 교육과 훈련을 실시함
- 개인정보보호 및 보안에 대한 방법을 마련해야 함

2-32. 관련 용어

19 20 22 *2 23

- 다음의 용어는 단답형으로 기출 되었음

Servitization	제조업과 서비스업의 융합을 나타내는 용어 예) 웅진 코웨이의 코디
CoE (Center of Excellence)	구성원들이 비즈니스 역량, IT 역량 및 분석 역량을 고루 갖추어야 하며, 협업부서 및 IT 부서와의 지속적인 커뮤니케이션을 수행하는 조직 내 분석 전문조직을 말함
ISP(정보전략계획)	기업의 경영목표 달성에 필요한 전략적 주요 정보를 포착하고, 주요 정보를 지원하기 위해 전사적 관점의 정보 구조를 도출하며, 이를 수행하기 위한 전략 및 실행 계획을 수립하는 전사적인 종합추진 계획
Sandbox	보안모델, 외부 접근 및 영향을 차단하여 제한된 영역 내에서만 프로그램을 동작시키는 것