11_빅데이터 수집과 통계의 기초

#1

1. 빅데이터 수집
데이터의 수집을 위해 데이터 수집자는 데이터 분석 목표를 이해하고, 비즈니스 도메인에 대한 이해를 바탕으로 원천 데이터를 탐색해야 합니다.

#2
가. 비즈니스 도메인 정보의 이해
비즈니스 도메인 정보를 습득하기 위해서는 비즈니스 모델, 비즈니스 용어집, 비즈니스 프로세스로부터 관련 정보를 습득합니다. 또 도메인 전문가 인터뷰를 통해 데이터의 종류, 유형, 특징 정보를 습득합니다.

- 비즈니스 모델: 비즈니스 모델은 비즈니스 전개를 위해 필요한 구성 요소 간의 상호 관계를 모델화한 것입니다.
- 비즈니스 용어집: 특정 비즈니스 영역에서 사용되는 신뢰할 수 있는 용어와 관계 사전입니다.
- 비즈니스 프로세스: 다양한 시스템과 비즈니스 유닛들에 넓게 분산되어 있고 커스터마이즈되어 있습니다. 복잡하고 역동적인 실체로서 고객에게 가치를 전달하는 데 필요한 모든 순차적이거나 병렬적인 활동들의 집합입니다.
- 도메인 전문가 인터뷰: 도메인 전문가는 도메인 분야에 경험과 깊이 있는 지식을 가진 사람입니다. 인터뷰를 통해 도메인에 사용되는 전문용어와 다른 의미로 통용되는 일상용어를 익힙니다. 또 해당 분야에서 다루어지는 데이터의 종류, 유형, 특징 정보를 습득합니다.

#3
나. 원천 데이터의 정보
데이터 분석에 필요한 대상의 원천 데이터의 수집 가능성과 데이터의 보안 및 정확성을 탐색합니다. 또 데이터 수집의 난이도, 수집 비용 등에 관련된 기초 자료를 수집할 수 있습니다.

- 데이터의 수집 가능성: 원천 데이터 수집의 용이성과 데이터의 발생 빈도를 탐색합니다. 그리고 데이터 활용에 있어 전처리 및 후처리 비용을 대략 산정할

수 있습니다.
- 데이터의 보안: 수집 대상 데이터의 개인정보 포함의 여부, 지적 재산권 존재의 여부를 판단해 데이터를 분석할 때 발생할 수 있는 문제를 예방합니다.
- 데이터의 정확성: 데이터를 분석하는 목적에 맞는 적절한 데이터 항목이 존재하고, 적절한 데이터 품질이 확보될 수 있는지를 탐색해야 합니다.
- 수집의 난이도: 원천 데이터의 존재 위치, 데이터의 유형, 데이터 수집 용량, 구축비용, 정제 과정의 복잡성을 고려해 데이터를 탐색합니다.
- 수집 비용: 데이터를 수집하기 위해 발생할 수 있는 데이터 획득 비용을 산정할 수 있습니다.

#4
질문자: 서비스 프로세스란 무엇인가요?
전문가: 서비스 프로세스는 서비스 주관자에서부터 서비스 소비자까지의 서비스 전달 절차나 활동의 흐름을 의미합니다. 비즈니스 도메인의 서비스 활용 범위를 식별하여 해당 서비스 활용 영역에서 발생하는 기초 데이터를 수집합니다. 여기서 서비스 활용 범위란, 해당 서비스를 이용해 특정 서비스 또는 메커니즘에 적용할 수 있는 활용의 범주를 말합니다. 서비스 활용 영역의 발생 데이터를 통해 서비스를 제공할 때 발생하는 다양한 기초 데이터를 수집할 수 있습니다.

#5
2. 세부 계획의 작성
가. 데이터 수집의 세부 계획 반영 활동
데이터 수집을 위한 세부적인 계획은 데이터 선정 이후에 수립합니다. 세부 계획 반영 활동이란, 데이터의 유형, 데이터의 위치, 데이터의 저장 방식, 데이터의 기술, 데이터의 보안 사항 등을 구체적으로 작성하는 활동입니다.

① 데이터의 유형
- 정형 데이터: 정형화된 스키마 구조를 가지고 DBMS에 내용이 저장될 수 있는 데이터를 의미합니다.
- 반정형 데이터: 데이터 내부에 데이터 구조에 대한 메타 정보를 가진 데이터를 의미합니다.
- 비정형 데이터: 수집 데이터 하나하나가 데이터의 객체로 구분될 수 있는 데이터를 의미합니다.

#6
② 데이터의 위치
- 내부 데이터: 데이터의 소스가 내부 시스템에 존재하는 데이터입니다. 이는 대부분 정형 데이터로 존재합니다. 내부 데이터의 수집은 담당자와 협의가 원활하고 수집의 난이도가 낮습니다.
- 외부 데이터: 데이터의 소스가 외부 시스템에 존재하거나 데이터의 세트로써 받을 수 있는 데이터입니다. 이는 대부분 반정형과 비정형 데이터로 존재합니다. 외부 데이터를 수집할 때는 외부의 소스 담당자와 의사소통이 어렵다는 단점이 있습니다. 게다가 대부분 추가적인 데이터의 가공 작업이 필요하고 수집의 난이도가 높습니다.

#7
③ 데이터의 저장 방식
- 파일 시스템: 데이터를 읽고 쓰며 찾기 위해 일정한 규칙으로 파일에 이름을 명명하고 파일의 위치를 지정하는 체계입니다.
- 관계형 DB: 데이터의 종류나 성격에 따라 여러 개의 칼럼을 포함하는 정형화된 테이블로 구성된 데이터 항목들의 집합체입니다.

④ 데이터의 수집 기술
- 정형 데이터 수집 기술: 관계형 데이터와 분산 환경 데이터 간의 전송 데이터입니다.
- 로그 데이터 수집 기술: 시스템 로그, IoT 센서 로그, 전산 장비 로그 데이터입니다.
- 웹 크롤링 및 소셜 데이터 수집 기술: 웹상에 존재하는 콘텐츠 데이터입니다.

#8
⑤ 데이터 확보 비용의 산정 및 이관 절차
데이터 확보 비용을 산정할 때는 데이터의 크기, 수집 주기, 수집 기술, 수집 방식, 대상 데이터의 가치를 고려합니다. 수집 대상 데이터의 조사, 데이터의 소유자와 이관 협의, 데이터의 이관 수행, 데이터의 검증을 거쳐 데이터를 이관합니다.

⑥ 데이터의 적절성 검증

수집 대상의 데이터가 제대로 수집되었는지는 데이터의 누락 여부, 소스 데이터와의 비교, 데이터의 정확성, 보안 사항, 저작권 사항, 대량 트래픽 발생 여부의 검증을 통해 알 수 있습니다.

#9
3. 내·외부 데이터의 수집
수집 데이터의 원천에 따라 내부 데이터와 외부 데이터로 구분합니다. 내부 데이터는 조직 내부의 데이터 담당자와 수집 주기와 방법을 협의해 데이터를 수집할 수 있습니다. 외부 데이터는 특정 기관의 담당자와의 협의를 통해 데이터를 수집할 수도 있고, 데이터를 제공하는 전문 업체를 통해 수집할 수도 있습니다.

#10
가. 내부 데이터
① 종류
내부 조직 간의 협의를 통해 데이터를 수집합니다. 주로 수집이 용이한 정형 데이터를 의미합니다. 비용 및 난이도는 외부 데이터 수집보다 유리하고 서비스의 수명 주기 관리가 용이합니다. 내부 데이터는 조직 내부의 서비스 시스템, 네트워크 및 서버 장비, 마케팅 관련 시스템 등으로부터 생성되는 데이터를 말합니다.

② 수집 특징
내부 데이터는 조직 내부에서 습득할 수 있는 데이터입니다. 네트워크 로그 데이터, 시스템 로그 데이터, 데이터베이스 관리 시스템(DBMS)의 저장 데이터가 있습니다. 일반적으로 내부 데이터는 실시간으로 수집해 분석할 수 있도록 합니다.

③ 수집 방법
내부 데이터는 일반적으로 기업이나 조직 내의 정보 시스템에 존재하는 정형화된 데이터를 말합니다. 내부 데이터를 수집하기 위해서는 조직 간의 협의를 거쳐야 합니다. 그리고 이와 같은 방법으로 수집된 내부 데이터는 분석에 적합한 정형화된 형식으로 수집됩니다. 이 데이터는 가공에 많은 노력을 기울이지 않아도 됩니다. 내·외부의 데이터 수집 방법은 같습니다.
- 업무 협의: 조직 내부의 협의를 거쳐 데이터를 수집합니다.

- 수집 경로: 인터페이스를 생성합니다.
- 수집 대상: 파일 시스템, DBMS, 센서 등을 대상으로 합니다.

#11
나. 외부 데이터
① 종류
외부 조직과의 협약, 데이터 구매, 웹상의 오픈 데이터를 통해 수집합니다. 주로 수집이 어려운 비정형 데이터를 의미합니다. 비용 및 난이도가 높습니다. 여건상 외부 환경에 대한 통제가 어려우므로 이에 따른 서비스 관리 정책을 수립할 필요가 있습니다. 외부 데이터의 종류는 다양한 소셜 데이터, 특정 기관의 데이터, M2M 데이터, 공공데이터(LoD: Linked Open Data) 등으로 나눌 수 있습니다.

② 수집 특징
외부 데이터는 일괄 수집 방식으로 수행할 것인지 일정 주기를 정해 데이터를 수집할 것인지를 결정하여 수집 데이터 관리 정책을 수립해야 합니다.

③ 수집 방법
외부 데이터는 분석 목표에 맞는 데이터를 탐색한 다음에 이를 수집하고 분석 목표에 맞게 수집 데이터를 변환하는 노력을 들여야 합니다.

#12
다. 내·외부 데이터의 수집 기술
① 정형 데이터
관계형 데이터베이스의 일반적인 형태인 정형 데이터는 주로 아파치 스쿱(Apache Sqoop)을 이용해 수집합니다.

아파치 스쿱 아키텍처에서 스쿱은 관계형 데이터베이스에서 읽어온 테이블을 하둡 분산 파일 시스템(HDFS)에서 파일 세트로 저장합니다. 병렬 처리 방식으로 적재하기 때문에 적재한 뒤에 HDFS에서 여러 개의 파일로 저장됩니다. 반대로 스쿱을 사용해 HDFS에 저장된 파일 세트를 읽고 관계형 데이터베이스로 적재하는 것도 가능합니다.

#13

② 로그 및 센서 데이터
빅데이터 분석에 있어 대표적인 데이터 유형인 로그 및 센서 데이터는 주로 아파치 플룸(Apache Flume), 페이스북 스크라이브(Facebook Scribe), 아파치 척와(Apache Chukwa)를 사용해 수집합니다. 다양한 오픈소스를 사용하지만 아파치 플룸이 가장 많이 사용됩니다. 플룸은 스트리밍 데이터 플로(Flow)를 기반으로 간단하고 유연한 아키텍처를 가지게 됩니다. 이는 데이터를 받는 소스, 소스와 싱크 사이에서 상호 연동을 지원하는 채널, 데이터를 저장하거나 전달하는 싱크로 구성됩니다.

#14
③ 텍스트, 이미지, 동영상, 웹 및 소셜 데이터
비정형 또는 반정형 데이터의 수집은 주로 FTP, API, 라이브러리(Library)를 이용하여 개발하고 크롤러를 이용해 수집합니다.

④ 스크래피(Scrapy) 아키텍처
비정형 또는 반정형 데이터를 수집할 때 스크래피 아키텍처를 이용해 크롤러를 만들 수 있습니다. 스크래피 아키텍처는 데이터 플로를 제어하는 스크래피 엔진, 수집 주기를 설정하는 스케줄러, 웹 페이지를 패치하는 다운로더, 커스텀 클래스인 스파이더, 아이템 프로세싱 역할을 하는 아이템 파이프라인의 다섯 가지 모듈로 구성되어 있습니다.

#15
4. 빅데이터 분석 5단계
가. 분석 기획: 비즈니스의 이해와 범위의 설정, 프로젝트의 정의와 계획 수립, 프로젝트의 위험 계획을 수립하는 단계입니다.
나. 데이터 준비: 필요 데이터의 정의, 데이터 스토어의 설계, 데이터의 수집과 정합성을 검증하는 단계입니다.
다. 데이터 분석: 분석용 데이터의 준비, 텍스트 분석, 탐색적 분석, 모델링, 모델의 평가 및 검증, 모델의 적용 및 운영 방안을 수립하는 단계입니다.
라. 시스템 구현: 설계 및 구현, 시스템의 테스트 및 운영하는 단계입니다.
마. 평가 및 전개: 모델 발전 계획의 수립, 프로젝트의 평가 및 보고하는 단계입니다.