Journal of The Korea Society of Computer and Information
Vol. 19, No. 3, March 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.3.073

빅데이터 분석 프로젝트 수행 방법론

김형래*, 전도홍**, 지승현*

Bigdata Analysis Project Development Methodology

Hyoungrae Kim*, Do-hong Jeon**, Sunghyun Jee*

요 약

기업 경쟁력 제고를 위해 빅데이터 분석의 중요성이 대두됨에 따라, 기업의 문제점을 체계적으로 파악하고 이를 해결하여 사업적 가치로 재평가하기 위해서는 통합적 빅데이터 프로젝트 수행 방법이 필요하다. 이에 따라 실무적 활용 용이성을 높이도록 소프트웨어 개발과 프로젝트 관리가 융합된 "과학적 데이터 분석 방법론(SDAD)"를 제안한다. SDAD는 프로젝트 수행 과정을 문제정의, 데이터준비, 모델설계, 모델구현, 결과평가, 서비스구현의 6단계를 구성한 후, 단계별 과업을 공정별(47개)로 세분화하고 산출물(93개)을 도출한다. SDAD 는 기존의 ISP, DW, SW 개발 방법론에서 빅데이터 분석과 관련된 부분을 통합하고 쉽게 결과물을 연동할 수 있도록 하였다. 또한, 다양한 분야의 전문가로 구성된 참여자 간에 의사소통의 효율성을 높이기 위해 RACI 챠트를 통해 공정별 책임자를 할당하는 방법과 표준화된 의사소통 절차를 제시한다. SDAD 방법론은 한국고용정보원에서 수행한 빅데이터 프로젝트에 적용하여 감리의 평가를 받은 결과 적정한 것으로 나타났다.

▸ Keywords : 빅데이터 프로젝트, 개발 방법론, 과학적 데이터 분석 방법론, 기업 경쟁력

Abstract

As the importance of big data analysis increases to improve the competitiveness of a corporate, a unified big data project development methodology is required in order to study the problem of a corporate in a systematic way and evaluate the problem w.r.t. a business value after solving the problem. This paper propose Scientific Data Anslysis and Development methodology(SDAD) which are integrated methodology of software development and project management for easier application into a field project. SDAD consisits of 6 stages(problem definition stage, data preparation stage, model design stage, model development stage, result extraction stage, service development state), each stages has detailed processes(47) and productions(93). SDAD, furthermore, unified previous

•제1저자 : 김형래 •교신저자 : 전도홍
•투고일 : 2014. 1. 16, 심사일 : 2014. 1. 27, 게재확정일 : 2014. 2. 6
* 한국고용정보원(Korea Employment Information Service)
** 관동대학교 컴퓨터과학과(Kwandong University)

ISP, DW, SW development methodologies in terms of the data analysis and can easily interchange the productions with them. This paper, lastly, introduces a way to assign responsible persons for each process and provide communication procedures in RACI chart to improves the efficiency of the interaction among professionals from different subjects. SDAD is applied to a Bigdata project in Korea Employment Information Services institution and the result turned out to be acceptable when evaluated by the supervision.

▸ Keywords : Bigdata project, Development methodology, Scientific data analysis & development methodology, Corporate competitiveness

I. 서 론

유최근 기업이 보유하고 있는 내·외부 데이터가 기업의 핵심 경쟁력을 제고 하는 주요 자원으로 인식됨에 따라, 빅데이터 분석 프로젝트는 사회적 이슈로 각광받고 있다. 빅데이터 활용의 확산에 따라 프로젝트 수행에 있어 안내역할을 하는 소프트웨어 개발방법론의 필요성도 함께 대두된다. 소프트웨어 개발 방법론은 프로젝트 수행에 있어 안내 역할을 하는 중요한 지식이기 때문이며, 프로젝트의 특성에 따라 다양한 개발 방법론에 대한 연구가 진행되어 왔다(Yoon et. al., 1999; Kuilboer and Ashrafi, 2000; Giunchiglia et al., 2003; Castro et al., 2001; Papazoglou and Heuvel, 2006). 기존의 ISO 표준 방법론은 최종적인 프로그램, 데이터베이스, 사용자인터페이스, 시스템 운용 환경을 정의하고 구현하는 기본 틀을 제공하는 장점이 있지만, 빅데이터 분석에서 나타나는 특성을 효과적으로 반영하지 못하는 단점이 있다. 프로젝트 수행 방법론은 프로젝트의 특성(공공, 민간 등) 또는 개발 분야(응용소프트웨어, 웹, 시뮬레이션 등)에 따라 다양하게 개발되고 있으며, 현 시점에서 빅데이터 분석 환경에 맞는 방법론이 필요한 실정이다.

빅데이터 분석 프로젝트에 적합한 방법론을 개발하기 위해 이의 특성과 어려움을 파악한 후 이를 하나씩 체계적으로 해결하는 방법으로 접근하고자 한다. 먼저, 빅데이터 분석 프로젝트의 범위가 기존 개발 방법론 보다 넓다는 특징이 있다. 범위는 기존의 SI 사업보다 기업 요구사항 분석, 데이터 수집 및 DW 구축, 운영계 시스템 변경 구축, 빅데이터 분석을 넓게 포함한다. 다음으로, 빅데이터 분석 프로젝트는 기존 프로젝트와 다른 특성을 가진다. 이러한 주요 특징을 데이터 측면과 분석 방법 측면에서 살펴본다. 데이터의 특성 측면에서는 대용량(Volume), 증가속도(Velocity), 다양성(Variety)을 들 수 있다. 빅데이터 분석이라고 할 때 반드시 대용량을 의미하진 않지만 일반적으로 데이터양에 있어 기존의 데이터베이스에 저장되는 양보다 훨씬 큰 규모인 경우가 많다. 빅데이터는 증가 속도가 기존 데이터보다 빠른 특성을 가지기도 한다. 가령 사용자 방문 로그는 데이터베이스의 정보보다 증가속도가 빠르다. 기존의 데이터 분석은 주로 데이터베이스의 정형화된 데이터를 다루었지만, 빅데이터는 형식이 텍스트, 웹로그, SNS, 이미지, 트윗 등 매우 다양하다. 데이터 분석 방법 측면에서의 특성은 정확성(Veracity)과 즉시성(Instancy)을 들 수 있다. 빅데이터는 현황 분석보다는 미래 예측 또는 판별 결과의 정확성에 초점을 맞춘다. 이와 함께 빅데이터 분석은 데이터 처리 속도에 있어 즉시성 있게 필요한 결과를 도출해야 한다. 마지막으로, 빅데이터의 특성은 프로젝트는 범위에 있어서도 차이를 보인다. 기업의 문제를 파악하기 위한 컨설팅 수행을 위해 산업분야에 대한 이해가 높은 전문가가 필요하며, 빅데이터를 수집하고 구축하기 위한 전산 전문가와 데이터를 분석하기 위한 분석 전문가가 필요하다. 이처럼 다양한 분야의 전문가가 참여하여 유기적으로 협력해야하는 특성이 있다.

프로젝트 수행 방법론은 개발 방법론(구축 효율화)과 관리 방법론(위험 관리)으로 구분되나 두 가지가 혼합된 형태도 사용된다. 개발 방법론은 소프트웨어 개발 효율성을 성공적으로 극대화하는 방향이고 관리 방법론은 위험을 관리하는 방향으로 두 방법에 차이가 있다. 프로세스 관리는 분석, 설계, 개발, 테스트 등 소프트웨어 프로세스를 중심으로 개발자 및 개발 조직의 역할을 다룬다(Light et al., 2005; Bresciani

et al., 2004). 프로세스가 올바르게 작동한다면 결과물은 자연히 좋아질 것으로 보는 관점이다(PMBOK, CMM/CMMI, ISO 등). 프로젝트 수행에 있어 구축과 관리의 두 가지 측면을 모두 고려하여야 하지만 실무적인 측면에서 사용이 편리하도록 이 두 가지를 혼합한 방법론이 국내에는 흔히 사용된다. 본 연구에서는 빅데이터의 특성을 고려하여 실무적인 측면에서 효용성을 높이기 위해 구축과 관리가 혼합된 방식의 방법론을 개발하고자 한다. 본 논문은 빅데이터 분석 프로젝트를 효과적으로 수행하기 위해 필요한 방법과 기능이 포괄된 "과학적 데이터 분석 방법론(SDAD: Scientific Data Analysis & Development Methodology)"을 소개한다.

본 연구의 기여는 빅데이터 분석 프로젝트의 특성과 문제점을 살펴보고 이를 고려한 개발 방법론을 제시하였다는 점에 의의가 있다. 방법론을 6개 단계로 구분 하고 단계별로 다양한 분야의 전문가가 효과적으로 의사소통하기 위한 체계적인 기반 방안을 제시하였다. 또한 이러한 방법론을 실무 빅데이터 분석 프로젝트에 적용하여 실무적 검증을 거친 결과 적합한 것으로 나타났다. 빅데이터 분석 프로젝트 산출물의 효용성을 높이기 위해 ISP 방법론, S/W개발방법론, DW 개발 방법론 등 기존 관련 프로젝트 방법론을 포괄하면서 연동이 용이하도록 하였다.

본 논문의 이하는 다음과 같이 구성되었다: 섹션 II는 프로젝트 수행 방법론 관련연구를 설명하고, 섹션 III는 빅데이터 분석의 특성과 이의 해결 방안을 도모하고, 섹션 IV는 프로젝트 수행 방법론을 구체화 하며, 섹션 V는 제시한 방법론을 검증하며, 섹션 VI는 본 논문의 연구 내용을 요약한다.

II. 관련 연구

SW 개발 방법론은 "분석-설계-개발-테스트"의 단계를 가지며, 각 단계별 수행 활동, 도구 기법, 산출물이 정리된다. 개발 방법론에는 구조적 방법론(Structured Development Methodology), 정보공학 방법론(IEM: Information Engineering Methodology), 객체지향 방법론(Object-oriented Development Methodology), 구성 요소 기반 방법론(CBD: Component Based Development Methodology)이 있다(Sommerville, 2010). 구조적 방법론은 프로세스 중심이며, 정보공학 방법론은 전사적 측면에서 프로젝트를 관리하고 데이터 흐름 측면을 중요시하고, 객체지향 방법론은 시스템구축과 소스코드의 재사용을 가능하게 하는 방법이며, CBD 방법론은 실행화일까지 재사용하도록 하는 방법론이다. 정보공학과 객체지향이 실무적으로 많이 사용되고 있다. 이외에도 Agile 방법론이 있다. 사용자 요구사항이 초반에 명확히 정리하기 힘든 경우나 여러 프로젝트가 결합되어 수행될 때 상기 방법론은 진척을 파악하기 복잡한 단점이 발생한다. 이를 보완한 Agile(SCRUM)은 사용자 요구사항에 유연하게 대처하도록 요구사항별로 프로젝트를 세분화하여 추진하는 방법론이다. XP(eXtreme Programming)는 Ken Beck(2004)가 만들었으며 작은 배포버전에서 시작하여 2주단위로 개발과 인수 테스트를 병행 진행하는 방법이다. Agile 과 XP는 국지적 활용에 유용하다.

관리 방법론은 통합관리, 범위, 인력, 일정, 품질, 위험, 예산, 조달구매, 의사소통의 9개 영역을 살피는 것이다. 관리 방법론 표준으로는 PMI(Project Management Institute)에서 발행하는 PMBOK(Project Management Body of Knowledge)가 있다. PMI 에서는 프로젝트 관리와 관련한 PMP(Project Management Professional) 등의 자격증을 발급하고 있다. 그외 관리 방법론과 관련된 인증 종류에는 PMP/PMI, PMBOK, ISO, CMMi 인증, SP인증 등이 있다(PMI, 2013).

한국의 정부에서 제공하는 개발 방법론은 개발과 관리 방법론이 혼합되는 특성이 있다. 국내에서 정보화 프로젝트 관리를 위해 사용되는 방법론을 살펴보면 한국정보화진흥원, 한국전산원, 한국소프트웨어진흥원에서 개발되어 보급되고 있다. 한국정보화진흥원에서 개발하여 배포하는 "CBD SW개발 표준 산출물 관리 가이드"는 객체지향 방법론과 CBD 방법론을 지원하며, 단계별 설명뿐만 아니라 산출물 양식을 모두 제공한다. 자체적인 방법론을 가지지 않은 중소 규모 개발 업체가 활용하기에도 용이하다. 방법론을 프로젝트에 적용하는 방법은 일반적으로 기본적인 방법론을 가지고 각 프로젝트의 용도에 맞추어 변형하여 사용한다. 이러한 변형 작업을 테일러링(tailoring)이라고 한다.

이외에도 프로젝트의 성격에 특화된 다양한 방법론이 연구되고 있다. 시뮬레이션 프로젝트를 성공적으로 수행하기 위한 가이드라인으로 실무적인 방법론이 제기되었다(Choi, 1999). 이는 자료수집, 과업 범위, 기대수준 관리, 팀 구성, 모델 개발, 변화 관리, 결과 보고, 기술 이전 등의 요소를 포함한다. CRM 고객센터의 구조와 구축을 위한 방법론(Jeong, 2002)과 웹 기반 시스템의 분석 및 설계를 위한 개발 방법론도 소개되었다(Jeong et. al., 2002). 본 논문은 빅데이터 분석 또한 이슈로 대두됨에 따라 이에 적합한 방법론에 대한 연구이다.

그간의 방법론이 일반 기업체를 대상으로 하였다면 공공기

관이란 조직의 특성을 반영한 방법론도 소개되고 있다. 공공기관은 정부3.0 추진 하에 IT 거버넌스와 빅데이터 도입에 대한 연구가 요구되고 있다(Kim, 2013; Lee, 2013). 공공기관의 특성을 고려한 정보화 프로젝트 관리 방법론이 서울시의 사례를 통해 제기 되었다(Yoon et. al., 2005). 공공 기관의 정보화 사업이 조직적으로 관리되고 통제되는 모형으로 전환하도록 조직 설계 관점에서 프로젝트 관리 조직을 구상함의 중요성을 강조하였다. 논문은 관리 조직까지 연구범위에 포함하지 않는다.

III. 빅데이터 분석의 특성

본 장에서는 빅데이터 분석 프로젝트의 특성을 프로젝트의 범위 측면, 데이터 특성, 기술적 측면, 참여자 측면에서 살펴보면서, 이러한 특성을 해결하기 위한 방안을 함께 제시한다.

3.1 빅데이터 분석 프로젝트 범위적 특성

빅데이터 분석 프로젝트의 특성은 프로젝트의 범위 측면에서 기업의 문제점을 분석하여 문제를 해결할 분석모델을 개발하고 기업 내 운영계 시스템에 적용하는 단계까지를 포함한다. 빅데이터 분석 프로젝트 범위는 기존의 SI 사업보다 넓다. 범위는 기업 요구사항 분석(KPIs 도출), 데이터 수집 및 DW 구축, 운영계 시스템 변경 구축, 빅데이터 분석을 포함한다(EMC, 2012).

기업의 문제를 체계적으로 분석하여 문제점을 파악하여야 하며, 이러한 문제점을 해결하였을 시의 사업적 가치를 평가하여야 하는데 이는 ISP와 BI에서 다룬다. 기업이 문제를 파악하여 제시하기 보다는 기업의 환경과 요구사항을 통해 문제를 파악할 수 있어야 한다. 기업 환경 분석은 기업 활동과 관련된 외부적인 시장 환경과 기업 내부적인 업무 환경을 분석하여, 이를 토대로 앞으로 나아갈 발전 방향을 발굴하는 과정을 거친다. 기업의 주요 이슈에 영향을 미치는 영향 요인을 분석하여, 기업의 현재 당면한 문제를 해결하기 위해 산업과 시장에서 고려해야 할 요소를 파악한다. 다음은 외부 시장 환경과 기업 내부 분석을 수행한 후 결과를 가지고 기업의 주요 핵심 전략을 발굴하는 과정을 살펴보아야한다.

데이터 수집 및 DW 구축은 빅데이터 분석과 밀접한 연관관계를 가진다. DW 방법론은 다양한 데이터를 운영계로부터 수집 적재하는 일련의 과정을 포함하고 있으므로 이를 빅데이터 수집에 응용하여 사용할 수 있다. DW는 빅데이터 분석과 성격이 다르지만 선행하여 구축되는 특성을 가진다. 기업이 만약 DW를 구축하지 않았다면, DW와 빅데이터 분석 프로젝트를 함께 발주할 가능성이 높다. 따라서 빅데이터 분석 방법론이 DW 구축 방법론을 포괄하도록 하는 것이 실무적인 측면에서 유용하다.

운영계 시스템 변경 구축과정은 빅데이터를 통해 분석된 결과를 반영하는 부분으로, 그 범위가 가장 크게 차이날 수 있는 부분이다. 분석 모델을 운영계 시스템에 적용하는 과정에 새로운 기술팀이 필요할 수도 있다. 먼저 운영계 시스템의 환경을 파악하기 위해 시스템 유지 관리하는 기술자가 참여하여야 할 것이다. 만약 시스템 과부하의 문제가 예상된다면 전체적인 시스템 구조를 살펴보거나 데이터베이스 재설계를 검토할 수도 있다. 이럴 경우 기간계 시스템 재구축이라는 새로운 사업(SI)이 진행되는 것이므로 과업 범위를 유의하여야 한다. 기간계 시스템의 인터페이스를 수정해야 할 경우 웹사용성(Web unability)을 고려하여야 하며, 인터페이스 설계의 변경이 클 경우 웹사용성 기술자의 도움이 필요하다. 소프트웨어 테스트(Software test) 기술을 통한 오류 점검과 개인정보 보호와 같은 자료 활용의 법적 타당성도 검토해야한다. 분석모델을 기간계 시스템에 적용하는 구현 절차 및 순서는 ISO/IEC 12207에서 정의한 공정과 유사하며, 프로그램, 데이터베이스, 사용자인터페이스, 소프트웨어를 운영하기위한 환경 등이 포함된다.

빅데이터 분석과 소프트웨어개발(SI) 방법의 차이점은 설계된 분석모델을 통해 원하는 수준의 결과를 도출하지 못할 경우 무수히 설계와 개발이 반복될 수 있다는 특징이 있다. 빅데이터 분석은 방대한 데이터에서 원하는 결과를 도출하기 위해 분석모델을 설계하고 구현하고 검증하는 과정을 가진다. 검증과정에서 분석 모델의 품질 및 위험을 관리할 수 있어야 한다. 빅데이터 분석 프로젝트 수행 방법론은 이러한 다양한 개발 분야를 포괄하여야 한다.

표 1. 포괄되는 다양한 개발 분야
Table 1. Composed Various Development Areas

범위	설명	해결 방안
산업 컨설팅	기업 요구사항 및 문제점 파악	ISP, BI 방법론을 포괄
DW 구축	데이터 이관 및 DW 구축	DW 구축 방법론을 포괄
운영계 시스템 반영	기존 운영계 시스템 개선 및 신규 구축	SI 방법론을 포괄
빅데이터 분석	분석 모델 개발 품질 및 위험 관리	특성에 맞추어 방법론을 개발

3.2 데이터 및 기술적 특성

빅데이터 분석 프로젝트의 특성을 과업 범위 측면이외에 데이터 특성과 기술적 특성을 살펴본다. 빅데이터 분석 프로젝트에서 다루는 데이터는 대용량이며, 증가속도가 빠르고, 다양한 구조를 가지므로 이를 감내할 수 있어야 한다(EMC, 2012). 따라서 수집할 대용량 데이터를 이관하고, 저장·분석하기 위한 별도의 분석시스템(샌드박스)을 구축해야 한다. 샌드박스 시스템은 별도로 구축되므로 기존 시스템 확장에서 오는 한계로부터 제약을 받지 않을 뿐 아니라, 데이터 처리속도, 다양한 구조의 데이터를 처리하도록 최적화 하는 것이 가능하다. 통상 샌드박스의 용량은 저장할 데이터의 약 10배정도로 준비한다. 샌드박스 구축결과보고서는 HW 현황, 네트워크 현황, SW 설치 현황을 포함한다. 데이터의 빠른 증가속도는 샌드박스의 구성을 병렬로 하여 데이터 수집 및 처리속도를 높이도록 구성하면 된다. 다양한 구조를 가지는 데이터를 처리하기 위해 정형, 비정형 데이터를 모두 저장하고 처리할 수 있도록 설계한다. 가령 관계형 데이터베이스, 하둡, 몽고디비, 파일 등 필요한 데이터구조를 처리할 수 있도록 구성하여야 한다.

빅데이터 분석을 위한 기술적 특성을 살펴보면 분석결과의 정확성과 즉시성을 들 수 있다. 빅데이터 분석은 기업체 내부의 관리자를 위한 사업 동향이나 추세를 분석하는 것뿐만 아니라, 분석 모델을 이용하여 업무 담당자의 판별을 돕는 서비스 시스템 구축을 포함한다. 업무처리의 실수는 기업의 수익과 직결되므로 분석 모델을 통한 업무적 판별의 정확성이 중요하다. 결과의 정확성은 기준이 되는 수준이 주어지게 되며 이는 검증을 통해 측정될 수 있으므로, 검증이 함께 고려된다는 점도 빅데이터의 특성이라고 할 수 있다.

즉시성은 기업 업무에서 필요로 하는 만큼 신속하게 분석 모델을 통한 판별 결과를 제시할 수 있어야 한다는 의미이다. 가령, 온라인 영화 추천서비스를 개발할 경우 사용자가 관심있어하는 정확한 결과를 빠르게 화면에 제공할 수 있어야 한다. 가령, 영화 추천 서비스를 위해서는 정확한 결과를 즉시성 있게 제공하여야 한다. 기존 DW를 통한 현황 통계 분석은 지난달의 결과를 몇 주 후에 산출하므로 차이를 보인다.

빅데이터 분석 프로젝트를 수행하기 위한 방법론은 이러한 정확성과 즉시성을 고려하여 분석 모델을 선별, 설계, 검증할 수 있도록 체계적인 절차가 필요하며 실험의 실패 과정까지 관리할 수 있어야 한다. 이러한 관리를 위해 본 연구에서 제시하는 방법론은 분석모델 설계와 분석보델 구축의 단계를 수시로 반복하면서 실험할 수 있도록 하였다. 뿐만 아니라 본 방법론은 이 두 단계를 분리하여 산출물을 통해 체계적으로 이력관리가 되도록 하여야 한다.

표 2. 데이터 및 기술적 특성 해결 방안
Table 2. Solution for Data & Technical Characteristics

구분	특성	해결 방안
데이터적 특성	- 대용량 - 증가속도 - 다양한 구조	별도의 데이터 분석 시스템(샌드박스)을 구축하여 해결
분석 방법적 특성	- 정확성 - 즉시성	분석모델의 정확성과 즉시성을 관리할 수 있도록 방법론 설계

3.3. 빅데이터 분석 참여자

빅데이터 분석 프로젝트는 폭넓은 범위만큼 다양한 분야의 전문분야가 긴밀히 협력해야 한다는 특징을 가진다. 빅데이터 분석 프로젝트는 산업컨설팅, DW 구축, 운영계 시스템 반영, 빅데이터 분석의 다양한 분야의 전문가가 참여하므로 이들의 역할과 협력 방안을 명확히 할 필요가 있다. 빅데이터 분석 프로젝트 참여자와 그의 역할을 살펴본다. 프로젝트는 발주기관과 수행 기관의 계약에 의해 진행되는데, 발주기관의 역할을 함께 정의하여 역할의 혼돈을 줄이고 원활한 프로젝트 진행을 돕고자 한다. 프로젝트 발주 기관에서는 발주 책임자와 최종 사용자가 관련되며, 프로젝트 수행 기관은 프로젝트 관리자(PM), 산업 전문가, 데이터 기술자, 데이터베이스 관리자, 데이터 과학자를 포함한다.

발주 책임자는 프로젝트가 성공하도록 자극을 가하며 사업적 핵심 문제를 제시한다. 또한 예산을 담당하고 최종 산출물을 업무에 적용하였을 때 가치를 최종 평가한다. 발주 책임자는 발주 기관의 대표, 부서장, 팀장, 기획/정보화 파트에서 사업을 기안하여 발주한 담당자 등이 모두 포함될 수 있다.

최종 사용자는 분석 산출물을 업무에 적용하여 업무를 수행한다. 최종 산출물의 가치와 업무에 활용하는 방법에 대해 자문을 줄 수 있는 매우 중요한 참여자이다. 프로젝트의 산출물인 분석 모델이 운영계 시스템에 도입됨으로 업무에 영향을 받는 모든 사람을 포함할 수 있다.

프로젝트 관리자(PM)는 기간 내에 최고의 품질을 내도록 자원과 일정을 관리하고 책임진다. 프로젝트의 방향과 범위를 발주기관과 협의하여 결정한다. 단계별 산출물을 관리하고 책임지는 역할을 담당한다.

산업 전문가는 발주 책임자 및 최종 사용자와의 면담을 통해 업무를 이해하고 사업적 문제점을 단순화하여 모호성을 줄이고 기업의 문제를 일련의 데이터 분석 문제로 변환 한다. 또한 최종 산출물을 발주 책임자에게 발표하는 역할을 담당한다. 프로젝트의 과업 범위를 산정하고, 핵심성과지표(KPIs)

를 찾아낸다.

데이터 기술자는 분석에 필요한 데이터를 추출하여 샌드박스로 이관 및 확인하는 역할을 담당한다. 분석 모델이 결정되면 운영계 시스템에 구축 도입한다. 데이터베이스 관리자는 분석 업무를 지원하기 위한 샌드박스 내 데이터베이스 환경을 구성하고 제공한다. 분석 모델로 데이터를 분석하기 위한 시스템 구축 및 실행 환경을 담당한다.

마지막으로, 데이터 과학자는 분석 모델을 개발하여 목적한 결과물을 도출하는 핵심 업무를 담당한다. 적합한 분석용 응용프로그램을 사용하여 사업 목적에 맞는 분석 모델을 설계하고 개발하여야 한다. 최종적으로 최종 사용자에게 산출물에 대한 기술적인 내용을 설명하는 역할을 담당한다.

산업 분야, 전산 처리, 데이터 분석 분야에서 다양한 분야의 전문가가 협력할 경우 바라보는 시각, 사용하는 용어, 상호 입장에 대한 이해의 어려움 등으로 인해 효과적으로 협력하기 위한 기본 방안이 필요하다. 프로젝트 참여자마다 각 과업을 바라보는 시간과 역할에 대한 인식이 매우 다를 수 있으며, 또한 시간이 지남에 따라 업무는 그대로인데도 사람의 인식은 바뀌기도 하므로 역할과 담당을 사전에 정리하는 것이 중요하다. 가령, 사업 중간에 특정 공정이 필요하다고 하면 참여자 간에 이의 필요성에 대한 이견이 발생할 수 있을 뿐만 아니라 누구도 업무 부하로 그 공정을 선뜻 맡을 수 없을 것이다. 빅데이터 분석 프로젝트 단계별 과업 범위와 역할을 명확히 하는데 RACI 차트[2]를 사용한다(Chung, 2007). RACI 차트는 누가 어떤 일을 어느 정도의 책임을 가지고 수행하는지를 규명할 뿐만 아니라, 의사소통(보고) 및 협력 체계(자문 지원, 협업 등)를 구체화함으로써 구성원 간의 유기적 협력과 효율적 업무 수행을 돕는다.

VI. 과학적 데이터 분석 방법론 설계

빅데이터 분석 프로젝트의 특성을 해결하기 위해 앞장에서 제시된 방안을 본 장에서는 구체화 한다. 필요한 공정은 단계별로 정리하며, 공정별 결과물은 단계별 산출물로 체계화 한다. 빅데이터 분석 프로젝트에 필요한 일련의 단계별 공정 및 산출물은 "과학적 데이터 분석 방법론(SDAD: Scientific Data Analysis & Development Methodology)"이란 이름을 부여한다. 앞장에서 제기된 참여자 간의 소통과 협력의 문제점은 RACI 챠트를 통해 기준 방안을 제시한다.

2) RACI 차트: 업무 프로세스 상의 부서/개인간 업무에 대한 역할과 책임 및 권한을 명확히 설명하는 표.

4.1. 과학적 데이터 분석 방법론 단계별 산출물

"과학적 데이터 분석 방법론"은 문제정의, 데이터준비, 모델설계, 모델구현, 결과평가, 서비스구현의 6단계로 구성된다. 여러 단계가 동시에 수행될 수도 있고, 필요시 이전 단계로 언제든 돌아갈 수 있다. 각 단계에서 필요한 정보를 충분히 얻었는지 그리고 충분한 진척이 있는지 판단하여 다음 단계로 진행한다. 그림 1에 보면 3단계와 4단계가 점선으로 묶어져 있고 이전 단계로 돌아갈 수 있도록 표시되었음을 볼 수 있는데, 이는 두 단계가 긴밀한 협조가 필요하다는 점을 강조하기 위함이다. 빅데이터 분석 실무에 활용하도록 방법론 단계별 산출물 목록을 정리한다. 문제정의단계는 ISP와 BI 방법론을, 데이터준비 단계는 DW 구축 방법론을, 서비스구현 단계는 SW 개발 방법론과 공유할 수 있도록 하여 기존의 타 프로젝트 산출물을 재사용 하거나, 본 프로젝트의 산출물을 타 관련 프로젝트에 사용하기 용이하다. 산출물 목록은 상기의 모든 산출물을 포괄하므로 이중 필요한 부분만 추출(tailoring)하여 사용하기 편리하다.

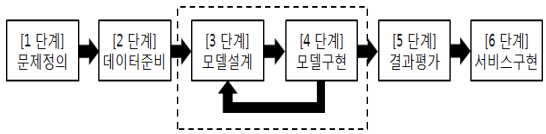

그림 1. 과학적 데이터 분석 방법론 6 단계
Fig. 1. Six stages of scientific data analysis & development methodology

① 문제정의단계: 발주 기관(이하 '기업'과 혼용하여 사용)의 사업적 핵심 문제점을 발견하고, 이러한 문제점을 해결했을 시 사업적 가치를 파악한다. 사업적 문제를 데이터 분석의 문제로 정형화하고 성공과 실패를 판별할 수 있도록 가설을 개발한다. 데이터 분석의 문제를 풀기 위해 필요한 데이터를 파악하고, 관련 설명서와 샘플 데이터를 요청하여 기초 분석을 수행한다. 선택한 분석 모델을 구현하기 위해 필요한 분석용 응용프로그램, 인원, 기술 등 자원을 파악한다. 프로젝트 내·외적으로 체계적인 의사소통이 되도록 참여자별 담당 업무와 책임을 명확히 한다.

문제정의단계는 환경분석, 현황분석, 목표정의의 세 부분으로 나누었다. 환경분석 부분은 외부환경과 내부환경을 분석하고 발전전략을 마련한다. 현황분석 부분은 업무현황, 정보시스템현황, IT 아키텍처분석, 데이터 현황 분석으로 구성된다. 목표정의 부분은 분석 목표를 수집하고 샌드박스 시스템 구조를 설계하며, 보고서 산출물의 형식 정의를 포함한다.

표 3. 문제정의 단계 산출물(1단계)
Table 3. Production lists of problem definition stage

단계	공정명	산출물명	설 명
문제정의	사업계획	사업 수행 계획서	프로젝트 목표 및 범위 확인 및 조직 구성(WBS, RACI 차트)
(환경분석)	외부환경분석	외부 환경 분석서	외부 환경 요인 분석 IT 동향 및 선진 사례 분석
	내부환경분석	내부 환경 분석서	문제점 및 내부 역량 분석
	발전전략마련	발전 전략 분석서	SWOT 분석, 핵심성공요소(CSFs) 및 핵심성과지표(KPIs) 도출
(현황분석)	업무현황분석	현행 업무 기능 절차도	조직 및 업무를 기능 중심으로 분석
		업무별 정보화 분석서	업무별 정보화 요구 사항 분석
		현행 업무 흐름도	업무 프로세스 분석
	정보시스템 현황분석	사용자 만족도 조사	
		기능 적정성 진단	
		사용성 분석	시스템 사용 편의성 조사
	IT아키텍처 분석	응용/데이터/기술 인프라/보안 현황 분석	기간계, EDW 현황 분석(ERD, 테이블명세서, 코드정의서, 용어정의서, 인터페이스 정의서, H/W, S/W, N/W, App 등)
		시스템 간 연계 구조 분석	
	데이터현황 분석	데이터 적재 대상 정의	데이터 수집 목록 및 활용 방안
		데이터 사전 분석	샘플 데이터를 통한 사전 분석
	개선사항도출	요구 사항 정의서	요구사항 분석 결과 정리 및 개선사항 도출(요구사항명세서/기술서/설명서)
		요구 사항 추적표	
(목표정의)	분석목표수립	투자 대비 효과 사전 분석서	
		데이터 분석 문제 정형화	업무적 문제를 데이터 분석 문제로 정형화(귀무가설 설계, 프로젝트 성공/실패 기준 설정)
		목표 시스템 정의	분석모델을 기간계에 반영 방향 마련
	샌드박스설계	샌드박스 아키텍처설계서	저장 공간, 속도, 구조 등 설계
	보고서표준화	보고서 표준 가이드	문서의 폰트, 좌우 여백 등

② 데이터준비단계: 프로젝트 수행기간 동안 분석 작업을 수행할 시스템(샌드박스)을 구축하고 분석환경을 마련한다. 자유롭게 다양한 분석을 수행하도록 분석에 필요한 모든 데이터를 샌드박스에 적재한다. 적재된 샌드박스 내 데이터의 기초분석, 오류정제, 표준화 등 데이터 품질을 높이는 작업을 수행한다. 비구조적 데이터는 분석 모델이 이용할 수 있도록 변형한다. 분석 모델에 사용가능한 유의한 속성을 도출한다.

③ 모델설계단계: 다양한 데이터 소스로부터 수집된 테이블의 속성들 간의 관계를 파악하여 업무 흐름과 데이터에 대한 개념적 이해를 명확히 한다. 필요한 경우 분석 모델을 이해하기 쉽도록 속성을 재생산한다. 분석 목적에 해당하는 속성(반응변수, 종속변수)과 상관관계가 높은 속성(설명변수, 독립변수)을 선별한다. 입력 변수와 출력 변수로 사용할 속성의 특성과 데이터양, 목적 등을 고려하여 분석 모델을 결정한다. 샘플 데이터를 이용하여 분석 모델의 실행 가능성을 점검한다.

표 4. 데이터준비단계 산출물(2단계)
Table 4. Production lists of data preparation stage

단계	공정명	산출물명	설명
데이터준비	샌드박스 구축	샌드박스 구축결과보고서	HW 현황, 네트워크 현황, SW 설치 현황
	매핑(표준)	매핑 표준 정의서	데이터 매핑을 위한 표준 정의
	매핑(설계)	소스 타겟 매핑 정의서	
	데이터이관 (표준)	개발 표준서	명명 규칙, ELT 개발 표준을 정리
	데이터이관 (설계)	데이터 이관 설계서	초기 데이터 이관 및 구축 계획
		ELT 흐름도	데이터 처리의 병목현상 표현
		ELT 프로그램 내역서	
		ELT 유니버스 명세서	
	데이터저장 (표준)	용어사전	테이블과 속성에 사용될 용어 정리
		DB 및 파일 설계표준	명명 규칙, SQL 작성 가이드, 속성 타입 정의 규칙(날짜와 Varchar 혼용 금지 등)
	데이터저장 (설계)	DB 및 파일 논리 설계서	데이터를 샌드박스에 적재 후 ERD, 테이블 정의서 등 설계
		디멘전 정의서	
		도메인 정의서	속성 내 값의 범위 정의(예, Y/N)
		코드 정의서	
		팩트 정의서	
	데이터 이관코드작성	ELT 프로그램 소스	
		DB 및 파일 물리 설계서	
	테스트	단위 테스트 결과서	
		이관 환경 구성 가이드	

표 5. 모델설계단계 산출물(3단계)
Table 5. Production lists of model design stage

단계	공정명	산출물명	설명
모델설계	분석모델 (표준)	개발 표준 가이드	명명 규칙, 주석 방법 등
		개발 환경 구성 가이드	로컬 개발 환경 구성 방법 등 설명서 작성
	분석모델 (설계서)	입출력 변수 명세서	입력 및 출력 변수 정의
		로직 설계서	로직 또는 알고리즘 설계서
	데이터설계	학습용 테이블 명세서	학습용 데이터
		검증용 테이블 명세서	검증용 데이터
	분석실행 시스템구조	분석 모델 실행 시스템 아키텍쳐 설계서	
	테스트	성능 평가 설계서	샘플 데이터로 수행한 결과 추가
		단위 시험 케이스	분석 모델 알고리즘 테스트

표 6. 모델구현단계 산출물(4단계)
Table 6. Production lists of model development stage

단계	공정명	산출물명	설명
모델구현	모델코드작성	프로그램 코드	분석 모델 프로그램 코드
	데이터구축	학습용 데이터	
		테스트용 데이터	
	테스트	단위 시험 결과서	분석 모델 코드 오류 검증 중심
		시스템 시험 결과서	시스템 성능 검증 중심
	모델평가	분석 모델 효과 평가 결과서	가설 검증 중심

④ 모델구현단계: 분석 모델을 실행하기 위해 학습용(training) 데이터셀과 검증용(testing) 데이터셀을 구성한다. 분석 모델을 실행하기 위한 최적의 시스템 환경을 구성하고, 분석 성능을 높일 수 있는 방안을 강구한다. 분석 모델을

실행하고 결과를 평가한다. 결과가 효과가 있는지 아니면 실패가 확실한지 판별될 때까지 모델설계단계와 반복 수행한다. 본 단계에서의 평가는 분석모델의 성능 및 유의성을 파악하는 것에 집중한다.

표 7. 결과도출단계 산출물(5단계)
Table 7. Production lists of result extraction stage

단계	공정명	산출물명	설명
결과보고	가치평가	사업적 가치 평가 결과서	사업적 가치 측면에서 분석 결과 재평가
	결과발표	관리자용 보고서	발주 책임자 등 의사결정자 대상
		개발자용 보고서	최종 사용자 등 실무자 중심
	도입방향설정	분석 모델 도입 방향 보고서	분석 모델이 업무에 미치는 영향 포함

표 8. 서비스구현단계 산출물(6단계)
Table 8. Production lists of service development state

단계	공정명	산출물명	설명
서비스구현	요구사항분석	유스케이스 명세서	시스템 화면 변경 시 변경 화면 정의 및 요구 사항 상세 내용 포함
(분석)	도입구축계획	개발 수행 계획서	분석 모델을 기간계에 반영 계획
(설계)	DB설계 (표준)	용어 사전	테이블과 속성에 사용될 용어 정리
		도메인 정의서	속성 값의 범위 정의(예, Y/N)
		모델링 가이드	DB 정규화 수준, ERD 작성 가이드, 매핑 가이드 등
		DB 설계 표준	
	DB설계 (설계서)	DB 설계서	
		객체 설계서	객체로 구분 가능한 테이블 정의
		데이터 이관 설계서	
		엔티티 관계 모형 설계서	논리모델 중심
	응용시스템 (표준)	웹 스타일 가이드	웹 화면의 디자인 가이드
		화면 표준 가이드	서비스화면이 있을 경우에 적용
		개발 표준 가이드	명명 규칙, 주석 방법 등
		개발 환경 구성 가이드	
		프레임웍 가이드	프레임웍 설명서
	응용시스템 (설계)	화면 설계서	
		메뉴 구성도	
		프로그램 명세서	
	연계/통합 (표준)	연계 표준 정의서	시스템 연계 또는 통합 표준 정의
	연계/통합 (설계)	인터페이스 설계서	기간계 시스템과 외부 빅데이터 연계 또는 통합을 위한 설계
	시스템 구조	아키텍처 설계서	
	테스트	통합 시험 시나리오	
		단위 시험 케이스	
(구현)	코드작성	프로그램 코드	
		DB 테이블	
		CRUD 매트릭스	물리 모델 중심(테이블, 프로그램 기준)
	단위테스트	단위 시험 결과서	
(검증)	통합테스트	통합 시험 결과서	
		시스템 시험 결과서	시스템 성능 중심
		시스템 설치 결과서	
		인수 시험 시나리오	
		인수 시험 결과서	
	매뉴얼	사용자 지침서	분석 모델을 기간계 시스템에 도입함에 따른 업무 변경 기술
		운영자 지침서	시스템의 변경된 옵션, 가동 절차 등을 모두 기록
	도입효과 재평가	서비스 재평가 보고서	분석모델 도입에 따른 사업적 효과와 시스템 안정성 재점검

⑤ 결과도출단계: 분석 모델을 실행하여 도출된 최종 결과물을 점검하고, 사업적 측면에서 결과의 가치를 재평가한다. 주요 발견 사항의 사업적 가치를 발주 기관의 관계자가 판단할 수 있도록 명확한 보고서와 시연(데모)을 준비한다. 발주 책임자에게 최종 결과물을 발표하고, 업무에 활용할 방안을 마련한다. 업무 관련자에게는 기술적 내용을 설명한다.

⑥ 서비스구현단계: 분석 모델을 파일럿 서비스를 통해 실 서비스에서 운영해 본 후, 안정적으로 확대하여 운영계 시스템에 구축한다. 일정 기간 분석 모델을 운영계 시스템에 운영한 후, 예상한 대로 수익이 증가하고 목표한 효과가 나타나는지 확인한다. 시스템 운영 상황도 재점검 한다. 프로젝트의 최종 산출물을 발주 기관에 전달한다.

서비스구현단계는 SW 개발 방법론에 따라 기본적으로 필요한 일련의 산출물을 종합적으로 정리하였다. 서비스구현단계는 분석 결과로 만들어진 분석 모델을 실 운영계 시스템에 적용하므로 또다시 분석, 설계, 구현, 시험의 과정이 필요하다. 산출물의 구성은 분석 단계에는 요구 사항 분석과 현황 분석이 중심이 된다. 설계 단계에는 DB, Application(App), 연계/통합, 시스템 구조, 테스트로 구분된다. 시스템 개발에 직접적으로 관련된 DB, App, 연계/통합 부분은 표준과 설계로 구분되어있다는 것을 유의한다. DB는 설계를 하기 전에 사용되는 용어와 설계 방법에 대한 표준을 정하고, 그 후 그러한 표준을 준수하면서 설계를 진행하여야 한다. App 또한 마찬가지로 디자인과 개발에 대한 표준을 만든 후, 이를 준수하면서 설계하여야 한다. 시스템 연계 및 통합 또한 표준을 정한 후 설계를 진행하는 것이 매우 중요하다. 시스템 구현 후 테스트 방안을 마련해야 한다. 또한 시스템 운영을 위한 매뉴얼을 작성하여야 한다.

4.2 참여자간 의사소통 방안(RACI 차트)

프로젝트 단계별 진행 과정에서 참여자의 책임 업무 및 의사소통 기준을 표 9에서와 같이 RACI 차트로 마련한다. 실 업무 특성에 맞춰 제시된 차트를 참고하여 사용한다면 프로젝트 관리가 용이할 수 있다. 챠트의 세로축은 각 단계별로 과업을 세분화하여 리스트하며 가로축은 프로젝트 참여자를 파트로 나타낸다. 차트에서 각 전문가 파트(예, 데이터 기술자 파트)가 한명이 아니고 주로 팀으로 구성되므로 담당은 개인이 아닌 파트 명을 기입한다. 기능별 역할 및 관계는 담당(Responsible), 책임(Accountable), 자문(Consult), 알림

(Inform)의 네 가지로 구분한다. 담당은 실제로 과업을 수행하여 완수하는 파트이며, 책임은 의사결정 권한을 가진 파트로서 동일한 과업에 대해서는 단 한 파트만이 책임을 가지도록 업무를 배분하며, 자문은 전문가로 자문을 제공하는 파트이며, 알림은 의사결정 또는 업무 완료 후 이러한 소식을 연락(보고)받는 파트이다. 파트별로 업무와 담당자를 더욱 세분화할 수 있다.

사업계획단계에서 산업전문가는 발주 책임자와 최종 사용자의 의견을 수렴하여 사업적 핵심 문제점을 발견하고, 데이터 분석 문제로 정형화하여 이를 데이터 과학자에게 알린다. 데이터 과학자는 산업 전문가와 함께 필요한 데이터를 파악하고 샘플 데이터를 통해 사전 분석을 수행한다. 이와 동시에 프로젝트 경영자는 데이터과학자와 협의하여 필요한 자원과 인력을 파악하여 일정 계획을 세우고, 그 결과를 모든 참여인원에게 알려 공유한다.

데이터준비단계에서는 데이터베이스 관리자가 샌드박스를 구성하고 데이터 기술자에게 알리면, 데이터 기술자가 필요한 데이터를 적재한 후 데이터 오류를 확인한다. 데이터 기술자는 데이터 과학자, 업무 분석가와 협의하여 데이터를 점검하고 유의한 속성을 도출한다. 데이터 적재와 기초적인 검증이 완료되면 데이터 설명서를 작성한다. 데이터 기술자는 산출물이 완성되면 데이터 과학자에게 알린다.

모델설계단계는 데이터 과학자가 업무 분석가와 협력하여 데이터 속성간의 관계를 깊게 탐구하여, 분석 모델에 사용할 변수(속성)를 선별하고 재생산한다. 데이터 과학자는 추출된 속성의 특성에 맞추어 분석 모델을 설계하고 로직을 개발한다. 실행해 볼만한 분석 모델과 이에 사용할 변수가 구상되면 이에 대한 데이터셋을 개발해야하므로 데이터 기술자와 데이터베이스 기술자에게 결과를 알린다.

분석 모델이 완성되면 데이터 과학자는 학습용과 검증용 실 데이터셋을 구축해야하며 이를 위해 데이터 기술자와 데이터베이스 관리자의 협조가 필요하다. 이후 데이터베이스 관리자가 프로젝트 수행을 위한 최적의 시스템 환경을 구축한 후 데이터 과학자와 데이터 기술자에게 알리면, 이들은 분석 모델을 실행한다. 데이터 과학자는 분석 모델을 실행하고 데이터 관점에서 결과의 타당성을 평가한 후 결과를 산업 전문가에게 알린다.

결과보고단계에는 산업 전문가가 분석 결과의 가치를 사업적 측면에서 재평가한다. 산업 전문가는 주요 핵심 사항을 중심으로 발주 책임자에게 결과를 발표한다. 데이터 과학자는 내·외부 사용자 및 운영계 시스템 관리자에게 보다 기술적인 측면에서 결과물을 설명한다. 교육과 함께 최종 산출물을 실업무에 반영하는 방안을 마련한 후, 결과를 데이터 기술자와 데이터베이스 관리자에게 알린다.

서비스구현단계에는 데이터 기술자가 데이터 과학자, 데이터베이스 관리자와 협력하여 분석 모델을 운영계 시스템에 구축한다. 일정 기간이 지난 후에 산업 전문가와 데이터 과학자는 분석 모델의 사업적 가치를 재평가하고, 평가 결과를 데이터 기술자와 데이터베이스 관리자에게 알린다. 데이터 과학자는 일정 기간 후 시스템의 안정적 운영 측면에서 재점검한다.

프로젝트 책임은 발주 책임자에 있으므로, 모든 진척 사항은 발주 책임자에게 보고되도록 작성되었다. 프로젝트 경영자는 진척 사항을 확인하고 단계별 산출물을 정리하는 책임을 지도록 명시되어있다. 사업의 관리는 프로젝트책임자(PM)을 별도로 두더라도 사업전반에 있어 데이터과학자가 분석모델 개발의 중심적 역할을 수행할 수 있도록 협력체계를 구성하였다. 발주기관의 사업적 문제 및 이를 분석모델을 통해 해결하였을 시의 사업적 가치는 산업전문가가 판별하도록 하여 산업분야의 전문성을 유지하였으며, 최종발표 또한 산업전문가가 담당하도록 하여 발주기관과의 의사소통을 원활히 하였다. 뿐만 아니라, 데이터 기술자와 데이터베이스 관리자를 별도로 두어 전산 기술을 안정적으로 지원받도록 하였다.

표 9. 과학적 데이터 분석을 위한 RACI 차트
Table 9. RACI chart for SDAD

코드	과업명	최종사용자	발주책임자	프로젝트경영자	산업전문가	데이터기술자	DB관리자	데이터과학자
A	**문제정의단계**							
A-001	사업적 핵심 문제점 및 가치 발견	C	CI	I	RA			C
A-002	사업적 문제를 데이터 분석 문제로 정형화		CI	I	RA			RI
A-003	필요한 데이터 파악 및 분석 모델 방향 설정		RI	I	R			RA
A-004	자원 파악 및 인력 구성		CI	RA	I	I	I	CI
A-005	단계별 산출물		I	RA				
B	**데이터준비단계**							
B-001	샌드박스 준비		I	I		I	RA	C
B-002	데이터 이관 적재		I	I		RA		I
B-003	데이터 점검 및 유의한 속성 도출	C	CI	I	R	RA		RI
B-004	단계별 산출물		I	RA				
C	**모델설계단계**							
C-001	데이터 탐구		I	I	C			RA
C-002	속성 선별 및 변형	C	CI	I	R			RA
C-003	분석 모델 결정 및 설계		I	I	C	I	I	RA
C-004	단계별 산출물		I	RA				
D	**모델구현단계**							
D-001	데이터셋 개발		I	I		RA	R	R
D-002	분석 모델 실행 환경 구축		I	I		I	RA	CI
D-003	분석 모델 실행 및 평가		I	I	I	R	R	RA
D-004	단계별 산출물		I	RA				
E	**결과보고단계**							
E-001	최종 산출물 가치 평가	C	I	I	RA			C

E-002	발주 책임자에게 최종 발표			I	RA			I
E-003	업무 담당자 및 최종 사용자에게 기술적 내용 발표		I	I	I	R	R	RA
E-004	최종 산출물을 업무에 활용하는 방안 마련	C	I	I	R	I	I	RA
E-005	단계별 산출물		I	RA				
F	**서비스구현단계**							
F-001	기간계 시스템에 분석 모델 적용		I	I	I	RA	R	R
F-002	도입 효과 및 안정성 재점검	C	I	I	R	C	C	RA
F-003	최종 산출물 전달		I	RA				

* 본 RACI 테이블은 프로젝트 특성에 맞추어 조정하여 사용

V. 방법론의 평가 및 결과

5.1. 방법론의 평가 방법

"과학적 데이터 분석 방법론"의 실무적 활용 타당성은 한국고용정보원의 고용보험과 산재보험을 연계 분석하는 프로젝트에 적용하여 평가하였다. 산재보험 업무를 분석하여 고용보험과 연계 범위를 설정하고, 사업장 및 개인을 연계할 수 있는 키를 개발 후 유효성을 실시간으로 검증하고, 이를 DW로 구축하여 동향 분석 시스템을 구축하는 과업이다. 본 사업의 성패를 결정할 가장 중요한 부분은 사업장과 개인을 식별할 수 있는 키를 찾아내어 매핑 후 실시간으로 키의 매핑 유의성을 평가하는 작업으로서, 본 과정을 충분히 검증이 필요하다. 사업장의 경우 1:1로 매핑되는 대표키가 존재하지 않고, 개인의 경우도 주민등록번호만을 이용하여 매핑 할 수 없는 상황에서 키 매핑을 위해서는 별도 분석모델이 개발되어야 한다. 또한 분석모델을 통해 매핑이 되지 않은 키에 대해서는 실시간으로 로그 기록을 제공하여야 하므로 빅데이터 분석 기술의 적용이 필요하다.

기 프로젝트는 정보화 컨설팅, 시스템 개발, 데이터웨어하우스 구축, 빅데이터 분석의 네 가지 성격을 모두 갖추어 본 방법론의 활용 타당성을 평가하기에 적합하다. 고용보험과 산재보험 업무를 분석하여 고용정책에 활용 가능한 지표를 찾아내는 작업은 정보화 컨설팅이며, 연계시스템 구축 및 현황 테이블 생성은 시스템 개발(SI) 사업이며, 데이터 마트를 구축하여 통계보고서 생성하는 작업은 데이터웨어하우스 구축 사업이며, 사업장과 개인을 식별할 수 있는 키를 찾아내어 유의성을 평가하는 작업은 빅데이터 분석 사업 성격을 가진다. 기 프로젝트의 기간은 12개월이며, LG C&S 와 밸텍컨설팅이 컨소시엄으로 프로젝트에 참여하였다. 정보연계 및 분석을 위해 전문 기술을 보유한 다수의 협력업체가 참여하였다. 산재보험 및 고용보험 업부 범위 분석 범위는 컨설팅 업체에 관련되며, 산재보험과 고용보험 데이터의 실시간 연계는 DW 및 SI 분야와 관련된 다수 기업이, 대표키를 개발 하고 검증하는 하는 분야는 빅데이터 관련 전문 기술을 보유한 기업이 참여하였다. 다양한 이질적 분야의 기업이 참여함으로 각 참여기업 간의 의사소통과 품질 관리를 위하여 활동 및 산출물들의 표준이 필요하였다. 사용자 요구사항은 제안서와 주관기관인 한국고용정보원의 관련 담당자와 회의에 의하여 수집하였다.

문제정의 단계에서는 정보시스템현황분석 공정에서 사용자 만족도 조사 및 사용성 분석은 제외되었다. 신규시스템 구축이므로 기존 사용자 만족도 조사는 불필요하다. 다만 기존 시스템의 성능을 고려하여 신규 시스템을 설계해야하므로 기능적적정성을 진단하였다. 데이터준비단계는 가장 시간이 많이 소요된 단계로서 지속적으로 데이터 검증과 사업기간 내내 데이터 정의서 산출물의 현행화를 점검하였다. 모델설계단계는 키 매핑로직을 설계하는 하였다. 모델구현단계는 매핑로직을 구현한 후 샘플 데이터를 이용하여 검증하였다. 매핑로직의 검증은 고객사와 회의를 통해 충분히 협의 후 결정하였다. 결과도출단계에서는 매핑로직은 실시간으로 검증하여 문제발생 시 경고(alert)를 내도록 도입방향을 정하였다. 서비스구현단계는 실시간 데이터 현황 테이블을 구축 후, DW 구축 및 통계 결과를 화면으로 개발하였다. 도입효과 재평가는 매핑로직을 통해 추출된 키 오류를 2개월간 사용 후 필요한 로직을 수정 보완하였다. 분석모델은 업무상의 보안 문제로 본 논문에서 상세하게 공개할 순 없지만, 객관적인 품질 평가와 방법론의 실무적 효용성은 감리 결과를 통해 평가할 수 있다.

5.2. 방법론의 검증 결과

본 사업에 적용된 방법론은 감리를 통해 실무적인 검증절차를 거치도록 하였다. 감리는 1차감리, 중간감리, 최종감리로 수행하였으며, 데이터 품질 검증을 위해 수시감리를 1회 실시하였다. 요구분석, 분석 결과 품질, 구축에 이르기 까지 전체적인 수행에 대한 감리결과는 전체 요구사항 23건에 대해 모두 "적합"한 것으로 최종 판정되었고, 산출물작성의 적정성 측면에서 운영자/사용자지침서에만 일부 보완이 필요한 것으로 나타났다. 도출된 개선 권고내용은 보환한 내용을 감리인에게 확인을 받아 감리가 최종 종료되었다. 다양한 이질적 기업이 참여함에도 불구하고 사업관리 및 의사소통의 부족에 대한 지적은 나타나지 않았다.

표 10. 감리 조치결과
Table 10. Results of Supervision

요구사항	검사기준	적/부 판정
산재보험-고용보험 사업장 관계식별 방안 분석 1건 **(정보화 컨설팅 분야)**	- 실제 데이터 분석을 통해서 사업장 간의 관계가 적절하게 식별되었는지 확인	적합
정보품질강화 1건 **(빅데이터 분석 사업 분야)**	- 매핑 키 기준에 부합하지 않는 데이터에 개한 로직이 구현되고 이의 검증이 이루어지면 적합 - 산재보험에서 연계되는 정보 중 건수, Key값등을 검증하여 타겟 컬럼 들이 누락없이 마스터데이터 모델에 적재되면 적합 - 산재보험과 고용보험 사업장, 피보험자 정보에 대한 비교 검토 실시를 하였고 해당 기준으로 데이터가 적재되었으면 적합	적합
산재보험정보 통합 마스터 데이터 모델 구축 관련 4건 **(연계시스템 구축 관련)**	- 고객핵심정보 및 코드정보는 마스터데이터베이스를 사용하여 관리하고 ETL 스케쥴에서 해당 데이터의 배포 기능이 구현되었으면 적합 - MDM 내부 Merge, 삭제 등의 데이터를 산재보험 데이터에 반영하고 있으면 적합	적합
산재보험정보 통합 DW 데이터 모델 구축 관련 17건 **(데이터웨어하우스 구축 관련)**	- 산재보험 사업장, 피보험자 정보가 DW에 구축되어 있으면 적합 - 산재보험 사업장 변경이력이 관리 가능하도록 데이터 모델링이 되었으면 적합 - 산재보험 사업장 변경이력이 테이블에 적재되면 적합	적합

* 적/부판정: 적합/부적합으로 판정하며 요구사항을 검사기준에 맞추어 달성 시 적합 판정

"과학적 데이터 분석 방법론(SDAD)"은 프로젝트에 간편하게 적용하도록 개발과 관리가 혼합된 방식이지만, 관리보다는 CBD SW 개발 방법론에 기반을 둔 개발 방법론에 가깝다. PMBOK(Project Management Body of Knowledge)과 비교하여 관리영역이면서 SDAD에 포함된 산출물을 살펴보면, 원가와 조달구매 영역을 제외한 대부분의 산출물이 포함됨을 알 수 있다.

표 11. PMBOK 방법론과 SDAD 비교
Table 11. Comparing SDAD with RACI

관리영역	PMBOK	SDAD
통합	사업수행계획서	사업수행계획서
범위	WBS, 요구사항 정의서	사업수행계획서, 요구사항정의서
인력	인력투입계획서, 공정별 인력투입현황표	사업수행계획서, 주간/월간보고서
품질	품질관리계획서, 테스트결과서	성능평가설계서, 단위시험케이스, 단위시험결과서, 시스템시험결과서
위험	위험관리계획서, 위험관리 등록부	주간보고, 월간보고, 수시보고
일정	WBS	사업수행계획서
의사소통	주간보고, 월간보고, 수시보고	주간보고, 월간보고, 수시보고
원가	실행예산계획서, 집행내역서	-
조달구매	조달구매계획서, 납품내역서	-

V. 결 론

빅데이터가 사회적 이슈로 대두됨에 따라 빅데이터 분석 프로젝트 수행에 필요한 절차를 정리한 "과학적 데이터 분석 방법론(SDAD)"을 제시하였다. "과학적 데이터 분석 방법론"은 문제정의, 데이터준비, 모델설계, 모델구현, 결과평가, 서비스구현의 6단계로 구성한다. 특히 모델설계단계와 모델구현단계는 긴밀한 협조가 필요하므로 반복 수행될 수 있도록 하였다. 각 단계에서 필요한 정보를 충분히 얻었는지 그리고 충분한 진척이 있는지 판단하여 다음 단계로 진행한다. 단계별 역할을 명확히 하기 위해 필요한 공정으로 구분하였다. 공정개수는 단계별로 12개, 8개, 5개, 4개, 3개, 15개를 각각 포함하며, 산물물은 단계별로 21개, 18개, 9개, 6개, 4개, 35개를 포함하여, 전체 47개 공정명과 93개의 산출물을 포함한다.

빅데이터 분석 실무에 용이하게 활용할 수 있을 뿐만 아니라 관련 방법론과 산출물의 공유를 원활히 하도록 기존의 ISP, DW, SW 개발 프로젝트와 쉽게 연계할 수 있도록 하였다. 문제정의단계는 ISP 방법론을, 서비스구현단계는 DW 구축 방법론과 SW 개발 방법론을 참조하였다. 산출물 목록에는 상기의 모든 산출물을 상세히 하였으므로, 이중 필요한 부분만 추출(tailoring)하여 사용하기 편리하도록 하였다. 기존의 실무 프로젝트 방법론이 "과학적 데이터 분석 방법론"에 단계별로 구분되어 적용되므로 기존 프로젝트 산출물을 그대로 가져다 사용할 수도 있으며, 반대로 빅데이터의 산물을 다른 프로젝트에 사용이 용이하다.

빅데이터 분석 프로젝트의 참여자와 그의 역할을 정리하여 책임과 권한을 명확히 하고 의사소통을 용이하게 하였다. 프로젝트 참여자는 프로젝트 발주기관과 수행기관 모두를 포함하여 의사소통의 현실성을 높였다. 프로젝트 발주 기관에서는 발주 책임자와 최종 사용자가 관련되며, 프로젝트 수행 기관은 프로젝트 관리자(PM), 산업 전문가, 데이터 기술자, 데이터베이스 관리자, 데이터 과학자를 포함한다. 이중 데이터 과학자는 프로젝트 전반에 기술적으로 관여하며, 분석 모델을 개발하여 목적한 결과물을 도출하는 핵심 업무를 담당하도록 하였다. 과업은 25개로 분류하고 발주기관을 포함한 7개의 프로젝트 참여자 간의 역할을 RACI 챠트로 정의하였다. 프로젝트 참여자의 기능별 역할 및 관계는 담당, 책임, 자문, 알림의 네 가지로 구분하였다. 이를 통해 과업의 담당자뿐만 아니라 진척 과정에서 의사소통 전달 책임 까지를 정의하였다.

본 연구에서 소개된 방법론을 한국고용정보원에서 발주한

빅데이터 프로젝트에 적용하여 감리를 통해 실무적 검증절차를 거치도록 하였다. 감리결과는 전체 요구사항 23건에 대해 모두 "적합"한 것으로 최종 판정되었고, 산출물작성의 적정성 측면에서 운영자/사용자지침서에만 일부 보완이 필요한 것으로 나타났다. 다양한 이질적 기업이 참여함에도 불구하고 사업관리 및 의사소통의 부족에 대한 지적은 나타나지 않아 SDAD 방법론이 실무에서도 활용이 가능한 것으로 검증되었다.

본 연구에서 제안한 SDAD 방법론을 실무에 적용하여 검증 과정을 거치고, 테일러링 기법을 통해 다양한 종류의 빅데이터 분석 프로젝트에 적용이 용이하도록 하였다. 하지만 빅데이터 분석 프로젝트 수행을 일반화하여 다양한 케이스를 고려한 객관적 타당성을 검증하였다는 점에는 한계가 있다. 이러한 한계에도 불구하고 빅데이터 프로젝트 수행 및 향후 관련 방법론 연구에 본 결과가 중요한 기초 자료가 될 것으로 사료된다.

본 논문은 지면의 한계 상 방법론의 산출물 형식을 추가하지 못하고 기본 구조만을 정리하였다. 향후 산출물 형식을 웹사이트 등을 통해 제공하여 본 방법론의 활용을 활대할 필요가 있다. 본 방법론은 개발 방법 측면에 치중하여 프로젝트 발주 및 유지보수에 관련한 부분이 생략되어있다. 빅데이터 분석 프로젝트는 다른 프로젝트와 달리 유지보수 과정에서 분석모델을 변경해야할 가능성이 높다. 따라서 이러한 빅데이터 분석 프로젝트의 특성을 고려하여 유지보수 관리 방안을 방법론으로 정형화할 필요가 있다. 또한 본 방법론을 통해 프로젝트를 수행 시 어느 정도의 효과가 있는지 정량적 지표를 개발할 필요가 있다.

참고문헌

[1] H. P. Andres, and R. W. Zmud, "A contingency approach to software project coordination," Journal of Management Information Systems, Vol. 18, No. 3, pp. 41-70, 2002.

[2] K. Beck, C. Andres, "Extreme Programming Explained: Embrace Change, 2nd Edition," Addison-Wesley, 2004.

[3] P. Bresciani, A. Perini, P. Giorgini, F. Giunchiglia and J. Mylopoulos, "Tropos: An Agent-Oriented Software Development Methodology," Autonomous Agents and Multi-Agent Systems, Vol. 8, No. 3, pp. 203-236, 2004.

[4] J. Castro, M. Kolp and J. Mylopoulos, "A Requirements-Driven Development Methodology," Lecture Notes in Computer Science, Advanced Information Systems Engineering, pp. 108-123, 2001.

[5] S. H. Choi, "Simulation project management methodology," Proc. of Korean Academic Society of Business Administration, Korean Academic Society of Business Administration, Vol. 1, pp. 289-289, 1999.

[6] S. Chung, "Service-Oriented Software Reengineering: SoSR," 40th Annual Hawaii International Conference on System Sciences, 2007.

[7] EMC, "Data Science and Big Data Analytics Student Guide," http://www.emc.com/, 2012

[8] F. Giunchiglia, J. Mylopoulos and A, Perini, "The Tropos Software Development Methodology: Processes, Models and Diagrams," Lecture Notes in Computer Science, Agent-Oriented Software Engineering III, pp. 162-173, 2003.

[9] B. K. Jeong, D. S. Kim, J. H. Song, J. S. Whang, "Development and Application of Analysis & Design Methodology for Web - based System," Communications of the Korean Institute of Information Scientists and Engineers, Korean Institute of Information Scientists and Engineers, Vol. 8, No. 2, pp. 155-166, 2002.

[10] K. J. Jeong, "Methodology for the structure of CRM customer centers and the development," Proc. of Korean Academic Society of Business Administration, Korean Academic Society of Business Administration, pp. 243-246, 2002.

[11] S. Kim, "The analysis of data governance model for business and IT alignment," Journal of The Korea Society of Computer and Information, Vol. 18, No. 7, 2013.

[12] J. P. Kuilboer and N. Ashrafi, "Software process and product improvement : an empirical assessment," Information and Software Technology, Vol. 42, pp. 27-34, 2000.

[13] S. Lee, "A Study on IT governance critical

success factors in korean government integrated data center," Journal of The Korea Society of Computer and Information, Vol. 18, No. 12, 2013.

〔14〕 M. Light, B. Rosser and S. Hayward, "Realizing the Benefits of Project and Portfolio Management," Gartner Research, pp. 19-21, 2005.

〔15〕 M. P. Papazoglou, W. Van and D. Heuvel, "Service-oriented design and development methodology," Journal International Journal of Web Engineering and Technology, Inderscience Publishers, Vol. 2, No. 4, pp. 412-442, 2006.

〔16〕 I. Sommerville, "Software Engineering(9th Edition)," Addison-Wesley, 2010.

〔17〕 PMI, "A Guide to the Project Management Body of Knowledge (PMBOK® Guide)—Fifth Edition," 2013.

〔18〕 B. K. Yoon, M. Y. Hong, Y. R. Choi, K. W. Jeong, "Architecture Establishment Method of Object - Oriented Development Methodology," Proc. of Korean Institute of Information Scientists and Engineers, Vol. 26, No. 1, pp. 584-586, 1999.

〔19〕 J. S. Yoon, O. N. Jeong, N. H. Jeong, S. I. Cha, W. S. Lee, "IT management methodology for public project: Seoul project management case study," Communications of the Korean Institute of Information Scientists and Engineers, Vol. 23, No. 12, pp. 77-85, 2005.

저 자 소 개

김 형 래

1997: 관동대학교 컴퓨터과학 학사
2005: Florida Inst. of Tech, 컴퓨터과학 석박사
현재: KEIS Research Fellow
관심분야: Data Mining
Email: goddoes8@gmail.com

전 도 홍

1985: Olahoma City Univ. 컴퓨터과학 학사
1987: Florida Inst. of Tech. 컴퓨터과학 석사
1990: Florida Inst. of Tech. 컴퓨터과학 박사
현재: 관동대학교 컴퓨터학과 교수
관심분야: Computer graphics, Data mining
Email : dhjeon@kd.ac.kr

지 승 현

1999: 충북대학교 전자계산학과 이학 박사
2000: 미주리 주립대학교 컴퓨터과학과 연구교수
현재: KEIS Research Fellow
관심분야: Data Mining
Email : jsh@keis.or.kr