한국빅데이터학회지
제2권 제2호, 2017, pp. 87-93

빅데이터 품질 확장을 위한 서비스 품질 연구

Applying Service Quality to Big Data Quality

박주석[1†] · 김승현[1] · 류호철[1] · 이준기[2] · 이장호[3] · 이준용[3]

경희대학교[1], 연세대학교[2], 투이컨설팅[3]

요 약

데이터 품질에 대한 연구는 오랜 기간 동안 수행되어 왔다. 하지만 이러한 데이터 품질관리 연구는 구조적 데이터를 대상으로 하였다. 최근에 디지털혁명 또는 4차산업혁명이 일어나면서 빅데이터에 대한 품질관리가 중요해 지고 있다. 본 논문에서는 기존 논문을 분석하여 빅데이터 품질 유형을 분류하고 비교 분석하였다. 요약하면, 빅데이터 품질 유형은 빅데이터 값, 빅데이터 구조, 빅데이터 품질 프로세스, 빅데이터 가치사슬 단계, 빅데이터 모형 성숙도 등으로 분류할 수 있다. 이러한 비교 연구를 바탕으로 본 논문에서는 새로운 기준을 제시하고자 한다.

■ 중심어 : 빅데이터, 품질, 품질 값, 품질 프로세스, 가치사슬, 성숙도모형, 서비스 품질

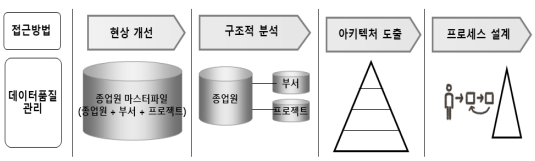
Abstract

The research on data quality has been performed for a long time. However, the research focused on structured data. With the recent digital revolution or the fourth industrial revolution, quality control of big data is becoming more important. In this paper, we analyze and classify big data quality types through previous research. The types of big data quality can be classified into value, data structure, process, value chain, and maturity model. Based on these comparative studies, this paper proposes a new standard, service quality of big data.

■ Keyword : Big Data, Quality, Maturity Model, Service Quality

Ⅰ. 연구의 필요성

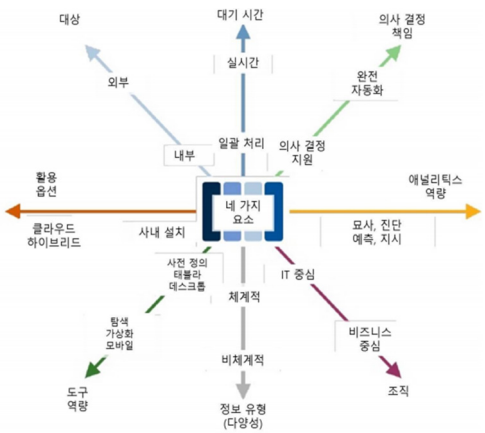
데이터 품질에 대한 연구는 오랜 기간 동안 수행되어왔다. 데이터 품질관리의 역사를 살펴보면 <그림 1>과 같다. 박주석(2009)에 의하면, 사용자(user) 관점에서 시작하여 모델(model) 관점으로, 아키텍처(architecture) 관점으로, 그리고 최근에는 거버넌스(governance) 관점으로 발전되었다. 하지만 이러한 데이터 품질관리 연구는 구조적 데이터를 대상으로 하였다.



〈그림 1〉 데이터 품질관리의 역사

2017년 12월 18일 접수; 2017년 12월 22일 수정본 접수; 2017년 12월 26일 게재 확정.
* 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No. 2017-0-00163, 빅데이터 품질 평가 도구 개발).
† 교신저자 jspark@khu.ac.kr

최근에 디지털혁명 또는 4차 산업혁명이 일어나면서 빅데이터에 대한 관심이 폭발적으로 증가하였다. 가트너(Gartner 2015)는 <그림 2>와 같이 빅데이터가 8가지 변화가 이루어지고 있고, 이러한 변화에 따라 데이터 품질관리도 재정립되어야 한다고 주장하였다. 카이(Cai, 2015)는 빅데이터로 인한 데이터 품질관련 주요 도전요소를 정리하였다. 제리(Jerry, 2016)는 빅데이터 품질 문제를 일으키는 원인으로 빅데이터 관리, 빅데이터 프로세싱, 빅데이터 서비스 품질을 제시하였다. 유엔(UN) 관련 기관이 빅데이터 공식 통계 활용을 위한 조사를 실시한 결과(UNSD, UNECE 2014), 빅데이터 수행 조직 2/3 이상이 빅데이터 품질평가 체계를 정의하고 있지 않았다.



〈그림 2〉 빅데이터의 8가지 변화

따라서 전통적인 데이터 품질관리를 넘어서 빅데이터 품질관리를 체계적으로 연구할 시점이 되었다. 본 연구는 그동안 수행되었던 빅데이터 품질연구를 분석하여 빅데이터 품질 유형을 분류하고 비교하여 빅데이터 품질관리 체계를 정립하는 것이다.

Ⅱ. 빅데이터 품질의 특성

빅데이터의 특성은 일반적으로 3V(Volume, Variety, Velocity)로 정의된다. 따라서 빅데이터 품질은 전통적인 데이터 품질보다 범위가 훨씬 확장된다. 소레스(Soares, 2012)는 전통적인 데이터 품질관리와 빅데이터 품질관리의 차이점을 <표 1>과 같이 제시하였다.

〈표 1〉 전통적 데이터 품질과 빅데이터 품질 관리 비교

관점	전통적 데이터 품질	빅데이터 품질
데이터 품질 프로세싱 빈도	배치 처리 방식	실시간과 배치처리 방식 혼용
데이터 다양성	구조화된 데이터	구조화, 반구조화, 비구조화
데이터 신뢰도 수준	데이터 분석을 위한 고품질 데이터 요구	심각한 오류(Noise) 제거 분석에 적합한 수준 요구
데이터 클린징 시점	데이터웨어하우스 로딩 전	원상태로 로딩 인메모리 분석, 스트리밍 적용
데이터 품질평가 항목	고객 주소 등 핵심 항목 중심 평가	분석 및 탐색주제에 따라 변경

하지만 맥킨지(2011)나 IDC(2011) 등의 빅데이터 정의에서도 알 수 있듯이, 빅데이터 활용이나 거버넌스도 빅데이터 품질에 포함할 필요가 있다.

따라서 빅데이터 품질은 전통적인 데이터 품질에 비해서 범위가 훨씬 확장된다. 이러한 빅데이터 품질 유형은 산업에 따라 매우 다르다.

Ⅲ. 빅데이터 품질 유형 분류 및 비교연구

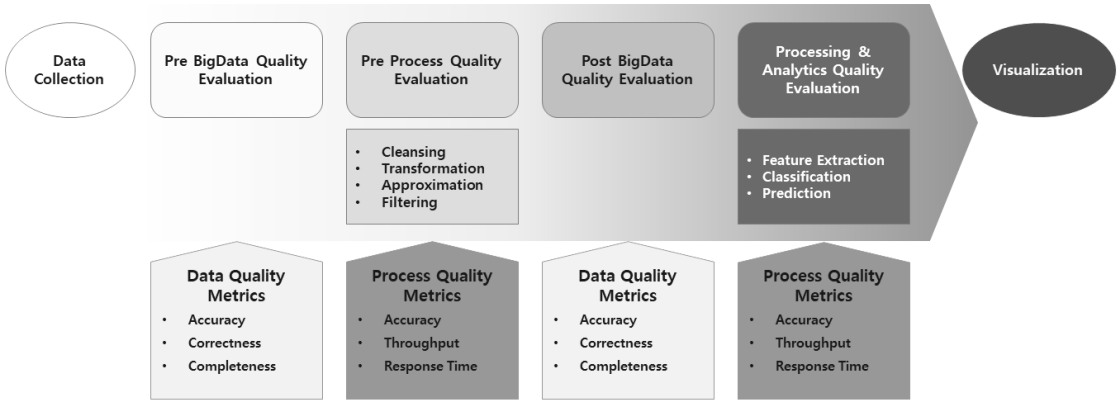
백커(Becker 2015)는 MIRTE에 의한 4가지 사례 연구를 통해서 빅데이터 품질을 연구하였다. 4가지 사례 연구결과는 빅데이터 환경에서 새로운 데이터 품질 문제를 발견할 수 없었고 빅데이터 품질은 빅데이터 종류와 적용기술에 따라 달라진다고 주장하였다.

〈표 2〉 빅데이터 품질 검증 투입 단계 영역별 구조

영역	품질 검증 요소	고려사항
출처	제도・사업적 환경	지속적 데이터 제공 가능성/신뢰도/투명성 해석 가능성
	개인 정보 보호・보안	관련 법률/데이터 제공자 데이터 소유자/이용 제한 사항
메타 데이터	복잡성	데이터 처리에 대한 기술적 제한 요소/데이터의 구조화 여부 /데이터의 가독성 /데이터의 계층 및 중첩 여부
	완전성	메타데이터의 활용가능성, 해석용이성, 완전성여부
	유용성	데이터 처리・저장・분석에 필요한 부가적인 자원/위험성 분석
메타 데이터	시간 관련 요소	시의성/주기성/시점별 개념 및 작성 방법의 변화
	연계가능성	연계변수의 존재 여부와 품질/연계 수준
	일관성	표준화된 주요 변수 개념 /주요 변수에 대한 메타데이터의 이용 가능 여부
	타당성	통계 산출 방법론과 과정의 투명성 및 공정성
데이터	정확성	총 표본오차/참고 데이터 세트/선택성
	연계가능성	연계변수의 질
	일관성	메타데이터와 관측된 데이터값의 일관성
	타당성	통계 산출 과정・방법과 관측된 데이터 값의 일관성

유엔(UNECE 2014)은 각 국의 빅데이터 현황 조사를 통해서 국가 통계 활용을 위한 빅데이터 품질검증 3단계를 제시하였다. 품질검증 3단계는 먼저 데이터를 수집하거나, 수집중인 상태를 의미하는 투입(input) 단계와 데이터의 가공, 분석 등이 수행되는 과정인 전환(throughput) 단계, 빅데이터에서 얻은 통계 품질을 검증하는 산출(output) 단계로 구분된다. 그 중 투입과 산출단계는 출처(source), 메타데이터(metadata), 데이터(data)로 구성된 3가지 영역으로 구분된다. 이러한 구성은 네덜란드 통계청에서 고안한 행정데이터 품질진단체계 개념에서 착안되었다.

세르하니(Serhani 2015)는 빅데이터 가치사슬 기반 데이터 값에 대한 평가와 빅데이터 프로세스 평가를 결합한 연구를 수행하였다. 빅데이터 가치사슬은 사전 프로세싱, 프로세싱, 분석, 시각화 단계로 구성된다. <그림 3>과 같이 가치사슬의 각 단계별로 데이터 품질평가를 위한 방안



〈그림 3〉 빅데이터가치사슬 품질평가를 위한 하이브리드 모델

을 제시하였다. 빅데이터 가치사슬 기반으로 품질 평가를 했다는 점에서 빅데이터 품질의 범위를 확대한 의미가 있다.

가오(Gao 2016)도 가치사슬 단계 별로 분석하였으나, 전 단계와 연계하여 빅데이터 유효성을 검사하였다. 즉, 빅데이터 유효성검사 프로세스를 데이터 수집, 데이터 클린징, 데이터 변환, 데이터 적재, 데이터 분석의 5단계로 구성하였다. 다수 원천인 경우 데이터 집계 단계를 추가하였다. 빅데이터 프로세스의 반복성을 품질에 고려했다는 점에서 의미가 있다.

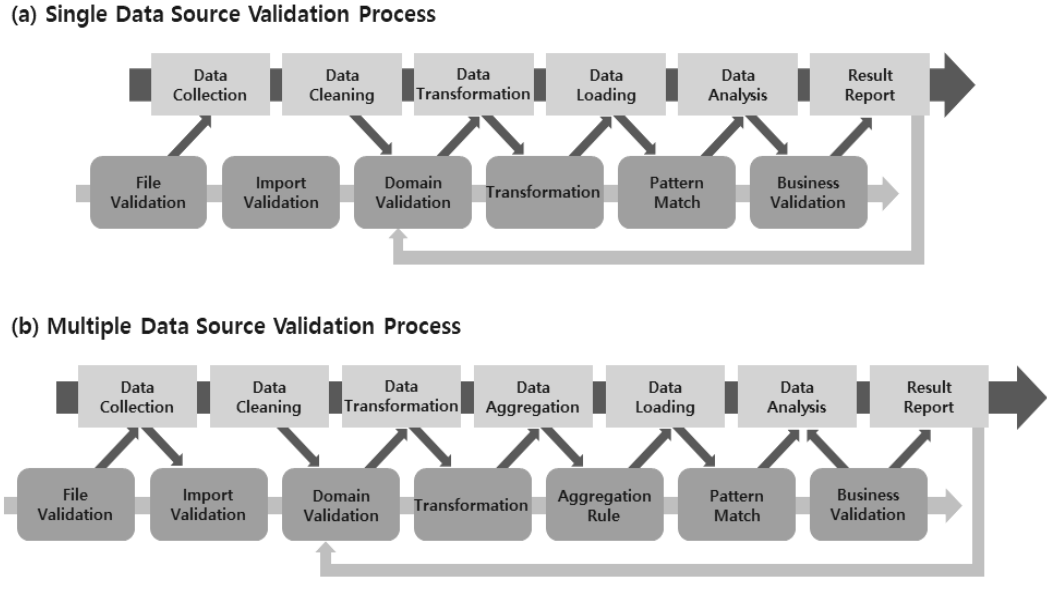
가트너그룹(Gartner Group)은 데이터 품질 성숙도를 인식(Aware), 대응(Reactive), 예방(Proactive), Managed(관리화), Optimized(최적화) 5단계로 정의하였다 이러한 성숙도 모델을 통해 데이터 품질의 정교함 수준을 평가 할 수 있으며, 공통지표 및 벤치마크를 통해 조직의 정보관리 능력을 향상시키는 여러 가지 개선 전략을 제공할 수 있다.

사사 바스카라다(Sasa Baskarada)는 데이터 품질 관리를 위해 TDQM 기반의 5단계 성숙도 모델을 제안하였다. 각 단계는 IQM(Information Quality Management)의 진화한 단계를 나타내며, 단계별 품질의 목표를 분리하고, 점진적으로 달성한다.

이러한 성숙도 모형은 최근에 다양한 빅데이터 성숙도 모형으로 확장되었다. 하지만 그 개념은 가트너나 사스바스카라다의 성숙도 모형 개념과 비슷하다.

오픈데이터의 대표적인 성숙도 모형이 팀버너스리의 “5 Star Open Data” 모형이다. 성숙도 1단계는 PDF 등 오픈라이선스 형식이고, 2단계는 Excel 등 기계가독형 형식이고, 3단계는 CSV 등 개방형 포맷 형식이고, 4단계는 URI 등 개체 식별 형식이며, 5단계는 LOD 등 네트워크로 연계된 데이터 형식이다. 오픈데이터는 텍스트, HTML, XML 등 반구조데이터가 중심이기 때문에 주목할 만한 빅데이터 성숙도 모형이라고 할 수 있다.

요약하면 빅데이터 품질 유형은 빅데이터 값, 빅데이터 구조, 빅데이터 품질 프로세스, 빅데이터 가치사슬 단계, 빅데이터 품질 성숙도 등으로 분류할 수 있다.



〈그림 4〉 빅데이터 검증 프로세스

Ⅳ. 빅데이터 라이프사이클에 따른 품질 유형

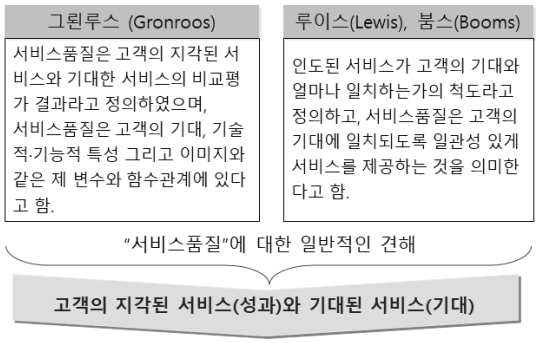
본 논문에서는 빅데이터 라이프사이클에 의한 데이터 품질 유형을 도출하고 새로운 데이터 품질 유형을 제안하고자 한다.

박주석(2016)은 빅데이터 환경이 도래하면서 데이터 생명주기가 DIA(Data, Insight, Action)로 변화되었다고 주장하였다. 즉, 데이터(Data)를 분석하여 통찰력(Insight)을 얻고, 통찰력을 프로세스에 내재화하여 조직이 바로 실행(Action)할 수 있다.

이러한 DIA 사이클을 데이터 품질 관점에서 살펴볼 수 있다. DIA 사이클의 데이터는 구조 데이터, 비구조 데이터, 그리고 오픈데이터를 모두 포함하고 있다. 3가지 형태의 데이터 품질을 고려해야 한다. DIA 사이틀의 통찰력은 반복적인 작업을 통한 분석알고리즘의 지속적인 개선을 포함한다. DIA 사이클의 실행은 프로세스 혁신을 추진하여 고객 또는 사용자에 대한 서비스 품질 향상을 의미 한다

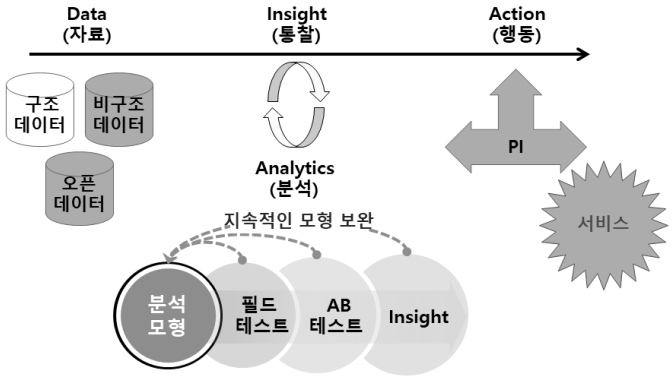
본 논문에서는 빅데이터환경하의 데이터서비스 품질 유형을 강조하고자 한다. 앞에서 언급했듯이 빅데이터 값보다는 빅데이터 프로세스나 가치사슬이 중요하다. 따라서 이러한 프로세스나 가치사슬의 궁극적인 가치인 서비스 품질을 관리하는 것이 더욱 중요하다.

데이터서비스 품질 모형을 도출하기 위해서 전통적인 서비스 품질 이론을 이해해야 한다. 전통적인 서비스 품질 이론은 주로 마케팅 분야에서 연구되었다. 일반적인 서비스 품질의 정의는 고객의 지각된 서비스(성과)와 기대된 서비스(기대)의 비교 평가 결과라고 얘기된다.



〈그림 6〉 일반적인 서비스 품질 이론

서비스 품질 측정 모형은 대표적으로 서브퀄(SERVQUAL) 모형과 서브퍼프(SERVPERF) 모형이 있다. 서브퀄 모형은 가장 보편적인 방법으로, 지각된 품질은 고객이 구매하기 전에 형성된 기대와 실제 고객이 경험한 후의 지각과의 차이라고 정의한다. 서브퀄 모형의 구성요인은 업체의 성격에 따라 업체가 속한 지역적, 문화적, 환경적 등에 의해 규명할 필요가 있다. 예를



〈그림 5〉 DIA 관점에서 본 데이터 품질 유형

들어 서비스 품질 측정도구인 5개의 품질차원과 22개의 항목을 개발되었다.

결론적으로 데이터서비스 품질은 사용자가 빅데이터 서비스를 받기 전에 형성된 기대와 실제 사용자가 경험한 후의 지각과의 차이라고 정의할 수 있다.

Ⅴ. 요약

본 논문에서는 기존 논문을 분석하여 빅데이터 품질 유형을 분류하고 비교 분석하였다.

기존의 빅데이터 품질 유형은 빅데이터 값, 빅데이터 구조, 빅데이터 품질 프로세스, 빅데이터 가치사슬 단계, 빅데이터 모형 성숙도 등으로 분류할 수 있다. 이러한 비교 연구를 바탕으로 본 논문에서는 새로운 기준을 제시하였다. 새로운 기준은 빅데이터 분석알고리즘 품질과 빅데이터 서비스 품질이다. 특히 본 논문에서는 빅데이터 서비스 품질을 강조하였다.

참 고 문 헌

[1] 김승현, 박주석, 박재홍, 김인현, "빅데이터 환경에서 분석자원이 기업성과에 미치는 영향", 한국빅데이터학회, 제1권, 제1호, 2016.

[2] 박주석, "데이터 중심의 공공 정보자원관리", 한국정보화진흥원, 연구보고서, 2016.

[3] 박주석, 김인현, "전통적 환경과 빅데이터 환경의 데이터자원관리 비교연구", 한국빅데이터학회, 제1권, 제2호, 2016.

[4] 박주석, 이형로, "진정한 데이터 품질관리의 조건은?", 2009년 데이터베이스 컨퍼런스, 한국데이터베이스진흥센터, 2009.

[5] 신덕호, 유비쿼터스 컴퓨팅 환경에서의 개인정보 정책 발전에 관한 연구, 단국대학교 석사학위논문, 2009.

[6] 이연우, 장현미, 홍승필, "빅데이터 환경 내 개인정보보호를 위한 대용량 개인정보 관리 모델 설계방안", 한국인터넷정보학회 추계학술발표대회논문집, 제1권, 제2호, pp.29-30, 2012.

[7] 진재현, 고금지, "유엔의 빅데이터 품질검증 기준과 시사점: 빅데이터의 국가통계 활용을 중심으로", 한국보건사회연구원, 2016.

[8] Becker, D., T.D. King, and B. McMullen, "Big Data-Big Data Quality Problem", *2015 IEEE International Conference on Big Data*, 2015.

[9] Caballero, I., B. Rivas, M. Serrano, M. Piattini, "A Data Quality in Use model for Big Data", *Future Generation Computer Systems*, Vol.63, pp.123-130, 2015.

[10] Cai, L. and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era", *Data Science Journal,* Vol.14, p.2, 2015, http://doi.org/10.5334/dsj-2015-002.

[11] Gao, J.C. Xie, and C. Xie, "Big Data Validation and Quality Assurance-Issuses, Challenges, and Needs".

[12] Gao, J., "Big Data Validation and Quality Assurance-Issuses, Challenges, and Needs", *IEEE 10th IEEE International Symposium on Service-Oriented System Engineering*, At Oxford, UK.

[13] Gartner, "Establish a Data Quality Program to Support Digital Business", Apr. 2015.

[14] Gartner, "Magic Quadrant for Data Quality Tools", Nov. 2016.

[15] Haryadi, F.A., J. Hulstijn, A. Wahyudi, H. van der Voort, and M. Janssen, "Antecedents of big data quality: An empirical examination in financial service organizations", In: *Proceedings of the IEEE International Conference on Big Data*, pp.116-121. DOI:10.1109/BigData.2016.7840595, 2016.

[16] Infromation Difference, “Data Quality Landscape”, Q1, 2017.

[17] Merino, J., I. Caballero, B. Rivas, M. Serrano, and M. Piattini, “Functions per proles in the 3cs data quality model”, 2015.

[18] Serhani, M.A., H.T. El Kassabi, I. Taleb, and A. Nujum, “An Hybrid Approach to Quality Evaluation Across Big Data Value Chain”, *2016 IEEE International Congress on Big Data*, pp.411-412, 2016.

저 자 소 개

박 주 석(Jooseok Park)

- 1981년 : 서울대학교 산업공학 (학사)
- 1983년 : 학국과학기술원 산업공학 (석사)
- 1990년 : University of California, Berkeley MIS (박사)
- 현재 : 경희대학교 경영대학 교수
- 관심분야 : 데이터베이스, 모델링, 아키텍처, 정보화전략 등

이 준 기(Zoonky Lee)

- 1985년 : 서울대학교 컴퓨터사이언스 (학사)
- 1991년 : Carnegie Mellon Univ. 사회심리학 (석사)
- 1999년 : Univ of Southern California 경영정보 (박사)
- 현재 : 연세대학교 정보대학원 교수
- 관심분야 : 디지털기업전략, 빅데이터 응용

이 장 호(Jangho Lee)

- 1998년 : 우석대학교 전산학과 (학사)
- 2009년 : 국민대학교 비즈니스 IT 컨설팅 (석사)
- 현재 : 투이컨설팅 Open-Data Center 이사
- 관심분야 : ISP, 빅데이터분석, AI, DQ 등

김 승 현(Seunghyun Kim)

- 2006년 협성대학교 경영학 (학사) (MIS 전공)
- 2008년 경희대학교 경영학 석사(MIS 전공)원 경영학과 MIS (석사)
- 2010년 경희대학교 경영학 (박사수료) (MIS 전공)
- 현재 : 경희대학교 빅데이터 연구센터
- 관심분야 : NCS, 정보화 관련 품질/성과, EA&ISP, Analytics 등

류 호 철(Hocheol Ryu)

- 2014년 : 청주대학교 컴퓨터정보공학 (학사)
- 현재 : 경희대학교 일반대학원 경영학과 MIS (석사)
- 관심분야 : 빅데이터분석, 스마트화 품질 및 인증서비스 연구 등

이 준 용(Junyong Lee)

- 2012년 연세대학교 경영정보학 (학사)
- 2016년 연세대학교 일반대학원 경영정보학 (석사)
- 현재 : 투이컨설팅 Open Data Center 선임
- 관심분야 : 데이터 품질관리, 공공데이터 개방, 프로세스혁신 등