韓國IT서비스學會誌
第13卷 第2號
2014年 6月, pp.185-194
http://dx.doi.org/10.9716/KITS.2014.13.2.185

빅데이터 처리 프로세스에 따른 빅데이터 위험요인 분석

이지은* · 김창재** · 이남용***

The Analyzing Risk Factor of Big Data : Big Data Processing Perspective

Ji-Eun Lee* · Chang-Jae Kim** · Nam-Yong Lee***

■ Abstract ■

Recently, as value for practical use of big data is evaluated, companies and organizations that create benefit and profit are gradually increasing with application of big data. But specifical and theoretical study about possible risk factors as introduction of big data is not being conducted. Accordingly, the study extracts the possible risk factors as introduction of big data based on literature reviews and classifies according to big data processing, data collection, data storage, data analysis, analysis data visualization and application. Also, the risk factors have order of priority according to the degree of risk from the survey of experts. This study will make a chance that can avoid risks by bid data processing and preparation for risks in order of dangerous grades of risk.

Keyword : Big Data, Risk Factor, Big Data Processing

논문투고일 : 2014년 02월 17일 논문수정완료일 : 2014년 06월 10일 논문게재확정일 : 2014년 06월 13일
* 숭실대학교 소프트웨어특성화대학원
** 숭실대학교 소프트웨어특성화대학원 교수
*** 숭실대학교 소프트웨어특성화대학원 교수, 교신저자

1. 서 론

최근 빅데이터를 통해 민간 부문과 공공 부문 등 사회 전 부문에 걸쳐 새로운 가치창출의 기회가 제공되고 있다. 또한 빅데이터 적용으로 인한 경쟁력 상승과 효율성이 제고되고 있다.

그러나 데이터 오남용, 개인 정보 유출, 관련 기술적 오류 등 빅데이터로 인한 위험도 기하급수적으로 증대하고 있다. 그럼에도 빅데이터의 위험성에 대한 인지와 대비가 미비하다. 현재까지 이론적 연구도 빅데이터의 위험에 관한 체계적인 위험요인분석이 이루어지지 않고 있으며, 위험의 분류에 초점을 맞추고 있을 뿐이다. 또한 빅데이터 처리 프로세스별 발생 위험요인에 대해서 구체적 연구가 되어 있지 않다.

따라서 본 논문은 빅데이터 전 부분에서 위험요인을 다루는 기존 논문들과 달리, 빅데이터 시스템 도입 시, 데이터 처리 프로세스에 따른 위험요인들에 대해 연구한다. 문헌 연구를 통해 추출한 빅데이터의 위험요인을 전문가들의 의견을 수렴하여 데이터 수집, 데이터 저장, 데이터 분석, 분석 데이터 가시화 및 활용의 각 프로세스별로 분류한다. 또한 위험요인의 위험도를 검증하기 위해 관련 분야의 전문가를 대상으로 설문지를 통해 위험도에 대한 위험 수치를 부여한다.

본 논문을 통해 첫째, 빅데이터 처리 프로세스별 위험요인을 제시함으로써 위험을 회피할 수 있는 기회를 만들고, 둘째, 빅데이터 처리 프로세스별 위험요인을 위험도에 따라 순위를 부여하여 각 프로세스별 우선 대비가 가능함을 기대한다.

2. 관련 연구

2.1 빅데이터 정의

빅데이터란 데이터를 수집, 통합, 저장, 관리, 분석 등의 데이터 처리 시, 기존 데이터베이스의 수용을 초과하는 대량의 정형 및 비정형 데이터 집합이다[13]. 또한 이러한 데이터로부터 새로운 핵심 가치를 추출하고 결과를 분석하는 기술을 포함한다[14].

현재, 빅데이터가 정부나 기업에서 분석되고 가공되어 가치를 생성한 결과가 증가함에 따라, 혁신적인 기술로 주목되고 있다. 따라서, 빅데이터를 이용하기 위한 기술적 및 환경적 발전도 지속되고 있다.

본 논문에서는 빅데이터란 데이터 자체의 대용량성과 다양성으로 처리 과정이 복잡한 반면, 생성에서 활용까지 빠른 속도로 가치를 창출하는 기술이라고 정의한다.

2.2 빅데이터 처리 프로세스

빅데이터 처리 프로세스는 크게 데이터 수집과 통합, 데이터 저장과 관리, 데이터 분석과 처리, 데이터 분석 가시화 프로세스로 나뉜다[4, 6, 13].

첫 번째 프로세스인 데이터 수집과 통합 단계에서 형태와 소재에 무관한 다양한 데이터를 수집하고, 데이터 저장과 관리 단계에서 수집된 대량의 데이터를 실시간으로 저장하고 관리할 수 있는 분산 컴퓨팅을 한다. 데이터 분석과 처리 단계에서 데이터에 내재된 가치를 추출하기 위한 분석을 수행한다. 마지막 프로세스인 데이터 분석 가시화 단계에서 IT 비전문가가 데이터 분석을 수행할 수 있는 환경을 제공하고 분석결과를 가시적으로 제공한다[4].

이러한 빅데이터 처리 프로세스에 조영임[4]은 비정형 스트림 데이터를 분석 가능한 형태로 구조화하여 분석의 정확성을 높이고 심층 분석이 가능한 데이터 전처리 단계를 추가하였다. 이재식[3]은 빅데이터 처리의 마지막 프로세스로 데이터 분석을 위해 이용될 데이터를 삭제하는 데이터 폐기 단계를 추가하였다.

이러한 선행 연구에 따라 본 논문은 빅데이터의 데이터 처리 프로세스를 데이터 수집, 데이터 저장, 데이터 분석, 분석 데이터 가시화 및 활용 프로세스로 구분하고 각 프로세스별 위험요인을 분류한다.

2.3 빅데이터 위험요인에 관한 기존 연구

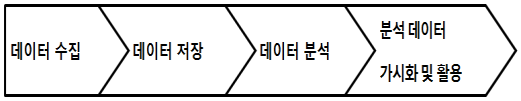
빅데이터 위험에 관한 연구는 국내외 학자들에 의해 수행되어왔다. 김기환[1]은 공공부문에서 빅데이터를 사용함으로 불필요한 데이터의 저장비용, 사회적 비용 증대, 빅데이터의 악용이나 오용, 계층간 상대적 박탈감 또는 데이터 불평등 등의 새로운 위험요인을 제안했다. 윤상오[2]는 빅데이터가 초래할 수 있는 위험을 기술적, 인적, 법제도적, 경제적, 사회문화적 위험으로 분류하였다. 기술적 위험 중에서 해킹이나 사이버테러, 천재지변과 사고, 법제도적 위험 중에서 개인정보 침해, 법제도적 충돌과 혼란, 경제적 위험에서 산업경쟁력 약화, 사회문화적 위험에서 빅데이터로 인한 사회적 병리현상 등이 정책적 대응 우선순위가 높게 나타났다. 조영임[4]은 빅데이터 도입관련 주요 쟁점으로 플랫폼 부재, 정보보안, 개인사생활 침해, 정보의 범위와 깊이 등을 제시하였다. Bollier[8]는 빅데이터의 기대와 위험을 논하며, 위험에는 빅데이터 남용과 오용, 개인정보 유출, 빅데이터 위험 인식 부족 등을 제안하였다. Boyd[9]는 잘못된 정보의 범람, 데이터 분석 오류 등을 빅데이터의 문제점으로 정의하였다. 그 밖에 Labrindidis and Jagadish[11], Chester[10], Manovich는 빅데이터의 해결 과제와 기회에 대해 논하였다,

본 연구에서는 빅데이터 위험요인에 대한 분류를 제안하기 위해 앞서 서술한 연구 외에 빅데이터 위험요인에 적용 가능하다 판단한 추병완[5]의 미래 인터넷 기술의 윤리 문제에 대한 문헌과 Paulraj Ponniah[15]의 데이터웨어하우징의 문제점을 빅데이터 위험요인에 보완하였다. 현재까지 빅데이터 위험요인을 제시하는 연구는 있었지만, 위험요인의 분류 체계가 광범위하여 요인의 세분화가 부족하고 상호배타적 분류가 되지 않았다. 따라서 본 논문은 빅데이터 시스템을 도입하기 위해 데이터 처리 프로세스에 따른 순차적 위험 분류를 하고 검증한다.

3. 빅데이터 위험요인 분석

3.1 위험요인 분류 프로세스 정의

본 연구에서는 [그림 1]과 같이 빅데이터의 데이터 처리 프로세스를 데이터 수집, 데이터 저장, 데이터 분석, 분석 데이터 가시화로 구분하여 위험요인을 분류한다. 빅데이터 처리 프로세스에 공통적으로 적용 가능한 위험요인은 '공통'으로 분류한다.



[그림 1] 위험요인 분류의 빅데이터 처리 프로세스

3.1.1 데이터 수집

데이터 수집에서는 빅데이터 시스템 구축을 위한 첫 번째 프로세스이다. 이 단계에서 대용량의 원천 데이터를 광범위한 영역에서 형태와 무관하게 수집한다. 그리고 수집한 데이터를 저장하기 위해 데이터를 통합하는 역할을 한다.

3.1.2 데이터 저장

데이터 저장 프로세스에서 수집된 데이터를 분산 저장하고, 축적된 데이터의 변형이나 유실 등을 대비해 보안을 유지한다. 필요시에 데이터를 수정하거나 삭제하고, 데이터를 읽기위해 접근 방법을 제공한다.

3.1.3 데이터 분석

데이터 분석 프로세스에서 저장되어있는 다양한 데이터를 목적에 따라 분석하여 내재된 가치를 추출하고 효율적인 처리를 위해 대규모의 심층적 통계처리를 한다.

3.1.4 분석 데이터 가시화 및 활용

빅데이터 처리의 마지막 프로세스인 분석 데이터 가시화 및 활용 단계에서는 빅데이터를 이해하

거나 수행할 수 있는 환경이 제공되며, 분석된 데이터의 결과를 함축적이고 직관적인 정보를 제공한다. 또한 분석된 데이터가 실질적으로 현실에 적용 및 활용되는 프로세스이다.

3.2 빅데이터 위험요인 분류 과정

본 논문에서는 빅데이터의 위험요인뿐 아니라 데이터웨어하우스, 미래 인터넷 기술 등 빅데이터와 관련된 분야의 위험이나 문제점을 다룬 10개의 문헌에서 위험요인을 도출한다. 도출된 위험요인은 아래 [그림 2]와 같이 각 위험요인을 빅데이터 처리 프로세스와 공통에 분류한다.

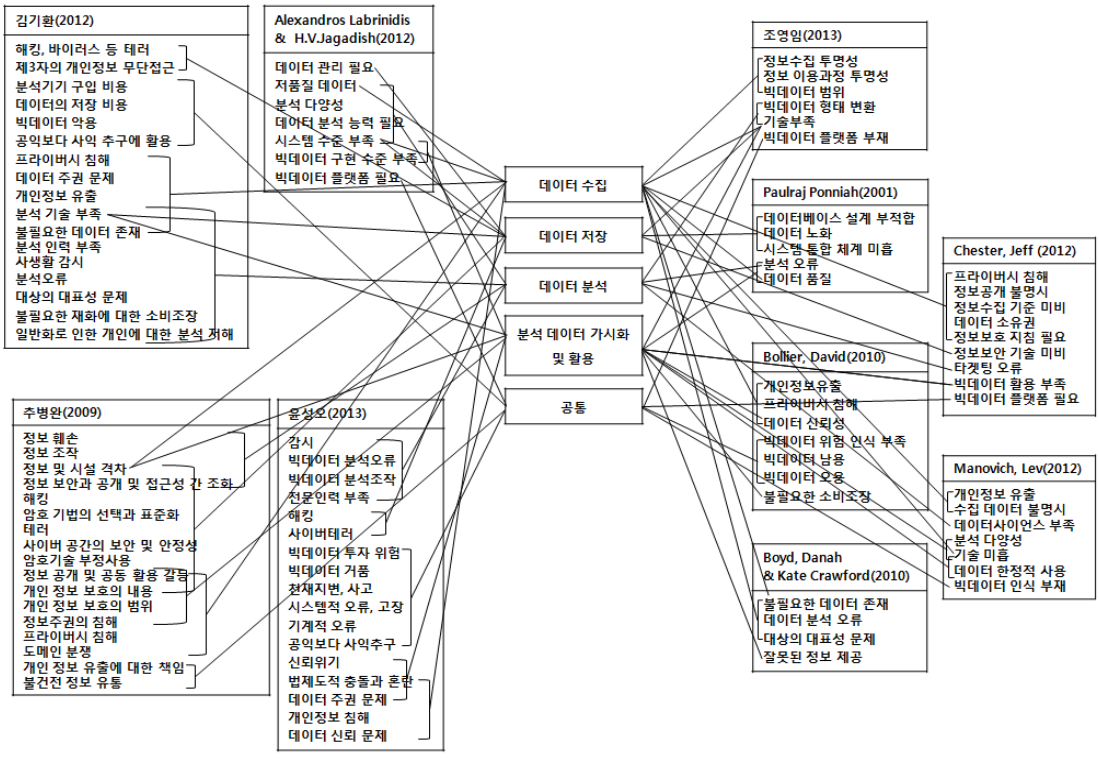
3.3 빅데이터 처리 프로세스별 위험요인 분석

본 연구의 위험요인은 김기환[1] 등의 빅데이터 위험요인과 Paulraj Ponniah[15]의 데이터웨어하우스 위험요인 연구를 통해 도출한 빅데이터 위험요인들을 근거로 한다. 전문가의 의견을 수렴하여 관련연구의 위험요인 중 유사한 개념의 요인들은 그룹화하고, 일반화된 요인들은 세분화하여 연구에 맞게 재설정하여 총 30개의 요인이 추출되었다. 설정된 위험요인은 데이터 수집, 데이터 저장, 데이터 분석, 데이터 분석 가시화 및 활용 프로세스, 공통에 분류한다. 본 연구의 위험요인은 <표 1>와 같이 각 요인별로 분류하고 조작적 정의한다.

4. 빅데이터 처리 프로세스에 따른 위험요소 검증

4.1 빅데이터 위험요인 검증 방법

본 논문의 위험요인 검증 방법은 빅데이터 위험



[그림 2] 문헌 연구의 위험 요소 도출 과정

〈표 1〉 빅데이터 처리 프로세스별 위험요인 분류

프로세스	위험요인	조작적 정의	선행 연구
데이터 수집	프라이버시 침해(c1)	개인 정보가 무단으로 유출되거나 생활 감시 등의 사생활 침해에 대한 위험	[1, 2, 5, 8, 10, 12]
	정보공개 기준 모호(c2)	공개되는 정보의 범위 기준 정립에 대한 위험	[5, 10]
	데이터 수집 기술 부족(c3)	원천 데이터를 수집 할 수 있는 기술 부족에 대한 위험	[4, 11]
	원천 데이터 신뢰성 부족(c4)	불필요한 데이터나 데이터의 품질이 낮음에 따라 분석 결과 저해 등의 위험이 발생 가능함에 생기는 위험	[2, 11]
	데이터 소유권 분쟁(c5)	데이터 수집의 통제구역이나 지식의 범위 구분이 불명확하기 때문에 발생하는 데이터 소유권 문제에 대한 위험	[2, 5, 10]
	정보 이용범위 불 명시(c6)	정보 소유 주체에게 정보 이용 범위에 대한 기준을 명시하지 않은 것에 대한 위험	[4, 10, 12]
	데이터 수집 범위 기준 모호(c7)	대상 데이터가 광범위함으로 효율적이고 가치 있는 데이터를 수집할 수 있는 기준이 모호함에 대한 위험	[4 10, 12]
데이터 저장	외부 공격(s1)	해킹이나 바이러스 등 네트워크의 취약한 보안망에 불법적으로 접근하거나 정보 시스템에 유해한 영향을 끼치는 행위로 인한 위험	[1, 2]
	외부 접근권한 통제 미흡(s2)	데이터 위변조를 예방하기 위한 접근 권한 통제가 미흡함에 대한 위험	[1, 2, 5, 10]
	빅데이터 암호화 표준정의 부재(s3)	빅데이터 암호화 표준이 미 확립됨에 따른 빅데이터 정보 위험	[5, 10]
	데이터 저장 기술 부족(s4)	전사 차원의 데이터베이스시스템 미비 등 빅데이터 저장 기술의 부족에 대한 위험	[4, 15]
	데이터 노화(s5)	오래된 데이터로 인해 트렌드에 맞는 분석에 지장이 발생함에 대한 위험	[15]
데이터 분석	내부 접근권한 통제 미흡(a1)	내부 분석 조작 등의 부정적 영향을 방지하기 위한 빅데이터 분석 접근 권한 통제가 미흡함에 대한 위험	[1, 2, 5]
	분석 전문 인력 부족(a2)	빅데이터를 분석할 수 있는 데이터 과학자가 부족함에 대한 위험	[1, 2, 11, 12,]
	분석 기술 부족(a3)	빅데이터를 분석하는 기술적 요건이 미흡함에 따른 위험	[1, 4, 5, 11, 12]
	불필요한 데이터 존재(a4)	정보 난립으로 가치 있는 정보의 발견에 지장이 있음에 대한 위험	[9]
	데이터 분석 조작(a5)	특정 목적을 위해 의도적으로 빅데이터 분석결과를 조작 및 왜곡함에 대한 위험	[1, 2, 5, 8, 11]
	분석 타겟팅 오류(a6)	빅데이터 적용 대상에 대한 명확한 타겟팅 분석에 대한 위험	[1, 9, 8, 10]
	분석 방법 기준의 다양성(a7)	빅데이터 형태 변환의 다양성에 따라 정보 가공법의 기준이 모호함에 대한 위험	[4, 11, 12]
분석 데이터 가시화 및 활용	데이터 가시화 기술 부족(v1)	빅데이터를 가시화 할 수 있는 기술이 부족함에 대한 위험	[4, 11, 12]
	빅데이터 분석결과 불신(v2)	고객이 빅데이터의 결과에 대한 신뢰가 부족함에 대한 위험	[2]
	잘못된 정보 제공(v3)	빅데이터 결과가 잘못된 정보로 제공됨에 대한 위험	[2]
	빅데이터 소유권 분쟁(v4)	가공된 데이터의 활용에 대한 데이터 소유권 문제에 대한 위험	[2, 5]
	부작용에 대한 책임 기준 모호(v5)	빅데이터 활용 시 부작용 발생에 대한 책임을 갖는 주체가 모호함에 대한 위험	[5]
	빅데이터 활용 격차(v6)	빅데이터를 이용할 줄 아는 특정 권력만 활용하는 데 있어 발생하는 위험	[1, 2, 5, 9, 10, 12]
	빅데이터 정보 남용(v7)	빅데이터를 활용할 수 있는 특정 권력이 본래의 목적이외의 용도로 함부로 사용함에 대한 위험	[1, 8, 12]
공통	빅데이터 도입 비용(e1)	데이터 과학자 육성, 분석기기, 데이터 저장 비용 등 빅데이터 시스템 활용 환경을 형성하기 위한 막대한 비용 지출에 대한 위험	[1]
	빅데이터 인식 부족(e2)	빅데이터의 도입 효과나 위험에 대한 인식이 부족하여 그에 대한 대안과 방안을 갖추지 못함에 대한 위험	[2, 8, 12]
	시스템적 오류(e3)	시스템 과부하, 기술적 오류나 결함 등으로 인한 시스템 자체의 고장에 대한 위험	[2]
	빅데이터 플랫폼 부재(e4)	분야별 통합된 데이터 분석을 위한 빅데이터 플랫폼이 제시되지 않아 효과적 활용을 할 수 없음에 대한 위험	[2, 10, 11]

과 관련된 선행연구를 기반으로 위험요인을 추출하여 빅데이터 처리 프로세스별 위험요인을 분류한다. 연구자에 의해 분류된 위험요인은 빅데이터 관련전문가에게 의뢰하여 수정 및 재분류하여 프로세스별 발생 가능한 위험요인 분류의 신뢰성을 확보하였다. 최종적으로 분류된 위험요인은 빅데이터를 활용하고 있는 기업의 전문가를 대상으로 방문 및 e-mail을 통해 설문 조사를 실시하여, 유효 데이터 80건을 확보하였다.

각 빅데이터 처리 프로세스별 위험요인은 위험성에 따라 5점 척도를 적용하여 요인 분석을 수행하고, 척도 평균값을 이용하여 요인별 위험도를 순위화하였다. 데이터는 통계 패키지인 SPSS v.20으로 분석하였다.

분석 요인은 고유값이 1 이상이고, 요인 분석 결과 요인적재량은 0.5 이상이며, 한 요인에서 나눠진 성분의 변수가 3개 이상 묶인 경우를 유의한 것으로 판단하였다. 설문구성의 신뢰성은 크론바흐 알파(Cronbach's alpha) 값이 모두 0.6을 초과하여 유효한 것으로 나타났다. 제시된 변수와 종속 변수의 강도는 동일한 것으로 가정하여 가중치는 1.0으로 동일하게 준다.

4.2 위험요인 분석 결과

관련연구를 통한 빅데이터 처리 프로세스에 따른 위험요인 총 30개를 <표 2>와 같이 요인 분석을 수행하여 총 27개의 유효 변수가 채택되었다.

데이터 수집 프로세스에서 총 7개의 변수의 분석 결과, '정보보안 측면', '원천 데이터 측면', '기술 측면'으로 3개의 하위 요인으로 분류되었다. 정보보안 측면의 변수는 프라이버시 침해, 정보 이용범위 불 명시, 정보공개 기준 모호와 같다. 즉, 데이터 수집 시 개인정보보호가 미흡함을 나타낸다. 원천 데이터 측면의 변수는 데이터 소유권 분쟁, 데이터 수집 범위 기준 모호, 원천 데이터 신뢰성 부족이다. 이는 데이터 소스의 출처 구분과 관련한 품질 문제를 나타낸다. 기술 측면으로 데이터 수집 기술 부족의 위험이 분류되었지만 추출된 변수가 1개로 공통 성분이 없으므로 기각한다.

데이터 저장 프로세스에서 외부 접근권한 통제 미흡, 외부 공격, 빅데이터 암호화 표준정의 부재, 데이터 저장 기술 부족은 해당 프로세스에서 발생 가능한 요인으로 분석되었고, 변수 '데이터 노화'는 나머지 변수 4개와 다른 성분으로 나타나 기각한다. 따라서 해당 프로세스에서 데이터를 저장 관리하는 기술적 위험요인이 도출되었고, 데이터 품질 문제와 같은 위험요인의 중요성은 상대적으로 낮게 판단된다.

데이터 분석 프로세스에서 7개의 변수, 분석 방법 기준의 다양성, 내부 접근권한 통제 미흡, 데이터 분석 조작, 분석 전문 인력 부족, 분석 기술 부족, 불필요한 데이터 존재, 분석 타겟팅 오류는 하나의 성분으로 추출되어, 해당 프로세스에서 발생 가능한 위험요인으로 분석된다.

분석 데이터 가시화 및 활용 프로세스는 총 7개의 변수를 분석한 결과, 2개의 하위 요인으로 분류되었다. 각 하위요인은 빅데이터 분석결과 불신, 부작용에 대한 책임 기준 모호, 빅데이터 소유권 분쟁, 잘못된 정보 제공과 같은 4개의 변수가 '빅데이터 활용 역기능 측면'으로 정의한다. 이는 빅데이터가 실질적으로 활용 후에 본래 의도와 다른 영향을 미치는 위험으로 설명된다. 또 다른 하나의 하위요인은 빅데이터 정보 남용, 데이터 가시화 기술 부족, 빅데이터 활용 격차로 3개의 변수가 '빅데이터 기술 복잡성 측면'으로 정의된다. 이는 빅데이터 활용이 용이하지 않고 기술적 보편화가 되지 못함에 대한 위험으로 설명된다.

공통 요인 4개 변수 중 '시스템적 오류'가 단일 성분으로 기각되었고, 빅데이터 인식 부족, 빅데이터 플랫폼 부재, 도입 비용의 요인이 한 성분으로 분석되었다. 이는 빅데이터 도입을 위한 빅데이터 활용 환경의 부족을 설명한다.

4.3 위험요인의 위험도에 따른 순위

본 논문의 요인분석에서 채택된 위험요인들 사이

〈표 2〉 빅데이터 처리 프로세스별 위험요인 분석 결과

구 분	변수		고유값	누적분산 비율	요인 적재량			크론바흐 알파 (Cronbach's α)	채택 여부
					1	2	3		
데이터 수집	정보 보안 측면	프라이버시 침해(c1)	2.059	29.412	**.822**	.040	-.113	.734	채택
		정보 이용범위 불 명시(c6)			**.752**	.141	.183		채택
		정보공개 기준 모호(c2)			**.722**	.326	.429		채택
	원천 데이터 측면	데이터 소유권 분쟁(c5)	1.808	55.241	.036	**.901**	.039	.661	채택
		데이터 수집 범위 기준 모호(c7)			.217	**.732**	.241		채택
		원천 데이터 신뢰성 부족(c4)			.472	**.550**	-.507		채택
	데이터 수집 기술 부족(c3)		1.271	73.395	.159	.176	**.851**	-	기각
데이터 저장	외부 접근권한 통제 미흡(s2)		2.733	54.664	**.928**	-.033	-	.849	채택
	외부 공격(s1)				**.834**	.124	-		채택
	빅데이터 암호화 표준정의 부재(s3)				**.817**	-.169	-		채택
	데이터 저장 기술 부족(s4)				**.711**	.513	-		채택
	데이터 노화(s5)		1.214	78.935	-.066	**.951**	-	-	기각
데이터 분석	분석 방법 기준의 다양성(a7)		4.539	64.848	**.969**	-	-	.895	채택
	내부 접근권한 통제 미흡(a1)				**.960**	-	-		채택
	데이터 분석 조작(a5)				**.803**	-	-		채택
	분석 전문 인력 부족(a2)				**.772**	-	-		채택
	분석 기술 부족(a3)				**.762**	-	-		채택
	불필요한 데이터 존재(a4)				**.748**	-	-		채택
	분석 타겟팅 오류(a6)				**.545**	-	-		채택
분석 데이터 가시화 및 활용	활용 역기능 측면	빅데이터 분석결과 불신(v2)	2.898	41.403	**.969**	.052	-	.872	채택
		부작용에 대한 책임 기준 모호(v5)			**.822**	.130	-		채택
		빅데이터 소유권 분쟁(v4)			**.819**	-.087	-		채택
		잘못된 정보 제공(v3)			**.756**	.477	-		채택
	기술 복잡성 측면	빅데이터 정보 남용(v7)	2.489	76.966	.104	**.978**	-	.817	채택
		데이터 가시화 기술 부족(v1)			-.042	**.809**	-		채택
		빅데이터 활용 격차(v6)			.168	**.790**	-		채택
공통	빅데이터 플랫폼 부재(e4)		2.348	58.702	**.976**	.126	-	.856	채택
	빅데이터 도입 비용(e1)				**.894**	-.268	-		채택
	빅데이터 인식 부족(e2)				**.772**	.409	-		채택
	시스템적 오류(e3)		1.168	87.910	.025	**.956**	-	-	기각

의 상대적 위험도 분석을 위하여 <표 3>과 같이 위험도 평균값에 따라 요인 내 순위와 전체 요인 순위를 제시하였다.

연구 결과, 데이터 수집 프로세스와 데이터 저장 프로세스에서 프라이버시 침해, 빅데이터 암호화 표준정의 부재 등 정보보안에 관한 위험요인이 상대적 높은 위험도를 나타냈다. 데이터 분석 프로세스에서는 분석 전문 인력의 부족과 분석 조작 등 전문적인 인적 자원의 부족이나 오류가 높은 위험 순위를 가졌다. 분석 데이터 가시화 및 활용 프로세스에서 빅데이터 활용의 부작용에 대한 책임 기준 모호, 빅데이터 소유권 분쟁 등의 요인이 높은 위험도를 나타냈다. 즉, 빅데이터가 가공된 후 활용 시에 사회적 부작용에 대한 문제가 위험

하다고 설명할 수 있다. 공통 요인 중에 빅데이터 인식 부족 등의 순으로 위험도가 높게 나타났다.

전체 위험요인 순위를 통해 프라이버시 침해, 정보공개 기준 모호 등 데이터 수집 프로세스에서의 정보 보안에 관한 위험요인이 심각성을 갖는다고 판단된다.

〈표 3〉 요인별 위험도 분석

프로세스	위험요인	평균	요인내 순위	전체 순위
데이터 수집	프라이버시 침해	4.23	1	1
	정보공개 기준 모호	4.18	2	2
	정보 이용범위 불 명시	4.1	3	4
	원천 데이터 신뢰성 부족	3.9	4	7
	데이터 소유권 분쟁	3.89	5	8
	데이터 수집 범위 기준 모호	3.24	6	20
데이터 저장	빅데이터 암호화 표준정의 부재	3.33	1	17
	외부 공격	3.26	2	18
	외부 접근권한 통제 미흡	3.24	3	20
	데이터 저장 기술 부족	3.08	4	23
데이터 분석	분석 전문 인력 부족	4.16	1	3
	데이터 분석 조작	3.91	2	6
	분석 방법 기준의 다양성	3.66	3	10
	분석 기술 부족	3.64	4	11
	내부 접근권한 통제 미흡	3.58	5	13
	불필요한 데이터 존재	3.15	6	21
	분석 타겟팅 오류	3.14	7	22
분석 데이터 가시화 및 활용	부작용에 대한 책임 기준 모호	3.83	1	9
	빅데이터 소유권 분쟁	3.61	2	12
	빅데이터 활용 격차	3.58	3	13
	빅데이터 분석결과 불신	3.55	4	14
	빅데이터 정보 남용	3.5	5	15
	잘못된 정보 제공	3.25	6	19
	데이터 가시화 기술 부족	3.15	7	21
공통	빅데이터 인식 부족	3.98	1	5
	빅데이터 도입 비용	3.55	2	14
	빅데이터 플랫폼 부재	3.48	3	16

5. 결 론

기존 연구를 분석한 결과 김기환[1] 윤성오[2], Boyd[9] 등의 연구에서 빅데이터의 위험요인을 제시하였지만 학술적 근거를 통한 위험요인 추출이 미비하였다. 또한 빅데이터 처리 프로세스라는 구체적 영역에서 객관적이고 신뢰성 있는 분류가 이루어지지 않았다. 이러한 점을 보완하여 본 연구에서는 빅데이터 도입에서 활용까지의 위험요인을 빅데이터 처리 프로세스인 데이터 수집, 데이터 저장, 데이터 분석, 분석 데이터 가시화 및 활용 프로세스에 따라 제시하였다. 본 연구를 통해 빅데이터 위험요인의 이해를 높이고, 빅데이터 처리 개선을 위해 빅데이터 처리 프로세스라는 세분화된 영역에서 위험에 대한 구체적인 학술적 연구의 기반이 될 것이다.

본 논문은 선행연구를 통해 설정된 빅데이터의 위험요인을 전문가 설문조사를 통해 각 요인의 위험도를 통계적으로 검증하였다. 그 결과 데이터 수집 프로세스 발생 가능한 위험요인은 프라이버시 침해 등 6개의 요인이 있으며 개인 정보 보안에 관련된 위험요인들이 위험도가 가장 높게 나타났다. 데이터 저장 프로세스에서는 4개의 위험요인이 발생 가능하며 이 또한 빅데이터 보안에 관련 위험요인이 높은 위험도를 나타냈다. 따라서 데이터 수집과 저장 프로세스에서는 위험요인을 대비하기 위해 빅데이터 개인정보보호 가이드라인 등과 같은 법적 지침을 이행하고, 데이터 필터링이나 접근 권한에 따른 통제 등과 같은 기술적 대안이 필요하다[3]. 데이터 사용에 있어 안전한 환경을 형성해야 한다. 데이터 분석 프로세스에서 분석 방법 기준의 다양성 등 7개의 위험요인이 분석되었고, 빅데이터 인적 자원에 관한 위험도가 높게 평가되었다. 즉, 해당 프로세스에서는 데이터 사이언스 양성 등 실력 있는 빅데이터 전문가를 육성하여 위험에 대비해야 한다. 분석 데이터 가시화 및 활용 프로세스에서 활용 부작용의 책임, 소유권 분쟁 등 7개의 위험요인이 발생 가능하며, 이를 대비하기 위해 출처가 명확한 원천데이터의 활용이 필요로 한다. 공통적으로 빅데이터 플랫폼 부재 등 3개의 위험요인이 유효했으며, 빅데이터 도입 환경의 미흡함을 나타내는 요인이 높은 위험순위를 가졌다.

따라서 빅데이터 활용의 가치와 위험 등 전반적인 인식 확산이 요구된다.

본 연구의 모델은 빅데이터 처리 프로세스에 따라 순차적으로 발생 가능한 위험요인을 파악하여 사전 대비를 통해 위험을 회피 할 수 있다. 또한 각 프로세스의 위험이 발생 시 위험도에 따라 요인별로 체계적으로 위험에 대응하여 효율적이고 효과적인 빅데이터 시스템에 적용이 가능할 것이다.

향후 연구는 본 연구의 위험요인을 바탕으로 각 프로세스별 위험요인을 회피하기 위해 기술적 또는 정책적인 대안을 연구 할 예정이다. 이로써 실제 빅데이터 활용 현장에서 위험요인에 따른 대책을 적용하여 위험을 회피할 수 있는 결과를 얻을 것이다.

참 고 문 헌

[1] 김기환, “공공부문 빅데이터의 활용성과 위험성”, 「정책분석평가학회보」, 제23권, 제2호(2012), pp.1-27.

[2] 윤상오, “빅데이터의 위험유형 분류에 관한 연구”, 「한국지역정보화학회지」, 제16권, 제2호(2013), pp.93-122.

[3] 이재식, “빅데이터 환경에서 개인정보보호를 위한 기술”, 「Internet and Security Focus」, (2013), pp.79-104.

[4] 조영임, “빅데이터의 이해와 주요 이슈들”, 「한국지역정보화학회지」, 제16권, 제3호(2013), pp.43-65.

[5] 추병완, “미래 인터넷 기술의 윤리 문제”, 「철학논집」, 제19권(2009), pp.65-95.

[6] 한국인터넷 진흥원, “빅데이터 기반 개인정보보호 기술수요 분석”, 2012.

[7] 홍성태 외, 「빅데이터와 위험 정보사회」, 커뮤니케이션북스, 2013.

[8] Bollier, D., *The Promise and peril of Big Data*, The Aspen Institute, Washington, DC, 2010.

[9] Boyd, D. and K. Crawford, *Critical questions for big data*, Information, Communication and Society, Massachusetts, 2012.

[10] Chester, J., *Cookie wars : How New Data Profiling and Targeting Techniques Threaten Citizens and Consumers in the ‘Big Data’ Era*, *In European Data Protection : In Good Health?*, Springer, 2012.

[11] Labrinidis, A. and H. V. Jagadish. “Challenges and Opportunities with Big Data”, *Proceedings of the VLDB Endowment*, Vol.5, No.12(2012) pp.2032-2033.

[12] Manovich, L., Trending : The Promises and the Challenges of Big Social Data in Debates in the Digital Humanities., *University of Minnesota Press*, 2012.

[13] Manyika, J. and M. Chui, “Big data : the next Frontier for Innovation, Competition, and Productivity”, *McKinsey Global Institute*, (2011), p.1.

[14] Gantz J. and D. Reinsel, *Extracting Value from Chaos*, IDC IVIEW, (2011), p.6.

[15] Paulraj Ponniah, *Data warehousing*, John Wiley and Sons, New Jersey, 2001.

◈ 저 자 소 개 ◈

이 지 은 (ptjn1@naver.com)

현재 숭실대학교 소프트웨어특성화대학원 석사과정에 재학 중이며, 서울여자대학교 컴퓨터학부를 졸업하였다. 주요 관심분야는 클라우드, 빅데이터, 데이터베이스, 품질보증 등이다.

김 창 재 (winchang@ssu.ac.kr)

숭실대학교 정보 과학대학원에서 석사 학위를 취득하고, 동 대학교 컴퓨터학부 박사 학위를 취득하였다. 현재 숭실대학교 소프트웨어특성화대학원 교수로 재직 중이다. 주요 관심분야는 소프트웨어공학, 소프트웨어 아키텍처, 데이터베이스, 빅데이터 등이다.

이 남 용 (nylee@ssu.ac.kr)

고려대학교 경영대학원에서 석사 학위를 취득하고, 미국 미시시피주립대학교 박사 학위를 취득하였다. 현재 숭실대학교 소프트웨어특성화대학원 교수로 재직 중이다. 주요 관심분야는 소프트웨어테스트, 품질보증, MIS, 정보보호 등이다.