Journal of Digital Contents Society Vol. 15 No. 4 Aug. 2014(pp. 533-545)
http://dx.doi.org/10.9728/dcs.2014.15.4.533

빅데이터 분산처리시스템의 품질평가모델

최승준*, 박제원**, 김종배***, 최재현****

요 약

IT기술이 발전함에 따라, 우리가 접하는 데이터의 양은 기하급수적으로 늘어나고 있다. 이처럼 방대한 데이터들을 분석하고 관리하기 위한 기술로 등장한 것이 빅데이터 분산처리시스템이다. 기존 분산처리시스템에 대한 품질평가는 정형 데이터 중심의 환경을 바탕으로 이루어져 왔다. 그러므로, 이를 비정형 데이터 분석이 핵심인 빅데이터 분산처리시스템에 그대로 적용시킬 경우, 정확한 품질평가가 이루어질 수 없다. 따라서, 빅데이터 분석 환경을 고려한 분산처리시스템의 품질평가모델에 대한 연구가 필요하다. 본 논문에서는 소프트웨어 품질에 관한 국제 표준인 ISO/IEC9126에 근거하여 빅데이터 분산처리시스템에서 요구되는 품질평가 요소를 도출하고, 이를 측정하기 위한 메트릭을 정의함으로써 새로이 품질평가모델을 제안한다.

키워드 : 분산처리시스템, 빅데이터, 품질평가모델, ISO/IEC9126

A Quality Evaluation Model for Distributed Processing Systems of Big Data

Seung-jun Choi*, Jea-Won Park**, Jong-Bae Kim***, Jae-Hyun Choi****

Abstract

According to the evolving of IT technologies, the amount of data we are facing increasing exponentially. Thus, the technique for managing and analyzing these vast data that has emerged is a distributed processing system of big data. A quality evaluation for the existing distributed processing systems has been proceeded by the structured data environment. Thus, if we apply this to the evaluation of distributed processing systems of big data which has to focus on the analysis of the unstructured data, a precise quality assessment cannot be made. Therefore, a study of the quality evaluation model for the distributed processing systems is needed, which considers the environment of the analysis of big data. In this paper, we propose a new quality evaluation model by deriving the quality evaluation elements based on the ISO/IEC9126 which is the international standard on software quality, and defining metrics for validating the elements.

Keywords : Distributed Processing Systems, Big Data, Quality Evaluation Model, ISO/IEC9126

1. 서론

PC시대를 지나 모바일시대로 접어들면서, 이제 IT는 일상생활 속에 자리 잡게 되었다. 개개인이 손 안에 PC를 휴대하는 새로운 IT패러다임은 소셜 네트워크 서비스의 활성화로 이어졌다[21]. 이로 인해 정보화 시대의 핵심이라 할 수 있는 '데이터'에 대한 인식도 변화하게 되었는데, 과거와 달리 일상생활의 기록 즉, '라이프로그(Life-log)' 등으로 대표되는 비정형 데이터의 가치가 화두가 되고 있다. 하루에도 수 없이 쏟아지는 데이터들로 인해 2007년부터 이미 그 저장 공간을 초월하기 시작하였으며, 2011년부터는 그 양이 제타 바이트 시대에 이르렀다[15].

이러한 정보의 홍수에서 이제는 기존의 데이터 분석과 관리 방식이 아닌 새로운 패러다임이

※ 교신저자(Corresponding Author): Jae-Hyun Choi
접수일:2014년 07월 10일, 수정일:2014년 08월 27일
완료일:2014년 08월 31일
* 숭실대학교 SW특성화대학원, csj0722@ssu.ac.kr
** 숭실대학교 SW특성화대학원, jwpark@ssu.ac.kr
*** 숭실대학교 SW특성화대학원, kjb123@ssu.ac.kr
**** 숭실대학교 SW특성화대학원, jaehyun@ssu.ac.kr

필요하게 되었으며 이에 주목 받고 있는 것이 바로 '빅데이터'이다. 실제, The Economist는 쏟아지는 데이터의 홍수를 빅데이터 시대의 하나의 원인으로 보았다[25]. 이미 하나의 트렌드로 자리 잡은 클라우드 컴퓨팅에 이어 이제는 대다수의 기업들이 빅데이터에 관심을 갖고 이를 활용하는 추세이다. 2011년 Gartner는 빅데이터를 앞으로 주목 받을 기술로 명시하며 현 단계를 '기술 발생 단계'로 언급하는 등, 향후 지속적인 발전을 할 것으로 전망하였다[11]. Mckinsey 또한, 빅데이터를 미래 비즈니스의 10대 핵심기술 중 하나로 선정하였다[19][20].

실제로 2004년, 구글에서 맵리듀스에 관한 논문이 발표된 이후, 하둡 등 오픈소스 프레임워크를 필두로 빅데이터의 기술적인 관심과 발전은 비약적으로 발전하고 있다. 그러나 이러한 기술적 발전에 비해 빅데이터 분산처리시스템의 품질에 대한 연구는 부족한 것이 현실이다. 분산처리시스템은 빅데이터 자원을 저장하고 활용하기 위한 대표적인 접근방식 중 하나이다[21]. 단순한 빅데이터 환경의 구축 및 분석보다는 분석의 목적을 갖는 것이 중요하지만[14] 더 나아가, 이를 가능하게 해주는 분산처리시스템에 관한 품질평가모델은 기업의 비즈니스적인 측면에서 볼 때 더욱 중요하다. 이에 본 논문에서는 국제 표준인 ISO/IEC9126을 기반으로 기존의 분산처리시스템은 물론, 빅데이터의 개념과 특성까지 고려한 품질평가 요소들을 도출한다. 또한, 이를 검증하기 위한 각 요소들에 대한 메트릭을 정의하여 최종적으로 빅데이터 분산처리시스템에 관한 품질평가모델을 제안하고자 한다.

2. 관련연구

2.1 빅데이터 분산처리시스템의 품질

분산처리란 여러 개의 데이터 프로세싱 센터 혹은 노드를 통해 나누어 돌아가는 프로그램들의 집합이라고 할 수 있다[22]. 이러한 특징으로 분산처리시스템은 시스템의 일부 고장에도 하위 시스템들 간의 데이터 공유가 가능하기 때문에, 단 한 번의 문제 발생에도 치명적 손실을 받을 수 있는 분야에서 널리 활용되고 있다[7]. 이처럼 분산시스템의 가장 큰 특징 중 하나는 시스템의 안정성에 있기에 다양한 연구에서 분산처리시스템의 성능평가 척도로 '신뢰성'을 고려하였다[5][6][7][13].

Brewer는 자신의 CAP이론에서 분산처리시스템이 지녀야 할 세 가지 특성으로 '일관성', '가용성', '생존성'을 제시하였는데, 생존성이란 네트워크상에서 일부 메시지에서의 문제에도 시스템은 정상동작 해야 한다는 특성이며, 운영체제와 네트워크 측면에서는 '가용성'이, 데이터베이스 측면에서는 '일관성'이 고려되어야 한다고 하였다[10].

장애상태에 초점을 맞추어 분산시스템이 지녀야 할 특성을 제시한 이러한 CAP이론을 보완하여 Abadi는 정상상태까지 고려한 PACELC을 제안하며 새로이 시스템의 응답시간을 고려하였다[8].

CAP와 PACELC 모두 분산처리시스템의 특징을 바탕으로 시스템이 지녀야 할 품질요소를 제시하였지만 빅데이터가 아닌, 기존 정형 데이터들을 대상으로 한다는 한계점이 존재한다.

기존 데이터 및 분산처리와는 달리, 빅데이터와 그 분산처리에 있어서의 특징 및 품질에 관한 연구는 아직까지 부족한 것이 현실이다.

2012년 O'Reilly에서는 빅데이터가 가지는 특성을 Volume, Variety, Velocity의 3V로 정의하였으며, 한국정보화진흥원(NIA)에서는 라이프-로그형태 등의 비정형 데이터들에 대한 실시간 처리와 이러한 다양한 형태의 대용량 데이터들로부터 야기될 수 있는 SNS나 앱 상에서의 개인 프라이버시문제를 빅데이터 활용에 있어서의 품질이슈로 언급하였다[21].

2013년 호주 정부에서는 빅데이터 분석에 있어서의 이슈로 프라이버시 및 보안성과 신뢰성, 데이터의 공유, 그리고 분석을 위한 기술을 언급하기도 하였다[2].

한편, 빅데이터 분산처리의 기술적 관점에서 현재 가장 대표적인 것이 바로 하둡과 NoSQL이다. 우선, 하둡에 있어서의 품질이슈로는 하둡의 맵리듀스를 통한 분산처리 시 발생할 수 있는 장애에 대한 대처 및 대용량 데이터에 대한 처리 성능 등이 있다 하였으며, 맵리듀스의 특징으로는 고장 허용성을 언급하였다[23].

또한, 오늘 날의 RDBMS가 분산 데이터 처리 환경에 적합하지 않게 되어 등장한 NoSQL의 특

징으로는 확장성과 대규모 클러스터 환경에서의 고장 허용성에 있다 하였으며, 품질이슈로는 신뢰성 및 접근성 향상을 언급하였다[23].

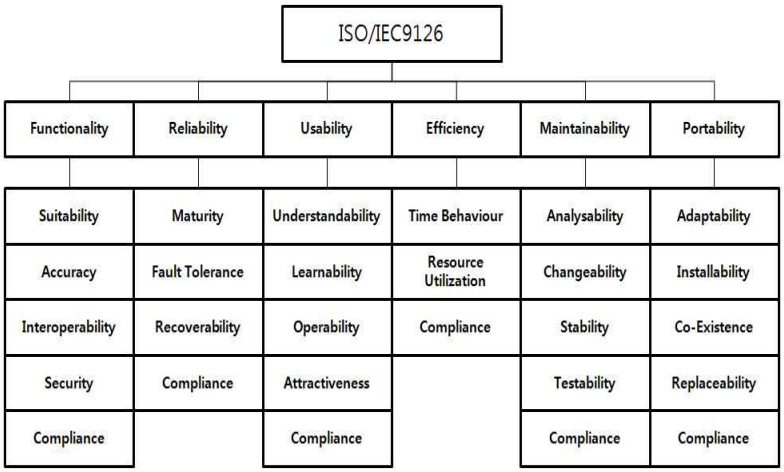
기존 분산처리시스템이 아닌, 새로이 빅데이터 분산처리시스템에 관한 품질평가를 위해서는 관련연구를 통해 언급된, 이러한 특징 및 이슈들을 함께 고려하여야 한다.

2.2 ISO/IEC9126 국제표준 품질평가모델

ISO/IEC9126은 소프트웨어 품질 평가를 위한 국제 표준 모델로, 6개의 주특성과 하위 세분화된 27개의 부특성을 정의하고 이를 측정하기 위한 메트릭 또한 정의한다[16][17]. (그림 1)은 ISO/IEC9126의 품질특성 및 부특성이다.

하지만, ISO/IEC9126 모델은 단일 소프트웨어를 대상으로 하기 때문에 본 논문에서 평가하고자 하는 빅데이터 분산처리시스템에 그대로 적용시키기에는 한계가 있다. 따라서 ISO/IEC9126 품질특성을 기반으로, 관련연구를 통해 도출된 특성들을 반영한 새로운 평가 항목에 대한 추가와 관련성이 낮은 항목에 대한 제외가 필요하며, 경우에 따라 일부 특성과 그에 대한 메트릭을 평가하고자 하는 시스템에 맞게 재정의 할 필요가 있다.

(그림 1) ISO/IEC9126 품질평가모델



(Figure 1) Quality evaluation model of ISO/IEC9126

3. 품질평가모델

3.1 빅데이터 분산처리시스템의 품질특성

이 장에서는 ISO/IEC9126에서 정의하는 품질특성과 부특성 그리고 메트릭 항목들을 바탕으로, 도출된 품질요인들을 매핑하여 빅데이터 분산처리시스템에 대한 품질평가모델을 제안한다. 이를 위해 우선, 앞서 관련연구를 통해 도출된 빅데이터 분산처리시스템의 품질특성에 부합하는, ISO/IEC9126 품질모델의 주특성 및 부특성을 식별하도록 한다.

3.1.1 기능성

기존의 분산처리시스템에서는 데이터 중복 시의 가용성을 높일 필요가 있으며, 이를 위해 일관성 또한 요구된다[1][4][12][24].

이러한 가용성은 또한, 분산처리의 핵심인 맵리듀스에서 요구되는 품질특성이기도 하다. 마스터에 해당하는 네임노드에서 모든 데이터에 관한 메타정보를 관리함으로 인하여, 네임노드에 장애가 발생 시, 하둡분산파일시스템(HDFS)으로의 데이터 저장이 불가하고, 네임노드의 데이터 유실 시, HDFS에서의 기존 데이터 조회 또한 불가하기 때문이다. 빅데이터 분산처리시스템에서의 가용성이란 결국 위에 언급된 작업 환경하에서, 시스템이 사용자에게 적절한 기능을 제

공하는지를 의미하므로 이는 적합성 항목에 해당한다. 또한, 그 과정에서 요구되는 일관성은 데이터뿐 아니라 데이터 처리에 있어서의 시스템의 정확성과 관련 있다고 할 수 있다.

또한, 네트워크 환경 상에서의 데이터 유실과 위조에 대한 보안성이 요구된다고 할 수 있는데[1][4][12][24], NIA에서도 빅데이터 분석 시 야기될 수 있는, SNS나 앱 상에서의 개인 프라이버시에 대한 보안성을 고려하였다[21].

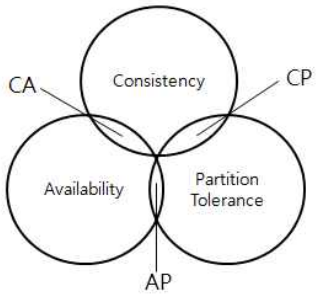
하지만 빅데이터 분산처리시스템의 품질 평가 시 생존성을 바탕으로 이기종 환경에서의 확장이나 접근성 측면에 초점을 맞추어야 하는 반면, 다른 시스템과의 상호운용 정도는 연관성이 낮으므로 상호운용성 항목은 제외하였다.

3.1.2 신뢰성

기존의 중앙 집중 형태의 프로세싱에 비해 분산처리시스템에서의 가장 큰 장점 중 하나는 신뢰성이라고 할 수 있다[9]. 이는 여러 개의 노드로 나누어 작업 시, 하나의 노드에 문제가 생겨도 그 노드에 국한되는 등, 데이터들이 하나의 노드에 편중되지 않기 때문이다[4]. 이러한 분산처리시스템에서의 신뢰성은 “분산환경 하에서의 주어진 서비스 또는 업무를 성공적으로 수행할 확률”이라 정의한다[7][13].

앞선 관련연구의 CAP이론에 따르면 이 이론의 핵심은 분산처리시스템이 3가지 중 2가지 특성을 선택해야 한다는 점에 있다. (그림 2)에서 볼 수 있듯이, CA, CP, AP의 세 가지 범주로 나뉘게 되는데, 여기서 CA는 기존의 RDBMS가 해당하는 범주이며, RDBMS는 트랜잭션 데이터 처리가 가장 핵심이 되기에 생존성을 포기하면서 일관성과 가용성을 취함을 의미한다[10]. 이 두 가지 요소는 앞서 기능성 항목에서 언급되었다. 반면 오늘 날의 대용량 데이터의 분산 처리 시, 네트워크의 완벽한 에러 방지는 불가능하다는 점에서 ‘P’라는 필수요소에 ‘C’ 또는 ‘A’의 선택 문제라 볼 수가 있는데, 여기서 P는 네트워크 에러 시의 성능 여부로, 말 그대로 결함허용성에 해당한다.

(그림 2) Brewer의 CAP이론



(Figure 2) Brewer's CAP Theorem

빅데이터 분산처리시스템에서의 신뢰성 평가의 핵심은 결함 발생을 전제로 하는 결함허용성이기 때문에, 사전의 결함 회피와 관련된 시스템의 성숙성이나 결함 발생 후 복구까지의 시간과 관련한 복구성은 제외함으로써 결함허용성 평가에 최대한 초점을 맞추도록 한다.

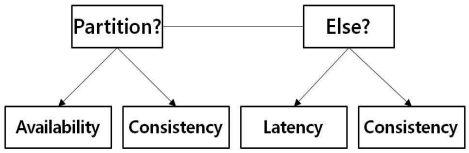
3.1.3 사용성

사용성은 시스템에 대한 사용자의 이해, 학습, 운용 그리고 선호와 관련한 항목이다[16]. 그러나 빅데이터 분산처리시스템은 사용자가 시스템에 대해 직접적으로 이해하거나 학습해야 할 필요성이 적으며, 사용자의 선호도 또한 빅데이터 분산처리시스템의 품질을 평가하기 위한 항목으로는 적합하지 않기 때문에 제외하도록 한다. 대신 사용자와 관련된 평가항목으로, 사용자의 시스템을 통한 빅데이터 자원 활용에 관한 새로운 항목인 ‘충분성‘을 통해 평가하도록 한다.

3.1.4 효율성

대용량 데이터를 효율적으로 처리하기 위해서는 시스템의 빠른 응답이 필수적이다. 관련연구를 통해 언급한 것처럼, CAP이론을 보완한 PACELC에서도 (그림 3)과 같이, Partition 상태 시는 기존의 CAP이론처럼 가용성과 일관성 사이에서의 선택이 요구되지만 새로이 정상상태 시에는 응답시간과 일관성 사이의 선택이 요구된다 하였으므로[8], 시간반응성은 품질 평가 항목으로 고려하였다.

(그림 3) Abadi의 PACELC



(Figure 3) Abadi's PACELC

반면에 사용하는 자원의 종류나 양에 따라 요구된 성능 제공과 관련 있는 자원효율성은[16], 본 논문에서 대상으로 하는 시스템에서 필요로 하는 품질특성들과 연관성이 낮기 때문에 제외하였다.

3.1.5 유지보수성

분산처리시스템은 관리상의 용이성을 특징으로 지닌다는 점에서[3][12] 유지보수성에 대한 품질평가가 필요하다. 특히 생존성 즉, 결함허용성이 반드시 고려되어야 한다는 점에서[10], 네트워크 등에서의 에러에 대한 원인 진단보다는 발생 후 취한, 시스템 변경에 대한 구현 능력인 변경성과 변경으로 인한 예기치 못한 영향을 최소화 할 수 있는 안정성이 중요하다. 안정성이란 결국, 수집된 데이터가 유실되지 않고 안정적으로 저장되는가를 의미한다[18]. 또한, 시스템의 실시간 데이터 처리 시 이러한 변경 사항을 즉시 확인하고 테스트할 필요가 있으므로 유지보수성의 부특성 항목 중에서는 분석성을 제외한 나머지 항목들을 평가한다.

3.1.6 이식성

이식성에서의 설치성 항목은 명세 된 환경에서의 소프트웨어 설치에 관한 품질항목이므로 별도의 설치를 요하는 프로그램이 아닌, 분산처리 환경에서의 시스템을 고려하는 본 논문의 품질평가모델에는 적합하지 않다. 또한, 단일 시스템을 대상으로 하기 때문에 공존 또는 대체와 관련된 나머지 항목들 또한 제외한다. 적응성 항목은 본 논문에서 필요로 하는 빅데이터 분산처리시스템의 확장정도를 평가하는데 필요한 개념 범위를 벗어나기 때문에, 별도로 확장성 항목을 분리하여 평가하도록 한다.

3.1.7 충분성

본 논문에서는 기존의 분산처리 환경과 다른 빅데이터 분산처리에서의 품질평가를 위해, '충분성'을 품질특성으로 추가하였다. 기존 데이터 품질에서의 가장 핵심적인 요소는 정확성이었지만, 빅데이터에 환경에서는 정확성보다는 충분성의 관점에서 접근해야 한다[21]. 충분성은 기존의 데이터품질의 관점에서는 활용성과 유용성에 속한 특성으로[18], 빅데이터 분산처리시스템 관점에서도 다양한 형태의 대용량 데이터들을 사용자가 유용하게 활용할 수 있는지에 대한 평가요소가 필요하기 때문에 충분성 항목을 통해 이를 측정 및 평가하도록 한다.

충분성의 첫 번째 부특성으로는 새로이 유연성 항목을 평가하도록 한다. 유연성이란, 빅데이터를 수집하기 위한 시스템은 3V 중 Variety 즉, 다양한 형태의 데이터 분석을 지원할 수 있어야 한다는 특성이다[21]. NIA에서는 빅데이터의 데이터 자원을 기존 정형 데이터 외에 크게 반정형, 비정형 데이터로 보았으며, 데이터 자원 확보 관점에서 이를 <표 1>과 같이 분류하였다. 따라서 유연성 항목에서는, 시스템의 이러한 다양한 형태의 데이터들에 대한 수용 정도를 평가하게 된다.

<표 1> 자원 확보 관점의 데이터 소스 구분

Criteria	Classification
Location	(1) Internal (2) External
Media	(1) Text (2) Audio (3) Video (4) Image (5) Mixed
State	(1) Analog (2) Digital

<Table 1> Classification of data sources at the viewpoint of resource securing

NIA에서는 '적시성'을 빅데이터 분석처리 시 고려해야 할 품질특성으로 분류하였는데[21], 적시성은 3V에서의 Velocity에 부합하는, 빅데이터의 실시간 라이프-로그 데이터와 같은 비정형 데이터 처리와 관련 있는 요소이다. 또한, 하둡 맵리듀스 등에서의 대용량 데이터 처리에 있어서도 시스템의 응답시간을 통한 빅데이터 자원의 적시적인 활용 정도에 대한 평가가 요구된다.

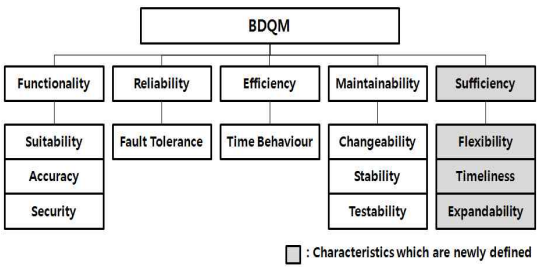
따라서 충분성의 두 번째 부특성으로 단순히 시스템의 응답시간과 관련된 시간반응성과는 다른, 적시성 항목을 추가하도록 한다.

세 번째로 확장성은 빅데이터를 수집하기 위한 시스템의 요건 중 하나로, 데이터 수집 대상이 되는 서버 대수를 무한히 확장 가능해야 한다는 것을 의미한다[21]. 또한, CAP이론에서의 결함허용성은 이기종 환경에서 고려될 경우 시스템의 확장성과 가장 밀접한 연관을 지닌 요소이기도 하다. 그 근거로, 기존의 데이터베이스에서도 분산데이터베이스가 존재하지만 높은 기능성에 비해 확장성이나 생존성이 현저히 떨어지는 반면, 빅데이터의 비정형 데이터에 대한 분산처리의 대표적 핵심 기술 중 하나인 NoSQL은 기존의 RDBMS와는 다르게 트랜잭션 처리보단 생존성을 바탕으로 사용 환경에 따라 CA가 아닌, CP나 AP범주로 특성이 분류된다는 점을 들 수 있다. 세부적으로 CP범주는 대용량 분산처리 시스템에 적합하면서 클러스터의 다중화에 고려된다면, AP범주의 대표적인 예는 바로 SNS라고 할 수 있다. 따라서 빅데이터 분산처리시스템에서 평가해야 할 품질특성으로 확장성 또한 필요하다.

마지막으로, ISO/IEC9126의 각 주특성에 속한 준수성 항목들은 품질평가 대상 시스템에 별도로 이에 관련된 규칙 또는 규정이 존재할 시 이를 준수하면 되기 때문에 제외하도록 하였다.

앞선 과정들을 통해 도출된 빅데이터 분산처리시스템에 대한 품질특성 및 부특성 항목들을 바탕으로, 본 논문에서 제안하는 '빅데이터 분산처리시스템에 대한 품질평가모델(BDQM: Big data - Distributed processing system's Quality Model)'은 다음과 같다.

(그림 4) BDQM

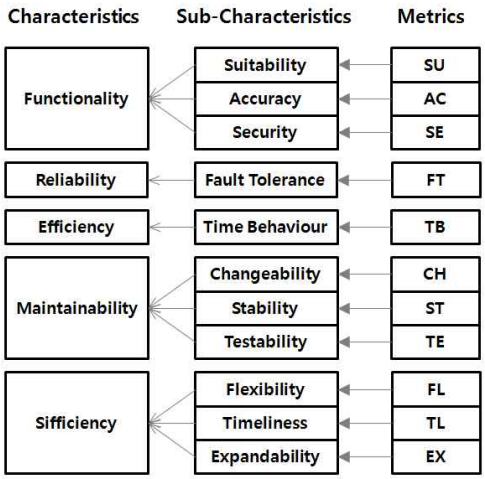


(Figure 4) BDQM

3.2 품질평가 메트릭

이 절에서는 앞서 정의된 품질속성들의 측정을 위한 메트릭을 정의하도록 한다. 각 메트릭은 계산을 위한 수식이나 값의 범위 등을 포함한다. 메트릭 명은 각 품질특성으로부터 도출하였으며 (그림 5)의 맵핑 관계로 표현된다.

(그림 5) 품질특성과의 메트릭 맵핑



(Figure 5) Mapping of the quality characteristics and metrics

3.2.1 기능성

기능성의 첫 번째 부특성인 적합성에 대한 메트릭은 시스템이 사용자가 필요한 기능을 적절하게 제공하는가를 평가하는 것이 그 목적이므로 다음과 같이 측정하도록 한다.

SU = 1 - {(측정 시 발견된 누락 기능 수) / (요구되는 시스템의 기능 수)}

적합성 메트릭 측정 값의 범위는 0과 1사이다. 누락된 수가 적을수록 측정 값은 1에 가까우며, 시스템의 적합성 또한 높다 할 수 있다.

정확성은 시스템의 데이터 처리 시의 일관성과 관련된 항목으로, 허용되는 기대치 범위를 벗어난 경우의 수를 통해 측정한다.

AC = 1 - {(허용되는 기대치 범위와 다른 경우의 수) / (데이터 처리 시도 횟수)}

정확성 메트릭은 0과 1사이의 값을 가지며 기대 범위 외 경우의 수가 적을수록 정확성은 높다고 할 수 있기 때문에, 1에 가까울수록 우수하다.

보안성에서는 빅데이터 분산처리시스템이 네트워크 상에서의 대용량 데이터 처리 시 데이터의 유실 또는 위조 등에 대비한 보안 기능을 측정하며 그 식은 다음과 같다.

SE = (문제 발생 시 제공되는 기능 수) / (문제 발생 시 필요한 기능 수)

이 식에서의 문제란, 앞서 언급한 것처럼 데이터의 유실이나 위조 등을 의미한다. 측정값은 0과 1사이의 값을 가지게 되며, 그 수치가 1에 가까울수록 보안성은 높다고 평가된다.

3.2.2 신뢰성

신뢰성에 대한 품질은 결함허용성을 통해 평가한다. 네트워크 오류 횟수에 대한 시스템 고장 횟수를 측정하도록 하며, 식은 다음과 같다.

FT = 1 - {(네트워크 오류로 인한 시스템 고장 횟수) / (네트워크 오류 발생 횟수)}

메트릭 측정값은 0에서 1사이의 값을 가지게 되며, 수치가 높을수록 즉, 1에 가까울수록 결함허용성은 우수하다고 할 수 있다.

3.2.3 효율성

효율성에서는 시간반응성 메트릭을 측정하며, 시스템에 기대되는 평균 응답시간 이내의 결과를 보인 결과 횟수를 통해 구하도록 한다.

TB = (평균 기대시간 이내 응답 횟수) / (응답시간 측정 횟수)

측정값은 0과 1사이로, 기대시간 이내로 측정된 응답 횟수가 많을수록 측정값이 1에 가까워지며 시간반응성은 우수하다고 할 수 있다.

3.2.4 유지보수성

유지보수성의 첫 번째 부특성인 변경성에 대한 메트릭은 데이터의 실시간 처리가 중요한 빅데이터 분산시스템에서, 변경사항이 얼마나 빠르게 반영될 수 있는가를 측정하게 된다.

CH = 1 - [{∑ (변경을 통한 해결 시간 - 문제 발견 시간)} / (문제 발생 수)]

변경성 메트릭은 계산식과 같이 시스템에서 발생한 문제들에 대한 변경사항 적용까지의 시간을 측정한 후 그 평균값 구하게 된다. 측정값은 0보다 크며, 값이 클수록 시스템의 변경성은 좋다고 할 수 있다.

안정성 메트릭은 시스템의 예기치 못한 변경으로 인한 영향을 최소화할 수 있는가를 측정하게 되며 다음과 같이 계산하도록 한다.

ST = 1 - {(변경사항 적용 후 발생한 문제 수) / (변경된 사항 수)}

메트릭 측정값은 0이상의 값을 가지며, 1에 가까울수록 발생한 문제 수는 적으며 안정성 또한 우수하다고 할 수 있다.

시험성 메트릭은 실시간 데이터 처리를 위해, 시스템에 대한 유지보수가 적절히 테스트될 수 있는지를 측정하게 된다.

TE = (사용할 수 있는 적절한 테스트 기능이 있는 경우의 수) / (전체 테스트 기회 수)

측정값은 0과 1사이 범위의 값을 가지며, 1에 가까울수록 시험성이 우수하다 할 수 있다.

3.2.5 충분성

충분성 항목에서는 시스템을 통한 빅데이터 활용 정도를 측정하게 된다. 첫 번째로, 유연성 메트릭은 다양한 형태로 수집되는 데이터들에 대한 시스템의 지원 정도에 대한 측정을 하게 되며 그 식은 다음과 같다.

FL = (시스템이 지원 가능한 데이터 형태 수) / (빅데이터 자원 형태 수)

유연성 측정값은 0과 1사이 값을 가지며, 시스템이 지원 가능한 형태가 다양할수록 측정값은 커지며 유연성 또한 높다고 할 수 있다.

빅데이터의 3V속성 중 앞서 유연성 항목이 다양한 형태의 데이터 처리 여부에 초점을 맞추었다면 적시성 메트릭에서는 대용량 데이터 처리 및 그 속도에 초점을 맞추도록 한다.

TL = [1/2 * {(목표 시간 내 데이터 처리 횟수) + (목표 데이터량 처리 횟수)}] / (데이터 처리 횟수)

적시성 측정값은 0과 1사이 값을 가지며 값이 클수록 적시성이 높다고 할 수 있는데, 이는 목표 시간 내 처리 및 처리량에 대한 성공 횟수가 많을수록, 측정값이 커지기 때문이다.

마지막으로 확장성은 시스템의 이기종 환경에서의 확장 정도를 평가하도록 한다.

EX = 1 - {(변화된 환경에서의 테스트 시 불완전한 기능 수) / (테스트된 기능 수)}

확장성 메트릭에서의 측정값은 0과 1사이 범위의 값을 가지며, 확장 시 불완전한 기능이 적을수록 1에 가깝기 때문에 결과적으로 측정값이 클수록 확장성은 좋다고 할 수 있다.

3.3 품질특성간 관계를 통한 가중치 적용

품질특성은 경우에 따라 다른 품질특성에 영

향을 미치기도 한다. 따라서 각각의 품질특성에 대한 보다 정확한 메트릭 적용을 위해서는 이러한 영향 관계를 바탕으로 가중치를 부여하여 그 값이 높은 특성 요인에 대한 평가 결과는 좀 더 비중 있게 반영될 수 있도록 해야 한다. 각 특성 간 영향 관계에서 세부적인 부특성 사이의 영향을 고려하여 가중치를 부여함으로써 가중치 적용 시 계산을 간편화 하도록 한다.

시스템이 빅데이터 분산처리에 적합한 기능을 제공하는지는 빅데이터를 적시적으로 활용할 수 있는지에 직접적인 영향을 미친다. 또한, 시스템의 보안 기능 제공 정도는 시스템의 네트워크 에러 시의 생존성을 높일 수 있으며, 데이터 유실을 막기 위한 안정성에도 영향을 미친다.

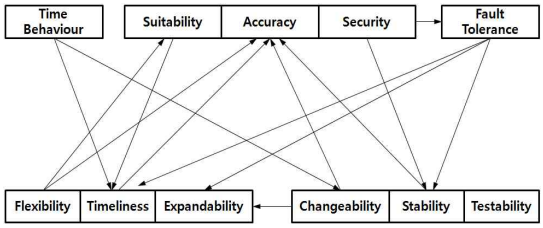
빅데이터 분산처리시스템에서의 결함허용성은 앞서 언급한 것처럼 시스템의 확장성과 연관이 있으며, 이러한 시스템의 네트워크 에러 상에서의 생존 능력은 빅데이터 자원의 실시간 활용과 안정적인 데이터 저장에 있어서도 중요하다.

시스템의 시간반응성은, 우수할수록 시스템의 변경사항에 대한 적용 또한 빨라지므로 변경성에 영향을 미친다 할 수 있다.

변경성이 우수할수록 시스템의 확장이 용이하면서 실시간 데이터 처리의 정확성 또한 높아지며, 안정성이 높으면 데이터가 유실되지 않고 저장될 수 있다는 점에서 변경성과 마찬가지로 정확성에 영향을 미친다 할 수 있다.

시스템의 다양한 데이터 형태에 대한 지원 여부는 곧 시스템의 데이터 분석에 대한 기능 제공과, 다양한 데이터 분석을 통한 정확한 분석에 영향을 미친다. 또한, SNS데이터와 같은 실시간 데이터들의 적시적 활용 여부는 곧 빅데이터 자원 분석의 정확도와 직결된다. 마지막으로, 적시성에는 시스템의 시간반응성 또한 영향을 미치게 된다.

(그림 6) 품질특성간 상호 영향 관계



(Figure 6) Mutual influence relationships of the quality characteristics

(그림 6)을 바탕으로 본 논문에서는 각 특성 간 영향 정도에 따라 <표 2>와 같이 가중치를 부여하도록 한다. 특성별 가중치는 각 부특성이 다른 부특성에 미치는 영향들에 대한 합에서 1을 더한 값으로, 이는 정확성이나 확장성 등과 같이 다른 부특성에 미치는 영향이 0인 항목의 가중치를 기본값인 1로 부여하기 위함이다. 특성별 가중치는 메트릭 적용을 통한 품질 평가 시, 비율에 따른 반영을 위해 총합이 1이 되도록 조정한다. 즉, 조정 가중치 총합인 1을 특성별 가중치 총합인 25로 나누어, 특성별 가중치 1당 0.04로 값을 조정한다.

<표 2> 품질특성별 조정 가중치 적용

Characteristic	Weight	Adjusted Weight
Suitability	2	0.08
Accuracy	1	0.04
Security	3	0.12
Fault Tolerance	4	0.16
Time Behaviour	3	0.12
Changeability	3	0.12
Stability	2	0.08
Testability	1	0.04
Flexibility	3	0.12
Timeliness	2	0.08
Expandability	1	0.04

<Table 2> Applying adjusted weight to the each quality characteristics

4. 사례연구

본 논문에서는 사례연구를 통해, 제안한 품질평가모델과 메트릭을 적용하여 빅데이터 분산처

리시스템에 대한 품질평가를 하도록 한다. 이를 위해, 가상의 분산시스템인 F와 M, 그리고 P를 가정하여 비교 측정하도록 한다. 각 시스템은 기능과 성능 면에서 서로 다른 특징을 지니고 있으며, 이는 두 가지 중 어떤 요소가 우수한 경우 시스템의 품질평가 결과에 더 큰 영향을 미쳤는지를 보다 명확하게 비교하기 위한 설정이다. 평가 항목 중 유지보수성의 경우, 대다수의 시스템에 공통적으로 적용될 수 있는 일반적인 품질특성이며, 정의된 메트릭 또한 기존 ISO/IEC9126에서 제시한 것과 동일하므로 본 논문에서의 사례연구에서는 제외하였다. 이는 빅데이터 분산처리시스템을 위해 새로이 추가되거나, 재정의된 항목들에 초점을 맞추기 위함이며, 유지보수성 항목에 대한 측정이 필요한 경우는 3장에서 제시한 메트릭과 가중치를 적용한다.

4.1 BDQM적용을 위한 시나리오 명세

“A회사에서는 마케팅을 통한 비즈니스적인 이윤 창출과 효율적인 고객 관리 등을 위하여 빅데이터 분산처리시스템을 도입하기로 결정하였으며, 대상 후보군 3곳은 ‘기능 제공에 특화된 F시스템’, ‘성능 면에서 특화된 P시스템’, 그리고 ‘기능과 성능 양측모두 중간 정도인 M시스템’이다. 회사에서는 이중 가장 품질이 우수한 시스템을 선정하기 위해 고려해야 할 테스트 요구사항을 작성하였다. 작성된 요구사항에는 선정시스템이 제공해야 하는 핵심기능 25가지와 보안기능 10가지에 대한 점검표가 체크리스트 형식으로 첨부되어 있었으며 시스템이 데이터에 대해 정확한 처리를 하는지에 대한 테스트를 언급하였다. 또한 회사의 업무 특성상, 네트워크 발생 시 얼마나 시스템이 정상 동작 가능한지가 중요하기 때문에 네트워크 에러를 인위적으로 발생시켜서라도 반드시 테스트해야 함을 명시하였다. 그 외 항목으로는 시스템의 응답시간 및 실시간 데이터 활용 정도에 대한 평가 결과 요구가 있었으며, 시스템의 확장능력은 앞서 첨부한 25가지의 기능 중 확장 시의 정상 적용 정도로 평가하도록 명시하였다. 마지막으로, 시스템이 얼마나 다양한 데이터 타입을 지원하는가를 고려하도록 당부하였다. 이러한 요구사항에 따라, 3개의 선정 후보 시스템에 대해 항목별 충족 정도를 조사한 결과는 다음과 같았다. 그렇다면 이 중 A회사에서 도입 대상으로 선정해야 할, 품질이 가장 우수한 시스템은 무엇 일까.”

(그림 7) 시나리오의 후보시스템 점검표

Test requirement checklist of the candidate systems				
* ( ) : Total number of functions		System ‘F’	System ‘P’	System ‘M’
Required essential functions (25)		25	24	24
Required essential functions - when the system is expanded (25)		23	19	21
Required security functions (10)		10	7	9
Supportable data source	Internal	O	O	O
	External	O	O	O
	Text	O	O	O
	Audio	O	△	△
	Video	△	△	△
	Image	O	O	O
	Analog	O	O	O
	Digital	O	O	O
* ( ) : Total number of trials		System ‘F’	System ‘P’	System ‘M’
System breakdown during a network error (16)		2	4	2
Data processing (25)	Result of processing within the range of expected value	21	24	23
	Response time of system within the range of expected time	21	24	22
	Data processing within the target time	19	24	21
	Data processing of the target data quantity	21	22	21

(Figure 7) Checklist of the candidate systems in scenario

4.2 시나리오에 대한 메트릭 적용

4.2.1 기능성

기능성에서는, 앞선 시나리오 명세를 바탕으로 적합성, 정확성, 보안성에 대한 메트릭을 적용한다. 적합성 항목의 경우 F시스템은 누락된 기능이 없기 때문에 측정값은 최대값인 1이 되지만 나머지 두 시스템은 총 25개 중 1개의 기능이 누락되었으므로 1 - (1/25)인 0.96의 값을 가진다. F시스템은 보안성 측정값 역시 1을 가지는 반면에 P시스템과 M시스템은 각각 10개의 기능 중 7개, 9개를 충족하므로 0.7, 0.9의 측정값을 가진다. 마지막으로 정확성 항목에서는 총 25회의 데이터 처리 시, F시스템은 4번의 허용 범위 외의 결과를 보여 1 - (4/25)으로 0.84의 값을 가지는 반면, P시스템은 1 - (1/25)으로 0.96, M시스템은 1 - (2/25)으로 0.92의 측정값을 가진다.

4.2.2 신뢰성

신뢰성은 결함허용성에 대한 메트릭으로 측정되며, 네트워크 에러 발생 시 시스템의 고장 횟수를 통해 구하도록 한다. 총 16회의 에러 발생

시, F시스템과 M시스템은 2회의 고장 횟수로, 1 - (2/16) 즉, 0.875의 측정값을 보이며 P시스템의 경우는 그보다 많은 4회의 고장으로 1 - (4/16) 즉, 0.75의 측정값을 가진다.

4.2.3 효율성

효율성은 시스템의 시간반응성 항목에 대한 측정으로 평가하며 F시스템의 경우, 25차례의 응답시간 측정 중 평균 기대시간을 21회 충족하여 21/25으로 0.84의 측정값을 보이며 P시스템은 그보다 높은 24/25으로 0.96의 측정값을 보인다. 마지막으로 M시스템은 22/25로 0.88의 측정값을 가진다.

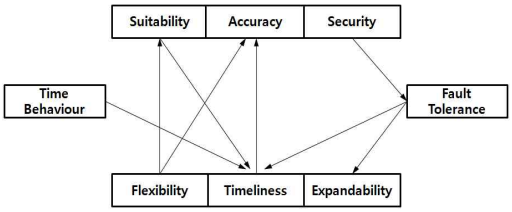
4.2.4 충분성

충분성 메트릭에서는 유연성, 적시성, 확장성의 세 항목에 대한 측정을 하게 된다. 첫 번째로 유연성의 경우, F시스템은 8개의 데이터 타입 중 비디오 부분을 제외한 7개 타입에 대한 지원을 충족하므로 측정값은 7/8 즉, 0.875로 나타났지만, 나머지 두 시스템의 경우는 그보다 적은 6/8로 0.75로 나타났다. 두 번째로 적시성 메트릭은 목표 시간 내 데이터 처리 횟수와 목표 데이터량 처리 횟수를 통해 측정하도록 한다. F시스템의 경우 [{1/2 * (19 + 21)}] / 25로 0.8의 값을 가지며 P시스템에서는 [{1/2 * (24 + 22)}] / 25로 0.96의 값을 가진다. M시스템에서는 [{1/2 * (21 + 21)}] / 25로 0.84의 값을 가지게 된다. 마지막으로 확장성 측정에서는 F시스템이 2개, P시스템이 6개, M시스템이 4개의 누락된 기능 수를 기록하여, 각각 1 - (2/25), 1 - (6/25), 1 - (4/25)으로 0.92, 0.76, 0.84의 값을 가진다.

4.3 가중치 부여

메트릭 적용을 통해 구한 측정값에는 (그림 6)에서의 품질특성간 영향 관계를 토대로, 가중치 부여를 위해 사례연구에서 제외한 유지보수성 항목 외의 나머지 특성들 간의 영향 관계를 고려하여 가중치를 부여하도록 한다.

(그림 8) 사례연구에서의 품질특성간 관계



(Figure 8) Mutual influence relationships of the quality characteristics in case study

(그림 8)의 품질특성간 관계를 바탕으로 한 특성별 가중치 총합은 16이며 <표 2>와 같은 방법으로 계산한, 특성별 가중치 1에 해당하는 조정 가중치는 1/16인 0.0625이다.

<표 3> 사례연구에서의 조정 가중치 적용

Characteristic	Weight	Adjusted Weight
Suitability	2	0.125
Accuracy	1	0.0625
Security	2	0.125
Fault Tolerance	3	0.1875
Time Behaviour	2	0.125
Flexibility	3	0.1875
Timeliness	2	0.125
Expandability	1	0.0625

<Table 3> Applying adjusted weight to the each quality characteristics in case study

4.4 결과값 분석

사례연구를 통해 비교한 각 시스템들에 대한 품질특성 메트릭 측정값과, 그에 대한 가중치 적용을 통한 최종적인 품질측정값은 (그림 9)와 같으며, 값은 소수점 3번째 자리까지 표기하였다.

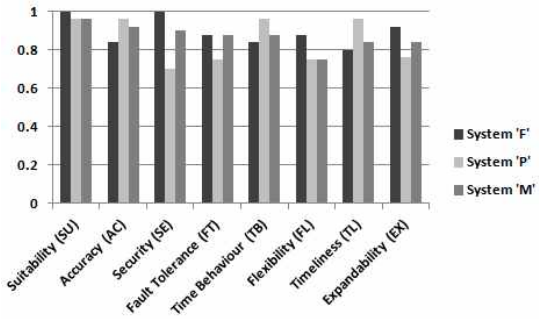
(그림 9) 후보시스템들의 BDQM 결과값

Metric	Measured value			Weight	Measured value * Weight		
	System 'F'	System 'P'	System 'M'		System 'F'	System 'P'	System 'M'
SU	1	0.96	0.96	0.125	0.125	0.12	0.12
AC	0.84	0.96	0.92	0.0625	0.052	0.06	0.057
SE	1	0.7	0.9	0.125	0.125	0.087	0.112
FT	0.875	0.75	0.875	0.1875	0.164	0.14	0.164
TB	0.84	0.96	0.88	0.125	0.105	0.12	0.11
FL	0.875	0.75	0.75	0.1875	0.164	0.14	0.14
TL	0.8	0.96	0.84	0.125	0.1	0.12	0.105
EX	0.92	0.76	0.84	0.0625	0.057	0.047	0.052
Sum					0.892	0.834	0.86

(Figure 9) BDQM result value of the candidate systems

또한, 그래프를 통한 세 시스템들의 각 특성별 측정 결과값 비교는 다음과 같다.

(그림 10) 결과값 비교 그래프



(Figure 10) Comparing graphs of the result value

기능 면에서 특화된 F시스템은 더 많은 보안 기능 제공으로, 네트워크 오류 발생 시에도 고장 발생 횟수가 낮음을 알 수 있다. 또한 이러한 생존 능력은 시스템의 확장 시에도 영향을 미쳤음을 확인할 수 있다. 반면에 분산처리 속도 및 성능에 초점을 맞춘 P시스템은 데이터 처리 시 보다 정확하고, 빠른 응답속도를 보였으며, 더 많은 목표시간 내 데이터 처리 및 목표 데이터량 처리 횟수를 기록하였음을 알 수 있다. 결론적으로, 기능적인 면에 특화된 F시스템은 BDQM 측정값이 0.892로 대략 89.2%의 품질평가 결과가 나왔다. F시스템보다 분산속도 등의 성능부분에서 장점을 보였던 P시스템은 측정값이 0.834로 83.4%의 품질평가 결과를 보였으며, M시스템의 결과는 86%였다. 따라서, F시스템이 가장 품질이 우수하다고 평가되었으며 후보군 간의 차이는 대략 3%에서 크게는 6%미만의 결과로 나타났다. 사례연구를 통한 품질평가모델 적용에서는 기능면에서 가장 우수한 빅데이터 분산처리시스템이 품질평가에서도 가장 우수한 결과를 보였다.

5. 결론

본 논문에서는 최근 가장 화두가 되고 있는 IT기술 중 하나인 빅데이터 분산처리에 있어, 이를 지원하는 시스템에 대한 품질평가모델을 제안하였다. 이를 위해, 국제 품질평가 표준인 ISO/IEC9126을 토대로 기존의 분산처리시스템에 요구되는 품질특성뿐만 아니라, 시스템에서의 빅데이터 분석 시 고려해야 할 품질특성 또한 도출하였다. 그 결과, 기존의 ISO/IEC9126으로는 평가할 수 없던 빅데이터 처리성능, 혹은 사용자의 빅데이터 활용 정도에 대한 시스템 품질평가가 가능하였다. 제안한 품질평가모델은 5개의 주특성과 11개의 부특성을 포함하며, 각각에 대한 측정 메트릭을 정의하고 사례연구를 통해 평가결과를 제시하였다.

본 논문을 통해 제안된 BDQM은 기존 분산처리시스템에서의 품질특성과 메트릭으로는 평가할 수 없는, 빅데이터 자원 분석 측면에서의 시스템에 대한 품질을 측정할 수 있다. 이러한 평가 결과는, 빅데이터 분산처리시스템의 구축 시 참고할 수 있으며, 기존 시스템의 보다 효율적인 운용을 위한 품질개선 방안 검토에도 활용 가능하다. 향후 연구로는 기업이나 기관 등의 실제 빅데이터 분산처리시스템에 대한 BDQM 적용과, 이를 통한 품질평가모델의 품질특성과 메트릭에 대한 공식화가 필요하다.

References

[1] Amajad Umar, “Distributed Computing”, Prentice-Hall,1993

[2] Australian Government (Department of Finance and Deregulation), “Big Data Strategy - Issues Paper”, (2013)

[3] B. M. Im, S. H. Hong, J. C. Song and M. H. Kim. (1995) “Development of A Storage System for Distributed Transaction Processing”, Proc, KIISE, 3-6

[4] C. H. Lee, “A Study of Distributed Data Processing System”, JournalofIndustrialScience&Technology115-126,(1980)

[5] Chang, M. S., Chen, D. J., Lin, M. S. and Ku, K. L. (2000), “The Distributed Program Reliability Analysis on Star Topologies”, ComputersandOperations Research,Vol.27,129-142.

[6] Chen, D. J. and Huang, T. H. (1992), “Reliability Analysis of Distributed Systems Based on a Fast Reliability Algorithm”, IEEETransactiononParallelandDistributedSystems,Vol.3,No.2,pp.139-154.

[7] Dai, Y. S., Xie, M., Poh, K. L.. and Liu, G. Q (2003), “A study of service reliability and availability for distributed systems”, ReliabilityEngineeringandSystemSafety,Vol.79,No.1,pp.103-112.

[8] Daniel, J. Abadi. “Consistency Tradeoffs in Modern Distributed Database System Design”, IEEEComputerSociety,2012

[9] Enslow, P. H., “What is a Distributed Data Processing System?”, Computer11,1(1978)

[10] Eric A. Brewer. “Toward robust distributed systems.” PrinciplesofDistributedComputing,Portland,Oregon,July,2000

[11] Gartner (2011). “Hype Cycle for Emerging Technologies 2011”

[12] Hesselgrave, M. R., “Consideration for Building Distributed Transaction Processing Systems on UNIX System V”, UniForumconference,1990

[13] Hsieh, C. and Hsieh, Y. (2003), “Reliability and cost optimization in distributed computing systems”, ComputersandOperationsResearch,Vol.30,No.8,pp.1103-1119.

[14] H. J. Lee, “Decombined Distributed Parallel VQ Codebook Generation Based on MapReduce”, Journal of Digital Contents Society Vol. 15 No. 3 Jun. 2014(pp. 365-371)

[15] IDC (2011). “The Digital Universe Study”

[16] ISO/IEC 9126-1. “Software Engineering-Product Quality-Part 1: Quality Model, 2001.

[17] ISO/IEC TR 9126-2. “Software Engineering-Product Quality-Part 2: Internal Metrics, 2003.

[18] Korea Database Promotion Center (2006). “Data Quality Management Maturity Model(Ver. 1.0)”

[19] Mckinsey and Company (2010). “Clouds, Big data and Smart assets: Ten tech-enabled business trends to watch”

[20] Mckinsey Global Institute (2011). “Big Data: The next frontier for innovation, competition and productivity”

[21] National Information Society Agency(NIA, 2012). “Data resource securing and Quality management plan in Big Data era” IT&FutureStrategy, No.5

[22] Scherr, A. L., “Distributed Data Processing”, IBMSystemsjournal17,4(1978)

[23] Seo et al. (2003). “Hadoop & NoSQL for the massive data analysis and processing, pp.145, 365”, Gilbut, ISBN 978-89-6618-503-0 03000

[24] Stevens, W. R., “Unix Network Programming”, Prentice-Hall,1990

[25] The Economist (2011). “Data, data everywhere”

최 승 준

2013년 2월 : 명지대학교 컴퓨터공학과 졸업 (공학사)
2013년 3월~현재: 숭실대학교 SW특성화대학원 석사과정

관심분야 : 데이터마이닝(Data Mining), 데이터 품질(Data Quality), 빅데이터 분석(Big data Analysis), 소셜 네트워크서비스(SNS) 등

박 제 원

2004년 2월 : 숭실대학교 컴퓨터학부(석사)
2011년 3월 : 숭실대학교 컴퓨터학부(박사)
2013년~현재 : 숭실대학교 SW특성화대학원 교수

관심분야 : 소프트웨어테스팅, 소프트웨어프로세스, 웹서비스, SOA/ESB, 프로젝트관리

최 재 현

2004년 2월 : 숭실대학교 컴퓨터학부(석사)
2011년 8월 : 숭실대학교 컴퓨터학부(박사)
2013년~현재 : 숭실대학교 SW특성화대학원 교수

관심분야 : SW공학, 클라우드, SW프레임워크, 서비스엔지니어링, 데이터마이닝

김 종 배

2002년: 숭실대학교 대학원 석사
2006년: 숭실대학교 대학원 박사
2004년~2006년: 남서울대학교 컴퓨터학과 겸임교수
2006년~현 재: 서울여자대학교 컴퓨터학부 겸임교수
2012년~현 재: 숭실대학교 SW특성화대학원 교수
관심분야 : 소프트웨어 개발 방법론, 에이전트 시스템, 오픈소스 SW 등