빅데이터의 효과적인 처리 및 활용을 위한 클라이언트-서버 모델 설계*

박대서
강원대학교 컴퓨터정보통신학과
(gentlevento@naver.com)

김화종
강원대학교 컴퓨터정보통신학과
(hjkim3@gmail.com)

최근 빅데이터 분석은 기업과 전문가뿐만 아니라 개인이나 비전문가들도 큰 관심을 갖는 분야로 발전하였다. 그에 따라 현재 공개된 데이터 또는 직접 수집한 데이터를 분석하여 마케팅, 사회적 문제 해결 등에 활용되고 있다. 국내에서도 다양한 기업들과 개인이 빅데이터 분석에 도전하고 있지만 빅데이터 공개의 제한과 수집의 어려움으로 문제이해를 부재에서부터 어려움을 겪고 있다. 본 논문에서는 빅데이터 공유를 방해하는 개인정보, 빅트래픽 등의 요소들에 대한 기존 연구와 사례들을 살펴보고 정책기반의 해결책이 아닌 시스템을 통해서 빅데이터 공유 제한 문제를 해결 할 수 있는 클라이언트-서버 모델을 이용해 빅데이터를 공개 및 사용 할 때 발생하 는 문제점을 해소하고 공유와 분석 활성화를 도울 수 있는 방안에 대해 기술한다. 클라이언트-서버 모델은 SPARK를 활용해 빅트래픽 문제점을 처리하며 Server Agent와 Client Agent로 구분해 데이터 제공자가 데이터를 공개할 때 서버 측의 프로세스와 데이터 사용자가 데이터를 사용하기 위한 클라이언트 측의 프로세스로 구분하여 설명한다. 특히, 빅데이터 공유, 분산 빅데이터 처리, 빅트래픽 문제에 초점을 맞추어 클라이언트- 서버 모델의 세부 모듈을 구성하고 각 모듈의 설계 방법에 대해 제시하고자 한다. 클라이언트-서버 모델을 통해 서 빅데이터 공유문제를 해결하고 자유로운 공유 환경을 구성하여 안전하게 빅데이터를 공개하고 쉽게 빅데이터를 찾는 이상적인 공유 서비스를 제공할 수 있다.

주제어 : 빅데이터, 클라이언트, 서버, 스파크, 사전분석

논문접수일 : 2016년 11월 22일  논문수정일 : 2016년 12월 18일  게재확정일 : 2016년 12월 21일
원고유형 : 일반논문  교신저자 : 김화종

1. 서론

ICT 분야는 물론 전 산업에 걸쳐 빅데이터에 대한 관심이 증가하면서, 시장의 궁정적 정회와 함께 다양한 활용 사례가 등장하고 있다. 세계 시장에서 빅데이터는 단순한 열풍에서 가치가 입증된 기술로 변모되고 있으며 빅데이터와 예측분석 솔루션을 이용하여 데이터의 가치를 최적화하려는 움직임이 확실하게 나타나고 있다 (Oh, 2015, p. 1). 또한, 빅데이터 분석을 지원하는 유무료 분석 도구들이 다양하게 소개되고 있으며 지속적인 기능 추가와 개선을 통해서 기관, 기업, 개인 등의 분석가들에게 쉽게 데이터를 다룰 수 있도록 지

* 본 연구는 한국정보화진흥원(NIA)의 미래네트워크 선도시험단 (KOREN) 사업 지원과제의 연구결과로 수행되었음 (16-951-00-001).
박대서 · 김화종

원하고 있다. 그러나 2008년 옥션 1,863만 건, 2009년 GS 칼텍스 1,125만 건을 비롯해 2014년 KT 1,170만 건 등 홈페이지 해킹 등에 의한 고객 정보 유출 사고 등 고객정보 유출 사건이 잇따라 일어나며 데이터의 공유 및 활용에 대한 고객들의 부정적 시선이 강해 공유와 유통이 차단되어 있는 것이 현실이다(Choi, 2016, p. 44).

현재 빅데이터의 활성화를 위한 제도 개선이나 빅데이터 공개 서비스 등이 국내외에서 다양하게 이루어지고 있으며 국내 정부 3.0(data.go.kr)과 같은 공공데이터를 개방하는 서비스가 주로 시행되고 있다. 정부 차원의 노력 이외에도 기업이나 개인이 보유한 데이터를 공유하는 서비스가 운영되고 있지만 공유되는 데이터의 양이 부족하여 유용한 데이터를 찾기가 어려운 상황이다. 또한, 공유되는 데이터에 대해서도 그 속성과 간단한 정보를 파악하기 위해서는 전자 데이터를 다운로드 받고 살펴보아야하기 때문에 빅트래픽 문제도 발생할 수 있다. 이러한 문제들을 해결해 빅데이터가 활성화 될 수 있도록 본 논문에서는 빅데이터 공유, 분산 빅데이터 처리, 빅트래픽의 3가지 문제에 초점을 맞추어 이를 해결할 수 있는 클라이언트-서버 모델의 모듈을 소개하고 설계 방법에 대하여 기술한다.

본 논문은 2. 이론적 배경에서 빅데이터 처리 및 활용에서의 문제 정의, 새로운 시스템의 필요성, 클라이언트-서버 모델 설계, 클라이언트-서버 모델의 효과에 대해 기술한다. 다음으로 3. 모델 설계에서 구체적인 클라이언트-서버 모델을 소개하고 4. 연구결과와 5. 결론을 통해 논문의 내용의 정리와 추후 연구 계획을 기술하였다.

2. 이론적 배경

빅데이터 활성화에 대한 노력이 지속적으로 추진되고 있지만 법적 장벽 등의 이유로 빅데이터 서비스는 많은 제한이 따르며 특히, 개인정보 보호와 관련한 법은 빅데이터 분석에 있어서 데이터의 수집 및 처리과정에 현실적인 장애 요소로 나타나고 있다(Hco, 2014, Judith, 2015). 이러한 맥락에서 개인정보 보호와 관련한 법은 빅데이터 활용을 제한하기 위해 여러 국가에서 정책 개선을 지속적으로 추진하고 있다(Judith, 2015).

문제들을 해소하고 국가적으로 빅데이터를 공유하고 활용하기 위해서 빅데이터에 대해 공유 수준을 정하고 공유 수준에 따라서 공유할 수 있도록 데이터의 범위를 설정할 수 있는 방법이 필요하다(Kim, 2014). 공유 시스템이 개인정보를 주고 받고 적절한 해결책이 구비되어 있지 않으면 사용자들은 시스템에 대한 부정적인 인식을 갖게 되며 빅데이터 활용화를 저해하게 될 것이다(Lee et al., 2013).

큰 이슈인 개인정보에 더하여 빅데이터는 그 자체를 네트워크상에서 주고 받음으로써 빅트래픽을 발생시킨다. 빅트래픽 문제는 나아가서 대규모 공유 시스템이 구축되었을 때 더욱 심각될 수 있으며 빅데이터를 공유 할 때 불규칙적인 지연을 발생 시키지 않아야 트래픽으로 인한 문제를 예방 할 수 있다(Jeon et al., 2013).

2.1 빅데이터 처리 및 활용 문제 정의

빅데이터 처리와 활용에 있어 3가지 문제를 다음과 같이 정의 할 수 있다. 첫째는 빅데이터 공유 문제로 빅데이터 공유 활성화를 어렵게 하
는 프라이버시, 공유 서비스 부족 등에 대한 문제이다. 최근 빅데이터를 이용한 성공적인 서비스 혁신 사례들 소개되면서 빅데이터가 개인 정보 유출이나 프라이버시 침해를 가속화 할 것이라는 우려도 커지고 있으며 빅데이터의 성공적인 활용은 사람들 신뢰를 필요로 하기 때문에 프라이버시 침해에 대한 안전성이 확보되어야 한다(Choi et al., 2013, p. 109). 또한, 빅데이터는 그 내용 자체가 방대하기 때문에 적절한 내용을 파악하는데 많은 시간이 소요된다. 빅데이터가 공개되면 사용자가 공개된 빅데이터를 빠르게 파악하여 다음받고 이용할 수 있어야 빅데이터 공유가 활성화 될 수 있다. 원본데이터를 그대로 다음받고 활용하기 전에 시스템 상에서 원본데이터를 전반적으로 이해하고 파악할 수 있는 요약, 샘플데이터 정보가 제공되어야 한다. 그렇지 않다면 다수의 빅데이터가 존재하더라도 사용자는 원하는 데이터를 찾기 어렵기 때문에 활용한 빅데이터 공유를 기대할 수 없다.

둘째는 분산 빅데이터 처리 문제로 데이터의 통합관리, 중복 데이터 처리, 실시간 분석에 대한 문제이다. 데이터 통합과 중복 데이터 처리, 실시간 분석 등의 전처리는 빅데이터의 특성상 많은 시간을 소요하게 된다. 때문에 빅데이터를 가공 처리 할 때에는 작업을 분할하여 여러 서버에서 처리 한 후에 다시 하나로 통합하는 것이 효율적이다.

셋째는 투트리픽 문제로 빅데이터 전송간에 트래픽 문제가 발생할 수 있으며 이것은 원시 데이터를 공유 할 때 더욱 심각해 될 수 있다. 또한, 공유된 빅데이터에 대해서 부실한 정보만 제공이 된다면 빅데이터를 탐색하는 사용자 입장에서 자세한 내용을 파악하기 위해 전체 빅데이터를 모두 다운로드 받아야만 한다. 이것은 한명의 사용자가 아닌 다수의 사용자에게 해당되는 문제로 그 만큼 트래픽 문제가 심각하게 나타나게 된다.

2.2 새로운 시스템의 필요성

앞서 정의한 문제들에 따라 빅데이터 처리 및 활용을 위한 새로운 시스템의 필요성을 정리할 수 있다.

첫째로 빅데이터 공유 문제를 해결하기 위한 방법으로 빅데이터 사전분석 기술이 필요하다. 사전 분석은 빅데이터 공유 문제를 해결하기 위해 본 논문에서 제안하는 개념으로 데이터를 사전에 미리 분석하여 생성된 결과물을 사용자에게 제공하는 것을 의미한다. 사전 분석을 통해서 데이터 사용자가 빅데이터를 검색 할 때 빅데이터의 속성이나 특성을 파악할 수 있는 정보를 제공함으로써 사용자의 빅데이터 활용성을 높여 줄 수 있다. 또한, 사전 분석을 통해 생성된 요약 데이터나 샘플 데이터를 공유함으로써 원본데이터를 공개할 때 발생할 수 있는 보안 문제를 해결하여 데이터 제공자와 데이터 사용자간의 빅데이터 공유를 활성화 할 수 있다.

둘째로는 스파크(Shoro, 2015) 등의 인메모리 기반 빅데이터 처리 도구를 활용한 빅데이터 분산처리를 통해서 원데이터에 대해 공유 개수 준이나 후 상태에 따라 적절한 전처리 결과를 빠르게 생성하고 결과를 사용자에게 제공하는 것이 필요하다(Bok et al., 2016). 디스크 기반의 환경에서는 대용량 데이터의 저장 시 큰 문제가 없지만 실제로 빅데이터를 이용해 분석을 수행하면 디스크 접근 횟수에 따라 큰 오버헤드를 발생시킨다. 디스크 기반의 빅데이터 분석의 문제점을 해결하기 위한 방법으로 현재 인메모리 환경에
서의 분석이 활발하게 이루어지고 있다. 인메모리 환경에서는 자주 사용하는 데이터 셋을 메모리에 유지시키며 접근하는 방식으로 디스크보다 빠른 성능을 보인다(Park et al., 2015).

세 번째로 빅트래픽 문제를 해결하기 위해 시스템에서는 실시간으로 네트워크 망의 트래픽을 감시하고 사용자가 요청한 데이터를 전처리할 때 현재 망에서 가용한 크기에 맞춰 전처리하여 사용자에게 전송함으로써 빅트래픽이 발생하지 않도록 하는 작업이 필요하다. 인터넷 구간의 네트워크 성능을 측정하여 시스템 이용기간에 발생할 수 있는 네트워크 문제를 미리 해소하며 여러 사용자들의 데이터 이용에 제약이 생기지 않도록 하는 기능이 필요하다(Kim, 2013).

또한, 사전 분석을 통해 공개 수준에 따라 다양한 크기의 데이터를 제공하도록 본 논문에서 제시하였다. 이러한 방식은 기존의 다수 시스템에서 원시데이터만을 공유하는 방식과 비교하였을 때 트래픽 발생량이 낮게 나타날 것으로 기대된다.

3. 모델 설계

3.1 클라이언트-서버 모델 설계

앞에서 언급된 빅데이터 처리 및 활용 시의 문제와 시스템의 필요성에 따른 기능을 <Table 1>과 같이 정리할 수 있다.

| problems                        | Function                                                                 |
|---------------------------------|--------------------------------------------------------------------------|
| Big Data Sharing                | Function aimed to expose data to a specified level using pre-analysis concepts for raw data with many restrictions on disclosure. |
| Distributed Big Data Processing | Work distributed processing function is order to solve most of the costs (time, system resources, etc.) related to the actual data analysis process. |
| Big Traffic                    | Traffic monitoring function designed to control the amount of data to be transmitted according to the state of the network |

3.1.1 클라이언트-서버 구성

시스템은 Server Agent와 Client Agent 두 가지 Agent로 구성되며 각 Agent는 Server와 Client 측에 배치되어 작업을 수행한다. <Table 2>는 각 Agent의 구성 모듈을 기술하고 있다.

| Agent          | Module                          |
|----------------|---------------------------------|
| Server Agent   | - Pre-Analysys Module           |
|                | - Job Distribution Module       |
|                | - Traffic Monitoring Module     |
| Client Agent   | - Data Browsing                 |
|                | - Gathering Data                |
|                | - Data Analysis                 |

Server Agent는 데이터 제공자에게 필요한 Agent로서 빅데이터의 사전 분석을 수행하여 Sample Data, Summary Data, Raw Data의 정보를 갖는 Data Descriptor를 생성한다. 또한, 빅데이터 분산처리를 통해서 빠르고 효율적인 빅데이터 전처리를 수행하며 네트워크 트래픽을 지속적으로 감시한다.

Client Agent는 데이터 사용자 측에 배치되는 Agent로 사전분석 결과인 Data Descriptor을 통해
서 비대이터를 검색하고 빠르게 데이터를 탐색 할 수 있다. 원하는 데이터는 서버에 요청하여 공개 수준에 따른 비대이터를 다운로드 할 수 있 다.

3.1.2 Agent별 모듈 설계

1) Server Agent

Table 3은 Pre-Analysis Module의 기능에 대해 간략화한 것으로 원시 데이터를 사전 분석하고 그 결과인 Data Descriptor을 생성한다.

| Module Function | Pre-Analysis Function |
|-----------------|-----------------------|
| Input           | Raw Data              |
| Output          | Data Descriptor       |

<Table 3>은 Pre-Analysis Module의 결과물인 Data Descriptor의 수준별 데이터를 나타낸다. Level 1의 Summary Data는 원시 데이터의 컬럼별 평균이나 데이터 값의 범위, 데이터 수집기간, 데이터 제공자 등 원시 데이터의 속성정보와 데이터의 요약 통계치를 갖는 데이터로 본 클라이언트-서버 모델에서 제공하는 가장 낮은 수준의 데이터이다.

Level 2의 Sample Data는 원시 데이터로부터 일정 크기로 랜덤 추출된 데이터로 원시 데이터의 일부에 해당하는 데이터이다. Sample Data를 만드는 과정에서 무작위로 크기를 설정하여 데이터를 추출하는 것이 아닌 원시 데이터를 분석하였을 때 결과와 가까운 결과를 보이는 크기로 데이터를 추출하여 Sample Data를 생성한다. 즉, Sample Data를 통한 분석 결과가 원시 데이터의 분석 결과를 대신 할 수 있을음을 의미한다.

Level 3의 원시 데이터는 데이터 제공자가 제공하는 원본 데이터를 의미하며 Data Descriptor에서는 원시 데이터에 접근 할 수 있는 링크를 생성하여 제공한다. 이를 통해 데이터 사용자는 원시 데이터를 보유한 제공자로부터 직접 원시 데이터를 얻을 수 있다.

Data descriptor는 앞에서 언급하였던 비대이터 공유 문제를 해결하는 핵심 개념으로 데이터 제공자는 문제가 되지 않는 수준으로만 데이터를 공개 할 수 있고 데이터 사용자는 공개된 데이터의 속성을 빠르게 파악 하도록 하여 비대이터 공유 활성에 큰 도움이 될 수 있다.

Pre-Analysis Module은 <Figure 2>와 같이 설계 할 수 있으며 데이터 제공자가 엽로드 한 원시데이터를 제공자가 원하는 수준으로 요약, 샘플링, 또는 원시 데이터 연결 작업을 통해 최종 결과물인 Data Descriptor을 생성한다. 생성된 Data Descriptors는 Internet을 통해 공개되며 사용자의 검색 결과로 활용되어 사용자의 분석 목적에 맞는 비대이터를 빠르게 찾을 수 있다.

작업을 여러 서버로 분산시켜 작업을 빠르게 처리하여 사용자 요청에 신속하게 응답할 수 있도록 하는 Job Distribution Module의 주요 기능.
온 <Table 4>와 같다. 또한, 트래픽 상황까지 반영해 데이터를 전송하기 때문에 빅데이터 문제를 해결하는 주요 모듈이다.

![데이터 분류 모듈](image)

*(Figure 2)* Process of Pre-Analysis Module

*Table 4* Job Distribution Module

| Module Function                                      | Job Distribution Function                                      |
|------------------------------------------------------|----------------------------------------------------------------|
| - Connection with Traffic Monitoring Module          | - Connection with Traffic Monitoring Module                    |
| - Big data conversion according to traffic condition | - Big data conversion according to traffic condition          |
| - Distributed processing for big data preprocessing (SPARK) | - Distributed processing for big data preprocessing (SPARK) |
| - User request data transmission                     | - User request data transmission                                 |
| Input                                                | Raw Data, Traffic Information                                  |
| Output                                               | Transformed Data                                               |

Server Agent의 Job Distribution Module은 SPARK를 사용하며 빅데이터를 분산 처리해 사용자 요청에 맞는 사전분석결과를 빠르게 제 생성하여 사용자가 요청한 데이터를 제공한다.

Pre-Analysis Module과의 차이점은 Pre-Analysis Module은 데이터 사용자의 검색과 탐색을 위한 Data descriptor를 생성해 게시하는 작업을 수행하고 Job Distribution Module은 사용자가 요청한 데이터를 네트워크 상태(트래픽)에 따라 원시 데이터로부터 빠르게 제 생성하여 사용자에게 제공한다. <Figure 3>과 같은 프로세스로 동작하며 원시 데이터와 트래픽 정보를 통해서 Transformed Data를 생성한다.

*Figure 3* Process of Job Distribution Module

Job Distribution Module의 핵심인 스파크는 UC Berkeley의 AMP 랩에서 개발하였으며, 하둡의 멀리듀스 작업에서 성능의 병목현상으로 지목되던 디스크 I/O 비용을 효율화하고 데이터 분석작업에 용이한 인터오라 컴퓨팅 기반의 데이터 분산처리 시스템이다. 스파크는 RDD(Zaharia, 2012) 단위로 데이터 연산을 수행하기 때문에 대용량 데이터를 처리하기에 용이하다. <Figure 4>는 스파크 클러스터 구조로 Driver Program은 여러 개의 병렬적인 작업을 Worker Node에 있는 Executor에서 실행하여 분산처리를 수행한다.

![데이터 분산처리 시스템](image)

*Figure 4* Process of Spark Clustering System
빅데이터의 효과적인 처리 및 활용을 위한 클라이언트-서버 모델 설계

<Table 5>의 Traffic Monitoring Module은 Job Distribution Module과 연계하여 네트워크 상황에 맞는 Transformed Data를 생성할 수 있도록 트래픽 정보를 생성한다. 트래픽 정보에는 현재 트래픽 정도와 가능한 데이터 전송량 등의 정보로 구성된다.

(Figure 4) Spark Cluster Structure

(Figure 5) Process of Traffic Monitoring Module

2) Client Agent

Client Agent는 세부 모듈로 작업을 구성하지 않고 <Table 6>의 기능과 같이 Client Agent 자체에서 빅데이터 검색과 다운로드요청을 처리한다. Client Agent는 데이터를 검색하고 검색한 데이터를 Server Agent에 요청하여 데이터를 수집할 수 있다. Client Agent는 웹 기반으로 구현하여 모바일, PC 등 기기에 구애받지 않고 인터넷에 접속가능한 모든 장치로 빅데이터를 탐색할 수 있다.

(Figure 6)은 Client Agent 프로세스를 나타낸다. Client Agent에서는 사용자가 원하는 데이터를 검색하면 인터넷 상에 게시된 Data Descriptor

| Module | Traffic Monitoring Function |
|--------|---------------------------|
| Module Function | - Network traffic monitoring<br>- Generation of the current traffic information and send it to Job Distribution Module |
| Input | Network Situation |
| Output | Traffic Information |

| Client Agent Function |
|-----------------------|
| Function | - Data Browsing<br>- Gathering Data |
| Input | Search Wolr / Data to be requested |
| Output | - Search result data list and information (Data Descriptor)<br>- Download requested data (Transformed Data) |

2) Client Agent

Client Agent는 세부 모듈로 작업을 구성하지 않고 <Table 6>의 기능과 같이 Client Agent 자체에서 빅데이터 검색과 다운로드 요청을 처리한다. Client Agent는 데이터를 검색하고 검색한 데이터를 Server Agent에 요청하여 데이터를 수집할 수 있다. Client Agent는 웹 기반으로 구현하여 모바일, PC 등 기기에 구애받지 않고 인터넷에 접속가능한 모든 장치로 빅데이터를 탐색할 수 있다.
를 통해서 빅데이터의 속성과 특성을 한눈에 파악하고 원하는 데이터를 빠르게 탐색하여 Server Agent에 데이터를 요청하게 된다. Server Agent는 요청을 받아 현재 네트워크 상황과 요청 데이터 수준에 따라 Transformed Data를 생성해 사용자에게 전송하고 사용자는 승인한 테이터를 이용해 데이터 분석을 수행하게 된다.

![Client Agent Diagram](image)

(Figure 6) Process of Client Agent

3) 모델 검증

많은 정보를 포함하고 있는 빅데이터에 대해서 Server Agent는 사전분석을 통해 요약, 샘플 데이터를 포함하는 Data Descriptor를 생성한다. 사전분석을 통한 효과를 보이기 위해 R 분석도구를 사용하여 원본데이터에 대한 요약, 샘플링 시 데이터 축소 효과와 요약 데이터를 통한 기초 분석 정보 제공의 효과를 검증한다. 사전 분석에 사용하는 데이터는 Kaggle에서 무료로 제공하는 데이터셋을 활용한다. 샘플데이터와 요약데이터를 생성하고 샘플데이터가 원본데이터와 비교되었을 때 정보의 차이가 어느 정도인가에 따라서, 본 연구의 모델이 적합한지를 판단하고자 한다. 빅데이터 샘플링은 기존의 랜덤, 충돌 추출 등이 존재하며 다양한 연구를 통해 지속적으로 기법이 개발되고 있다(Jun, 2011; Kim, 2015). 여기서는 랜덤 샘플링(Kim, 2015)을 통해 원본 데이터를 축소한 데이터를 생성하고 그 샘플 데이터의 통계치와 원본 데이터의 통계치를 비교한다. <Table 7>은 Global Land Temperature(Kaggle) 데이터를 랜덤 샘플링하여 데이터 크기를 축소하고 크기에 따른 수치 데이터(AverageTemperature)의 통계량을 계산한 결과이다. Sample 1과 Sample 2를 보면 데이터 크기는 큰 폭으로 줄어들지만 통계량은 대체로 약간에 변동이 일어나는 것을 확인할 수 있다. 빅데이터는 방대한 정보로부터 의미있는 결과를 찾아내는 목적인지만 향상 모든 데이터를 분석하는 것은 아니다. 분석 목적에 따라서 데이터를 줄여야 할 수도 있고 적절한 샘플링에 의해 훈련 데이터, 검증데이터로 원본데이터를 나누어서 분석하기도 한다. 본 연구의 클라이언트-서버 모<Table 7> Global Land Temperature Data Statistic By Size(Sample)

| Data                | Size  | Rows   | AverageTemperature |
|---------------------|-------|--------|--------------------|
|                     |       |        | Min    | 1st Qu | Median | Mean  | 3rd Qu | Max   |
| GlobalLand Temperature | 520MB | 8235082 | -42.70 | 10.30  | 18.83  | 16.73 | 25.21  | 39.65 |
| Sample 1            | 72MB  | 1000000| -38.40 | 10.29  | 18.84  | 16.73 | 25.20  | 39.16 |
| Sample 2            | 7MB   | 100000 | -35.51 | 10.36  | 18.92  | 16.80 | 25.23  | 38.63 |
빅데이터의 효과적인 처리 및 활용을 위한 클라이언트-서버 모델 설계

모델에서는 사전분석을 통해 원본데이터의 정보를 해치지 않는 수준에서 적절한 샘플링을 수행하고 그 샘플 데이터를 사용자에게 제공한다. 사용자는 하나의 데이터에 대한 여러 샘플을 얻을 수 있으며 서버는 트래픽 문제가 발생했을 때 작은 크기의 샘플데이터를 사용자에게 제공함으로써 빅트래픽에도 대응할 수 있다. 또한, 요약 데이터를 생성하여 데이터를 탐색하는 사용자에게 제공함으로써 검색 시간 단축, 불필요한 트래픽 방지 등이 가능하다. <Figure 7>은 원본 데이터에 대한 요약 정보를 생성한 결과이다. 사용자는 요약 정보로 데이터를 직접 다운받고 전처리 및 탐색 없이도 브라우저 화면상에서 빠르게 확인이 가능하다.

4. 연구결과

Server Agent와 Client Agent의 모듈별 프로세스와 기능, 결과물들을 통해서 빅데이터 처리 및 활용을 위한 클라이언트-서버 모델 설계에 필요한 요소와 모듈들을 제시하였다. 제시한 모델을 기반으로 설계된 시스템을 통해서 데이터를 수집하든 사용자는 원하는 방향으로 데이터를 분석하거나 새롭게 가공하게 되며 새롭게 가공된 데이터를 Server Agent를 통해 사전분석 함으로써 데이터 사용자는 데이터 제공자로 역할을 알게 되어 타인에게 데이터를 제공하게 된다. 데이터 제공자 또한 자신이 공개한 데이터의 Data Descriptor로부터 유용한 통계정보를 얻을 수 있으며 Sample Data를 이용해 새로운 분석을 수행하는 데이터 사용자가 될 수 있다. 이렇게 원시 데이터를 처리하고 처리된 빅데이터를 사용자가 활용하게 됨으로써 자연스러운 공유 환경이 구현되며 데이터 제공자와 데이터 사용자의 역할이 구분되지 않고 누구나 제공자와 사용자가 될 수 있는 이상적인 공유 서비스를 제공한다.

5. 결론

빅데이터는 산업 전반에 걸쳐 다양한 형태로 수집되고 있으며 기술의 발달과 관심의 증가로 수집되는 빅데이터 또한, 폭발적으로 증가하고
있다. 하지만, 이러한 현상과 반대로 개인과 개인, 개인과 기업 등 상호간의 빅데이터 공유는 프라이버시와 같은 문제로 인해 제한적으로 이루어지고 있다. 물론, 공공데이터의 경우 점점 많이 공개되고 있는 추세이지만 개인이나 기업의 데이터를 공유하는 서비스를 살펴보면 공개 정도가 현저히 떨어지는 것을 볼 수 있다. 공유에 따르는 문제들을 해결하고 데이터 제공자와 사용자들을 장려하기 위해서 본 논문에서는 데이터 공유, 분산 데이터 처리, 빅데이터 문제에 대해서 살펴보고 문제 해결을 위해 필요한 두 가지 Agent와 세부 모듈을 설계하였다. 데이터 사용자가 빠르게 데이터를 찾고 탐색할 수 있도록 사전 분석 개념을 적용하였고 사용자가 요청한 데이터를 빠르게 생성하여 제공할 수 있는 SPARK를 이용한 분산 처리, 네트워크 상태의 지속적 감시를 수행하도록 모델을 설계하고 제시하였다.

본 논문에서 제시한 모델은 빅데이터 공유 서비스를 운영하는 시스템에 세부 모듈로 구성되거나 세롭게 개발되는 시스템의 기초가 될 수 있고 향후 빅데이터의 자유로운 공유가 가능하기를 기대한다.

추후 연구에서는 본 논문에서 설계한 클라이언트-서버 모델의 구축 시에 발생할 수 있는 문제점과 극복 방안에 대해 연구하여 보다 안정적인 시스템으로 발전시킬 것이다.

참고문헌(References)

Bok, K. S, M. S. Yook, Y. W. Noh, J. E. Han, Y. W. Kim, J. T. Lim, J. S. Yoo, “In-Memory Based Incremental Processing Method for Stream Query Processing in Big Data Environments”, Journal of The Korea Contents Association, Vol. 16, No. 2(2016), 163–173.

Choi, J. K, “Current status and implications of big data analysis at home and abroad”, Korea Institute of S&T Evaluation and Planning, 2016. Available at http://www.kistep.re.kr (Downloaded 16 November, 2016).

Choi, K, H. J. Kim, “A Suggestion on the Strategy for Common Sharing of Big-DATA”, Korea Institute of Information & Telecommunication Facilities Engineering, (2013), 108–114.

Jeon, Y. H. J. S. Jang,“Big Data Networking Considerations and Cisco Case Studies”, Korean Institute of Information Technology Magazine, Vol. 10, No. 3(2012), 11–16.

Judith, R, “Round and Round the Garden? Big Data, Small Government and the Balance of Power in the Information Age”, Journal of Law & Economic Regulation, Vol. 8, No. 1(2015), 49–61.

Jun, S. H, “A New Statistical Sampling Method for Reducing Computing time of Machine Learning Algorithms”, Korean Institute of Intelligent Systems, Vol. 21, No. 2(2011), 171–177.

Kim, H. J, “ODI-based data access framework for spread Big data”Information and Communications Magazine, vol. 31, No. 11(2014), 67–71.

Kim, T. H, “A Hybrid Under-sampling Approach for Better Bankruptcy Prediction”, Journal of Intelligence and Information Systems, Vol. 21, No. 2(2015), 173–190.

Kim, Y. S, “Agile Network Delay Time Modeling of Web Traffic”, Journal of KIIT, Vol. 11,
Lee, H. S, D. W. Lim, H. J. Zo, “Personal Information Overload and User Resistance in the Big Data Age”, *Journal of Intelligence and Information Systems*, Vol. 19, No. 1(2013), 125–139.

Heo, S. W, “Big Data Legal Issues in Korea”, *Journal of Law & Economic Regulation*, vol. 7, No. 2(2014), 7–21.

Oh, J. H, “Big Data Industry Top 10 News & Issues in 2015”, National Information Society Agency(NIA), 2016. Available at http://www.nia.or.kr (Downloaded 15 November, 2016).

Park, J. H, H. J. Kim, S. W. Choi, S. R. Yoon, “Comparative Performance Analysis of Logistic Regression on Apache Spark Framework”, *Korea Computer Congress*, (2015), 1531–1533.

Shoro, A. G, T. R Soomro, “Big Data Analysis: Apache Spark Perspective”, *Global Journal of Computer Science and Technology*, Vol. 15, No. 1-C(2015), 7–14.

Um, J. H, T. H. Kim, S. W. Lee, C. H. Jung and H. M, Jung, “Next-generation real-time big data distribution system trend”, Institute for Information & communications Technology Promotion, 2014. Available at http://www.itfind.or.kr/itfind (Downloaded 16 November, 2016).

Zaharia, M, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica, “Resilient Distributed Datasets: A fault-tolerant abstraction for in-memory cluster computing”, *Networked Systems Design and Implementation(NSDI)*, Vol. 12, No. 4(2012), 15–28.
Abstract

**Design of Client-Server Model For Effective Processing and Utilization of Bigdata**

Dae Seo Park* · Hwa Jong Kim**

Recently, big data analysis has developed into a field of interest to individuals and non-experts as well as companies and professionals. Accordingly, it is utilized for marketing and social problem solving by analyzing the data currently opened or collected directly. In Korea, various companies and individuals are challenging big data analysis, but it is difficult from the initial stage of analysis due to limitation of big data disclosure and collection difficulties.

Nowadays, the system improvement for big data activation and big data disclosure services are variously carried out in Korea and abroad, and services for opening public data such as domestic government 3.0 (data.go.kr) are mainly implemented. In addition to the efforts made by the government, services that share data held by corporations or individuals are running, but it is difficult to find useful data because of the lack of shared data. In addition, big data traffic problems can occur because it is necessary to download and examine the entire data in order to grasp the attributes and simple information about the shared data.

Therefore, We need for a new system for big data processing and utilization. First, big data pre-analysis technology is needed as a way to solve big data sharing problem. Pre-analysis is a concept proposed in this paper in order to solve the problem of sharing big data, and it means to provide users with the results generated by pre-analyzing the data in advance. Through preliminary analysis, it is possible to improve the usability of big data by providing information that can grasp the properties and characteristics of big data when the data user searches for big data. In addition, by sharing the summary data or sample data generated through the pre-analysis, it is possible to solve the security problem that may occur when the original data is disclosed, thereby enabling the big data sharing between the data provider and the data user.

Second, it is necessary to quickly generate appropriate preprocessing results according to the level

* Dept. of Computer and Communications Engineering, Kangwon National University
** Corresponding Author: Hwa Jong Kim
Dept. of Computer and Communications Engineering, Kangwon National University
1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do, 24341 Republic of Korea
Tel: +82-2-250-6323 Fax: +82-2-259-5677, E-mail: hjkim3@gmail.com
of disclosure or network status of raw data and to provide the results to users through big data distribution processing using spark.

Third, in order to solve the problem of big traffic, the system monitors the traffic of the network in real time. When preprocessing the data requested by the user, preprocessing to a size available in the current network and transmitting it to the user is required so that no big traffic occurs. In this paper, we present various data sizes according to the level of disclosure through pre-analysis. This method is expected to show a low traffic volume when compared with the conventional method of sharing only raw data in a large number of systems.

In this paper, we describe how to solve problems that occur when big data is released and used, and to help facilitate sharing and analysis. The client-server model uses SPARK for fast analysis and processing of user requests. Server Agent and a Client Agent, each of which is deployed on the Server and Client side. The Server Agent is a necessary agent for the data provider and performs preliminary analysis of big data to generate Data Descriptor with information of Sample Data, Summary Data, and Raw Data. In addition, it performs fast and efficient big data preprocessing through big data distribution processing and continuously monitors network traffic. The Client Agent is an agent placed on the data user side. It can search the big data through the Data Descriptor which is the result of the pre-analysis and can quickly search the data. The desired data can be requested from the server to download the big data according to the level of disclosure. It separates the Server Agent and the client agent when the data provider publishes the data for data to be used by the user. In particular, we focus on the Big Data Sharing, Distributed Big Data Processing, Big Traffic problem, and construct the detailed module of the client-server model and present the design method of each module.

The system designed on the basis of the proposed model, the user who acquires the data analyzes the data in the desired direction or preprocesses the new data. By analyzing the newly processed data through the server agent, the data user changes its role as the data provider. The data provider can also obtain useful statistical information from the Data Descriptor of the data it discloses and become a data user to perform new analysis using the sample data. In this way, raw data is processed and processed big data is utilized by the user, thereby forming a natural shared environment. The role of data provider and data user is not distinguished, and provides an ideal shared service that enables everyone to be a provider and a user. The client-server model solves the problem of sharing big data and provides a free sharing environment to securely big data disclosure and provides an ideal shared service to easily find big data.

**Key Words** : Big Data, Client, Server, Spark, Pre-Analysis

Received : November 22, 2016  Revised : December 18, 2016  Accepted : December 21, 2016
Publication Type : Regular Paper  Corresponding Author : Hwa Jong Kim
저 자 소개

박 대서
현재 강원대학교 컴퓨터정보통신공학과 석사과정에 재학 중이며 강원대학교 컴퓨터정보통신공학과에서 학사학위를 취득하였다. 주요 관심분야는 데이터마이닝, 빅데이터, 공유, 빅데이터 플랫폼 등이다.

김 화종
현재 강원대학교 컴퓨터정보통신공학과 교수로 재직 중이다. 서울대 전자공학과를 졸업하고 KAIST 전자공학과에서 통신공학 석사와 데이터통신 박사 학위를 취득하였다. 2013년부터 현재까지 강원대학교 데이터분석센터의 센터장을 겸임하고 있다. 주요 관심분야는 빅데이터, 데이터통신, 컴퓨터 네트워크 등이다.