# 시뮬레이션-프로비넌스 데이터 분석 서비스 프레임워크 설계 및 구현

# SPA: Design and Implementation of Simulation-Provenance Data Service Analytics Framework


요 약

최근에 하드웨어의 성능의 상향, 데이터 수집 기술의 발달, 공공데이터의 개방 등의 요인으로 계산과학 분야에서 대량의 데이터가 축적되었으며, 이를 바탕으로 다양한 데이터 분석 활동을 진행한다. 하지만 대부분 분석 활동의 결과물들이 띄는 형태는 실제 사용자들이 이용하기에는 불편함을 초래한다. SPA (Simulation Provenance data service Analytics framework) 시스템은 이러한 불편함을 해소하기 위해 사용자에게 시뮬레이션 작업을 보조하는 환경을 제공하려 한다. SPA 시스템은 REST API 를 이용하여 유연한 결합성과 다양한 시뮬레이션 보조 기능의 제공을 통하여 사용자에게 편의성과 다양성을 제공할 수 있는 프레임워크를 제공한다.


## 1. 서 론

최근에 보조기억장치를 비롯한 다양한 하드웨어의 성능이 향상되었다. 또한 다양한 크롤링(crawling) 라이브러리의 발달, IOT 기기를 통한 데이터 수집, 공공데이터의 활성화, 다양한 플랫폼에서 제공되는 데이터에 의해 계산과학 분야의 연구자들은 어느때보다 손 쉽게 데이터를 수집하여 관리 할 수 있게 되었다.

이렇게 축적된 방대한 데이터, 즉 빅데이터와 R[1]과 같은 통계 계산을 위한 프로그래밍 언어 혹은 많은 종류의 데이터 분석 및 머신 러닝 라이브러리를 보유한 Python[2]과 같은 도구를 이용하여 연구자들은 수월하게 데이터 분석 활동을 진행할 수 있게 되었다.

하지만 연구자들이 분석 활동을 지속하여 만들어낸 결과물들은 대부분이 문서 혹은 소스 파일의 형태로 되어있을 뿐이다. 결과적으로 프로그램의 형태를 띄지 못하기 때문에 실제로 사용자들이 이용하기 위해선 직접 작업 환경을 구축하는 등 많은 불편함과 제약이 따른다.

이러한 문제점을 보완하기 위해 본 논문은 EDISON[3] 플랫폼으로부터 수집한 데이터를 기반으로 사용자에게 시뮬레이션 작업 환경을 제공해주는 데이터 분석 프레임워크인 SPA (Simulation Provenance data service Analytics framework) 시스템을 제안한다.

SPA 시스템은 프레임워크 구현을 위해 다음과 같은 기능을 중점적으로 설계하였다. REST API[4]와 JSON 을 이용하였다. 이를 통해 통합 인터페이스 환경을 지원하여 다양한 플랫폼과 연계될 수 있도록 유연한 결합성을 제공한다.

또한 프레임워크내에서 다양한 시뮬레이션 툴들을 작동할 수 있도록 하는 환경을 제공함에 따라 사용자 편의에 따라 다양한 활동을 진행 할 수 있다

본 논문의 구성은 다음과 같다. 다음 장은 본 연구와 연관된 관련 연구를 복기한다. 이어, 3 장은 SPA 시스템에 대해 서술한다. 4 장에서 사례 연구 결과를 서술한다. 끝으로 5 장에서 본 연구의 결론 및 향후 연구에 관해 기술한다.

## 2. 관련 연구

해당 연구[5]는 스마트 폰으로부터 데이터를 추출하여 서버로 전송한 후 서버에서 데이터를 수집 및 분석하는 기능을 한다. 본 연구와는 클라이언트-서버의 형태를 가진 프레임워크라는 점에서 비슷하나, 해당 연구는 프레임워크내에서 규정한 시뮬레이션에 한하여 시뮬레이션을 진행할 수 있다는 점, 인터페이스를 모바일 애플리케이션의 형태로 제공하는 점, 데이터를 분석하여 통계에 그친다는 점, API 를 제공하지 않는다는 점에서 본 연구와 다르다.

해당 연구[6]은 IOT 을 통해 수집한 데이터를 분석하기 위한 프레임워크에 대해 다루고 있다. 본 연구와는 데이터를 관리하는 프레임워크라는 점에서 비슷하나, 해당 연구는 대량의 데이터를 수집하기 위해 클라우딩 컴퓨팅을 사용했다는 점에서 본 연구와는 다르다.

EDISON 플랫폼은 계산과학공학 시뮬레이션을 온라인으로 실행해 줄 수 있도록 하는 시스템이다. EDISON 플랫폼과는 같은 데이터 사용한다는 점에서 비슷하나, 플랫폼내에서 규정한 시뮬레이션만 제공한다는 점이 다르다.

본 논문에선 위 글에서 언급한 다른 점을 프레임워크내에 Wrapper 기능을 구현하여 사용자가 원하는 시뮬레이션 소스 코드 혹은 툴을 wrapping 하여 사용자에게 여러 가지 시뮬레이션을 제공할 수 있다는 점, REST API 를 이용하여 모든 플랫폼에서 사용가능한 느슨한 결합 형태를 제공하는 점을 이용하여 연구를 진행하였다.

## 3. SPA 시스템 설계 및 구현

SPA 시스템은 프레임워크 내에서 스크립트 언어를 통한 시뮬레이션 제공 및 REST API 서버 구축, 웹 인터페이스를 구현 하기 위한 Node js[7]와 데이터를 수집 및 관리하기 위한 MongoDB 을 기반으로 한, 그림 1 과 같은 구조로 되어 있다. 시스템은 크게 (1) Data Manager, (2) REST API Request Handler, (3) Simulation Data Base, (4) Simulation Estimator, (5) Simulation Query Interface 로 구성되어 있다.

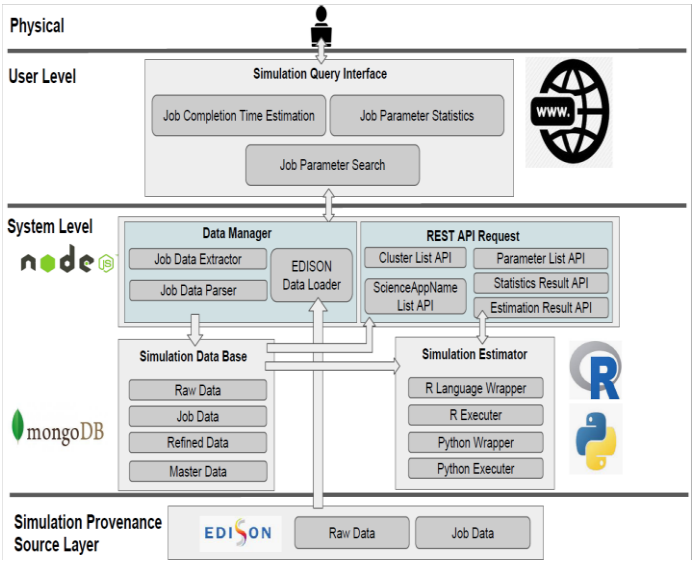


그림 1. SPA 구조

### 3.1 Data Manager & Simulation Data Base

'Data Manager'은 크게 Job Data Extractor, Job Data Parser, EDISON Data Loader 로 구성된다. EDISON Data Loader 는 Simulation Provenance Source Layer 에 있는 EDISON 플랫폼으로부터 시뮬레이션 데이터인 Raw Data 와 Job Data 를 수집한다. 수집한 데이터는 Simulation Data Base 에 Raw Data, Job Data 의 형태로 저장된다. 이어 Job Data Extractor 은 이렇게 수집한 데이터로부터 가장 최근 날짜, 즉 최신 버전의 Parameter Set 을 추출하여 Simulation Data Base 즉 MongoDB 에 Master Data 로 저장하게 된다. 저장된 Master Data 는 향후 들어올 데이터들의 기준 정보이므로 최신 버전을 유지하게 된다. Job Data Parser 는 MongoDB 에 저장된 Master Data 를 기준으로 하여 EDISON Data Loader 를 통해 받아온 데이터를 Parsing 하여 표 1 의 스키마에 서술된 데이터의 형태로 추출하여 Refined Data 로 저장하게 된다.

| | |
|---|---|
| cluster: String | EDISON Simulation 소프트웨어 |
| scienceAppName: String | Cluster 에 속해 잇는 ScienceAppName |
| simulationUuid: String | Simulation 고유 Id |
| jobExecTime: String | Simulation 수행 시간 |
| jobStatus: String | Simulation 진행 상태 |
| parameter: Array | ScienceAppName 에 따른 Parameter List |
| values: Array | Simulation Parameter 에 해당하는 값 |

표 1. EDISON 데이터 저장 및 관리를 위한 스키마(JSON 기반)

### 3.2 REST API Request Handler

'REST API Request Handler 는 Cluster List API, ScienceAppName List API, Parameter List API, Statistics Result API, Estimation Result API 로 구성된다. 위의 API 들에 대한 설명은 표 2 을 통해 대체하겠다.

| | |
|---|---|
| Cluster List API | Simulation Data Base 에 저장된 Master Data 를 기준으로 EDISON 플랫폼에서 지원하는 EDISON Simulation 소프트웨어를 JSON 의 형태로 제공한다. |
| ScienceAppName List API | Simulation Data Base 에 저장된 Master Data 를 기준으로 EDISON 플랫폼에서 지원하는 Cluster 에 속해 있는 ScienceAppName 을 Json 의 형태로 제공한다. |
| Parameter List API | Simulation Data Base 에 저장된 Master Data 를 기준으로 ScienceAppName 에 따른 Parameter List 를 JSON 의 형태로 제공한다. |

| | |
|---|---|
| Statistics Result API | 사용자가 입력한 Input Data : Cluster, ScienceAppName 을 기반으로 하여 Refined Data 에 축적된 데이터를 분석하여 현재 EDISON 플랫폼에서 많이 사용된 Parameter Set Top 10 을 JSON 의 형태로 제공한다. |
| Estimation Result API | Simulation Query Interface 에서 사용자가 입력한 Input Data : Cluster, ScienceAppName, Parameter List , 이에 해당하는 Refined Data collection 기반으로 하여, Simulation Estimator 의 R, Python Executor 를 통해 Simulation 의 수행시간을 추정하여 ms 단위로 제공한다. |

표 2. REST API 설명

### 3.3 Simulation Estimator

'Simulation Estimator'는 R 언어와 Python 에 해당하는 Wrapper, Executor 로 구성된다. Wrapper 는 Executor 즉 R 혹은 Python 으로 실행되는 시뮬레이션 툴과 Node Js 기반 서버를 연결해주는 인터페이스이다. Wrapper 의 구현 로직은 다음과 같다. Estimation Result API 가 호출 시 Child Process 를 서버내에서 생성하여 사용자가 입력한 Input Data 을 List 형태로 변화한 다음 Executor 를 호출한 뒤 Executor 에 변환된 List Data 를 제공하며 Executor 에서 반환한 값을 서버에 다시 넘겨주게 된다. Executor 는 Wrapper 로 붙어 받은 Data 를 기반으로 추정한 Simulation 의 수행시간을 Output 으로 가지며 Output 은 Wrapper 로 반환하게 된다.

### 3.4 Simulation Query Interface

'Simulation Query Interface'는 Job Completion Time Estimation, Job Parameter Statistics, Job Parameter Search 로 구성된다. Job Completion Time Estimation 은 사용자에게 Cluster, ScienceAppName, Parameter Value 를 입력 받아 Estimation Time 을 웹 인터페이스로 제공한다. Job Parameter Statistics 는 사용자에게 Cluster, ScienceAppName 을 입력 받아 해당하는 Simulation 앱 내에서 많이 사용된 Parameter Set 을 분석하여 Parameter Set Top 10 을 웹 인터페이스로 제공한다. Job Parameter Search 는 사용자가 ScienceAppName 을 입력할 경우 해당하는 Simulation 앱에서 필요한 Parameter 정보를 웹 인터페이스로 제공한다.

## 4. 사례 연구

현재 SPA 시스템에서 구현된 시뮬레이션 시간 추정 서비스와 파라미터 랭킹 시스템을 통해 SPA 시스템의 실행 시나리오를 살펴 볼 것이다.

### 4.1 시뮬레이션 시간 추정 서비스

시뮬레이션 시간 추정 서비스는 SPA 시스템내의 Simulation Query Interface 중 Job Completion Time 인터페이스를 통해 서비스를 제공한다. 시뮬레이션 시간 추정 서비스에는 Request Handler 에서 Cluster API List, ScienceAppName List, Parameter List, Estimation Result API 가 사용된다. 그림 2 와 같이 현재 구현이 되어 있다.

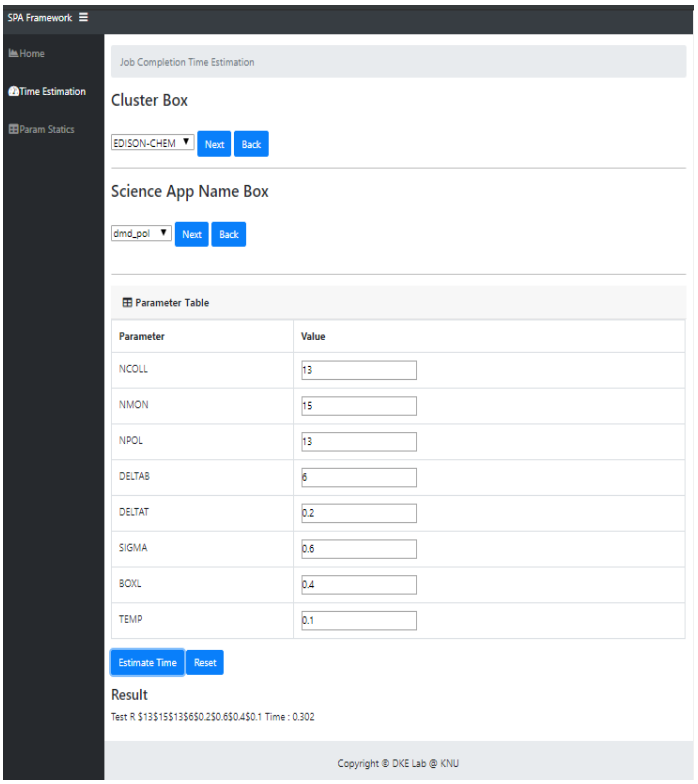


그림 2. SPA 시스템에서 제공하는 시뮬레이션 시간 추정 서비스의 View

그림 2 에서 Cluster Box 에는 Cluster List API 로부터 제공받은 EDISON 플랫폼의 Simulation 소프트웨어의 종류가 리스트에 출력된다. 그후 Cluster Box List 의 Next 버튼을 클릭 시 Cluster 에 해당하는 ScienceAppName List 가 ScienceAppName List API 로부터 제공받아 리스트에 종류가 출력된다. 원하는 ScienceAppName 을 입력한 뒤 Science App Name Box 의 Next 버튼을 클릭 시 ScienceAppName 에 해당하는 Parameter Set 이 Parameter List API 로부터 제공받아 테이블의 형태로 Parameter 셋과 value 를 입력할 input box 가 제공된다. 원하는 Value 를 입력한 뒤 Estimation Time 버튼을 클릭 시 Estimation Result API 가 호출되어 Input Data 를 서버에 전송 후 R Wrapper, R Executor 을 통해 수행 시간을 추정 후 결과 값을 하단의 Result 라벨에 출력하게

된다. Reset 버튼을 누르면 모든 상태 값들이 초기화되며, 각 Box 에 해당하는 Back 버튼을 클릭 시 이전 상태로 돌아가게 된다.

### 4.2 파라미터 랭킹 서비스

파라미터 랭킹 서비스는 SPA 시스템내의 Simulation Query Interface 중 Job Parameter Statistics 인터페이스를 통해 서비스를 제공한다. 파라미터 랭킹 시스템 서비스에는 Request Handler 에서 Cluster API List, ScienceAppName List, Statistics Result API 가 사용된다. 그림 3 과 같이 현재 구현이 되어 있다.

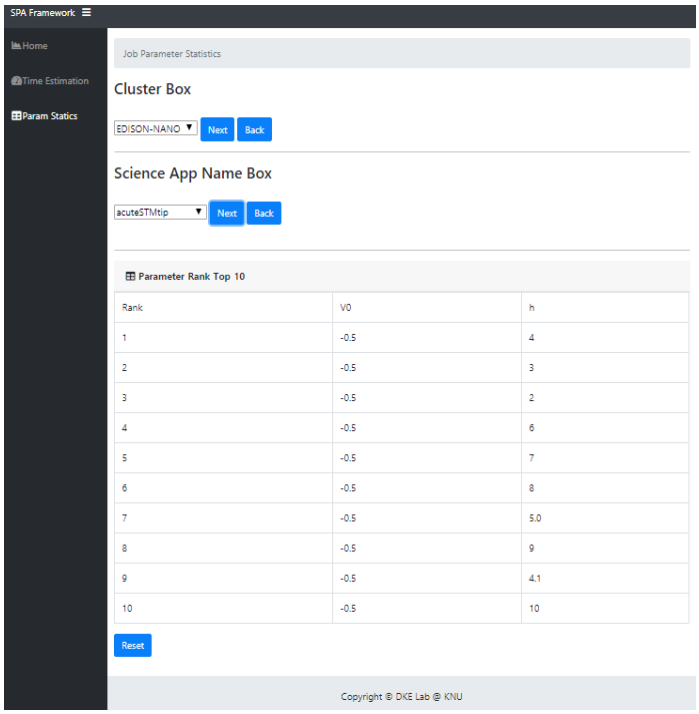


그림 3. SPA 시스템에서 제공하는 파라미터 랭킹 서비스의 View

그림 3 에서 Cluster Box 와 Science App Name Box 까지의 시나리오는 시뮬레이션 추정 서비스와 동일한 로직을 사용한다. Science App Name Box 의 Next 를 클릭 시 Input Data 를 기반으로 Statistics Result API 호출하여 사용자가 선택한 시뮬레이션에 대해 많이 사용된 Parameter Set Rank 10 을 테이블의 형태로 제공한다. Reset 및 Back 클릭 시 일어나는 시나리오 또한 시뮬레이션 추정 서비스와 동일한 로직을 사용한다.

## 5. 결론 및 향후 연구

본 논문은 Node Js, Mongodb 를 이용하여 사용자들에게 유연하고 편리한 데이터 분석 서비스 시뮬레이션 프레임워크인 SPA 에 대해 서술하였다. 기존 연구와는 다르게 프레임워크내에 여러 가지 시뮬레이션을 제공할 수 있다는 점, REST API 를 이용하여 다양한 인터페이스에서 사용가능한 것이 장점인 느슨한 결합 형태를 제공하는 점은 SPA 가 가진 장점이다. 이러한 장점을 바탕으로 SPA 가 시뮬레이션 보조 프레임워크로서 계산과학공학 연구자들에게 많은 도움을 줄 수 있을 것이라 기대한다.

향후 연구로는 Python, R 뿐만 아니라 Matlab[8]과 같이 다양한 계산공학 언어들을 프레임워크내에서 지원하도록 하여 사용자들에게 더 많은 선택지를 제공하고자 한다. 또한 시뮬레이션 보조 프레임워크로서 기능 뿐만 아니라 SPA 시스템을 이용하여 연구 데이터를 축적하여, 이를 바탕으로 다양한 분석 활동을 진행 할 수 있도록 할 것이다.

## 참 고 문 헌


[1] R, https://www.r-project.org/ (2018 년 10 월 확인)

[2] Python, https://www.python.org/ (2018 년 9 월 확인)

[3] EDISON, https://www.edison.re.kr/ (2018 년 8 월 확인)

[4] Roy T. Fielding, Richard N. Taylor, Justin R. Erenkrantz, Michael M. Gorlick, Jim Whitehead, Rohit Khard, "**Reflections on the REST Architectural Style and 'Principle Design of the Modern Web Architecture**", ESEC/FSE'17, September 4–8, 2017, Paderborn, Germany.

[5] 강동주, 조성현, 강신진 "**A Framework for Usage Pattern Analysis of Smartphone Applications**", 정보과학회 논문지: 컴퓨팅의 실제 및 레터, 19 권, 10 호, 7 쪽, 2013-10.

[6] 김병희, "**A Study on the IOT(Internet of things) Framework based on Cloud Computing**", 한국통신학회 2014 년도 추계종합학술발표회. 2 쪽, 2014-10.

[7] Node js, https://nodejs.org/ko/ (2018 년 7 월 확인)

[8] Matlab, https://www.mathworks.com (2018 년 8 월 확인)