계산과학공학 시뮬레이션의 효율화를 위한 시뮬레이션 데이터 관리 시스템 설계

이기용*, 신윤재*, 최연정*, 서영균**, 사정환**, 조금원**
*숙명여자대학교 컴퓨터과학부
**한국과학기술정보연구원 계산과학공학연구실
e-mail : kiyonglee@sookmyung.ac.kr

Design of a Simulation Data Management System for Efficient Computational Science and Engineering Simulations

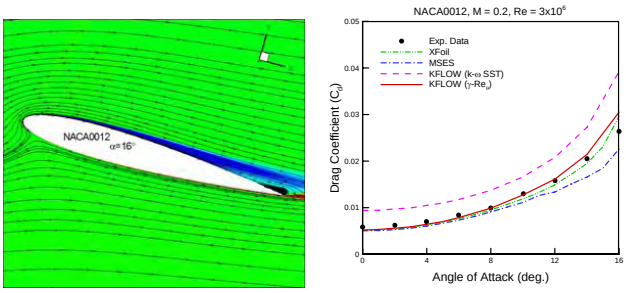
Ki Yong Lee*, Yoonjae Shin*, Yeonjung Choi*, Young-kyoon Suh**, Jeonghwan Sa**, Kum Won Cho**
*Division of Computer Science, Sookmyung Women's University
**Computational Science and Engineering Lab, Korea Institute of Science and Technology Information

요 약

계산과학공학 및 여러 과학분야에서 컴퓨터 시뮬레이션을 통한 연구가 활발히 수행되고 있다. 하지만 정확도 및 복잡도가 높아짐에 따라 시뮬레이션을 수행하는 비용도 크게 증가하고 있다. 특히 여러 입력 변수를 변화시켜가며 시뮬레이션을 수행하는 경우, 시뮬레이션 수행 비용은 더욱 커진다. 하지만 지금까지는 이전에 수행된 시뮬레이션 결과를 재활용하는 연구는 거의 이루어지지 않았다. 본 논문에서는 이와 관련된 연구들을 살펴보고, 기존 시뮬레이션 결과를 사용하여 반복 요청된 시뮬레이션에 대한 결과를 즉시 반환하거나 유사 시뮬레이션에 대한 결과를 예측하는 시스템을 설계한다. 본 논문에서 설계한 시스템을 통해, 사용자는 시뮬레이션을 수행하지 않고도 반복 또는 유사 시뮬레이션에 대한 결과를 빠르게 얻을 수 있다.

1. 서론

최근 계산과학공학 및 여러 과학분야에서 컴퓨터 시뮬레이션을 통한 연구가 매우 활발히 수행되고 있다[1]. 하지만 시뮬레이션에 요구되는 정확도 및 복잡도 증가에 따라 시뮬레이션을 수행하는 비용 자체도 크게 증가하고 있다. 특히 시뮬레이션 수행을 위한 입력 변수가 여러 개이고, 각 변수 값들을 변화시켜가며 시뮬레이션을 수행하는 경우 시뮬레이션을 수행하는 비용은 몇 시간에서 며칠 이상 소요될 수도 있다. 아래 그림은 다양한 받음각(attack of angle)에 대해 공기역학 시뮬레이션을 수행하는 모습을 나타낸다.



(그림 1) 다양한 받음각에 대한 공기 역학 시뮬레이션의 수행 예

하지만 지금까지는 이전에 수행된 시뮬레이션 결과를 활용하여 새로 요청된 시뮬레이션을 수행하는 연구는 거의 이루어지지 않았다. 대부분의 시스템은 기존 시뮬레이션 결과를 활용하지 않거나, 활용하더라도 필요한 데이터를 사용자가 직접 검색하여 사용하는 기능만을 제공한다.

본 논문에서는 기생성된 시뮬레이션 데이터의 활용과 관련된 기존 연구 및 시스템들을 살펴보고, 기존 시뮬레이션 결과를 사용하여 (1) 반복 요청된 시뮬레이션에 대한 결과를 즉시 반환하거나, (2) 유사 시뮬레이션에 대한 결과를 예측하는 시스템을 설계한다. 본 논문에서 설계한 시스템을 통해, 사용자는 시뮬레이션을 수행하지 않고도 반복 또는 유사 시뮬레이션에 대한 결과를 빠르게 얻을 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 기존 시뮬레이션 및 과학 데이터 활용에 대한 기존 연구 및 시스템들을 소개한다. 3 장에서는 이전에 수행된 시뮬레이션 데이터를 재활용하여, 새로 요청된 시뮬레이션을 효율적으로 처리하는 시스템을 설계한다. 4 장에서는 추후 연구를 기술하고 결론을 맺는다.

2. 관련 연구

본 장에서는 시뮬레이션 결과를 포함하여 기존 과

학 데이터를 활용하기 위해 개발된 기존 시스템들을 살펴보고, 기존 시뮬레이션 결과를 활용하는 관련 연구들을 살펴본다.

2.1 과학 데이터 저장 및 공유 서비스

유전체, 천체, 기후 등 다양한 과학 데이터를 저장하여 제공하는 시스템은 이미 다수 존재한다. 하지만 대부분은 미리 정의된 검색 항목들을 사용자가 직접 입력하고, 필요한 데이터를 파일 형태로 내려받을 수 있는 기능만을 제공하며, 프로그램이 자동으로 기존의 데이터를 활용하는 기능은 제공하지 않는다. 이러한 시스템의 예는 다음과 같다.

① EBI ArrayExpress (www.ebi.ac.uk/arrayexpress)
- 기능 유전체에 대한 실험 결과를 Zip 파일 형태로 제공

② Sloan Digital Sky Survey (www.sdss.org)
- 천체 이미지, 광학 스펙트럼, 적외선 스펙트럼 등에 대한 과학 데이터 제공

③ National Centers for Environmental Information (www.ncdc.noaa.gov)
- 날씨 및 기후에 대한 여러 측정 데이터 제공

④ National Snow & Ice Data Center (www.nsidc.org)
- 얼음, 눈, 온도, 빙하 등에 대한 데이터 및 관련 분석, 이미지 도구 제공

⑤ National Aeronautics and Space Administration (www.nasa.gov)
- 천제물리 관측데이터, 대기 데이터, 태양계 관측 데이터 등 다양한 데이터와 관련 분석 및 시각화 도구 제공

⑥ National Center for Biotechnology Information (www.ncbi.nlm.nih.gov)
- 생물의학, 유전체 데이터를 FTP 로 제공하는 한편, 관련 API, 라이브러리, 분석 도구 제공

2.2 온라인 과학 시뮬레이션 서비스

웹을 통해 시뮬레이션을 수행할 수 있도록 하는 시스템도 이미 다수 존재한다. 하지만 대부분 기존 데이터를 활용하는 대신 1 회성 수행이거나 일정 기간 후 결과를 삭제한다.

① PhET (phet.colorado.edu)
- 교육용 물리, 화학, 생물, 지구과학 시뮬레이션 기능 제공

② ALF (alfsim.org)
- 루트 게놈(root genome)을 관련된 여러 게놈으로 시뮬레이션하는 기능 제공

③ BiDaS (bioserver-3.bioacademy.gr/Bioserver/BiDaS)
- 대규모의 몬테 카를로 시뮬레이션 시퀀스 및 수치 계산 결과 데이터 생성 기능 제공

④ WebArrayDB (www.webarraydb.org)
- 마이크로어레이 데이터 분석 기능 제공

⑤ Simtk (simtk.org/xml/index.xml)
- 여러 시뮬레이션 S/W 및 시뮬레이션 모델 공유 사이트

2.3 시뮬레이션 데이터 공유 시스템

본 논문에서 설계하는 시스템과 목적이 가장 유사한 시스템으로서, 이미 수행된 시뮬레이션 결과를 다른 곳에서 활용할 수 있도록 하는 것이 목적이다.

① DataSpaces[2]
- 시뮬레이션 데이터에 대한 공유 저장 공간을 제공하며, 시뮬레이션 데이터 접근을 위한 프로그래밍 API, 색인, 질의 처리 기능 제공

② BIGNASim[3]
- 분자 시뮬레이션 결과 및 관련 정보를 NoSQL 인 Cassandra 와 MongoDB 에 각각 저장하여, 추후에 검색 및 재활용 가능

③ SciDrive[4]
- FITS, TIFF, YT, ASCII, Excel 등 다양한 과학 데이터 파일을 Dropbox 형태로 공유하고, 자동으로 메터데이터를 추출하여 다양한 형태로 검색 가능

④ DCMS[5]
- 분자 시뮬레이션 결과를 관계형 DBMS 에 저장함으로써 기존의 SQL, 색인, 질의 처리 기술을 이용하여 시뮬레이션 결과 검색 가능

⑤ iBIOMES[6]
- 바이오분자 시뮬레이션 결과 및 계산화학 데이터를 분산 파일 시스템에 저장하고, 그에 대한 메타데이터를 MySQL 에 저장하여 검색 기능 제공

⑥ SciBox[7]
- 시뮬레이션 데이터 및 과학 측정 데이터에 대한 클라우드 저장 시스템으로서, 사용자가 필요로 하는 데이터만 클라우드에 저장하고 원하는 데이터 전송

⑦ MongoChem[8]
- 다양한 형태의 화학 데이터를 스키마가 존재하지 않는 NoSQL 인 MongoDB[9]에 저장하여 검색 기능 제공

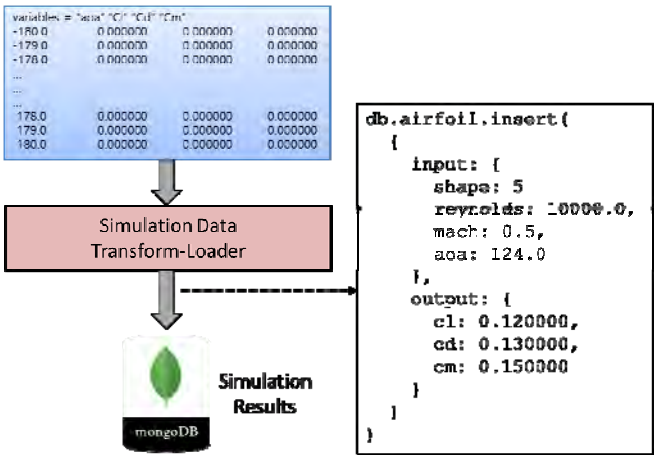
3. 시뮬레이션 데이터 관리 시스템 설계

본 장에서는 새로 요청된 시뮬레이션을 기존에 수행한 시뮬레이션 결과를 사용하여 효율적으로 처리하는 시스템을 설계한다. 특히 제안 시스템은 여러 시뮬레이션 소프트웨어의 결과를 활용할 수 있도록 한다. 제안 시스템은 크게 (1) 시뮬레이션 데이터를 데이터베이스에 적재하는 모듈, (2) 시뮬레이션 데이터를 검색하는 모듈, (3) 각 시뮬레이션 데이터에 대한 메타데이터를 관리하는 모듈 등으로 나뉘며, 아래에서 각 모듈에 대해 설명한다.

3.1 시뮬레이션 데이터 적재 모듈

시뮬레이션 데이터를 추후에 검색할 수 있게 하기 위해 시뮬레이션 데이터를 데이터베이스에 저장하는 모듈이다. 시뮬레이션 소프트웨어의 출력 형태는 서로 다르기 때문에, 각 출력 형태에 대한 Simulation Data Transform-Loader 가 필요하다. Simulation Data

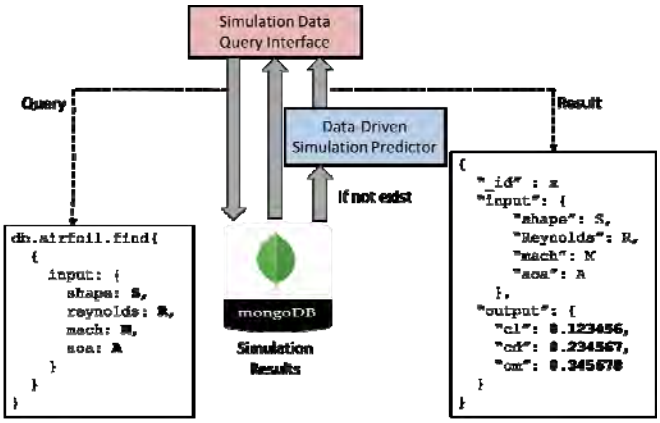
Transform-Loader 는 출력된 시뮬레이션 데이터를 변환하여 데이터베이스에 적재한다. 본 연구에서는 다양한 형태의 시뮬레이션 데이터를 지원하기 위해 스키마가 없는 대표적인 document-store NoSQL 인 MongoDB 를 사용하였다. Simulation Data Transform-Loader 는 시뮬레이션 출력 결과를 읽어 MongoDB 에 삽입가능한 document 형태로 변환한 뒤, 이를 MongoDB 에 적재한다. (그림 2)는 에어포일(airfoil) 시뮬레이션 결과를 MongoDB 에 적재하는 예이다.



(그림 2) 시뮬레이션 데이터 적재 모듈

3.2 시뮬레이션 데이터 검색 모듈

사용자가 새로운 시뮬레이션을 요청한 경우, 해당 시뮬레이션의 결과가 이미 존재하는지 MongoDB 에서 검색한다. 만약 존재하면 해당 결과를 반환하고, 존재하지 않으면 기존 시뮬레이션 결과들을 활용하여 요청한 시뮬레이션에 대한 결과를 예측한다. 본 연구에서는 시뮬레이션 결과를 예측하기 위해, 기존의 측정 데이터를 바탕으로 측정되지 않은 데이터를 예측하는데 주로 사용되는 방법인 다중 회귀 분석을 사용한다. 구체적인 예측 방법은 본 논문의 범위를 벗어나므로 본 논문에서는 생략한다.



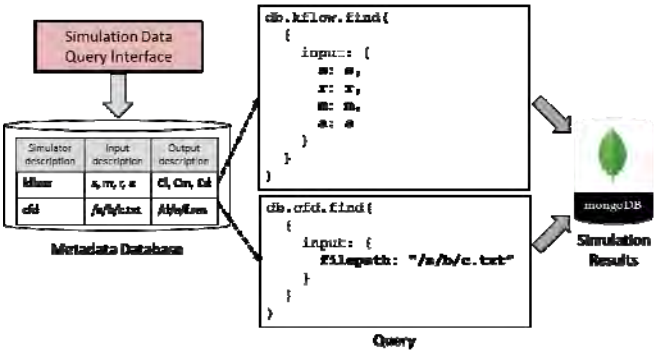
(그림 3) 시뮬레이션 데이터 검색 모듈

3.3 시뮬레이션 데이터 메타데이터 모듈

사용자들이 사용하는 시뮬레이션 소프트웨어는 여러 종류일 수 있으며, 각 시뮬레이션 소프트웨어는 서로 다른 입력 형태와 서로 다른 출력 형태를 가진다. 이를 관리하기 위해 Metadata database 를 둔다. Metadata database 는 각 시뮬레이터에 대해 크게 다음 세 종류 정보를 관리한다.

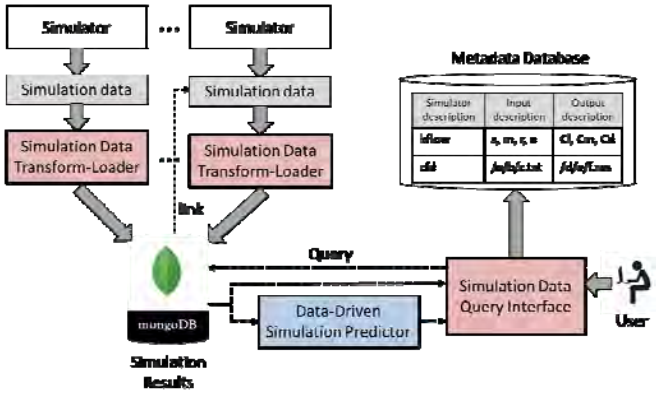
① 소프트웨어에 관한 정보: 소프트웨어, 버전, OS, 수행관련 환경 변수 등
② 입력 정보: 시뮬레이션 수행에 사용된 입력 값에 대한 정보. 변수-값 쌍이 될 수도 있고, 입력 값을 저장하고 있는 파일 경로를 나타낼 수도 있음
③ 출력 정보: 시뮬레이션 결과에 포함된 내용에 관한 정보. 어떤 값들을 가지고 있는지를 나타낼 수도 있고, 결과를 저장하고 있는 파일 경로를 나타낼 수도 있음

(그림 4)는 시뮬레이션 데이터 검색 모듈이 Metadata database 를 참고하여 특정 시뮬레이터에 대한 입력 정보를 찾고, 찾은 입력 정보에 따라 MongoDB 에 대한 질의를 생성하여 검색하는 예를 나타낸다.



(그림 4) 시뮬레이션 데이터 메타데이터 모듈

3.4 전체 시스템 아키텍처



(그림 5) 전체 시스템 아키텍처

(그림 5)는 지금까지 설명한 시뮬레이션 데이터 관리 시스템의 전체 아키텍처를 나타내는 그림이다. 3.1 절에서 설명한 바와 같이 사용자는 다양한 시뮬레이션 소프트웨어(혹은 시뮬레이터)를 사용할 수 있으며, 각각은 서로 다른 입력 형태와 출력 형태를 가진다.

각 시뮬레이터가 출력한 시뮬레이션 결과는 각 시뮬레이터에 대해 별도로 구현된 Simulation Data Transform Loader 에 의해 JSON 형태의 document 로 변환되어 MongoDB 에 저장된다. 이 때 Metadata Database 는 각 시뮬레이터에 대한 정보(소프트웨어정보, 입력정보, 출력정보)를 저장한다.

사용자 또는 어떤 프로그램이 시뮬레이션의 수행을 새로 요청하면, 이 요청은 Simulation Data Query Interface 에 전달된다. Simulation Data Query Interface 는 먼저 Metadata Database 를 참조하여, 요청된 시뮬레이션에 대한 입력 형태를 파악한다. 그리고 파악된 입력 형태에 맞춰 MongoDB 에 대한 질의를 구성한 뒤 이를 MongoDB 에 제출하여 검색을 수행한다. 만약 검색된 결과가 있으면 이를 바로 사용자에게 반환한다. 만약 검색 결과가 없으면 별도의 Data-Driven Simulation Predictor 를 호출한다. Data-Driven Simulation Predictor 는 요청된 시뮬레이션에 대한 결과를 MongoDB 에 저장된 기존의 시뮬레이션 데이터를 분석하여 예측한 후 해당 예측 결과를 반환한다.

따라서 사용자는 시뮬레이션을 새로 수행하지 않고도 기존 결과 혹은 예측 결과를 빠르게 전달받을 수 있다.

4. 결론 및 추후 연구

본 논문에서는 기존에 생성된 시뮬레이션 데이터를 활용하여, 새로 요청된 시뮬레이션을 효율적으로 처리하는 시스템을 설계하였다. 이를 위해 우선 시뮬레이션 데이터를 포함하여 다양한 과학 데이터를 재사용 및 공유하는 시스템들을 조사하였다. 그리고 스키마가 자유로운 MongoDB 를 사용하여 시뮬레이션 데이터를 저장하고 검색하는 시뮬레이션 데이터 관리 시스템을 설계하였다. 제안 시스템은 다양한 시뮬레이터가 출력한 다양한 형태의 시뮬레이션 데이터를 검색할 수 있는 한편, 기존의 시뮬레이션 수행 결과가 없다면 비슷한 시뮬레이션 수행 결과를 사용하여 시뮬레이션 수행 결과를 예측하는 기능을 제공한다. 이러한 시스템을 통해, 사용자는 시뮬레이션을 새로 수행하지 않고도 매우 빠르게 원하는 시뮬레이션 결과를 얻을 수 있다.

추후 연구로는 본 논문의 설계 내용을 기반으로 실제 시뮬레이션 데이터 관리 시스템을 구현하고, 감소된 수행 시간을 측정하여 유용성을 평가할 것이다. 또한 시뮬레이션 결과 예측 모듈을 구현하여 예측 결과의 정확도를 실제 수행 결과와 비교하여 평가할 것이다. 마지막으로 제안 시스템이 점차 더 다양한 환경에서 범용적으로 사용될 수 있도록 기능을 확장해 나갈 예정이다.

Acknowledgement

이 논문은 2015 년도 정부 (미래창조과학부)의 재원으로, 한국연구재단 첨단사이언스 및 교육 허브 개발 사업 (EDISON)의 지원을 받아 수행된 연구임 (No. NRF-2011-0020576)

참고문헌

[1] Angela B. Shiflet and George W. Shiflet, “Introduction to Computational Science: Modeling and Simulation for the Sciences,” 2nd edition, Princeton University Press, 2014.

[2] Cipran Docan, Manish Parashar, Scott Klasky, “DataSpaces: an interaction and coordination framework for coupled simulation workflows,” Cluster Computing, vol. 15, no. 2, pp. 163-181, 2012.

[3] Adam Hospital, Pau Andrio, Cesare Cugnasco, Laia Codo, Yolanda Becerra, Pablo D. Dans, Federica Battistini, Jordi Torres, Ramon Goni, Modesto Orozco, and Josep Ll. Gelpı, “BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data,” Nucleic Acids Research, vol. 44, 2016.

[4] D. Mishin, D. Medvedev, A. S. Szalay, R.Plante, and M. Graham, “Data Sharing and Publication Using the SciDrive Service,” Astronomical Data Analysis Software and Systems, vol. 485, 2014.

[5] Anand Kumar, Vladimir Grupcev, Meryem Berrada, Joseph C Fogarty, Yi-Cheng Tu, Xingquan Zhu, Sagar A Pandit and Yuni Xia, “DCMS: A data analytics and management system for molecular simulation,” Journal of Big Data, vol. 1, no. 9, 2014.

[6] Julien C. Thibault, Julio C. Facelli and Thomas E. Cheatham, III, “iBIOMES: Managing and Sharing Biomolecular Simulation Data in a Distributed Environment,” Journal of Chemical Information and Modeling, vol. 53, pp. 726-736, 2013.

[7] Jian Huang, Xuechen Zhang, Greg Eisenhauer, Karsten Schwan, Matthew Wolf, Stephane Ethier, Scott Klasky, “Scibox: Online Sharing of Scientific Data via the Cloud,” In Proceedings of the 28th IEEE International Parallel & Distributed Processing Symposium, pp. 145-154, 2014.

[8] Sanjoy Singh Ningthoujam, Manabendra Dutta Choudhury, Kumar Singh Potsangbam, Pankaj Chetia, Lutfun Nahar, Satyajit D. Sarker, Norazah Basar and Anupam Das Talukdar, “NoSQL Data Model for Semi-automatic Integration of Ethnomedicinal Plant Data from Multiple Sources,” Phytochemical Analysis, vol. 25, no. 6, pp. 495-507, 2014.

[9] MongoDB, https://www.mongodb.org/.