Analysis of Biomedical Big Data with Cloud-based In-Memory Database and Dynamic Querying
A 3 Hour Hands-on Experience with Real-world Data

Mengling Feng [1], Mohammad Ghassemi [1], Thomas Brennan [1], John Ellenberger [2], Ishrar Hussain [2], Roger Mark [1]

[1]*Massachusetts Institute of Technology,* [2]*SAP Research*
[1]{*mfeng, ghassemi, tpb, rgmark*}*@mit.edu,* [2]{*john.ellenberger, ishrar.hussain*}*@sap.com*

Abstract

Analyzing Biomedical Big Data (BBD) is computationally expensive due to high dimensionality and large data volume. Performance and scalability issues of traditional database management systems (DBMS) often limit the usage of more sophisticated and complex data queries and analytic models. Moreover, in the conventional setting, data management and analysis use separate software platforms. Exporting and importing large amounts of data across platforms require a significant amount of computational and I/O resources, as well as potentially putting sensitive data at a security risk. In this tutorial, we will explore an in-memory DBMS as a solution to BBD analysis. The participants will learn the advantages of in-memory DBMS through hands-on exercises with the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) database, a large, multi-dimensional Electronic Health Record (EHR) database for over 60,000 ICU admissions. In addition, we seek to educate the participants on effective analytic methods for BBD, including dynamic queries and statistical analysis with R.

1 Introduction

Due to advances in data capturing technologies and the recent effort to digitize medical record data, large amounts of biomedical data have been generated from genetic research studies, Electronic Health Record (EHR) systems, and physiological sensors. Analyzing Biomedical Big Data (BBD), as well as other “Big Data” archives, is computationally expensive due to the high dimensionality and large volume of the data. Performance and scalability issues of traditional DBMS often limit the usage of more sophisticated and complex data queries and analytic models.

BBD analysis is complex and often involves complicated statistical inference tasks, such as hypothesis testing, association studies and causal inferences. To accomplish these tasks, one needs to include multiple variables, the interactions among these variables and their transformed features. In addition, the analysis may also require sophisticated machine learning methods, such as kernel methods, boosting methods and Hidden Markov Models (HMM). Standard online analytical processing (OLAP) tools cannot handle the complex analysis of BBD. Therefore, for BBD analysis, the data extraction and analysis are traditionally carried on separate software platforms. However, moving between platforms, especially when the data are large and complex, presents two major challenges: firstly, it can place sensitive data at a secutiry risk, and secondly it requires significant processing resources.

In-memory database management system (DBMS) is a potential solution for more effective analysis of BBD. In contrast to traditional DBMS, in-memory DBMS are capable of storing large amount of data in the main memory to enable fast and dynamic querying. Since the internal optimization processes are simpler and indexing methods are more efficient [8], in-memory DBMS often outperform traditional DBMS by an order of magnitude. The in-memory nature also allows the in-memory DBMS (e.g. HANA from SAP) to be easily integrated with statistical analysis and machine learning programing languages and libraries. Thanks to the rise of affordable memory technologies [10], in-memory DBMS are now a viable option for BBD management and analysis.

Technical issues are not the only limiting factor in effective analysis of BBD. It is also important to note that although the relational database approach has the advantage of reducing data repetitions, minimizing storage space usage and simplifying data management, it comes at the expense of architectural complexity. As a relational database increases in size, the complexity grows and so does the need for analyst's enhanced SQL skills and familiarity with database-specific idiosyncrasies. Even for an experienced database analyst, the task of discerning the subtle intricacies of a database is a non-trivial task.

In this tutorial, the participants will learn the difference between in-memory DBMS and traditional DBMS through hands-on exercises using SAP's cloud-based HANA [4] in-memory DBMS in conjunction with the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) dataset. MIMIC [9] is an open-access critical care EHR archive (over 4TB in size) and consists of structured, unstructured and waveform data. Furthermore, this tutorial will seek to educate the participants on how a combination of dynamic querying, and in-memory DBMS may enhance the management and analysis of complex clinical data.

The learning objectives of the tutorial are:

- to learn the differences between in-memory DBMS and traditional DBMS through hands-on exercises,
- to demonstrate the advantage of in-memory DBMS for BBD management and analyses,

- to create a better understanding of dynamic querying and analysis,
- to introduce in-memory DBMS and dynamic querying tools that are available for the research community, and
- to introduce MIMIC, a valuable open-access BBD resource, to the VLDB community.

2 Targeted Audience

The tutorial is suitable for a wide spectrum of researchers. In particular, we aim to attract data scientists, who are interested in:

- knowledge discovery from BBD,
- in-memory DBMS,
- dynamic querying and data analysis, and
- cloud-based data management and analytic solutions.

3 Previous Related Tutorials

We conducted a tutorial at the ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM-BCB) in September 2013 in Washington, DC [5] to introduce the content of the MIMIC dataset. In that tutorial we guided participants through the MIMIC schema and explained the interpretation and significance of the data.

4 Outline of Tutorial

4.1 MIMIC: An Open-Access Biomedical Big Data (45 minutes)

The data generated in the process of medical care has historically been drastically under-utilized. This was due in part to the difficulty of accessing, organizing and utilizing data entered on paper charts. In addition, the notable variability in clinical documentation methods and quality made secondary analysis of clinical data even more challenging [6]. Our lab and partners have built and maintained the MIMIC database since 2003. MIMIC is the largest open-access BBD resource for critical care, and it is currently used by over 1,000 academic and industrial investigators from over 32 countries. MIMIC holds clinical data from over 60,000 ICU patient admissions. The data has been meticulously de-identified and is shared online with the research community via the PhysioNet portal [3].

In this demonstration, we will guide the participants in navigating through both the clinical component (400 GB) and the time-series components (4 TB) of MIMIC, hosted in the cloud-based HANA platform. The clinical component includes structured data, such as demographic data, lab & microbiology test results, medication & fluid records, diagnoses, hourly nurse-verified vital sign readings and patient mortality data, as well as rich unstructured text data, including physician notes, imaging reports and discharge summaries. The time-series component of MIMIC consists of continuous, high-resolution (8-bit, 125Hz) physiological waveform data, as well as vital sign trend data sampled at 1 data point per minute. Using our SQL query cookbook [2], we will run through a number of exercises to allow participants to familiarize themselves with the HANA in-memory DBMS and MIMIC's data content.

4.2 In-Memory Database Management Systems for Data Science (90 minutes)

It is common knowledge that BBD has been growing in both dimensionality and complexity. Performance and scalability issues often become critical when using conventional DBMS for large, complex data models. Training and/or testing these models with BBD sources involve extensive query times making them unfeasible for near-real time applications. In this tutorial, we will explore how today's in-memory DBMS can be used to combat these problems. We will be focusing on BBD, as it highlights the benefits of combining fast, dynamic queries with inline analysis with R.

The in-memory DBMS' that have been introduced in the recent years (e.g. Oracle TimesTen, SAP HANA, Microsoft SQL Server etc.) are capable of storing large volumes of data in main memory, instead of on disk drives. They are capable of dramatically reducing the time needed to perform data queries by eliminating the disk read time [7]. For our exercises, we will be using SAP HANA as our in-memory DBMS and PostgreSQL as our traditional DBMS. Both platforms will be populated with the MIMIC database and will be hosted in a cloud server.

The SAP HANA database platform includes built-in data analytic capabilities, which allows both data extraction and analysis under a single platform. The HANA platform comes integrated with the R libraries, which consist of rich collections of analytical, statistical and machine learning algorithms. With the cloud-based HANA system, participants will be guided to conduct data analytic tasks with the MIMIC data.

Overview of the session

1. **Introduction to the tools:** A brief introduction to the in-memory and conventional database platforms and the associated tools used in these exercises
2. **Performance comparison:** Comparison of the query speed of the in-memory and traditional DBMS. Some example exercises using the MIMIC data include:

- build histograms for variables, such as age, height and length of hospital stay,
- compute and generate the scatter plot of patients' total fluid intake during the entire ICU stays and their 28-day mortality, and
- calculate correlations between the hourly recorded urine output and mean blood pressure

The participants will have a chance to try out the integrated R libraries in some of these exercises.

3. **Statistical inference & machine learning exercises:** We will guide the participants over a number of statistical studies and also demonstrate the application of some machine learning and clustering algorithms:

   - compute t-test and analysis of variance (ANOVA) for hypothesis testing,
   - conduct survival analysis to investigate the differences in long-term mortality rates among various ICU units,
   - apply support vector machine and random forest to predict patients' in-hospital mortality, and
   - apply K-means clustering to cluster similar patients into groups

4.3 Dynamic Querying and Analysis with In-Memory DBMS (45 minutes)

Challenges in querying complex databases
Effective querying a complex database requires an intimate knowledge of the database schema. It follows that building familiarity with the data source is a common bottleneck for study design and implementation. The real-word nature of clinical databases, like MIMIC, are especially prone to these issues, given the hetergoneity of healthcare systems, and therefore not easily understood from inspection of the schema alone. For example, the EHR system behind the MIMIC data did not enforce a strict coding system and allowed free text input for medical concepts. As a consequence, medical concepts, such as heart rate, or blood pressure readings, were coded with multiple identification IDs in the database. Attempting to query MIMIC without this knowledge could easily lead a novice user to extract an incomplete subset of the available data, or produce inaccurate findings.

Effective query coding with dynamic SQL
Dynamic SQL is simply a way of structuring queries that allows them to applied to be a more general purpose class of inquires. It is synonymous to writing functions, in other programming languages. A dynamic query could,

for instance, perform a fixed set of operations on a table which is specified by the user at run-time. The main advantage of dynamic querying is that it enables database users to effectively share knowledge about querying techniques and database idiosyncrasies. This significantly reduces the barrier to entry for more novice users of the database.

DynaMIMIC, a public library of dynamic SQL queries for MIMIC Writing effective dynamic SQL queries requires an intimate familiarity with the database schema. We have developed DynaMIMIC, a dynamic query toolkit for the MIMIC database. Our decision to develop DynaMIMIC comes from our experiences at previous MIMIC related workshops, where we found novice users struggling to develop an intuition on how to query the database. We will showcase DynaMIMIC and show users how to access it for their own research purposes.

5 Hardware & Software Requirements

The software tools and DBMS platforms required in this tutorial will be hosted remotely on a cloud server. Thus, participants are not required to download or install any software packages. All that is required is a laptop with a web browser. A demo version of the MIMIC database will be used for this tutorial. Since the demo version contains only the data of patients who died, the data is open without any restrictions.

6 Biographies of Instructors

Roger G. Mark earned the SB and PhD degrees in electrical engineering from Massachusetts Institute of Technology and the MD degree from Harvard Medical School and trained in internal medicine with the Harvard Medical Unit at Boston City Hospital. At present Dr. Mark is Distinguished Professor of Health Sciences and Technology, and Professor of Electrical Engineering at MIT. His current research activities include *Integrating Data, Models, and Reasoning in Critical Care* (http://mimic.mit.edu) and *PhysioNet* (http://www.physionet.org), both of which involve physiological signal processing, cardiovascular modeling, and the development, use and open distribution of large physiological and clinical databases.

John Ellenberger currently works in the "Chairman's Special Projects" group at SAP Labs where he serves as the liaison to MIT where he focuses on Big Data Topic including medical analytics, machine learning and data privacy and architecture. John has worked in SAP's research function for 10 years in both strategic and technical research roles. Previous to that John spent 30 years managing software development teams working on the

1st generation products in the areas of multimedia services, online product and catalog management, network-based speech services, software tools and VLSI layout/verification. Most recently he led the engineering team at Nokia that developed the first generation of multimedia messaging services.

Mengling Feng is currently a visiting scholar at MIT in Harvard-MIT Health Sciences and Technology Division. He is also one of the faculty members of MIT course HST 936 “Global Health Informatics”. Dr. Feng’s PhD study focused on developing data mining methods to discover meaningful knowledge that impacts real life practices. Before his current affliction at MIT, Dr. Feng joined the Data Analytic Department of Institute for Infocomm Research (I2R) as a research scientist. Dr. Feng was awarded the Ministry of Education Scholarship for his undergraduate studies and the A*STAR Graduate Scholarship for his PhD study. His work was also recognized with the *Bi-annual Best Paper Award* from the Institute for Infocomm Research. Dr. Feng’s research focus is to develop data mining and machine learning methods to discover or infer casual phenomenon among real-life practice and strategic planning.

Thomas Brennan is currently a post-doctoral Research Engineer in the Laboratory of Computation Physiology, MIT. With a background in electrical and computer engineering, Thomas Brennan was awarded the Rhodes Scholarship from South Africa in 2004. In 2009 he completed his D.Phil. in Biomedical Engineering from the University of Oxford. He then worked for the Vodafone Foundation developing a mobile-based platform to monitor and support community health workers in South Africa. In 2010, he accepted a Wellcome Trust post-doctoral research fellowship at the Institute of Biomedical Engineering in Oxford to develop and assess mobile health solutions for monitoring chronic disease in resource-constrained settings. He has over 10 years experience in cardiac modeling and biomedical signal processing, with a special focus on machine learning.

Mohammad Ghassemi is a PhD student in Electrical and Computer Engineering at the Massachusetts Institute of Technology with a research interest in statistical signal processing and medical informatics. In 2010, Mohammad received the Gates-Cambridge Scholarship to fund his MPhil at the University of Cambridge in Information Engineering. He was also awarded the Goldwater scholarship while pursuing two undergraduate degrees in Electrical Engineering and Applied Mathematics. He holds two patents, and has several years of experience working in both research and industrial settings in North America, the Middle East and Europe. Mohammad’s prior research experience spans machine learning, signal processing and neuroscience.

Ishrar Hussain is working as a Machine Learning Researcher at SAP Research Labs, Montreal, and developing state-of-the-art machine learning solutions for Big Data applications using the SAP HANA’s in-memory DBMS. He received his Master’ degree in Computer Science in 2007 from

Concordia University in Montreal, Canada. He is now also a Ph.D. Candidate at the same university and is expecting to graduate in 2014. Ishrar has been continuing his research in the fields of Natural Language Processing, Machine Learning and Requirements Engineering for the last eight years. He is proficient in different platforms for linguistic analysis, e.g. GATE, UIMA, Stanford NLP etc., and has developed several standalone data science applications. Ishrar is also the author of nine research papers.

References

[1] MIMIC Homepage. `http://physionet.org/mimic2/mimic2_clinical_flatfiles.shtml`, 2014. [Online; accessed 15-March-2014].

[2] MIMIC Query Cookbook. `http://criticaldata.mit.edu/resources/docs/MIMIC_Cookbook_HANAedition.pdf`, 2014. [Online; accessed 15-March-2014].

[3] PhysioNet. `www.physionet.org`, 2014. [Online; accessed 15-March-2014].

[4] SAP HANA In-memory Computing. `http://www.sap.com/pc/tech/in-memory-computing-hana/software/overview/index.html`, 2014. [Online; accessed 15-March-2014].

[5] ACM-BCB. ACM-BCB 2013 Tutortial: Hands-on Experience with the MIMIC II Database: An Open-Access Database for Knowledge Discovery and Reasoning in Critical Care. `http://www.cse.buffalo.edu/ACM-BCB2013/acceptedtutorial.html`, 2014. [Online; accessed 15-March-2014].

[6] L. A. Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery. Big data in the intensive care unit. closing the data loop. *American Journal of Respiratory and Critical Care Medicine*, 187(11):1157–1160, 2013.

[7] H. Garcia-Molina and K. Salem. Main memory database systems: An overview. *IEEE Trans. on Knowl. and Data Eng.*, 4(6):509–516, Dec. 1992.

[8] T. J. Lehman and M. J. Carey. Query processing in main memory database management systems. In *Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data*, SIGMOD '86, pages 239–250, New York, NY, USA, 1986. ACM.

[9] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960, May 2011.

[10] M. Vizard. The Rise of In-Memory Databases. `http://slashdot.org/topic/datacenter/the-rise-of-in-memory-databases/`, 2014. [Online; accessed 15-March-2014].