

ADAM Genomics Schema - extension for precision medicine research*

Extended Abstract

Fodil Belghait
École de Technologie Supérieure
1100 Notre-Dame West
Montreal, Canada
fodil.belghait.1@ens.etsmtl.ca

Beatriz Kanzki
École de Technologie Supérieure
1100 Notre-Dame West
Montreal, Canada
beatriz.kanzki@ens.etsmtl.ca

Alain April
École de Technologie Supérieure
1100 Notre-Dame West
Montreal, Canada
alain.april@etsmtl.ca

ABSTRACT

High-throughput sequencing technologies have made research on precision medicine possible. Precision medicine treatments will be effective for individual patients based on their genomic, environmental, and lifestyle factors. This requires integrating this data to find one, or a combination of, single nucleotide polymorphisms (SNPs) linked to a disease or treatment [1]. In 2013, the University of California Berkeley's AmpLab created the ADAM genomic format that allows the transformation, analysis and querying of large amounts of genomics data by using a columnar file format. However, while ADAM addresses the issue of processing large genomics data; it lacks the ability to link the patients' clinical and demographical data, which is crucial in precision medicine research. This paper presents an ADAM genomic schema extension to support clinical and demographical data by automating the addition of data items to the currently available ADAM schema. This extension allows for clinical, demographical and epidemiological analysis at large scale as initially intended by the AmpLab.

CCS CONCEPTS

• **Database Theory** → **Database structures and algorithms**;
Data integration

KEYWORDS

Database, Genomics, Precision Medicine, Bioinformatics

*Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org

DH'18, April 23-26, 2018 Lyon, France

© 2018 Copyright is held by the owner/authors(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6493-5/18/04...\$15.00

<https://doi.org/10.1145/3194658.3194669>

1 INTRODUCTION

Next generation sequencing (NGS), has dominated the space in genomics research and has progressively entered the clinical practice [4]. One of the many efforts to leverage this type of technology is ADAM, a set of formats, APIs, and processing stage implementations for genomic data [3]. ADAM proposes a scalable pipeline for processing genomic data on top of high performance distributed computing frameworks. It uses Spark [6] as a compute engine and Parquet for fast data access [3]. The way these two technologies accomplish very high performance and efficiency are as follow:

1. Spark is an open-source fast compute engine for large-scale data processing, which provides an interface for programming entire clusters with implicit data parallelism and fault tolerance [5]
2. Parquet is also an open-source columnar data storage format available to Hadoop and Spark. When accessing a large amount of data, Parquet has offers interoperability, space and query efficiency.

While the ADAM format scales well when processing genomics data, its current schema (shown as the left part of figure 1) does not currently include the patient clinical and demographical data fields. We know that the availability of these data fields are crucial for research in the field of precision medicine [6]. To allow ADAM's processing pipeline for NGS technologies in this field, it is essential that these data entities be included. Precision medicine [7], aims at tailoring healthcare treatments for patients by using clinically actionable genomic mutations in guiding treatment and prevention of diseases. As shown in the current ADAM schema, ADAM includes genomic data [8]. To maximize discovery, NGS technologies coupled with the ADAM processing pipeline will have to be adapted to include the clinical needs as well. This paper describes the extension proposed to the current ADAM genomic schema to leverage its current framework and toolset to perform large-scale clinical analysis for precision medicine researchers. One design characteristic of our proposal is that individual data items can be dynamically added. This easily allows the extended schema to be used for different research topics aiming at different goals in the field of precision medicine.

2 METHOD

2.1 Model Description

In most genomic sequencing pipelines, the endpoint is an interpretation phase, which is typically done by the analysis and visualization of the results. The objective of this extended schema is to allow researchers, especially in precision medicine, to collocate all their study data for more efficient analysis. This includes patient data, such as genomics, clinical and demographical data, as well as the research analysis meta data provided by the ADAM framework. The intention is to have all this data placed into the same database schema. We intend to release this extended schema as well as the source code for efficient loading using Apache Spark and Parquet file format in ADAM. This will help promote BigData analysis in precision medicine as it removes the current issue of dealing with heterogeneous data types and sources for the patient clinical data outside of ADAM.

2.2 Extended Schema

The extended ADAM genomics schema is currently composed of three sub-schemas in Fig. 1 : Genomics, Clinical and Analysis. Each of those is subdivided into categories and classes, which will be further explained.

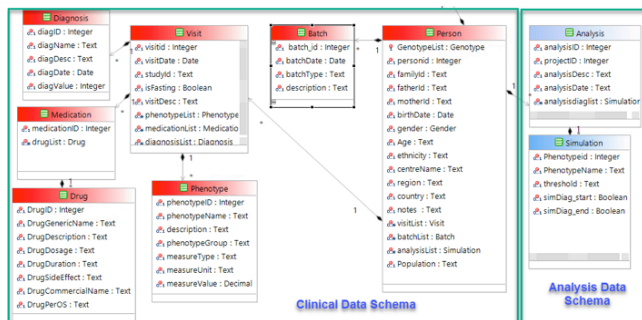


Figure 1 : ADAM genomics extension schema

2.3 Genomics Data Schema

The genomics schema is the core schema of the ADAM framework [3]. Currently, the schema cannot be altered by extensions in order to ensure its compatibility with existing ADAM frameworks' APIs.

2.4 Clinical Data Schema

The clinical data schema, presented in the left part of figure 1, provides precision medicine researchers with the data entities that can contain the clinical and demographical data of the patient. This is especially useful to allow cross analysis and data mining along with the genomics data of the patient. It contains the following five data classes: personal, visit and diagnostic history as well as phenotype and medical treatment.

2.4.1 *Personal history.* The personal history plays the role of a main index to all patient information. Medical history is directly stored in this class. However, complex data such as genomic and simulation data will be stored in a separate class, and the person class will store a reference to the latter. Here are examples of data categories stored in this class:

1. Demographic data such as age, ethnicity, population group, and information regarding patients' location: his country and the region from where he originated;
2. Reference to patients' clinical and genomic data, along with the reference to all analysis performed with the data can be found.

2.4.2 *Visit history.* The visit history class stores all the information required to keep track of each medical visit during the clinical experiment period along with all the major clinical information gathered during these visits such as: identified phenotypes, diagnostics' list, prescribed therapy and the list of medications.

2.4.3 *Diagnostic history.* The diagnostic history class contains the most relevant information from each diagnostic evaluated for the patient such as the diagnosis' name, description, and a binary value to specify if the patient has been diagnosed positive or negative for a specific health problem.

2.4.4 *Phenotype.* This class contains test results such as blood, urine test and other tests to diagnose health problems. The physician often uses set thresholds to locate specific issues. Having access to this information will allow for different diagnostic research based on varying threshold values and allow for more precise analysis of this data linked to a patients' genomic data.

2.4.5 *Medical Treatment.* This class is used to store information about the medication prescribed to the patients. It includes: medication generic and commercial names, treatment duration, dosage and secondary effect and can be combined with patients demographical and genomics data to enable precision medicine.

Overall, the clinical data schema proposed is intended to provide precision medicine researchers with the basic data items required to conduct a large-scale analysis using one single data store that can take advantage of recent BigData computing power and resources.

2.5 Analysis Data Schema

We have seen that our goal is to enable large-scale, complex and detailed analysis patient data analysis using BigData and machine learning algorithms, across these three collocated data schemas. This collocation will allow precision medicine researchers to use the power of open source BigData technology easily to identify potential correlations such as candidate genes responsible for specific diseases and impact of therapies and medications on a patients' health. This impact is typically evaluated using patients' category such as: age, gender, ethnicity, weight, etc. The objective of this schema extension, located at the right of Fig. 1., is to allow

researchers to better organize, track and reproduce their analysis results. It is composed of the following two main data classes:

2.5.1 *Analysis*. This contains general information that identifies the project with which the analysis is associated, along with a reference to all the simulations that have been done for each analysis.

2.5.2 *Simulation*. For each analysis, researchers can make several simulations for every phenotype with different thresholds and compare the results. This class stores all the information that identifies each simulation. The phenotype analyzed along with the threshold measurement used. This information will allow for the adjustment of the threshold values and the re-execution of the same simulation using different sets of data.

3 SCHEMA EXTENSION PROCESS

3.1 Extended schema creation process flow

We also designed an automated process to extend the current ADAM schema using the following workflow, shown in Fig. 2:

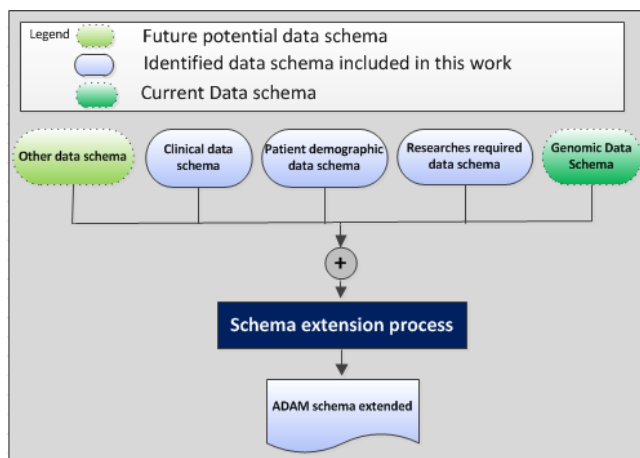


Figure 2: ADAM schema extension workflow

The schema extension workflow executes five sequential steps as follows:

1. Identify all the data fields required for your specific precision medicine analysis;
2. Verify if the current ADAM schema or one that has already been extended recently in the community has all the required data items you are looking for;
3. If some data fields are missing, write the needed fields definition, using the Avro format shown in Fig. 3, so that they can be added in an extended schema;
4. Modify the schema extension script, to add the Avro definition of these new data fields;
5. Execute the script that will generate the new ADAM extended schema composed of the existing ADAM data fields structure, in Avro format with the Parquet files.

Figure 3: Example of AVRO record definition

```

record diagnosis {
  /** id diagnosis */
  union { null, integer } diagID = null;

  /** nameDiagnosis */
  union { null, string } diagName = null ;

  /** descDiagnosis */
  union { null, date } diagDesc = null ;

  /** date Diagnosis */
  union { null, date } diagDate = null ;

  /** value Diagnosis, binary value to indicate if
  the patient has been diagnosed positive or
  negative for the current phenotype */
  union { null, integer } diagValue = null ;
}
    
```

Following these simple steps will allow you to use this personalized extended schema without creating any impact on the existing ADAM APIs that are currently available to load data using many different genomics file formats. However, now you will need to create your own APIs in order to populate your added data fields.

4 SCHEMA EXTENSION PROCESS DESCRIPTION

To use our automated extension script, you will require the following software tools:

1. Apache Maven: an popular open source tool used to build and manage Java-based projects [9];
2. Apache Hadoop: the well known BigData open source framework used to support processing and storage of very large datasets across distributed clusters of computers designed to scale up from single servers to thousands of machines easily [10];
3. Apache Avro: a recent open source framework offering remote procedure call communication protocol and data serialization. Avro uses the JSON format for data types and protocols definition and serializes the data in a compact binary file. Our automated extension script uses the following two components of Avro [11] as well as the ADAM schema:
 - a. avrotools.jar: can be obtained from avro.apache.org ;
 - b. avro2parquet.jar: can be obtained from github at /tispartick/avro2parquet ;
 - c. bdg.avdl: which is the current ADAM schema that can be obtained from github at bigdatagenomics.

Once you have installed the needed toolset you are now ready to run our automated script: /ExtendAdamSchema.sh and it will

generate an extended ADAM schema. The script executes the following 3 steps sequentially:

1. First, it prepares the execution environment: It downloads and installs Apache Maven;
2. Second, it executes the schema conversion steps to generate the new ADAM Avro schema;
3. Third, it converts the new ADAM schema into Parquet files ready to receive your data.

After a successful execution, the script will generate, for each schema, a number of files: Analysis.avsc, Diagnosis.avsc, Medication.avsc, Phenotype.avsc, Simulation.avsc, Variant.avsc, Batch.avsc, Drug.avsc, Visit.avsc, Person.avsc and Genotype.avsc.

4 FUTURE RESEARCH

It is planned that in addition to this automated process and tool that easily extend the ADAM schema without impacting its current APIs, we will be designing other open source tools like:

4.1.1 *Large scale data loader.* The new schemas need to be loaded efficiently with large amounts of data. In this research project, we are currently designing an automated and scalable process to efficiently merge and load data efficiently into an extended ADAM genomics schema using Spark clusters. The challenge here is to merge and load very large quantities of heterogeneous data quickly and easily as clinical data is typically located in some existing relational database and the genomic data come from a number of very large gen/sample files.

4.1.2 *Machine learning APIs.* Once the data is quickly loaded into the extended ADAM genomics format, precision medicine researchers will want to consider different types of analysis in order to identify different patterns. In this research project, we are building APIs to allow easy integration of the extended ADAM genomics schema to open source BigData machine learning frameworks like H2O.ai and TensorFlow.

4.1.3 *Case study using Advance data.* The ADAM schema extension script, the large scale data loader as well as a machine learning API for H2O.ai will be experimented in a next phase of our research project. We will be using large amount of data originating from the Advance diabetes clinical data, their demographical data as well as the patients' genetic data. This data was provided to us by the Centre Hospitalier de l'Université de Montréal. This test of a large scale precision medicine study using Spark clusters and H2O will be performed hopefully during the fall of 2018.

4 LIMITATIONS

Our script was designed and tested on AWS using original versions of all open source software libraries. This is a choice we made and this is by no means an out-of-the-box solution and will require you to master a number of open source BigData

frameworks. Your results may vary if you use our script Cloudera, Hortonworks or MapR.

4 CONCLUSIONS

This paper has presented an open source process and toolset to extend the data elements of the current ADAM schema in order to add the patients' clinical and demographical data required for precision medicine research involving very large amount of data. A first advantage of the approach is not to disturb ADAM current APIs by not modifying the ADAM schema. A second advantage is that as the ADAM schema evolves, there is minimal impact to your extended schema, as you only have to rerun it against the new version of ADAM schema. The final advantage of collocating this data is that it now allows for large-scale precision medicine investigations that involve machine learning on the whole genome.

ACKNOWLEDGMENTS

This research is done without any funding. Researchers and students in software engineering at École de Technologie Supérieure (ÉTS) conduct this work freely during their master degree capstone project and accept to release the results as part of the University Berkeley ADAM project. Diabetes researchers and bioinformatics staff at Dr. Pavel Hamet research unit located at the Centre Hospitalier de l'Université de Montréal (CHUM) support us in understanding the data with the hope of finding an early predictor for early diagnosis and treatment of diabetes type 2.

REFERENCES

- [1] T. C. Carter and M. M. He, "Challenges of Identifying Clinically Actionable Genetic Variants for Precision Medicine," *Journal of Healthcare Engineering*, vol. 2016, pp:1–14, 2016.
- [2] F. A. Nothaft, M. Massie, T. Danford et al., "Rethinking Data-Intensive Science Using Scalable Analytics Systems", *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, May31-June 4, Melbourne, Australia, pp:631–646, 2015.
- [3] M. Massie, F. Nothaft, C. Hartl et al., "ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing", *Technical Report No. UCB/EECS-2013-207*, University California Berkeley, 2013, 22p. [Online] available at: <https://pdfs.semanticscholar.org/2228/b4208c5ea6754df6edcae805038f3e47857c.pdf> (Accessed: March/10/2018).
- [4] R. Gullapalli, M. Lyons-Weiler et al., "Clinical integration of next-generation sequencing technology", *Clinics in laboratory medicine*, vol. 32, no. 4, pp:585–599, 2012.
- [5] E. A. Ashley, "Towards precision medicine", *Nature Reviews Genetics*, vol. 17, no. 9, pp:507–522, 2016.
- [6] M. Zaharia, M. J. Franklin, A. Ghodsi, et al., "Apache Spark: a unified engine for big data processing", *Communications of the ACM*, vol. 59, no. 11, pp:56–65, 2016.
- [7] A.D.M. Roden and R.F. Tyndale, "Genomic Medicine, Precision Medicine, Personalized Medicine : What's in a Name? *Clinical Pharmacology & Therapeutics*, vol. 94, no. 2, pp:169–172, 2013.
- [8] B. Louie, P. Mork, F. Martin-Sanchez, et al., "Data integration and genomic medicine", *Journal of Biomedical Informatics*, vol. 40, no. 1, pp:5–16, 2007.
- [9] The Apache Software Foundation, "Apache Maven Project", [Online]. Available: <https://maven.apache.org> (Accessed: March/10/2018).
- [10] The Apache Software Foundation, "Apache Hadoop", 2014. [Online] available at: <http://hadoop.apache.org/> (Accessed: March/10/2018).
- [11] The Apache Software Foundation, "Apache Avro™ 1.8.1 Documentation", 2018. [Online] available at: <http://avro.apache.org/> (Accessed: March/10/2018).