T. Grust et al. (Hrsg.): Datenbanksysteme für Business, Technologie und Web (BTW 2019),
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 507

MSDataStream — Connecting a Bruker Mass Spectrometer to the Internet

Roman Zoun[1], Kay Schallert[2], David Broneske[1], Wolfram Fenske[1], Marcus Pinnecke[1], Robert Heyer[2], Sven Brehmer[3], Dirk Benndorf[2], Gunter Saake[1]

Abstract: Metaproteomics is the biological research of proteins of whole communities comprised of thousands of species using tandem mass spectrometry. But still it follows a sequential non parallelizable workflow. Hence, researchers have to wait for hours or even days until the measurement data are available. In our demo, we show a way to decrease the smallest unit of the workflow to a minimum to realize a near real time stream processing system on a fast data architecture.

Keywords: Internet of Things; Bioinformatics; Mass Spectrometry; SMACK Stack; Streaming

1 Introduction

Metaproteomics is the biological research of proteins of whole communities with thousands of species (e.g. from ocean samples or the human gut) [Ma07]. Metaproteomics is very important for diagnosing diseases, optimizing biogas plants, etc. [Ma07, PF17]. This research relies on a mass spectrometer, that is figuratively speaking a huge scale for tiny particles [Ma07]. The mass spectrometer measures the mass of parts (peptides) of a protein complex. Proteomics and metaproteomics follow a similar workflow, as shown in Figure 1. The workflow encompasses the following steps: (1) The biological sample gets purified, in a way that only proteins are left in the sample. (2) The proteins are split into smaller parts, called peptides. (3) The peptides are measured in the mass spectrometer. (4) The mass spectrometer transforms all peptides into a digital signal, one peptide is represented as one spectrum (~2 hours) [Ma07]. (5) Conversion of the digital signal into a readable format, such as MGF (~1 hours) [Ki10]. (6) Compare the experimental spectra with real-world proteins to identify the experimental data (~2 hours) [Mi13]. (7) The identified spectra need a validation to remove false positive results (~2 hours) [El10]. (8) Biological researchers analyze the experimental results. All these steps are sequentially executed and it takes hours to complete. The workflow has not changed for years and newer devices increase the amount of data produced by a mass spectrometer (~20GB per experiment, ~45000 spectra,

[1] University of Magdeburg, Working Group Databases and Software Engineering, Germany, {firstname.lastname@ovgu.de}
[2] University of Magdeburg, Chair of Bioprocess Engineering, Germany, {firstname.lastname@ovgu.de}
[3] Bruker Daltonik GmbH, 28359 Bremen, Germany, Sven.Brehmer@bruker.com

doi:10.18420/btw2019-33

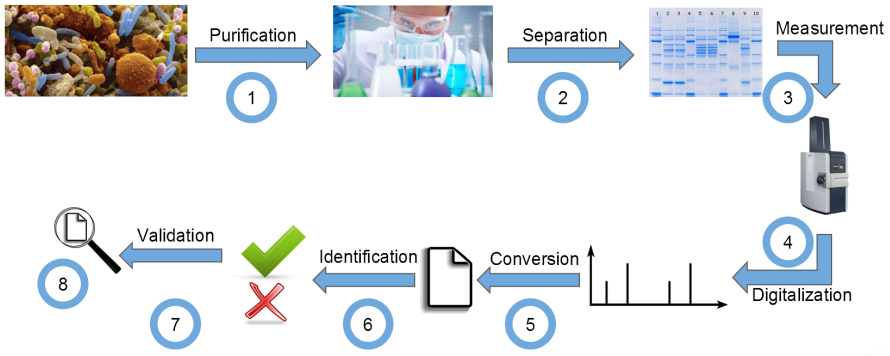

Fig. 1: The current (meta-)proteomics research workflow.

~6 spectra/second) and the current file-based workflow is outdated [He17]. Especially in clinical use cases, the goal is to process the data in real-time, but this is infeasible with the current approach. The steps 1, 2 and 3 are realized in a laboratory, step 4 is done by a mass spectrometer, but the other steps are completely digital and can be optimized using IT. Currently, the smallest parallelizable unit is a whole experiment. Since the mass spectrometer measurement and digitalization duration (~2 hours) cannot be avoided, we shrink the smallest unit after the fourth step to a single spectrum instead of a whole experiment. That means, we connect a mass spectrometer to the cloud for outsourcing the calculation and overlap mass spectrometer processing (3-4) with data processing (5-8) by using a streaming-based architecture [Wa16]. We choose a Fast data architecture, since we need near real time processing of huge amounts of mass spectrometry data [Wa16]. Specifically, we deployed a SMACK Stack[4]. In this demo, we present a cornerstone of our architecture, our tool MSDataStream. The tool is responsible for grabbing the single spectrum data from the mass spectrometer as it arrives, converting it into a readable format and streaming the data to the cloud for processing. In the cloud each spectrum needs a comparison to all peptides in our database (~27 millions) and validation of the matched results using machine learning classification.

2 System Architecture

For our development, we collaborate with Bruker Daltonik GmbH, a mass spectrometer company from Bremen, Germany. Each of their devices are connected via a digitizer to a computer. The digital signal is collected in a proprietary RAW file format. Additionally, the measurement software provides structure and meta data to index spectrum data that belongs together. Since each manufacturer uses their own RAW file format, Bruker provided us with a library (DLL file) to access the digital spectrum data. The index data is stored in an SQLite database and provides the location of spectrum data in the RAW file. The index structure consists of 16 tables that describe meta-information and the spectrum location

[4] pipeline of Big Data technologies: Spark, Mesos, Akka, Cassandra, Kafka

for each single spectrum. MSDataStream reads the meta-information for a single spectrum from the SQLite index and extracts the spectrum data from the RAW file via the DLL. Afterwards, MSDataStream converts the spectrum data into a readable format. Then, several pre-processing methods increase the quality of the spectrum data. Finally, MSDataStream sends the data to the cloud via Kafka broker for further processing or to write the data into a file. Summarized, MSDataStream checks periodically (e.g. every 2 seconds) for new measured data, collects it and produces messages. The system architecture is shown in Figure 2. MSDataStream requires several parameters for the execution. We implemented

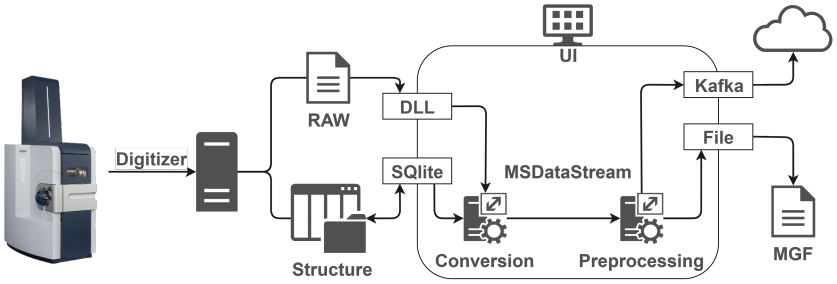

Fig. 2: The flow of the mass spectrometer data through the MSDataStream software

a JavaFX interface, which queries MSDataStream tasks and helps the user to configure the parameters. After sending the first data messages, the user waits until the results are generated from the fast data system and shown in a live updated visualization.

3 Streaming Workflow

The streaming data processes work for each spectrum separately (Figure 3). (1) Each spectrum, which arrives at the cloud system first gets prepared for the identification step with several filters on the data. (2) This step requires real-world protein data (sequence database), which is already in the system [Zo18a]. (3) The comparison of the experimental spectrum with the real world data produces peptide spectrum matches [Zo18b]. (4) Each of these matches needs validation, which is based on a machine learning classifier [Zo18c]. (5) Validated matches are stored as results in our database. The database stores all the experiment data such as validated results, spectra data and the proteins.

4 Demo Walkthrough

Our tool MSDataStream works with live measured data or with already finished measurements. For the demonstration, we will use a laggy version of the second method with time delays, which simulates live measurement. After a demo of the measurement process on a model of a mass spectrometer, the user can use the MSDataStream UI to configure and start a streaming process. During the process, the user will be shown the throughput of the system in the Apache Spark UI. Furthermore, the user can take a closer look at the components of the SMACK stack deployment in Marathon and mesos webUI. Afterwards, we show a live and continuously updated visualization of the experiment results, i.e., protein list.

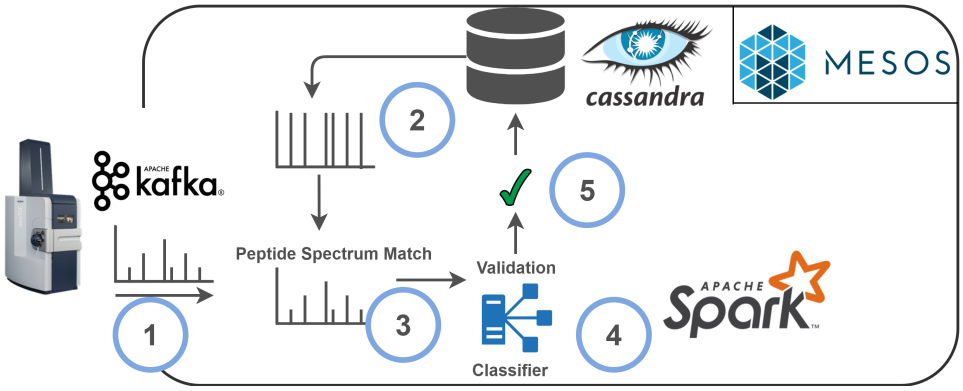

Fig. 3: The metaproteomics data analysis workflow on a SMACK stack.

Acknowledgments The authors sincerely thank Xiao Chen, Gabriel Campero Durand, Sebastian Krieter and Andreas Meister for their support and advice. This work is partly funded by the de.NBI Network (031L0103), the European Regional Development Fund (grant no.: 11.000sz00.00.0 17 114347 0), the DFG (grant no.: SA 465/50-1), by the German Federal Ministry of Food and Agriculture (grants no.: 22404015) and dedicated to the memory of Mikhail Zoun.

References

[El10] Elias, Joshua et al: Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. Methods in Molecular Biology, 604:55–71, 2010.

[He17] Heyer, Robert et al: Challenges and perspectives of metaproteomic data analysis. Journal of Biotechnology, 261(Supplement C):24 – 36, 2017.

[Ki10] Kirchner, Marc et al: MGFp: An Open Mascot Generic Format Parser Library Implementation. Journal of Proteome Research, 9(5):2762–2763, 2010. PMID: 20334363.

[Ma07] Maron, Pierre-Alain et al: Metaproteomics: A New Approach for Studying Functional Microbial Ecology. Microbial Ecology, Volume 53:486—493, 2007.

[Mi13] Millioni, Renato et al: Pros and cons of peptide isolectric focusing in shotgun proteomics. Journal of chromatography. A, 1293:1—9, June 2013.

[PF17] Petriz, Bernardo A.; Franco, Octávio L.: Metaproteomics as a Complementary Approach to Gut Microbiota in Health and Disease. Front Chem, 2017.

[Wa16] Wampler, Dean: Fast Data Architectures for Streaming Applications. O'Reilly Media, 1005 Gravenstein Highway North, Sebastopol, CA 95472., first edition, sep 2016.

[Zo18a] Zoun, Roman: Internet of Metaproteomics - Optimizing the Metaproteomics Workflow Using Fast Data on the SMACK Stack. Phd Symposium, ICDE, 2018.

[Zo18b] Zoun, Roman et al: Protein Identification as a Suitable Application for Fast Data Architecture. In: BIOKDD workshop, DEXA. Springer, Cham, pp. 168–178, 2018.

[Zo18c] Zoun, Roman et al: Streaming FDR Calculation for Protein Identification. In: New Trends in Databases and Information Systems. Springer, Cham, pp. 80–87, 2018.